An Infinitely Large Napkin

https://web.evanchen.cc/napkin.html

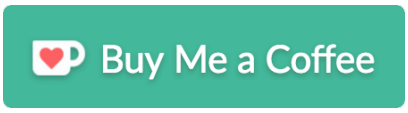
Evan Chen

Version: v1.6.20250519

When introduced to a new idea, always ask why you should care.
Do not expect an answer right away, but demand one eventually.

— *Ravi Vakil [Va17]*

If you like this book and want to support me,
please consider buying me a coffee!



https://ko-fi.com/evanchen/

For Brian and Lisa, who finally got me to write it.

This is (still!) an **incomplete draft**. Please send corrections, comments, pictures of kittens, etc. to evan@evanchen.cc, or pull-request at https://github.com/vEnhance/napkin.

Last updated May 19, 2025.

Preface

The origin of the name "Napkin" comes from the following quote of mine.

> I'll be eating a quick lunch with some friends of mine who are still in high school. They'll ask me what I've been up to the last few weeks, and I'll tell them that I've been learning category theory. They'll ask me what category theory is about. I tell them it's about abstracting things by looking at just the structure-preserving morphisms between them, rather than the objects themselves. I'll try to give them the standard example Grp, but then I'll realize that they don't know what a homomorphism is. So then I'll start trying to explain what a homomorphism is, but then I'll remember that they haven't learned what a group is. So then I'll start trying to explain what a group is, but by the time I finish writing the group axioms on my napkin, they've already forgotten why I was talking about groups in the first place. And then it's 1PM, people need to go places, and I can't help but think:
>
> *"Man, if I had forty hours instead of forty minutes, I bet I could actually have explained this all".*

This book was initially my attempt at those forty hours, but has grown considerably since then.

About this book

The *Infinitely Large Napkin* is a light but mostly self-contained introduction to a large amount of higher math.

I should say at once that this book is not intended as a replacement for dedicated books or courses; the amount of depth is not comparable. On the flip side, the benefit of this "light" approach is that it becomes accessible to a larger audience, since the goal is merely to give the reader a feeling for any particular topic rather than to emulate a full semester of lectures.

I initially wrote this book with talented high-school students in mind, particularly those with math-olympiad type backgrounds. Some remnants of that cultural bias can still be felt throughout the book, particularly in assorted challenge problems which are taken from mathematical competitions. However, in general I think this would be a good reference for anyone with some amount of mathematical maturity and curiosity. Examples include but certainly not limited to: math undergraduate majors, physics/CS majors, math PhD students who want to hear a little bit about fields other than their own, advanced high schoolers who like math but not math contests, and unusually intelligent kittens fluent in English.

Source code

The project is hosted on GitHub at https://github.com/vEnhance/napkin. Pull requests are quite welcome! I am also happy to receive suggestions and corrections by email.

Philosophy behind the Napkin approach

As far as I can tell, higher math for high-school students comes in two flavors:

- Someone tells you about the hairy ball theorem in the form "you can't comb the hair on a spherical cat" then doesn't tell you anything about why it should be true, what it means to actually "comb the hair", or any of the underlying theory, leaving you with just some vague notion in your head.

- You take a class and prove every result in full detail, and at some point you stop caring about what the professor is saying.

Presumably you already know how unsatisfying the first approach is. So the second approach seems to be the default, but I really think there should be some sort of middle ground here.

Unlike university, it is *not* the purpose of this book to train you to solve exercises or write proofs,[1] or prepare you for research in the field. Instead I just want to show you some interesting math. The things that are presented should be memorable and worth caring about. For that reason, proofs that would be included for completeness in any ordinary textbook are often omitted here, unless there is some idea in the proof which I think is worth seeing. In particular, I place a strong emphasis over explaining why a theorem *should* be true rather than writing down its proof. This is a recurrent theme of this book:

Natural explanations supersede proofs.

My hope is that after reading any particular chapter in Napkin, one might get the following out of it:

- Knowing what the precise definitions are of the main characters,

- Being acquainted with the few really major examples,

- Knowing the precise statements of famous theorems, and having a sense of why they *should* be true.

Understanding "why" something is true can have many forms. This is sometimes accomplished with a complete rigorous proof; in other cases, it is given by the idea of the proof; in still other cases, it is just a few key examples with extensive cheerleading.

Obviously this is nowhere near enough if you want to e.g. do research in a field; but if you are just a curious outsider, I hope that it's more satisfying than the elevator pitch or Wikipedia articles. Even if you do want to learn a topic with serious depth, I hope that it can be a good zoomed-out overview before you really dive in, because in many senses the choice of material is "what I wish someone had told me before I started".

More pedagogical comments and references

The preface would become too long if I talked about some of my pedagogical decisions chapter by chapter, so Appendix A contains those comments instead.

In particular, I often name specific references, and the end of that appendix has more references. So this is a good place to look if you want further reading.

[1] Which is not to say problem-solving isn't valuable; I myself am a math olympiad coach at heart. It's just not the point of this book.

Historical and personal notes

I began writing this book in December 2014, after having finished my first semester of undergraduate at Harvard. It became my main focus for about 18 months after that, as I became immersed in higher math. I essentially took only math classes (gleefully ignoring all my other graduation requirements), and merged as much of it as I could (as well as lots of other math I learned on my own time) into the Napkin.

Towards August 2016, though, I finally lost steam. The first public drafts went online then, and I decided to step back. Having burnt out slightly, I then took a break from higher math, and spent the remaining two undergraduate years[2] working extensively as a coach for the American math olympiad team, and trying to spend as much time with my friends as I could before they graduated and went their own ways.

During those two years, readers sent me many kind words of gratitude, many reports of errors, and many suggestions for additions. So in November 2018, some weeks into my first semester as a math PhD student, I decided I should finish what I had started. Some months later, here is what I have.

Acknowledgements

I am indebted to countless people for this work. Here is a partial (surely incomplete) list.

- Thanks to all my teachers and professors for teaching me much of the material covered in these notes, as well as the authors of all the references I have cited here. A special call-out to [Ga14], [Le14], [Sj05], [Ga03], [Ll15], [Et11], [Ko14], [Va17], [Pu02], [Go18], which were especially influential.
- Thanks also to dozens of friends and strangers who read through preview copies of my draft, and pointed out errors and gave other suggestions. Special mention to Andrej Vuković and Alexander Chua for together catching over a thousand errors. Thanks also to Brian Gu and Tom Tseng for many corrections. (If you find mistakes or have suggestions yourself, I would love to hear them!) Thanks also to Royce Yao and `user202729` for their contributions of guest chapters to the document.
- Thanks to Jenny Chu and Lanie Deng for the cover artwork.
- I'd also like to express my gratitude for many, many kind words I received during the development of this project. These generous comments led me to keep working on this, and were largely responsible for my decision in November 2018 to begin updating the Napkin again.

Finally, a huge thanks to the math olympiad community, from which the Napkin (and me) has its roots. All the enthusiasm, encouragement, and thank-you notes I have received over the years led me to begin writing this in the first place. I otherwise would never have the arrogance to dream a project like this was at all possible. And of course I would be nowhere near where I am today were it not for the life-changing journey I took in chasing my dreams to the IMO. Forever TWN2!

[2] Alternatively: " ... and spent the next two years forgetting everything I had painstakingly learned". Which made me grateful for all the past notes in the Napkin!

Advice for the reader

§1 Prerequisites

As explained in the preface, the main prerequisite is some amount of mathematical maturity. This means I expect the reader to know how to read and write a proof, follow logical arguments, and so on.

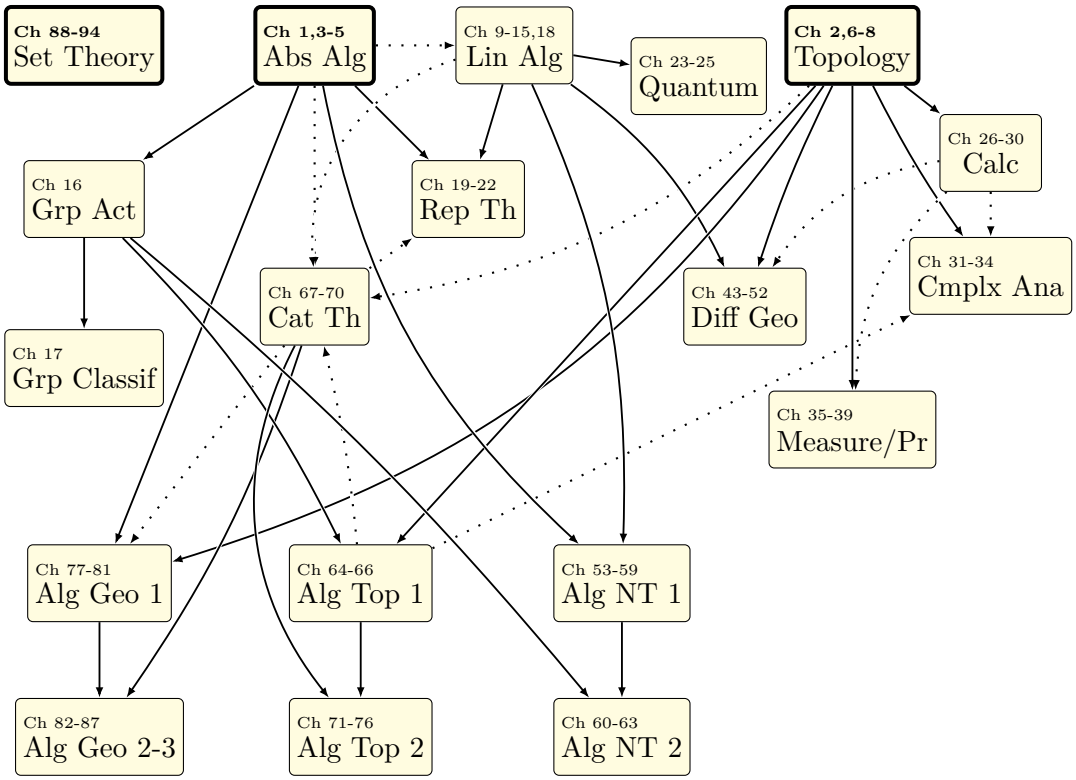
I also assume the reader is familiar with basic terminology about sets and functions (e.g. "what is a bijection?"). If not, one should consult Appendix E.

§2 Deciding what to read

There is no need to read this book in linear order: it covers all sorts of areas in mathematics, and there are many paths you can take. In Chapter 0, I give a short overview for each part explaining what you might expect to see in that part.

For now, here is a brief chart showing how the chapters depend on each other; again see Chapter 0 for details. Dependencies are indicated by arrows; dotted lines are optional dependencies. **I suggest that you simply pick a chapter you find interesting, and then find the shortest path.** With that in mind, I hope the length of the entire PDF is not intimidating.

(The text in the following diagram should be clickable and links to the relevant part.)



§3 Questions, exercises, and problems

In this book, there are three hierarchies:

- An inline **question** is intended to be offensively easy, mostly a chance to help you internalize definitions. If you find yourself unable to answer one or two of them, it probably means I explained it badly and you should complain to me. But if you can't answer many, you likely missed something important: read back.
- An inline **exercise** is more meaty than a question, but shouldn't have any "tricky" steps. Often I leave proofs of theorems and propositions as exercises if they are instructive and at least somewhat interesting.
- Each chapter features several trickier **problems** at the end. Some are reasonable, but others are legitimately difficult olympiad-style problems. Harder problems are marked with up to three chili peppers (🌶🌶🌶), like this paragraph.

  In addition to difficulty annotations, the problems are also marked by how important they are to the big picture.

  – **Normal problems**, which are hopefully fun but non-central.

  – **Daggered problems**, which are (usually interesting) results that one should know, but won't be used directly later.

  – **Starred problems**, which are results which will be used later on in the book.[1]

Several hints and solutions can be found in Appendices B and C.

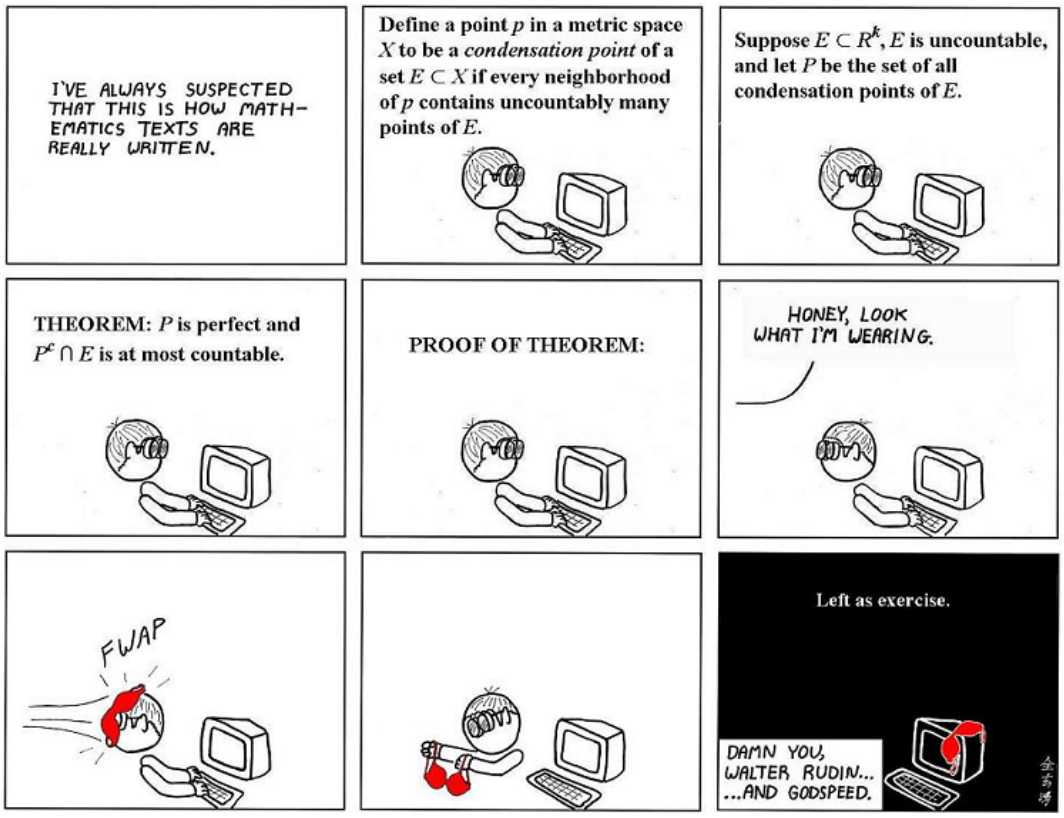


Image from [Go08]

[1] This is to avoid the classic "we are done by PSet 4, Problem 8" that happens in college sometimes, as if I remembered what that was.

§4 Paper

At the risk of being blunt,

> **Read this book with pencil and paper.**

Here's why:

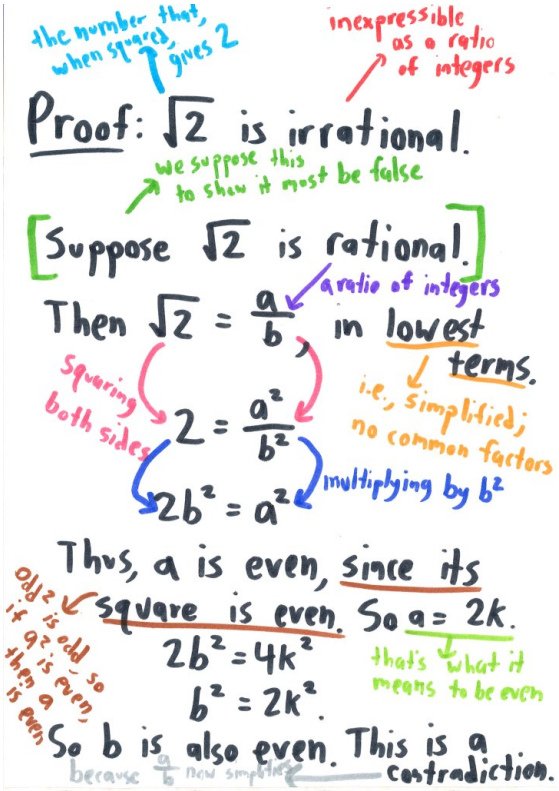


Image from [Or]

You are not God. You cannot keep everything in your head.[2] If you've printed out a hard copy, then write in the margins. If you're trying to save paper, grab a notebook or something along with the ride. Somehow, some way, make sure you can write. Thanks.

§5 On the importance of examples

I am pathologically obsessed with examples. In this book, I place all examples in large boxes to draw emphasis to them, which leads to some pages of the book simply consisting of sequences of boxes one after another. I hope the reader doesn't mind.

I also often highlight a "prototypical example" for some sections, and reserve the color red for such a note. The philosophy is that any time the reader sees a definition or a theorem about such an object, they should test it against the prototypical example. If the example is a good prototype, it should be immediately clear why this definition is intuitive, or why the theorem should be true, or why the theorem is interesting, et cetera.

Let me tell you a secret. Whenever I wrote a definition or a theorem in this book, I would have to recall the exact statement from my (quite poor) memory. So instead I often consider the prototypical example, and then only after that do I remember what

[2] See also https://blog.evanchen.cc/2015/03/14/writing/ and the source above.

the definition or the theorem is. Incidentally, this is also how I learned all the definitions in the first place. I hope you'll find it useful as well.

§6 Conventions and notations

This part describes some of the less familiar notations and definitions and settles for once and for all some annoying issues ("is zero a natural number?"). Most of these are "remarks for experts": if something doesn't make sense, you probably don't have to worry about it for now.

A full glossary of notation used can be found in the appendix.

§6.i Natural numbers are positive

The set $\mathbb{N}$ is the set of *positive* integers, not including 0. In the set theory chapters, we use $\omega = \{0, 1, \dots\}$ instead, for consistency with the rest of the book.

§6.ii Sets and equivalence relations

This is brief, intended as a reminder for experts. Consult Appendix E for full details.

An **equivalence relation** on a set X is a relation $\sim$ which is symmetric, reflexive, and transitive. An equivalence relation partitions X into several **equivalence classes**. We will denote this by $X/\!\sim$. An element of such an equivalence class is a **representative** of that equivalence class.

I always use $\cong$ for an "isomorphism"-style relation (formally: a relation which is an isomorphism in a reasonable category). The only time $\simeq$ is used in the Napkin is for homotopic paths.

I generally use $\subseteq$ and $\subsetneq$ since these are non-ambiguous, unlike $\subset$. I only use $\subset$ on rare occasions in which equality obviously does not hold yet pointing it out would be distracting. For example, I write $\mathbb{Q} \subset \mathbb{R}$ since "$\mathbb{Q} \subsetneq \mathbb{R}$" is distracting.

I prefer $S \setminus T$ to $S - T$.

The power set of S (i.e., the set of subsets of S), is denoted either by 2^S or $\mathcal{P}(S)$.

§6.iii Functions

This is brief, intended as a reminder for experts. Consult Appendix E for full details.

Let $X \xrightarrow{f} Y$ be a function:

- By $f^{\text{pre}}(T)$ I mean the **pre-image**
$$f^{\text{pre}}(T) \coloneqq \{x \in X \mid f(x) \in T\}.$$
This is in contrast to the $f^{-1}(T)$ used in the rest of the world; I only use f^{-1} for an inverse *function*.

  By abuse of notation, we may abbreviate $f^{\text{pre}}(\{y\})$ to $f^{\text{pre}}(y)$. We call $f^{\text{pre}}(y)$ a **fiber**.

- By $f^{\text{img}}(S)$ I mean the **image**
$$f^{\text{img}}(S) \coloneqq \{f(x) \mid x \in S\}.$$
Almost everyone else in the world uses $f(S)$ (though $f[S]$ sees some use, and $f''(S)$ is often used in logic) but this is abuse of notation, and I prefer $f^{\text{img}}(S)$ for emphasis. This image notation is *not* standard.

- If $S \subseteq X$, then the **restriction** of f to S is denoted $f\restriction_S$, i.e. it is the function $f\restriction_S : S \to Y$.
- Sometimes functions $f : X \to Y$ are *injective* or *surjective*; I may emphasize this sometimes by writing $f : X \hookrightarrow Y$ or $f : X \twoheadrightarrow Y$, respectively.

§6.iv Cycle notation for permutations

Additionally, a permutation on a finite set may be denoted in *cycle notation*, as described in say https://en.wikipedia.org/wiki/Permutation#Cycle_notation. For example the notation $(1\ 2\ 3\ 4)(5\ 6\ 7)$ refers to the permutation with $1 \mapsto 2$, $2 \mapsto 3$, $3 \mapsto 4$, $4 \mapsto 1$, $5 \mapsto 6$, $6 \mapsto 7$, $7 \mapsto 5$. Usage of this notation will usually be obvious from context.

§6.v Rings

All rings have a multiplicative identity 1 unless otherwise specified. We allow $0 = 1$ in general rings but not in integral domains.

All rings are commutative unless otherwise specified. There is an elaborate scheme for naming rings which are not commutative, used only in the chapter on cohomology rings:

	Graded	Not Graded
1 not required	graded pseudo-ring	pseudo-ring
Anticommutative, 1 not required	anticommutative pseudo-ring	N/A
Has 1	graded ring	N/A
Anticommutative with 1	anticommutative ring	N/A
Commutative with 1	commutative graded ring	ring

On the other hand, an *algebra* always has 1, but it need not be commutative.

§6.vi Choice

We accept the Axiom of Choice, and use it freely.

§7 Further reading

The appendix Appendix A contains a list of resources I like, and explanations of pedagogical choices that I made for each chapter. I encourage you to check it out.

In particular, this is where you should go for further reading! There are some topics that should be covered in the Napkin, but are not, due to my own ignorance or laziness. The references provided in this appendix should hopefully help partially atone for my omissions.

Contents

I

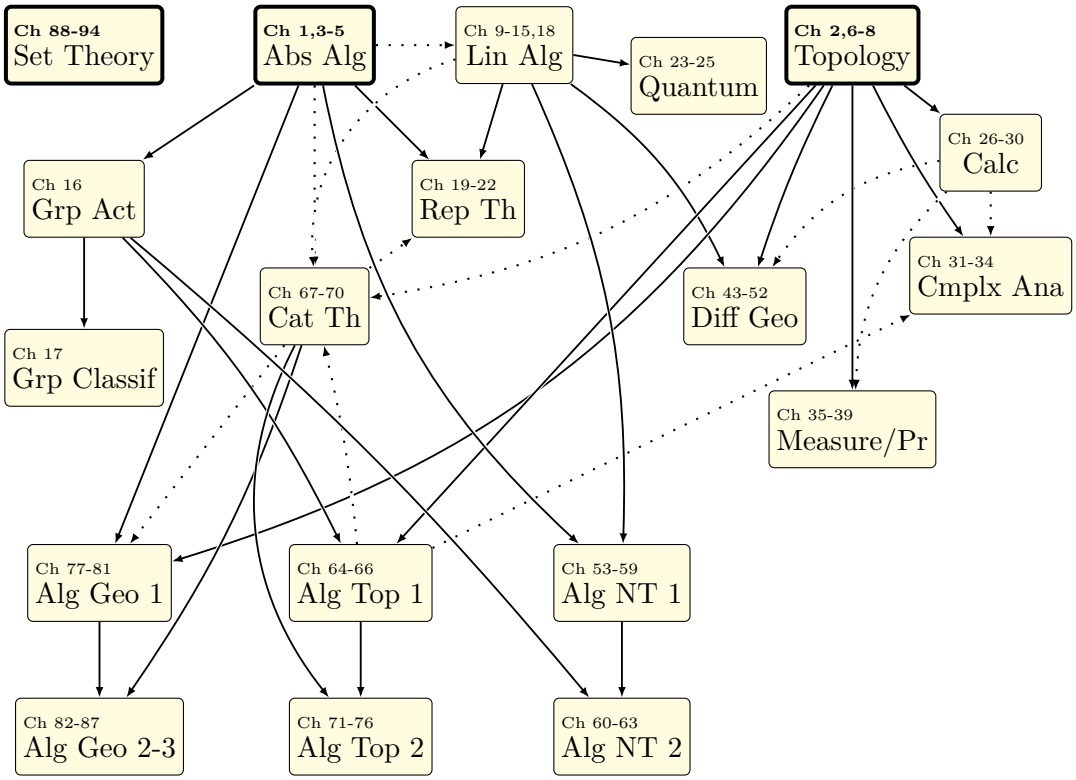
Starting Out

Part I: Contents

0 Sales pitches

This chapter contains a pitch for each part, to help you decide what you want to read and to elaborate more on how they are interconnected.

For convenience, here is again the dependency plot that appeared in the frontmatter.



§0.1 The basics

I. Starting Out.

I made a design decision that the first part should have a little bit of both algebra and topology: so this first chapter begins by defining a **group**, while the second chapter begins by defining a **metric space**. The intention is so that newcomers get to see two different examples of "sets with additional structure" in somewhat different contexts, and to have a minimal amount of literacy as these sorts of definitions appear over and over.[1]

II. Basic Abstract Algebra.

The algebraically inclined can then delve into further types of algebraic structures: some more details of **groups**, and then **rings** and **fields** — which will let you generalize $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. So you'll learn to become familiar with all sorts of other nouns that appear in algebra, unlocking a whole host of objects that one couldn't talk about before.

[1] In particular, I think it's easier to learn what a homeomorphism is after seeing group isomorphism, and what a homomorphism is after seeing continuous map.

We'll also come to **ideals**, which generalize the GCD in $\mathbb{Z}$ that you might know of. For example, you know in $\mathbb{Z}$ that any integer can be written in the form $3a+5b$ for $a, b \in \mathbb{Z}$, since $\gcd(3,5) = 1$. We'll see that this statement is really a statement of ideals: "$(3,5) = 1$ in $\mathbb{Z}$", and thus we'll understand in what situations it can be generalized, e.g. to polynomials.

III. Basic Topology.

The more analytically inclined can instead move into topology, learning more about spaces. We'll find out that "metric spaces" are actually too specific, and that it's better to work with **topological spaces**, which are based on the so-called **open sets**. You'll then get to see the buddings of some geometrical ideals, ending with the really great notion of **compactness**, a powerful notion that makes real analysis tick.

One example of an application of compactness to tempt you now: a continuous function $f : [0,1] \to \mathbb{R}$ always achieves a *maximum* value. (In contrast, $f : (0,1) \to \mathbb{R}$ by $x \mapsto 1/x$ does not.) We'll see the reason is that $[0,1]$ is compact.

§0.2 Abstract algebra

IV. Linear Algebra.

In high school, linear algebra is often really unsatisfying. You are given these arrays of numbers, and they're manipulated in some ways that don't really make sense. For example, the determinant is defined as this funny-looking sum with a bunch of products that seems to come out of thin air. Where does it come from? Why does $\det(AB) = \det A \det B$ with such a bizarre formula?

Well, it turns out that you *can* explain all of these things! The trick is to not think of linear algebra as the study of matrices, but instead as the study of *linear maps*. In earlier chapters we saw that we got great generalizations by speaking of "sets with enriched structure" and "maps between them". This time, our sets are **vector spaces** and our maps are **linear maps**. We'll find out that a matrix is actually just a way of writing down a linear map as an array of numbers, but using the "intrinsic" definitions we'll de-mystify all the strange formulas from high school and show you where they all come from.

In particular, we'll see *easy* proofs that column rank equals row rank, determinant is multiplicative, trace is the sum of the diagonal entries. We'll see how the dot product works, and learn all the words starting with "eigen-". We'll even have a bonus chapter for Fourier analysis showing that you can also explain all the big buzz-words by just being comfortable with vector spaces.

V. More on Groups.

Some of you might be interested in more about groups, and this chapter will give you a way to play further. It starts with an exploration of **group actions**, then goes into a bit on **Sylow theorems**, which are the tools that let us try to *classify all groups*.

VI. Representation Theory.

If G is a group, we can try to understand it by implementing it as a *matrix*, i.e. considering embeddings $G \hookrightarrow \mathrm{GL}_n(\mathbb{C})$. These are called **representations** of G; it turns out that they can be decomposed into **irreducible** ones. Astonishingly we

will find that we can *basically characterize all of them*: the results turn out to be short and completely unexpected.

For example, we will find out that there are finitely many irreducible representations of a given finite group G; if we label them V_1, V_2, ..., V_r, then we will find that r is the number of conjugacy classes of G, and moreover that

$$|G| = (\dim V_1)^2 + \cdots + (\dim V_r)^2$$

which comes out of nowhere!

The last chapter of this part will show you some unexpected corollaries. Here is one of them: let G be a finite group and create variables x_g for each $g \in G$. A $|G| \times |G|$ matrix M is defined by setting the (g,h)th entry to be the variable $x_{g \cdot h}$. Then this determinant will turn out to *factor*, and the factors will correspond to the V_i we described above: there will be an irreducible factor of degree $\dim V_i$ appearing $\dim V_i$ times. This result, called the **Frobenius determinant**, is said to have given birth to representation theory.

VII. Quantum Algorithms.

If you ever wondered what **Shor's algorithm** is, this chapter will use the built-up linear algebra to tell you!

§0.3 Real and complex analysis

VIII. Calculus 101.

In this part, we'll use our built-up knowledge of metric and topological spaces to give short, rigorous definitions and theorems typical of high school calculus. That is, we'll really define and prove most everything you've seen about **limits**, **series**, **derivatives**, and **integrals**.

Although this might seem intimidating, it turns out that actually, by the time we start this chapter, *the hard work has already been done*: the notion of limits, open sets, and compactness will make short work of what was swept under the rug in AP calculus. Most of the proofs will thus actually be quite short. We sit back and watch all the pieces slowly come together as a reward for our careful study of topology beforehand.

That said, if you are willing to suspend belief, you can actually read most of the other parts without knowing the exact details of all the calculus here, so in some sense this part is "optional".

IX. Complex Analysis.

It turns out that **holomorphic functions** (complex-differentiable functions) are close to the nicest things ever: they turn out to be given by a Taylor series (i.e. are basically polynomials). This means we'll be able to prove unreasonably nice results about holomorphic functions $\mathbb{C} \to \mathbb{C}$, like

- they are determined by just a few inputs,
- their contour integrals are all zero,
- they can't be bounded unless they are constant,
- ....

We then introduce **meromorphic functions**, which are like quotients of holomorphic functions, and find that we can detect their zeros by simply drawing loops in the plane and integrating over them: the famous **residue theorem** appears. (In the practice problems, you will see this even gives us a way to evaluate real integrals that can't be evaluated otherwise.)

X. Measure Theory.

Measure theory is the upgraded version of integration. The Riemann integration is for a lot of purposes not really sufficient; for example, if f is the function equals 1 at rational numbers but 0 at irrational numbers, we would hope that $\int_0^1 f(x)\,dx = 0$, but the Riemann integral is not capable of handling this function f.

The **Lebesgue integral** will handle these mistakes by assigning a *measure* to a generic space Ω, making it into a **measure space**. This will let us develop a richer theory of integration where the above integral *does* work out to zero because the "rational numbers have measure zero". Even the development of the measure will be an achievement, because it means we've developed a rigorous, complete way of talking about what notions like area and volume mean — on any space, not just $\mathbb{R}^n$! So for example the Lebesgue integral will let us integrate functions over any **measure space**.

XI. Probability (TO DO).

Using the tools of measure theory, we'll be able to start giving rigorous definitions of **probability**, too. We'll see that a **random variable** is actually a function from a measure space of worlds to $\mathbb{R}$, giving us a rigorous way to talk about its probabilities. We can then start actually stating results like the **law of large numbers** and **central limit theorem** in ways that make them both easy to state and straightforward to prove.

XII. Differential Geometry.

Multivariable calculus is often confusing because of all the partial derivatives. But we'll find out that, armed with our good understanding of linear algebra, that we're really looking at a **total derivative**: at every point of a function $f\colon \mathbb{R}^n \to \mathbb{R}$ we can associate a *linear map* Df which captures in one object the notion of partial derivatives. Set up this way, we'll get to see versions of **differential forms** and **Stokes' theorem**, and we finally will know what the notation dx really means. In the end, we'll say a little bit about manifolds in general.

§0.4 Algebraic number theory

XIV. Algebraic NT I: Rings of Integers.

Why is $3+\sqrt{5}$ the conjugate of $3-\sqrt{5}$? How come the norm $\left\|a+b\sqrt{5}\right\| = a^2-5b^2$ used in Pell's equations just happens to be multiplicative? Why is it we can do factoring into primes in $\mathbb{Z}[i]$ but not in $\mathbb{Z}[\sqrt{-5}]$? All these questions and more will be answered in this part, when we learn about **number fields**, a generalization of $\mathbb{Q}$ and $\mathbb{Z}$ to things like $\mathbb{Q}(\sqrt{5})$ and $\mathbb{Z}[\sqrt{5}]$. We'll find out that we have unique factorization into prime ideals, that there is a real *multiplicative norm* in play here, and so on. We'll also see that Pell's equation falls out of this theory.

XV. Algebraic NT II: Galois and Ramification Theory.

All the big buzz-words come out now: **Galois groups**, the **Frobenius**, and friends. We'll see quadratic reciprocity is just a shadow of the behavior of the Frobenius element, and meet the **Chebotarev density theorem**, which generalizes greatly the Dirichlet theorem on the infinitude of primes which are $a \pmod n$. Towards the end, we'll also state **Artin reciprocity**, one of the great results of **class field theory**, and how it generalizes quadratic reciprocity and cubic reciprocity.

§0.5 Algebraic topology

XVI. **Algebraic Topology I: Homotopy.**

What's the difference between an annulus and disk? Well, one of them has a "hole" in it, but if we are just given intrinsic topological spaces it's hard to make this notion precise. The **fundamental group** $\pi_1(X)$ and more general **homotopy group** will make this precise — we'll find a way to define an abelian group $\pi_1(X)$ for every topological space X which captures the idea there is a hole in the space, by throwing lassos into the space and seeing if we can reel them in.

Amazingly, the fundamental group $\pi_1(X)$ will, under mild conditions, tell you about ways to cover X with a so-called **covering projection**. One picture is that one can wrap a real line $\mathbb{R}$ into a helix shape and then project it down into the circle S^1. This will turn out to correspond to the fact that $\pi_1(S^1) = \mathbb{Z}$ which has only one subgroup. More generally the subgroups of $\pi_1(X)$ will be in bijection with ways to cover the space X!

XVII. **Category Theory.**

What do fields, groups, manifolds, metric spaces, measure spaces, modules, representations, rings, topological spaces, vector spaces, all have in common? Answer: they are all "objects with additional structure", with maps between them.

The notion of **category** will appropriately generalize all of them. We'll see that all sorts of constructions and ideas can be abstracted into the framework of a category, in which we *only* think about objects and arrows between them, without probing too hard into the details of what those objects are. This results in drawing many **commutative diagrams**.

For example, any way of taking an object in one category and getting another one (for example π_1 as above, from the category of spaces into the category of groups) will probably be a **functor**. We'll unify $G \times H$, $X \times Y$, $R \times S$, and anything with the $\times$ symbol into the notion of a product, and then even more generally into a **limit**. Towards the end, we talk about **abelian categories** and talk about the famous **snake lemma**, **five lemma**, and so on.

XVIII. **Algebraic Topology II: Homology.**

Using the language of category theory, we then resume our adventures in algebraic topology, in which we define the **homology groups** which give a different way of noticing holes in a space, in a way that is longer to define but easier to compute in practice. We'll then reverse the construction to get so-called **cohomology rings** instead, which give us an even finer invariant for telling spaces apart.

§0.6 Algebraic geometry

XIX. **Algebraic Geometry I: Classical Varieties.**

We begin with a classical study of classical **complex varieties**: the study of intersections of polynomial equations over $\mathbb{C}$. This will naturally lead us into the geometry of rings, giving ways to draw pictures of ideals, and motivating **Hilbert's nullstellensatz**. The **Zariski topology** will show its face, and then we'll play with **projective varieties** and **quasi-projective varieties**, with a bonus detour into **Bézout's theorem**. All this prepares us for our journey into schemes.

XX. Algebraic Geometry II: Affine Schemes.

We now get serious and delve into Grothendieck's definition of an **affine scheme**: a generalization of our classical varieties that allows us to start with any ring A and construct a space $\operatorname{Spec} A$ on it. We'll equip it with its own Zariski topology and then a sheaf of functions on it, making it into a **locally ringed space**; we will find that the sheaf can be understood effectively in terms of **localization** on it. We'll find that the language of commutative algebra provides elegant generalizations of what's going on geometrically: prime ideals correspond to irreducible closed subsets, radical ideals correspond to closed subsets, maximal ideals correspond to closed points, and so on. We'll draw lots of pictures of spaces and examples to accompany this.

§0.7 Set theory

XXI. Set Theory I: ZFC, Ordinals, and Cardinals.

Why is **Russell's paradox** such a big deal and how is it resolved? What is this **Zorn's lemma** that everyone keeps talking about? In this part we'll learn the answers to these questions by giving a real description of the **Zermelo-Frankel** axioms, and the **axiom of choice**, delving into the details of how math is built axiomatically at the very bottom foundations. We'll meet the **ordinal numbers** and **cardinal numbers** and learn how to do **transfinite induction** with them.

XXII. Set Theory II: Model Theory and Forcing.

The **continuum hypothesis** states that there are no cardinalities between the size of the natural numbers and the size of the real numbers. It was shown to be *independent* of the axioms — one cannot prove or disprove it. How could a result like that possibly be proved? Using our understanding of the ZF axioms, we'll develop a bit of **model theory** and then use **forcing** in order to show how to construct entire models of the universe in which the continuum hypothesis is true or false.

1 Groups

A group is one of the most basic structures in higher mathematics. In this chapter I will tell you only the bare minimum: what a group is, and when two groups are the same.

§1.1 Definition and examples of groups

Prototypical example for this section: The additive group of integers $(\mathbb{Z}, +)$ and the cyclic group $\mathbb{Z}/m\mathbb{Z}$. Just don't let yourself forget that most groups are non-commutative.

A group consists of two pieces of data: a set G, and an associative binary operation $\star$ with some properties. Before I write down the definition of a group, let me give two examples.

Example 1.1.1 (Additive integers)
The pair $(\mathbb{Z}, +)$ is a group: $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ is the set and the associative operation is *addition*. Note that

- The element $0 \in \mathbb{Z}$ is an *identity*: $a + 0 = 0 + a = a$ for any a.
- Every element $a \in \mathbb{Z}$ has an additive *inverse*: $a + (-a) = (-a) + a = 0$.

We call this group $\mathbb{Z}$.

Example 1.1.2 (Nonzero rationals)
Let $\mathbb{Q}^\times$ be the set of *nonzero rational numbers*. The pair $(\mathbb{Q}^\times, \cdot)$ is a group: the set is $\mathbb{Q}^\times$ and the associative operation is *multiplication*.
Again we see the same two nice properties.

- The element $1 \in \mathbb{Q}^\times$ is an *identity*: for any rational number, $a \cdot 1 = 1 \cdot a = a$.
- For any rational number $x \in \mathbb{Q}^\times$, we have an inverse x^{-1}, such that
$$x \cdot x^{-1} = x^{-1} \cdot x = 1.$$

From this you might already have a guess what the definition of a group is.

Definition 1.1.3. A **group** is a pair $G = (G, \star)$ consisting of a set of elements G, and a binary operation $\star$ on G, such that:

- G has an **identity element**, usually denoted 1_G or just 1, with the property that
$$1_G \star g = g \star 1_G = g \text{ for all } g \in G.$$
- The operation is **associative**, meaning $(a \star b) \star c = a \star (b \star c)$ for any $a, b, c \in G$. Consequently we generally don't write the parentheses.
- Each element $g \in G$ has an **inverse**, that is, an element $h \in G$ such that
$$g \star h = h \star g = 1_G.$$

Remark 1.1.4 (Unimportant pedantic point) — Some authors like to add a "closure" axiom, i.e. to say explicitly that $g \star h \in G$. This is implied already by the fact that $\star$ is a binary operation on G, but is worth keeping in mind for the examples below.

Remark 1.1.5 — It is not required that $\star$ is commutative ($a \star b = b \star a$). So we say that a group is **abelian** if the operation is commutative and **non-abelian** otherwise.

Example 1.1.6 (Non-Examples of groups)

- The pair $(\mathbb{Q}, \cdot)$ is NOT a group. (Here $\mathbb{Q}$ is rational numbers.) While there is an identity element, the element $0 \in \mathbb{Q}$ does not have an inverse.
- The pair $(\mathbb{Z}, \cdot)$ is also NOT a group. (Why?)
- Let $\text{Mat}_{2\times 2}(\mathbb{R})$ be the set of 2×2 real matrices. Then $(\text{Mat}_{2\times 2}(\mathbb{R}), \cdot)$ (where $\cdot$ is matrix multiplication) is NOT a group. Indeed, even though we have an identity matrix
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
we still run into the same issue as before: the zero matrix does not have a multiplicative inverse.

  (Even if we delete the zero matrix from the set, the resulting structure is still not a group: those of you that know some linear algebra might recall that any matrix with determinant zero cannot have an inverse.)

Let's resume writing down examples. Here are some more **abelian examples** of groups:

Example 1.1.7 (Complex unit circle)

Let S^1 denote the set of complex numbers z with absolute value one; that is
$$S^1 \coloneqq \{z \in \mathbb{C} \mid |z| = 1\}.$$
Then $(S^1, \times)$ is a group because

- The complex number $1 \in S^1$ serves as the identity, and
- Each complex number $z \in S^1$ has an inverse $\frac{1}{z}$ which is also in S^1, since $|z^{-1}| = |z|^{-1} = 1$.

There is one thing I ought to also check: that $z_1 \times z_2$ is actually still in S^1. But this follows from the fact that $|z_1 z_2| = |z_1|\,|z_2| = 1$.

Example 1.1.8 (Addition mod n)

Here is an example from number theory: Let $n > 1$ be an integer, and consider the

residues (remainders) modulo n. These form a group under addition. We call this the **cyclic group of order** n, and denote it as $\mathbb{Z}/n\mathbb{Z}$, with elements $\overline{0}, \overline{1}, \ldots$. The identity is $\overline{0}$.

Example 1.1.9 (Multiplication mod p)

Let p be a prime. Consider the *nonzero residues modulo* p, which we denote by $(\mathbb{Z}/p\mathbb{Z})^\times$. Then $((\mathbb{Z}/p\mathbb{Z})^\times, \times)$ is a group.

Question 1.1.10. Why do we need the fact that p is prime?

(Digression: the notation $\mathbb{Z}/n\mathbb{Z}$ and $(\mathbb{Z}/p\mathbb{Z})^\times$ may seem strange but will make sense when we talk about rings and ideals. Set aside your worry for now.)

Here are some **non-abelian examples**:

Example 1.1.11 (General linear group)

Let n be a positive integer. Then $\mathrm{GL}_n(\mathbb{R})$ is defined as the set of $n \times n$ real matrices which have nonzero determinant. It turns out that with this condition, every matrix does indeed have an inverse, so $(\mathrm{GL}_n(\mathbb{R}), \times)$ is a group, called the **general linear group**.
(The fact that $\mathrm{GL}_n(\mathbb{R})$ is closed under $\times$ follows from the linear algebra fact that $\det(AB) = \det A \det B$, proved in later chapters.)

Example 1.1.12 (Special linear group)

Following the example above, let $\mathrm{SL}_n(\mathbb{R})$ denote the set of $n \times n$ matrices whose determinant is actually 1. Again, for linear algebra reasons it turns out that $(\mathrm{SL}_n(\mathbb{R}), \times)$ is also a group, called the **special linear group**.

Example 1.1.13 (Symmetric groups)

Let S_n be the set of permutations of $\{1, \ldots, n\}$. By viewing these permutations as functions from $\{1, \ldots, n\}$ to itself, we can consider *compositions* of permutations. Then the pair $(S_n, \circ)$ (here $\circ$ is function composition) is also a group, because

- There is an identity permutation, and
- Each permutation has an inverse.

The group S_n is called the **symmetric group** on n elements.

Example 1.1.14 (Dihedral group)

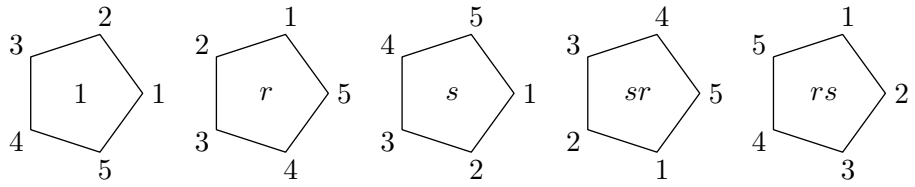
The **dihedral group of order** $2n$, denoted D_{2n}, is the group of symmetries of a regular n-gon $A_1A_2 \ldots A_n$, which includes rotations and reflections. It consists of

the $2n$ elements

$$\left\{1, r, r^2, \dots, r^{n-1}, s, sr, sr^2, \dots, sr^{n-1}\right\}.$$

The element r corresponds to rotating the n-gon by $\frac{2\pi}{n}$, while s corresponds to reflecting it across the line OA_1 (here O is the center of the polygon). So rs means "reflect then rotate" (like with function composition, we read from right to left). In particular, $r^n = s^2 = 1$. You can also see that $r^k s = sr^{-k}$.

Here is a picture of some elements of D_{10}.



Trivia: the dihedral group D_{12} is my favorite example of a non-abelian group, and is the first group I try for any exam question of the form "find an example…".

More examples:

Example 1.1.15 (Products of groups)

Let $(G, \star)$ and $(H, *)$ be groups. We can define a **product group** $(G \times H, \cdot)$, as follows. The elements of the group will be ordered pairs $(g, h) \in G \times H$. Then

$$(g_1, h_1) \cdot (g_2, h_2) = (g_1 \star g_2, h_1 * h_2) \in G \times H$$

is the group operation.

Question 1.1.16. What are the identity and inverses of the product group?

Example 1.1.17 (Trivial group)

The **trivial group**, often denoted 0 or 1, is the group with only an identity element. I will use the notation $\{1\}$.

Exercise 1.1.18. Which of these are groups?

(a) Rational numbers with odd denominators (in simplest form), where the operation is addition. (This includes integers, written as $n/1$, and $0 = 0/1$).

(b) The set of rational numbers with denominator at most 2, where the operation is addition.

(c) The set of rational numbers with denominator at most 2, where the operation is multiplication.

(d) The set of nonnegative integers, where the operation is addition.

§1.2 Properties of groups

Prototypical example for this section: $(\mathbb{Z}/p\mathbb{Z})^\times$ *is possibly best.*

Abuse of Notation 1.2.1. From now on, we'll often refer to a group $(G, \star)$ by just G. Moreover, we'll abbreviate $a \star b$ to just ab. Also, because the operation $\star$ is associative, we will omit unnecessary parentheses: $(ab)c = a(bc) = abc$.

Abuse of Notation 1.2.2. From now on, for any $g \in G$ and $n \in \mathbb{N}$ we abbreviate

$$g^n = \underbrace{g \star \cdots \star g}_{n \text{ times}}.$$

Moreover, we let g^{-1} denote the inverse of g, and $g^{-n} = (g^{-1})^n$.

In mathematics, a common theme is to require that objects satisfy certain minimalistic properties, with certain examples in mind, but then ignore the examples on paper and try to deduce as much as you can just from the properties alone. (Math olympiad veterans are likely familiar with "functional equations" in which knowing a single property about a function is enough to determine the entire function.) Let's try to do this here, and see what we can conclude just from knowing Definition 1.1.3.

It is a law in Guam and 37 other states that I now state the following proposition.

Fact 1.2.3. Let G be a group.

(a) The identity of a group is unique.

(b) The inverse of any element is unique.

(c) For any $g \in G$, $(g^{-1})^{-1} = g$.

Proof. This is mostly just some formal manipulations, and you needn't feel bad skipping it on a first read.

(a) If 1 and $1'$ are identities, then $1 = 1 \star 1' = 1'$.

(b) If h and h' are inverses to g, then $1_G = g \star h \implies h' = (h' \star g) \star h = 1_G \star h = h$.

(c) Trivial; omitted. □

Now we state a slightly more useful proposition.

Proposition 1.2.4 (Inverse of products)
Let G be a group, and $a, b \in G$. Then $(ab)^{-1} = b^{-1}a^{-1}$.

Proof. Direct computation. We have

$$(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = aa^{-1} = 1_G.$$

Similarly, $(b^{-1}a^{-1})(ab) = 1_G$ as well. Hence $(ab)^{-1} = b^{-1}a^{-1}$. □

Finally, we state a very important lemma about groups, which highlights why having an inverse is so valuable.

Lemma 1.2.5 (Left multiplication is a bijection)
Let G be a group, and pick a $g \in G$. Then the map $G \to G$ given by $x \mapsto gx$ is a bijection.

Exercise 1.2.6. Check this by showing injectivity and surjectivity directly. (If you don't know what these words mean, consult Appendix E.)

Example 1.2.7

Let $G = (\mathbb{Z}/7\mathbb{Z})^\times$ (as in Example 1.1.9) and pick $g = 3$. The above lemma states that the map $x \mapsto 3 \cdot x$ is a bijection, and we can see this explicitly:

$$\begin{aligned}
1 &\xmapsto{\times 3} 3 \pmod 7 \\
2 &\xmapsto{\times 3} 6 \pmod 7 \\
3 &\xmapsto{\times 3} 2 \pmod 7 \\
4 &\xmapsto{\times 3} 5 \pmod 7 \\
5 &\xmapsto{\times 3} 1 \pmod 7 \\
6 &\xmapsto{\times 3} 4 \pmod 7.
\end{aligned}$$

The fact that the map is injective is often called the **cancellation law**. (Why do you think so?)

Abuse of Notation 1.2.8 (Later on, sometimes the identity is denoted 0 instead of 1)**.** You don't need to worry about this for a few chapters, but I'll bring it up now anyways. In most of our examples up until now the operation $\star$ was thought of like multiplication of some sort, which is why $1 = 1_G$ was a natural notation for the identity element.

But there are groups like $\mathbb{Z} = (\mathbb{Z}, +)$ where the operation $\star$ is thought of as addition, in which case the notation $0 = 0_G$ might make more sense instead. (In general, whenever an operation is denoted $+$, the operation is almost certainly commutative.) We will eventually start doing so too when we discuss rings and linear algebra.

§1.3 Isomorphisms

Prototypical example for this section: $\mathbb{Z} \cong 10\mathbb{Z}$.

First, let me talk about what it means for groups to be isomorphic. Consider the two groups

- $\mathbb{Z} = (\{\dots, -2, -1, 0, 1, 2, \dots\}, +)$.
- $10\mathbb{Z} = (\{\dots, -20, -10, 0, 10, 20, \dots\}, +)$.

These groups are "different", but only superficially so – you might even say they only differ in the names of the elements. Think about what this might mean formally for a moment.

Specifically the map

$$\phi \colon \mathbb{Z} \to 10\mathbb{Z} \text{ by } x \mapsto 10x$$

is a bijection of the underlying sets which respects the group operation. In symbols,

$$\phi(x + y) = \phi(x) + \phi(y).$$

In other words, ϕ is a way of re-assigning names of the elements without changing the structure of the group. That's all just formalism for capturing the obvious fact that $(\mathbb{Z}, +)$ and $(10\mathbb{Z}, +)$ are the same thing.

Now, let's do the general definition.

Definition 1.3.1. Let $G = (G, \star)$ and $H = (H, *)$ be groups. A bijection $\phi\colon G \to H$ is called an **isomorphism** if

$$\phi(g_1 \star g_2) = \phi(g_1) * \phi(g_2) \quad \text{for all } g_1, g_2 \in G.$$

If there exists an isomorphism from G to H, then we say G and H are **isomorphic** and write $G \cong H$.

Note that in this definition, the left-hand side $\phi(g_1 \star g_2)$ uses the operation of G while the right-hand side $\phi(g_1) * \phi(g_2)$ uses the operation of H.

Example 1.3.2 (Examples of isomorphisms)
Let G and H be groups. We have the following isomorphisms.

(a) $\mathbb{Z} \cong 10\mathbb{Z}$, as above.

(b) There is an isomorphism
$$G \times H \cong H \times G$$
by the map $(g, h) \mapsto (h, g)$.

(c) The identity map $\mathrm{id}\colon G \to G$ is an isomorphism, hence $G \cong G$.

(d) There is another isomorphism of $\mathbb{Z}$ to itself: send every x to $-x$.

Example 1.3.3 (Primitive roots modulo 7)
As a nontrivial example, we claim that $\mathbb{Z}/6\mathbb{Z} \cong (\mathbb{Z}/7\mathbb{Z})^\times$. The bijection is

$$\phi(a \bmod 6) = 3^a \bmod 7.$$

- This map is a bijection by explicit calculation:
$$(3^0, 3^1, 3^2, 3^3, 3^4, 3^5) \equiv (1, 3, 2, 6, 4, 5) \pmod 7.$$
(Technically, I should more properly write $3^{0 \bmod 6} = 1$ and so on to be pedantic.)
- Finally, we need to verify that this map respects the group operation. In other words, we want to see that $\phi(a + b) = \phi(a)\phi(b)$ since the operation of $\mathbb{Z}/6\mathbb{Z}$ is addition while the operation of $(\mathbb{Z}/7\mathbb{Z})^\times$ is multiplication. That's just saying that $3^{a+b \bmod 6} \equiv 3^{a \bmod 6} 3^{b \bmod 6} \pmod 7$, which is true.

Example 1.3.4 (Primitive roots)
More generally, for any prime p, there exists an element $g \in (\mathbb{Z}/p\mathbb{Z})^\times$ called a **primitive root** modulo p such that $1, g, g^2, \ldots, g^{p-2}$ are all different modulo p. One can show by copying the above proof that
$$\mathbb{Z}/(p-1)\mathbb{Z} \cong (\mathbb{Z}/p\mathbb{Z})^\times \text{ for all primes } p.$$
The example above was the special case $p = 7$ and $g = 3$.

Exercise 1.3.5. Assuming the existence of primitive roots, establish the isomorphism $\mathbb{Z}/(p-1)\mathbb{Z} \cong (\mathbb{Z}/p\mathbb{Z})^\times$ as above.

It's not hard to see that $\cong$ is an equivalence relation (why?). Moreover, because we really only care about the structure of groups, we'll usually consider two groups to be the same when they are isomorphic. So phrases such as "find all groups" really mean "find all groups up to isomorphism".

§1.4 Orders of groups, and Lagrange's theorem

Prototypical example for this section: $(\mathbb{Z}/p\mathbb{Z})^\times$.

As is typical in math, we use the word "order" for way too many things. In groups, there are two notions of order.

Definition 1.4.1. The **order of a group** G is the number of elements of G. We denote this by $|G|$. Note that the order may not be finite, as in $\mathbb{Z}$. We say G is a **finite group** just to mean that $|G|$ is finite.

Example 1.4.2 (Orders of groups)
For a prime p, $|(\mathbb{Z}/p\mathbb{Z})^\times| = p-1$. In other words, the order of $(\mathbb{Z}/p\mathbb{Z})^\times$ is $p-1$. As another example, the order of the symmetric group S_n is $n!$ and the order of the dihedral group D_{2n} is $2n$.

Definition 1.4.3. The **order of an element** $g \in G$ is the smallest positive integer n such that $g^n = 1_G$, or ∞ if no such n exists. We denote this by $\operatorname{ord} g$.

Example 1.4.4 (Examples of orders)
The order of -1 in $\mathbb{Q}^\times$ is 2, while the order of 1 in $\mathbb{Z}$ is infinite.

Question 1.4.5. Find the order of each of the six elements of $\mathbb{Z}/6\mathbb{Z}$, the cyclic group on six elements. (See Example 1.1.8 if you've forgotten what $\mathbb{Z}/6\mathbb{Z}$ means.)

Example 1.4.6 (Primitive roots)
If you know olympiad number theory, this coincides with the definition of an order of a residue mod p. That's why we use the term "order" there as well. In particular, a primitive root is precisely an element $g \in (\mathbb{Z}/p\mathbb{Z})^\times$ such that $\operatorname{ord} g = p-1$.

You might also know that if $x^n \equiv 1 \pmod{p}$, then the order of $x \pmod{p}$ must divide n. The same is true in a general group for exactly the same reason.

Fact 1.4.7. If $g^n = 1_G$ then $\operatorname{ord} g$ divides n.

Also, you can show that any element of a finite group has a finite order. The proof is just an olympiad-style pigeonhole argument. Consider the infinite sequence $1_G, g, g^2, \dots$, and find two elements that are the same.

Fact 1.4.8. Let G be a finite group. For any $g \in G$, $\operatorname{ord} g$ is finite.

What's the last property of $(\mathbb{Z}/p\mathbb{Z})^\times$ that you know from olympiad math? We have Fermat's little theorem: for any $a \in (\mathbb{Z}/p\mathbb{Z})^\times$, we have $a^{p-1} \equiv 1 \pmod{p}$. This is no coincidence: exactly the same thing is true in a more general setting.

Theorem 1.4.9 (Lagrange's theorem for orders)
Let G be any finite group. Then $x^{|G|} = 1_G$ for any $x \in G$.

Keep this result in mind! We'll prove it later in the generality of Theorem 3.4.1.

§1.5 Subgroups

Prototypical example for this section: $\mathrm{SL}_n(\mathbb{R})$ *is a subgroup of* $\mathrm{GL}_n(\mathbb{R})$.

Earlier we saw that $\mathrm{GL}_n(\mathbb{R})$, the $n \times n$ matrices with nonzero determinant, formed a group under matrix multiplication. But we also saw that a subset of $\mathrm{GL}_n(\mathbb{R})$, namely $\mathrm{SL}_n(\mathbb{R})$, also formed a group with the same operation. For that reason we say that $\mathrm{SL}_n(\mathbb{R})$ is a subgroup of $\mathrm{GL}_n(\mathbb{R})$. And this definition generalizes in exactly the way you expect.

Definition 1.5.1. Let $G = (G, \star)$ be a group. A **subgroup** of G is exactly what you would expect it to be: a group $H = (H, \star)$ where H is a subset of G. It's a **proper subgroup** if $H \neq G$.

Remark 1.5.2 — To specify a group G, I needed to tell you both what the set G was and the operation $\star$ was. But to specify a subgroup H of a given group G, I only need to tell you who its elements are: the operation of H is just inherited from the operation of G.

Example 1.5.3 (Examples of subgroups)

(a) $2\mathbb{Z}$ is a subgroup of $\mathbb{Z}$, which is isomorphic to $\mathbb{Z}$ itself!

(b) Consider again S_n, the symmetric group on n elements. Let T be the set of permutations $\tau \colon \{1, \dots, n\} \to \{1, \dots, n\}$ for which $\tau(n) = n$. Then T is a subgroup of S_n; in fact, it is isomorphic to S_{n-1}.

(c) Consider the group $G \times H$ (Example 1.1.15) and the elements $\{(g, 1_H) \mid g \in G\}$. This is a subgroup of $G \times H$ (why?). In fact, it is isomorphic to G by the isomorphism $(g, 1_H) \mapsto g$.

Example 1.5.4 (Stupid examples of subgroups)
For any group G, the trivial group $\{1_G\}$ and the entire group G are subgroups of G.

Next is an especially important example that we'll talk about more in later chapters.

Example 1.5.5 (Subgroup generated by an element)
Let x be an element of a group G. Consider the set

$$\langle x \rangle = \left\{ \dots, x^{-2}, x^{-1}, 1, x, x^2, \dots \right\}.$$

This is also a subgroup of G, called the subgroup generated by x.

Exercise 1.5.6. If $\operatorname{ord} x = 2015$, what is the above subgroup equal to? What if $\operatorname{ord} x = \infty$?

Finally, we present some non-examples of subgroups.

Example 1.5.7 (Non-examples of subgroups)
Consider the group $\mathbb{Z} = (\mathbb{Z}, +)$.

(a) The set $\{0, 1, 2, \dots\}$ is not a subgroup of $\mathbb{Z}$ because it does not contain inverses.

(b) The set $\{n^3 \mid n \in \mathbb{Z}\} = \{\dots, -8, -1, 0, 1, 8, \dots\}$ is not a subgroup because it is not closed under addition; the sum of two cubes is not in general a cube.

(c) The empty set $\varnothing$ is not a subgroup of $\mathbb{Z}$ because it lacks an identity element.

§1.6 Groups of small orders

Just for fun, here is a list of all groups of order less than or equal to ten (up to isomorphism, of course).

1. The only group of order 1 is the trivial group.
2. The only group of order 2 is $\mathbb{Z}/2\mathbb{Z}$.
3. The only group of order 3 is $\mathbb{Z}/3\mathbb{Z}$.
4. The only groups of order 4 are
   - $\mathbb{Z}/4\mathbb{Z}$, the cyclic group on four elements,
   - $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, called the Klein Four Group.
5. The only group of order 5 is $\mathbb{Z}/5\mathbb{Z}$.
6. The groups of order six are
   - $\mathbb{Z}/6\mathbb{Z}$, the cyclic group on six elements.
   - S_3, the permutation group of three elements. This is the first non-abelian group.

   Some of you might wonder where $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ is. All I have to say is: Chinese remainder theorem!

   You might wonder where D_6 is in this list. It's actually isomorphic to S_3.
7. The only group of order 7 is $\mathbb{Z}/7\mathbb{Z}$.
8. The groups of order eight are more numerous.

- $\mathbb{Z}/8\mathbb{Z}$, the cyclic group on eight elements.
- $\mathbb{Z}/4\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.
- $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.
- D_8, the dihedral group with eight elements, which is not abelian.
- A non-abelian group Q_8, called the *quaternion group*. It consists of eight elements ± 1, $\pm i$, $\pm j$, $\pm k$ with $i^2 = j^2 = k^2 = ijk = -1$.

9. The groups of order nine are
   - $\mathbb{Z}/9\mathbb{Z}$, the cyclic group on nine elements.
   - $\mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$.
10. The groups of order 10 are
    - $\mathbb{Z}/10\mathbb{Z} \cong \mathbb{Z}/5\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ (again Chinese remainder theorem).
    - D_{10}, the dihedral group with 10 elements. This group is non-abelian.

§1.7 Unimportant long digression

A common question is: why these axioms? For example, why associative but not commutative? This answer will likely not make sense until later, but here are some comments that may help.

One general heuristic is: Whenever you define a new type of general object, there's always a balancing act going on. On the one hand, you want to include enough constraints that your objects are "nice". On the other hand, if you include too many constraints, then your definition applies to too few objects.

So, for example, we include "associative" because that makes our lives easier and most operations we run into are associative. In particular, associativity is required for the inverse of an element to necessarily be unique. However we don't include "commutative", because examples below show that there are lots of non-abelian groups we care about. (But we introduce another name "abelian" because we still want to keep track of it.)

Another comment: a good motivation for the inverse axioms is that you get a large amount of *symmetry*. The set of positive integers with addition is not a group, for example, because you can't subtract 6 from 3: some elements are "larger" than others. By requiring an inverse element to exist, you get rid of this issue. (You also need identity for this; it's hard to define inverses without it.)

Even more abstruse comment: Problem 1F† shows that groups are actually shadows of symmetric groups. This makes rigorous the notion that "groups are very symmetric".

§1.8 A few harder problems to think about

Problem 1A. What is the joke in the following figure? (Source: [Ge].)

Problem 1B. Prove Lagrange's theorem for orders in the special case that G is a finite abelian group.

Problem 1C. Show that $D_6 \cong S_3$ but $D_{24} \ncong S_4$.

Problem 1D*. Let p be a prime. Show that if G is a group of order p then $G \cong \mathbb{Z}/p\mathbb{Z}$.

Problem 1E (A hint for Cayley's theorem)**.** Find a subgroup H of S_8 which is isomorphic to D_8, and write the isomorphism explicitly.

Problem 1F†. Let G be a finite group.[1] Show that there exists a positive integer n such that

(a) (Cayley's theorem) G is isomorphic to some subgroup of the symmetric group S_n.

(b) (Representation Theory) G is isomorphic to some subgroup of the general linear group $\mathrm{GL}_n(\mathbb{R})$. (This is the group of invertible $n \times n$ matrices.)

Problem 1G. Find the smallest integer n such that the symmetric group S_n has a subgroup isomorphic to the dihedral group D_{2018} of order 2018.

Problem 1H (IMO SL 2005 C5)**.** There are n markers, each with one side white and the other side black. In the beginning, these n markers are aligned in a row so that their white sides are all up. In each step, if possible, we choose a marker whose white side is up (but not one of the outermost markers), remove it, and reverse the closest marker to the left of it and also reverse the closest marker to the right of it.

Prove that if $n \equiv 1 \pmod 3$ it's impossible to reach a state with only two markers remaining. (In fact the converse is true as well.)

Problem 1I. Let p be a prime and $F_1 = F_2 = 1$, $F_{n+2} = F_{n+1} + F_n$ be the Fibonacci sequence. Show that $F_{2p(p^2-1)}$ is divisible by p.

[1] In other words, permutation groups can be arbitrarily weird. I remember being highly unsettled by this theorem when I first heard of it, but in hindsight it is not so surprising.

2 Metric spaces

At the time of writing, I'm convinced that metric topology is the morally correct way to motivate point-set topology as well as to generalize normal calculus.[1] So here is my best attempt.

The concept of a metric space is very "concrete", and lends itself easily to visualization. Hence throughout this chapter you should draw lots of pictures as you learn about new objects, like convergent sequences, open sets, closed sets, and so on.

§2.1 Definition and examples of metric spaces

Prototypical example for this section: $\mathbb{R}^2$*, with the Euclidean metric.*

Definition 2.1.1. A **metric space** is a pair (M, d) consisting of a set of points M and a **metric** $d \colon M \times M \to \mathbb{R}_{\geq 0}$. The distance function must obey:

- For any $x, y \in M$, we have $d(x,y) = d(y,x)$; i.e. d is symmetric.
- The function d must be **positive definite** which means that $d(x,y) \geq 0$ with equality if and only if $x = y$.
- The function d should satisfy the **triangle inequality**: for all $x, y, z \in M$,

$$d(x,z) + d(z,y) \geq d(x,y).$$

Abuse of Notation 2.1.2. Just like with groups, we will abbreviate (M, d) as just M.

Example 2.1.3 (Metric spaces of $\mathbb{R}$)

(a) The real line $\mathbb{R}$ is a metric space under the metric $d(x,y) = |x-y|$.

(b) The interval $[0,1]$ is also a metric space with the same distance function.

(c) In fact, any subset S of $\mathbb{R}$ can be made into a metric space in this way.

Example 2.1.4 (Metric spaces of $\mathbb{R}^2$)

(a) We can make $\mathbb{R}^2$ into a metric space by imposing the Euclidean distance function

$$d\left((x_1,y_1),(x_2,y_2)\right) = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2}.$$

(b) Just like with the first example, any subset of $\mathbb{R}^2$ also becomes a metric space after we inherit it. The unit disk, unit circle, and the unit square $[0,1]^2$ are special cases.

[1] Also, "metric" is a fun word to say.

Example 2.1.5 (Taxicab on $\mathbb{R}^2$)
It is also possible to place the **taxicab distance** on $\mathbb{R}^2$:

$$d\left((x_1, y_1), (x_2, y_2)\right) = |x_1 - x_2| + |y_1 - y_2|.$$

For now, we will use the more natural Euclidean metric.

Example 2.1.6 (Metric spaces of $\mathbb{R}^n$)
We can generalize the above examples easily. Let n be a positive integer.

(a) We let $\mathbb{R}^n$ be the metric space whose points are points in n-dimensional Euclidean space, and whose metric is the Euclidean metric

$$d\left((a_1, \dots, a_n), (b_1, \dots, b_n)\right) = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}.$$

This is the n-dimensional **Euclidean space**.

(b) The open **unit ball** B^n is the subset of $\mathbb{R}^n$ consisting of those points $(x_1, \dots, x_n)$ such that $x_1^2 + \cdots + x_n^2 < 1$.

(c) The **unit sphere** S^{n-1} is the subset of $\mathbb{R}^n$ consisting of those points $(x_1, \dots, x_n)$ such that $x_1^2 + \cdots + x_n^2 = 1$, with the inherited metric. (The superscript $n-1$ indicates that S^{n-1} is an $n-1$ dimensional space, even though it lives in n-dimensional space.) For example, $S^1 \subseteq \mathbb{R}^2$ is the unit circle, whose distance between two points is the length of the chord joining them. You can also think of it as the "boundary" of the unit ball B^n.

Example 2.1.7 (Function space)
We can let M be the space of continuous functions $f \colon [0,1] \to \mathbb{R}$ and define the metric by $d(f,g) = \int_0^1 |f - g| \, dx$. (It admittedly takes some work to check $d(f,g) = 0$ implies $f = g$, but we won't worry about that yet.)

Here is a slightly more pathological example.

Example 2.1.8 (Discrete space)
Let S be any set of points (either finite or infinite). We can make S into a **discrete space** by declaring

$$d(x,y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y. \end{cases}$$

If $|S| = 4$ you might think of this space as the vertices of a regular tetrahedron, living in $\mathbb{R}^3$. But for larger S it's not so easy to visualize...

Example 2.1.9 (Graphs are metric spaces)

Any connected simple graph G can be made into a metric space by defining the distance between two vertices to be the graph-theoretic distance between them. (The discrete metric is the special case when G is the complete graph on S.)

Question 2.1.10. Check the conditions of a metric space for the metrics on the discrete space and for the connected graph.

Abuse of Notation 2.1.11. From now on, we will refer to $\mathbb{R}^n$ with the Euclidean metric by just $\mathbb{R}^n$. Moreover, if we wish to take the metric space for a subset $S \subseteq \mathbb{R}^n$ with the inherited metric, we will just write S.

§2.2 Convergence in metric spaces

Prototypical example for this section: The sequence $\frac{1}{n}$ *(for* $n = 1, 2, \dots$*) in* $\mathbb{R}$.

Since we can talk about the distance between two points, we can talk about what it means for a sequence of points to converge. This is the same as the typical epsilon-delta definition, with absolute values replaced by the distance function.

Definition 2.2.1. Let $(x_n)_{n \geq 1}$ be a sequence of points in a metric space M. We say that x_n **converges** to x if the following condition holds: for all $\varepsilon > 0$, there is an integer N (depending on ε) such that $d(x_n, x) < \varepsilon$ for each $n \geq N$. This is written

$$x_n \to x$$

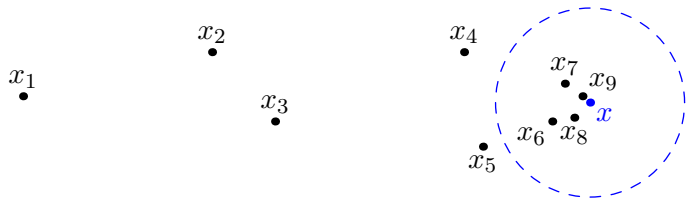
or more verbosely as

$$\lim_{n \to \infty} x_n = x.$$

We say that a sequence converges in M if it converges to a point in M.

You should check that this definition coincides with your intuitive notion of "converges".

Abuse of Notation 2.2.2. If the parent space M is understood, we will allow ourselves to abbreviate "converges in M" to just "converges". However, keep in mind that convergence is defined relative to the parent space; the "limit" of the space must actually be a point in M for a sequence to converge.



Example 2.2.3

Consider the sequence $x_1 = 1$, $x_2 = 1.4$, $x_3 = 1.41$, $x_4 = 1.414$, ...

(a) If we view this as a sequence in $\mathbb{R}$, it converges to $\sqrt{2}$.

(b) However, even though each x_i is in $\mathbb{Q}$, this sequence does NOT converge when we view it as a sequence in $\mathbb{Q}$!

Question 2.2.4. What are the convergent sequences in a discrete metric space?

§2.3 Continuous maps

In calculus you were also told (or have at least heard) of what it means for a function to be continuous. Probably something like

> A function $f\colon \mathbb{R} \to \mathbb{R}$ is continuous at a point $p \in \mathbb{R}$ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $|x - p| < \delta \implies |f(x) - f(p)| < \varepsilon$.

Question 2.3.1. Can you guess what the corresponding definition for metric spaces is?

All we have to do is replace the absolute values with the more general distance functions: this gives us a definition of continuity for any function $M \to N$.

Definition 2.3.2. Let $M = (M, d_M)$ and $N = (N, d_N)$ be metric spaces. A function $f\colon M \to N$ is **continuous** at a point $p \in M$ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$d_M(x, p) < \delta \implies d_N(f(x), f(p)) < \varepsilon.$$

Moreover, the entire function f is continuous if it is continuous at every point $p \in M$.

Notice that, just like in our definition of an isomorphism of a group, we use the metric of M for one condition and the metric of N for the other condition.

This generalization is nice because it tells us immediately how we could carry over continuity arguments in $\mathbb{R}$ to more general spaces like $\mathbb{C}$. Nonetheless, this definition is kind of cumbersome to work with, because it makes extensive use of the real numbers (epsilons and deltas). Here is an equivalent condition.

Theorem 2.3.3 (Sequential continuity)
A function $f\colon M \to N$ of metric spaces is continuous at a point $p \in M$ if and only if the following property holds: if $x_1, x_2, \ldots$ is a sequence in M converging to p, then the sequence $f(x_1), f(x_2), \ldots$ in N converges to $f(p)$.

Proof. One direction is not too hard:

Exercise 2.3.4. Show that ε-δ continuity implies sequential continuity at each point.

Conversely, we will prove if f is not ε-δ continuous at p then it does not preserve convergence.

If f is not continuous at p, then there is a "bad" $\varepsilon > 0$, which we now consider fixed. So for each choice of $\delta = 1/n$, there should be some point x_n which is within δ of p, but which is mapped more than ε away from $f(p)$. But then the sequence x_n converges to p, and $f(x_n)$ is always at least ε away from $f(p)$, contradiction. □

Example application showcasing the niceness of sequential continuity:

Proposition 2.3.5 (Composition of continuous functions is continuous)
Let $f\colon M \to N$ and $g\colon N \to L$ be continuous maps of metric spaces. Then their composition $g \circ f$ is continuous.

Proof. Dead simple with sequences: Let $p \in M$ be arbitrary and let $x_n \to p$ in M. Then $f(x_n) \to f(p)$ in N and $g(f(x_n)) \to g(f(p))$ in L, QED. □

Question 2.3.6. Let M be any metric space and D a discrete space. When is a map $f\colon D \to M$ continuous?

§2.4 Homeomorphisms

Prototypical example for this section: The unit circle S^1 is homeomorphic to the boundary of the unit square.

When do we consider two groups to be the same? Answer: if there's a structure-preserving map between them which is also a bijection. For metric spaces, we do exactly the same thing, but replace "structure-preserving" with "continuous".

Definition 2.4.1. Let M and N be metric spaces. A function $f\colon M \to N$ is a **homeomorphism** if it is a bijection, and both $f\colon M \to N$ and its inverse $f^{-1}\colon N \to M$ are continuous. We say M and N are **homeomorphic**.

Needless to say, homeomorphism is an equivalence relation.

You might be surprised that we require f^{-1} to also be continuous. Here's the reason: you can show that if ϕ is an isomorphism of groups, then ϕ^{-1} also preserves the group operation, hence ϕ^{-1} is itself an isomorphism. The same is not true for continuous bijections, which is why we need the new condition.

Example 2.4.2 (Homeomorphism $\neq$ continuous bijection)

(a) There is a continuous bijection from $[0, 1)$ to the circle, but it has no continuous inverse.

(b) Let M be a discrete space with size $|\mathbb{R}|$. Then there is a continuous function $M \to \mathbb{R}$ which certainly has no continuous inverse.

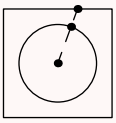
Note that this is the topologist's definition of "same" – homeomorphisms are "continuous deformations". Here are some examples.

Example 2.4.3 (Examples of homeomorphisms)

(a) Any space M is homeomorphic to itself through the identity map.

(b) The old saying: a doughnut (torus) is homeomorphic to a coffee cup. (Look this up if you haven't heard of it.)

(c) The unit circle S^1 is homeomorphic to the boundary of the unit square. Here's one bijection between them, after an appropriate scaling:



Example 2.4.4 (Metrics on the unit circle)

It may have seemed strange that our metric function on S^1 was the one inherited from $\mathbb{R}^2$, meaning the distance between two points on the circle was defined to be the length of the chord. Wouldn't it have made more sense to use the circumference of the smaller arc joining the two points?

In fact, it doesn't matter: if we consider S^1 with the "chord" metric and the "arc" metric, we get two homeomorphic spaces, as the map between them is continuous. The same goes for S^{n-1} for general n.

Example 2.4.5 (Homeomorphisms really don't preserve size)

Surprisingly, the open interval $(-1, 1)$ is homeomorphic to the real line $\mathbb{R}$! One bijection is given by

$$x \mapsto \tan(x\pi/2)$$

with the inverse being given by $t \mapsto \frac{2}{\pi}\arctan(t)$.

This might come as a surprise, since $(-1, 1)$ doesn't look that much like $\mathbb{R}$; the former is "bounded" while the latter is "unbounded".

§2.5 Extended example/definition: product metric

Prototypical example for this section: $\mathbb{R} \times \mathbb{R}$ *is* $\mathbb{R}^2$.

Here is an extended example which will occur later on. Let $M = (M, d_M)$ and $N = (N, d_N)$ be metric spaces (say, $M = N = \mathbb{R}$). Our goal is to define a metric space on $M \times N$.

Let $p_i = (x_i, y_i) \in M \times N$ for $i = 1, 2$. Consider the following metrics on the set of points $M \times N$:

$$\begin{aligned} d_{\max}(p_1, p_2) &\coloneqq \max\left\{d_M(x_1, x_2), d_N(y_1, y_2)\right\} \\ d_{\text{Euclid}}(p_1, p_2) &\coloneqq \sqrt{d_M(x_1, x_2)^2 + d_N(y_1, y_2)^2} \\ d_{\text{taxicab}}(p_1, p_2) &\coloneqq d_M(x_1, x_2) + d_N(y_1, y_2). \end{aligned}$$

All of these are good candidates. We are about to see it doesn't matter which one we use:

Exercise 2.5.1. Verify that

$$d_{\max}(p_1, p_2) \le d_{\text{Euclid}}(p_1, p_2) \le d_{\text{taxicab}}(p_1, p_2) \le 2d_{\max}(p_1, p_2).$$

Use this to show that the metric spaces we obtain by imposing any of the three metrics are homeomorphic, with the homeomorphism being just the identity map.

Definition 2.5.2. Hence we will usually simply refer to *the* metric on $M \times N$, called the **product metric**. It will not be important which of the three metrics we select.

Example 2.5.3 ($\mathbb{R}^2$)
If $M = N = \mathbb{R}$, we get $\mathbb{R}^2$, the Euclidean plane. The metric d_{Euclid} is the one we started with, but using either of the other two metric works fine as well.

The product metric plays well with convergence of sequences.

Proposition 2.5.4 (Convergence in the product metric is by component)
We have $(x_n, y_n) \to (x, y)$ if and only if $x_n \to x$ and $y_n \to y$.

Proof. We have $d_{\max}((x,y),(x_n,y_n)) = \max\{d_M(x,x_n), d_N(y,y_n)\}$ and the latter approaches zero as $n \to \infty$ if and only if $d_M(x,x_n) \to 0$ and $d_N(y,y_n) \to 0$. □

Let's see an application of this:

Proposition 2.5.5 (Addition and multiplication are continuous)
The addition and multiplication maps are continuous maps $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

Proof. For multiplication: for any n we have

$$\begin{aligned} x_n y_n &= (x + (x_n - x))(y + (y_n - y)) \\ &= xy + y(x_n - x) + x(y_n - y) + (x_n - x)(y_n - y) \\ \implies |x_n y_n - xy| &\le |y|\,|x_n - x| + |x|\,|y_n - y| + |x_n - x|\,|y_n - y|. \end{aligned}$$

As $n \to \infty$, all three terms on the right-hand side tend to zero. The proof that $+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is continuous is similar (and easier): one notes for any n that

$$|(x_n + y_n) - (x + y)| \le |x_n - x| + |y_n - y|$$

and both terms on the right-hand side tend to zero as $n \to \infty$. □

Problem 2C covers the other two operations, subtraction and division. The upshot of this is that, since compositions are also continuous, most of your naturally arising real-valued functions will automatically be continuous as well. For example, the function $\frac{3x}{x^2+1}$ will be a continuous function from $\mathbb{R} \to \mathbb{R}$, since it can be obtained by composing $+$, $\times$, $\div$.

§2.6 Open sets

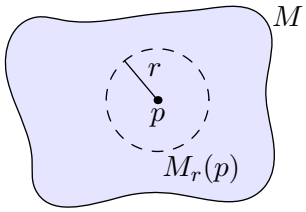
Prototypical example for this section: The open disk $x^2 + y^2 < r^2$ in $\mathbb{R}^2$.

Continuity is really about what happens "locally": how a function behaves "close to a certain point p". One way to capture this notion of "closeness" is to use metrics as we've done above. In this way we can define an r-neighborhood of a point.

Definition 2.6.1. Let M be a metric space. For each real number $r > 0$ and point $p \in M$, we define

$$M_r(p) \coloneqq \{x \in M : d(x,p) < r\}.$$

The set $M_r(p)$ is called an ***r*-neighborhood** of p.



We can rephrase convergence more succinctly in terms of r-neighborhoods. Specifically, a sequence (x_n) converges to x if for every r-neighborhood of x, all terms of x_n eventually stay within that r-neighborhood.

Let's try to do the same with functions.

Question 2.6.2. In terms of r-neighborhoods, what does it mean for a function $f\colon M \to N$ to be continuous at a point $p \in M$?

Essentially, we require that the pre-image of every ε-neighborhood has the property that some δ-neighborhood exists inside it. This motivates:

Definition 2.6.3. A set $U \subseteq M$ is **open** in M if for each $p \in U$, some r-neighborhood of p is contained inside U. In other words, there exists $r > 0$ such that $M_r(p) \subseteq U$.

Abuse of Notation 2.6.4. Note that a set being open is defined *relative to* the parent space M. However, if M is understood we can abbreviate "open in M" to just "open".

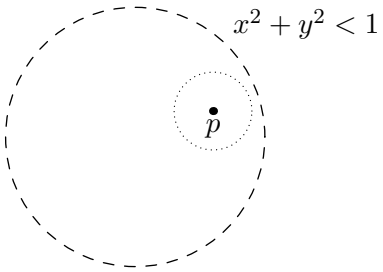


Figure 2.1: The set of points $x^2 + y^2 < 1$ in $\mathbb{R}^2$ is open in $\mathbb{R}^2$.

Example 2.6.5 (Examples of open sets)

(a) Any r-neighborhood of a point is open.

(b) Open intervals of $\mathbb{R}$ are open in $\mathbb{R}$, hence the name! This is the prototypical example to keep in mind.

(c) The open unit ball B^n is open in $\mathbb{R}^n$ for the same reason.

(d) In particular, the open interval $(0, 1)$ is open in $\mathbb{R}$. However, if we embed it in $\mathbb{R}^2$, it is no longer open!

(e) The empty set $\varnothing$ and the whole set of points M are open in M.

Example 2.6.6 (Non-examples of open sets)

(a) The closed interval $[0, 1]$ is not open in $\mathbb{R}$. There is no ε-neighborhood of the point 0 which is contained in $[0, 1]$.

(b) The unit circle S^1 is not open in $\mathbb{R}^2$.

Question 2.6.7. What are the open sets of the discrete space?

Here are two quite important properties of open sets.

Proposition 2.6.8 (Intersections and unions of open sets)

(a) The intersection of finitely many open sets is open.

(b) The union of open sets is open, even if there are infinitely many.

Question 2.6.9. Convince yourself this is true.

Exercise 2.6.10. Exhibit an infinite collection of open sets in $\mathbb{R}$ whose intersection is the set $\{0\}$. This implies that infinite intersections of open sets are not necessarily open.

The whole upshot of this is:

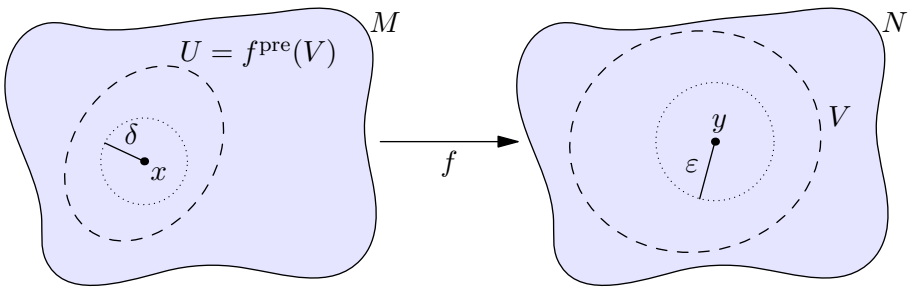
Theorem 2.6.11 (Open set condition)

A function $f \colon M \to N$ of metric spaces is continuous if and only if the pre-image of every open set in N is open in M.

Proof. I'll just do one direction...

Exercise 2.6.12. Show that δ-ε continuity follows from the open set condition.

Now assume f is continuous. First, suppose V is an open subset of the metric space N; let $U = f^{\text{pre}}(V)$. Pick $x \in U$, so $y = f(x) \in V$; we want an open neighborhood of x inside U.



As V is open, there is some small ε-neighborhood around y which is contained inside V. By continuity of f, we can find a δ such that the δ-neighborhood of x gets mapped by f

into the ε-neighborhood in N, which in particular lives inside V. Thus the δ-neighborhood lives in U, as desired. □

§2.7 Closed sets

Prototypical example for this section: The closed unit disk $x^2 + y^2 \le r^2$ in $\mathbb{R}^2$.

It would be criminal for me to talk about open sets without talking about closed sets. The name "closed" comes from the definition in a metric space.

Definition 2.7.1. Let M be a metric space. A subset $S \subseteq M$ is **closed** in M if the following property holds: let $x_1, x_2, \dots$ be a sequence of points in S and suppose that x_n converges to x in M. Then $x \in S$ as well.

Abuse of Notation 2.7.2. Same caveat: we abbreviate "closed in M" to just "closed" if the parent space M is understood.

Here's another way to phrase it. The **limit points** of a subset $S \subseteq M$ are defined by

$$\lim S \coloneqq \{p \in M : \exists (x_n) \in S \text{ such that } x_n \to p\}.$$

Thus S is closed if and only if $S = \lim S$.

> **Exercise 2.7.3.** Prove that $\lim S$ is closed even if S isn't closed. (Draw a picture.)

For this reason, $\lim S$ is also called the **closure** of S in M, and denoted $\overline{S}$. It is simply the smallest closed set which contains S.

Example 2.7.4 (Examples of closed sets)

(a) The empty set $\varnothing$ is closed in M for vacuous reasons: there are no sequences of points with elements in $\varnothing$.

(b) The entire space M is closed in M for tautological reasons. (Verify this!)

(c) The closed interval $[0,1]$ in $\mathbb{R}$ is closed in $\mathbb{R}$, hence the name. Like with open sets, this is the prototypical example of a closed set to keep in mind!

(d) In fact, the closed interval $[0,1]$ is even closed in $\mathbb{R}^2$.

Example 2.7.5 (Non-examples of closed sets)

Let $S = (0,1)$ denote the open interval. Then S is not closed in $\mathbb{R}$ because the sequence of points

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots$$

converges to $0 \in \mathbb{R}$, but $0 \notin (0,1)$.

I should now warn you about a confusing part of this terminology. Firstly, **"most" sets are neither open nor closed**.

Example 2.7.6 (A set neither open nor closed)
The half-open interval $[0, 1)$ is neither open nor closed in $\mathbb{R}$.

Secondly, it's **also possible for a set to be both open and closed**; this will be discussed in Chapter 7.

The reason for the opposing terms is the following theorem:

Theorem 2.7.7 (Closed sets are complements of open sets)
Let M be a metric space, and $S \subseteq M$ any subset. Then the following are equivalent:

- The set S is closed in M.
- The complement $M \setminus S$ is open in M.

Exercise 2.7.8 (Great)**.** Prove this theorem! You'll want to draw a picture to make it clear what's happening: for example, you might take $M = \mathbb{R}^2$ and S to be the closed unit disk.

§2.8 A few harder problems to think about

Problem 2A. Let $M = (M, d)$ be a metric space. Show that

$$d \colon M \times M \to \mathbb{R}$$

is itself a continuous function (where $M \times M$ is equipped with the product metric).

Problem 2B. Consider $\mathbb{Q}$ and $\mathbb{N}$ as metric spaces (each with the obvious metric $d(x, y) = |x - y|$). Are these spaces homeomorphic?

Problem 2C (Continuity of arithmetic continued)**.** Show that subtraction is a continuous map $- \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and division is a continuous map $\div \colon \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}$.

Problem 2D. Exhibit a function $f \colon \mathbb{R} \to \mathbb{R}$ such that f is continuous at $x \in \mathbb{R}$ if and only if $x = 0$.

Problem 2E. Prove that a function $f \colon \mathbb{R} \to \mathbb{R}$ which is strictly increasing must be continuous at some point.

Problem 2F. Someone on the Internet posted the question "is $1/x$ a continuous function?", leading to great controversy on Twitter. How should you respond?

II

Basic Abstract Algebra

Part II: Contents

3 Homomorphisms and quotient groups

§3.1 Generators and group presentations

Prototypical example for this section: $D_{2n} = \langle r, s \mid r^n = s^2 = 1 \rangle$

Let G be a group. Recall that for some element $x \in G$, we could consider the subgroup

$$\left\{ \dots, x^{-2}, x^{-1}, 1, x, x^2, \dots \right\}$$

of G. Here's a more pictorial version of what we did: **put x in a box, seal it tightly, and shake vigorously**. Using just the element x, we get a pretty explosion that produces the subgroup above.

What happens if we put two elements x, y in the box? Among the elements that get produced are things like

$$xyxyx, \quad x^2y^9x^{-5}y^3, \quad y^{-2015}, \quad \dots$$

Essentially, I can create any finite product of x, y, x^{-1}, y^{-1}. This leads us to define:

Definition 3.1.1. Let S be a subset of G. The subgroup **generated** by S, denoted $\langle S \rangle$, is the set of elements which can be written as a finite product of elements in S (and their inverses). If $\langle S \rangle = G$ then we say S is a set of **generators** for G, as the elements of S together create all of G.

Exercise 3.1.2. Why is the condition "and their inverses" not necessary if G is a finite group? (As usual, assume Lagrange's theorem.)

Example 3.1.3 ($\mathbb{Z}$ is the infinite cyclic group)

Consider 1 as an element of $\mathbb{Z} = (\mathbb{Z}, +)$. We see $\langle 1 \rangle = \mathbb{Z}$, meaning $\{1\}$ generates $\mathbb{Z}$. It's important that -1, the inverse of 1 is also allowed: we need it to write all integers as the sum of 1 and -1.

This gives us an idea for a way to try and express groups compactly. Why not just write down a list of generators for the groups? For example, we could write

$$\mathbb{Z} \cong \langle a \rangle$$

meaning that $\mathbb{Z}$ is just the group generated by one element.

There's one issue: the generators usually satisfy certain properties. For example, consider $\mathbb{Z}/100\mathbb{Z}$. It's also generated by a single element x, but this x has the additional property that $x^{100} = 1$. This motivates us to write

$$\mathbb{Z}/100\mathbb{Z} = \left\langle x \mid x^{100} = 1 \right\rangle.$$

I'm sure you can see where this is going. All we have to do is specify a set of generators and **relations** between the generators, and say that two elements are equal if and only if you can get from one to the other using relations. Such an expression is appropriately called a **group presentation**.

Example 3.1.4 (Dihedral group)
The dihedral group of order $2n$ has a presentation

$$D_{2n} = \left\langle r, s \mid r^n = s^2 = 1, rs = sr^{-1} \right\rangle.$$

Thus each element of D_{2n} can be written uniquely in the form r^α or sr^α, where $\alpha = 0, 1, \ldots, n-1$.

Example 3.1.5 (Klein four group)
The **Klein four group**, isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, is given by the presentation

$$\left\langle a, b \mid a^2 = b^2 = 1, ab = ba \right\rangle.$$

Example 3.1.6 (Free group)
The **free group on n elements** is the group whose presentation has n generators and no relations at all. It is denoted F_n, so

$$F_n = \langle x_1, x_2, \ldots, x_n \rangle.$$

In other words, $F_2 = \langle a, b \rangle$ is the set of strings formed by appending finitely many copies of a, b, a^{-1}, b^{-1} together.

Question 3.1.7. Notice that $F_1 \cong \mathbb{Z}$.

Abuse of Notation 3.1.8. One might unfortunately notice that "subgroup generated by a and b" has exactly the same notation as the free group $\langle a, b \rangle$. We'll try to be clear based on context which one we mean.

Presentations are nice because they provide a compact way to write down groups. They do have some shortcomings, though.[1]

Example 3.1.9 (Presentations can look very different)
The same group can have very different presentations. For instance consider

$$D_{2n} = \left\langle x, y \mid x^2 = y^2 = 1, (xy)^n = 1 \right\rangle.$$

(To see why this is equivalent, set $x = s$, $y = rs$.)

§3.2 Homomorphisms

Prototypical example for this section: The "mod out by 100" map, $\mathbb{Z} \to \mathbb{Z}/100\mathbb{Z}$.

How can groups talk to each other?

[1] Actually, determining whether two elements of a presentation are equal is undecidable. In fact, it is undecidable to even determine if a group is finite from its presentation.

Two groups are "the same" if we can write an isomorphism between them. And as we saw, two metric spaces are "the same" if we can write a homeomorphism between them. But what's the group analogy of a continuous map? We simply drop the "bijection" condition.

Definition 3.2.1. Let $G = (G, \star)$ and $H = (H, *)$ be groups. A **group homomorphism** is a map $\phi\colon G \to H$ such that for any $g_1, g_2 \in G$ we have

$$\phi(g_1 \star g_2) = \phi(g_1) * \phi(g_2).$$

(Not to be confused with "homeomorphism" from last chapter: note the spelling.)

Example 3.2.2 (Examples of homomorphisms)
Let G and H be groups.

(a) Any isomorphism $G \to H$ is a homomorphism. In particular, the identity map $G \to G$ is a homomorphism.

(b) The **trivial homomorphism** $G \to H$ sends everything to 1_H.

(c) There is a homomorphism from $\mathbb{Z}$ to $\mathbb{Z}/100\mathbb{Z}$ by sending each integer to its residue modulo 100.

(d) There is a homomorphism from $\mathbb{Z}$ to itself by $x \mapsto 10x$ which is injective but not surjective.

(e) There is a homomorphism from S_n to S_{n+1} by "embedding": every permutation on $\{1, \dots, n\}$ can be thought of as a permutation on $\{1, \dots, n+1\}$ if we simply let $n+1$ be a fixed point.

(f) A homomorphism $\phi : D_{12} \to D_6$ is given by $s_{12} \mapsto s_6$ and $r_{12} \mapsto r_6$.

(g) Specifying a homomorphism $\mathbb{Z} \to G$ is the same as specifying just the image of the element $1 \in \mathbb{Z}$. Why?

The last two examples illustrate something: suppose we have a presentation of G. To specify a homomorphism $G \to H$, we only have to specify where each generator of G goes, in such a way that the relations are all satisfied.

Important remark: the right way to think about an isomorphism is as a "bijective homomorphism". To be explicit,

Exercise 3.2.3. Show that $G \cong H$ if and only if there exist homomorphisms $\phi\colon G \to H$ and $\psi\colon H \to G$ such that $\phi \circ \psi = \mathrm{id}_H$ and $\psi \circ \phi = \mathrm{id}_G$.

So the definitions of homeomorphism of metric spaces and isomorphism of groups are not too different.

Some obvious properties of homomorphisms follow.

Fact 3.2.4. Let $\phi\colon G \to H$ be a homomorphism. Then $\phi(1_G) = 1_H$ and $\phi(g^{-1}) = \phi(g)^{-1}$.

Proof. Boring, and I'm sure you could do it yourself if you wanted to. □

Now let me define a very important property of a homomorphism.

Definition 3.2.5. The **kernel** of a homomorphism $\phi\colon G \to H$ is defined by

$$\ker \phi \coloneqq \{g \in G : \phi(g) = 1_H\}.$$

It is a *subgroup* of G (in particular, $1_G \in \ker \phi$ for obvious reasons).

Question 3.2.6. Verify that $\ker \phi$ is in fact a subgroup of G.

We also have the following important fact, which we also encourage the reader to verify.

Proposition 3.2.7 (Kernel determines injectivity)
The map ϕ is injective if and only if $\ker \phi = \{1_G\}$.

To make this concrete, let's compute the kernel of each of our examples.

Example 3.2.8 (Examples of kernels)

(a) The kernel of any isomorphism $G \to H$ is trivial, since an isomorphism is injective. In particular, the kernel of the identity map $G \to G$ is $\{1_G\}$.

(b) The kernel of the trivial homomorphism $G \to H$ (by $g \mapsto 1_H$) is all of G.

(c) The kernel of the homomorphism $\mathbb{Z} \to \mathbb{Z}/100\mathbb{Z}$ by $n \mapsto \overline{n}$ is precisely

$$100\mathbb{Z} = \{\dots, -200, -100, 0, 100, 200, \dots\}.$$

(d) The kernel of the map $\mathbb{Z} \to \mathbb{Z}$ by $x \mapsto 10x$ is trivial: $\{0\}$.

(e) There is a homomorphism from S_n to S_{n+1} by "embedding", but it also has trivial kernel because it is injective.

(f) A homomorphism $\phi\colon D_{12} \to D_6$ is given by $s_{12} \mapsto s_6$ and $r_{12} \mapsto r_6$. You can check that

$$\ker \phi = \left\{1, r_{12}^3\right\} \cong \mathbb{Z}/2\mathbb{Z}.$$

(g) Exercise below.

Exercise 3.2.9. Fix any $g \in G$. Suppose we have a homomorphism $\mathbb{Z} \to G$ by $n \mapsto g^n$. What is the kernel?

Question 3.2.10. Show that for any homomorphism $\phi : G \to H$, the image $\phi^{\text{img}}(G)$ is a subgroup of H. Hence, we'll be especially interested in the case where ϕ is surjective.

§3.3 Cosets and modding out

Prototypical example for this section: Modding out by n: $\mathbb{Z}/(n \cdot \mathbb{Z}) \cong \mathbb{Z}/n\mathbb{Z}$.

The next few sections are a bit dense. If this exposition doesn't work for you, try [Go11].
Let G and Q be groups, and suppose there exists a *surjective* homomorphism

$$\phi\colon G \twoheadrightarrow Q.$$

In other words, if ϕ is injective then $\phi\colon G \to Q$ is a bijection, and hence an isomorphism. But suppose we're not so lucky and $\ker\phi$ is bigger than just $\{1_G\}$. What is the correct interpretation of a more general homomorphism?

Let's look at the special case where $\phi\colon \mathbb{Z} \to \mathbb{Z}/100\mathbb{Z}$ is "modding out by 100". We already saw that the kernel of this map is

$$\ker\phi = 100\mathbb{Z} = \{\dots, -200, -100, 0, 100, 200, \dots\}.$$

Recall now that $\ker\phi$ is a subgroup of G. What this means is that ϕ **is indifferent to the subgroup** $100\mathbb{Z}$ **of** $\mathbb{Z}$:

$$\phi(15) = \phi(2000 + 15) = \phi(-300 + 15) = \phi(700 + 15) = \dots.$$

So $\mathbb{Z}/100\mathbb{Z}$ is what we get when we "mod out by 100". Cool.

In other words, let G be a group and $\phi\colon G \twoheadrightarrow Q$ be a surjective homomorphism with kernel $N \subseteq G$.

We claim that Q should be thought of as the quotient of G by N.

To formalize this, we will define a so-called **quotient group** G/N in terms of G and N only (without referencing Q) which will be naturally isomorphic to Q.

For motivation, let's give a concrete description of Q using just ϕ and G. Continuing our previous example, let $N = 100\mathbb{Z}$ be our subgroup of G. Consider the sets

$$\begin{aligned} N &= \{\dots, -200, -100, 0, 100, 200, \dots\} \\ 1 + N &= \{\dots, -199, -99, 1, 101, 201, \dots\} \\ 2 + N &= \{\dots, -198, -98, 2, 102, 202, \dots\} \\ &\vdots \\ 99 + N &= \{\dots, -101, -1, 99, 199, 299, \dots\}. \end{aligned}$$

The elements of each set all have the same image when we apply ϕ, and moreover any two elements in different sets have different images. Then the main idea is to notice that

We can think of Q as the group whose *elements* are the *sets* above.

Thus, given ϕ we define an equivalence relation $\sim_N$ on G by saying $x \sim_N y$ for $\phi(x) = \phi(y)$. This $\sim_N$ divides G into several equivalence classes in G which are in obvious bijection with Q, as above. Now we claim that we can write these equivalence classes very explicitly.

Exercise 3.3.1. Show that $x \sim_N y$ if and only if $x = yn$ for some $n \in N$ (in the mod 100 example, this means they "differ by some multiple of 100"). Thus for any $g \in G$, the equivalence class of $\sim_N$ which contains g is given explicitly by

$$gN \coloneqq \{gn \mid n \in N\}.$$

Here's the word that describes the types of sets we're running into now.

Definition 3.3.2. Let H be any subgroup of G (not necessarily the kernel of some homomorphism). A set of the form gH is called a **left coset** of H.

Remark 3.3.3 — Although the notation might not suggest it, keep in mind that g_1N is often equal to g_2N even if $g_1 \neq g_2$. In the "mod 100" example, $3 + N = 103 + N$. In other words, these cosets are *sets*.
This means that if I write "let gH be a coset" without telling you what g is, you can't figure out which g I chose from just the coset itself. If you don't believe me, here's an example of what I mean:

$$x + 100\mathbb{Z} = \{\dots, -97, 3, 103, 203, \dots\} \implies x = ?.$$

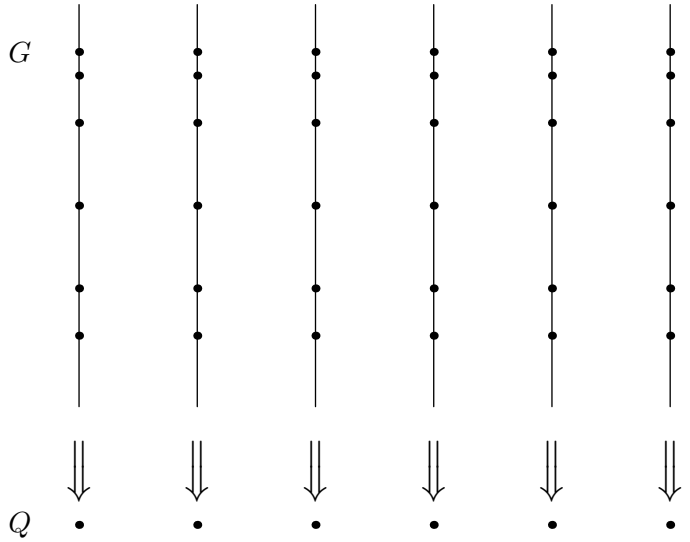
There's no reason to think I picked $x = 3$. (I actually picked $x = -13597$.)

Remark 3.3.4 — Given cosets g_1H and g_2H, you can check that the map $x \mapsto g_2g_1^{-1}x$ is a bijection between them. So actually, all cosets have the same cardinality.

So, long story short,

Elements of the group Q are naturally identified with left cosets of N.

In practice, people often still prefer to picture elements of Q as single points (for example it's easier to think of $\mathbb{Z}/2\mathbb{Z}$ as $\{0, 1\}$ rather than $\{\{\dots, -2, 0, 2, \dots\}, \{\dots, -1, 1, 3, \dots\}\}$). If you like this picture, then you might then draw G as a bunch of equally tall fibers (the cosets), which are then "collapsed" onto Q.



Now that we've done this, we can give an *intrinsic* definition for the quotient group we alluded to earlier.

Definition 3.3.5. A subgroup N of G is called **normal** if it is the kernel of some homomorphism. We write this as $N \trianglelefteq G$.

Definition 3.3.6. Let $N \trianglelefteq G$. Then the **quotient group**, denoted G/N (and read "G mod N"), is the group defined as follows.

- The elements of G/N will be the left cosets of N.

- We want to define the product of two cosets C_1 and C_2 in G/N. Recall that the cosets are in bijection with elements of Q. So let q_1 be the value associated to the coset C_1, and q_2 the one for C_2. Then we can take the product to be the coset corresponding to q_1q_2.

  Quite importantly, **we can also do this in terms of representatives of the cosets**. Let $g_1 \in C_1$ and $g_2 \in C_2$, so $C_1 = g_1N$ and $C_2 = g_2N$. Then $C_1 \cdot C_2$ should be the coset which contains g_1g_2. This is the same as the above definition since $\phi(g_1g_2) = \phi(g_1)\phi(g_2) = q_1q_2$; all we've done is define the product in terms of elements of G, rather than values in H.

  Using the gN notation, and with Remark 3.3.3 in mind, we can write this even more succinctly:
  $$(g_1N) \cdot (g_2N) \coloneqq (g_1g_2)N.$$

And now you know why the integers modulo n are often written $\mathbb{Z}/n\mathbb{Z}$!

Question 3.3.7. Take a moment to digest the above definition.

By the way we've built it, the resulting group G/N is isomorphic to Q. In a sense we think of G/N as "G modulo the condition that $n = 1$ for all $n \in N$".

§3.4 (Optional) Proof of Lagrange's theorem

As an aside, with the language of cosets we can now show Lagrange's theorem in the general case.

Theorem 3.4.1 (Lagrange's theorem)
Let G be a finite group, and let H be any subgroup. Then $|H|$ divides $|G|$.

The proof is very simple: note that the cosets of H all have the same size and form a partition of G (even when H is not necessarily normal). Hence if n is the number of cosets, then $n \cdot |H| = |G|$.

Question 3.4.2. Conclude that $x^{|G|} = 1$ by taking $H = \langle x \rangle \subseteq G$.

Remark 3.4.3 — It should be mentioned at this point that in general, if G is a finite group and N is normal, then $|G/N| = |G|/|N|$.

§3.5 Eliminating the homomorphism

Prototypical example for this section: Again $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/n\mathbb{Z}$.

Let's look at the last definition of G/N we provided. The short version is:

- The elements of G/N are cosets gN, which you can think of as equivalence classes of a relation $\sim_N$ (where $g_1 \sim_N g_2$ if $g_1 = g_2n$ for some $n \in N$).
- Given cosets g_1N and g_2N the group operation is
  $$g_1N \cdot g_2N \coloneqq (g_1g_2)N.$$

Question: where do we actually use the fact that N is normal? We don't talk about ϕ or Q anywhere in this definition.

The answer is in Remark 3.3.3. The group operation takes in two cosets, so it doesn't know what g_1 and g_2 are. But behind the scenes, **the normal condition guarantees that the group operation can pick any g_1 and g_2 it wants and still end up with the same coset.** If we didn't have this property, then it would be hard to define the product of two cosets C_1 and C_2 because it might make a difference which $g_1 \in C_1$ and $g_2 \in C_2$ we picked. The fact that N came from a homomorphism meant we could pick any representatives g_1 and g_2 of the cosets we wanted, because they all had the same ϕ-value.

We want some conditions which force this to be true without referencing ϕ at all. Suppose $\phi\colon G \to K$ is a homomorphism of groups with $H = \ker \phi$. Aside from the fact H is a group, we can get an "obvious" property:

Question 3.5.1. Show that if $h \in H$, $g \in G$, then $ghg^{-1} \in H$. (Check $\phi(ghg^{-1}) = 1_K$.)

Example 3.5.2 (Example of a non-normal subgroup)
Let $D_{12} = \langle r, s \mid r^6 = s^2 = 1, rs = sr^{-1} \rangle$. Consider the subgroup of order two $H = \{1, s\}$ and notice that

$$rsr^{-1} = r(sr^{-1}) = r(rs) = r^2 s \notin H.$$

Hence H is not normal, and cannot be the kernel of any homomorphism.

Well, duh – so what? Amazingly it turns out that this is the *sufficient* condition we want. Specifically, it makes the nice "coset multiplication" we wanted work out.

Remark 3.5.3 (For math contest enthusiasts) — This coincidence is really a lot like functional equations at the IMO. We all know that normal subgroups H satisfy $ghg^{-1} \in H$; the surprise is that from the latter seemingly weaker condition, we can deduce H is normal.

Thus we have a new criterion for "normal" subgroups which does not make any external references to ϕ.

Theorem 3.5.4 (Algebraic condition for normal subgroups)
Let H be a subgroup of G. Then the following are equivalent:

- $H \trianglelefteq G$.
- For every $g \in G$ and $h \in H$, $ghg^{-1} \in H$.

Proof. We already showed one direction.

For the other direction, we need to build a homomorphism with kernel H. So we simply *define* the group G/H as the cosets. To put a group operation, we need to verify:

Claim 3.5.5. If $g_1' \sim_H g_1$ and $g_2' \sim_H g_2$ then $g_1'g_2' \sim_H g_1g_2$.

Proof. Boring algebraic manipulation (again functional equation style). Let $g_1' = g_1h_1$

and $g_2' = g_2 h_2$, so we want to show that $g_1 h_1 g_2 h_2 \sim_H g_1 g_2$. Since H has the property, $g_2^{-1} h_1 g_2$ is some element of H, say h_3. Thus $h_1 g_2 = g_2 h_3$, and the left-hand side becomes $g_1 g_2 (h_3 h_2)$, which is fine since $h_3 h_2 \in H$. ■

With that settled we can just *define* the product of two cosets (of normal subgroups) by

$$(g_1 H) \cdot (g_2 H) = (g_1 g_2) H.$$

Thus the claim above shows that this multiplication is well-defined (this verification is the "content" of the theorem). So G/H is indeed a group! Moreover there is an obvious "projection" homomorphism $G \to G/H$ (with kernel H), by $g \mapsto gH$. □

Example 3.5.6 (Modding out in the product group)
Consider again the product group $G \times H$. Earlier we identified a subgroup

$$G' = \{(g, 1_H) \mid g \in G\} \cong G.$$

You can easily see that $G' \trianglelefteq G \times H$. (Easy calculation.)
Moreover, you can check that

$$(G \times H)/(G') \cong H.$$

Indeed, we have $(g, h) \sim_{G'} (1_G, h)$ for all $g \in G$ and $h \in H$.

Example 3.5.7 (Quotients and products don't necessarily cancel)
It is not necessarily true that $(G/H) \times H \cong G$. For example, consider $G = \mathbb{Z}/4\mathbb{Z}$ and the normal subgroup $H = \{0, 2\} \cong \mathbb{Z}/2\mathbb{Z}$. Then $G/H \cong \mathbb{Z}/2\mathbb{Z}$, but $\mathbb{Z}/4\mathbb{Z} \ncong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. (Footnote: the precise condition for this kind of "canceling" is called the Schur-Zassenhaus lemma.)

Example 3.5.8 (Another explicit computation)
Let $\phi : D_8 \to \mathbb{Z}/4\mathbb{Z}$ be defined by

$$r \mapsto \overline{2}, \quad s \mapsto \overline{2}.$$

The kernel of this map is $N = \{1, r^2, sr, sr^3\}$.
We can do a quick computation of all the elements of D_8 to get

$$\phi(1) = \phi(r^2) = \phi(sr) = \phi(sr^3) = \overline{0} \text{ and } \phi(r) = \phi(r^3) = \phi(s) = \phi(sr^2) = \overline{2}.$$

The two relevant fibers are

$$\phi^{\text{pre}}(\overline{0}) = 1N = r^2 N = srN = sr^3 N = \{1, r^2, sr, sr^3\}$$

and

$$\phi^{\text{pre}}(\overline{2}) = rN = r^3 N = sN = sr^2 N = \{r, r^3, s, sr^2\}.$$

So we see that $|D_8/N| = 2$ is a group of order two, or $\mathbb{Z}/2\mathbb{Z}$. Indeed, the image of ϕ is

$$\{\overline{0}, \overline{2}\} \cong \mathbb{Z}/2\mathbb{Z}.$$

Question 3.5.9. Suppose G is abelian. Why does it follow that any subgroup of G is normal?

Finally here's some food for thought: suppose one has a group presentation for a group G that uses n generators. Can you write it as a quotient of the form F_n/N, where N is a normal subgroup of F_n?

§3.6 (Digression) The first isomorphism theorem

One quick word about what other sources usually say.

Most textbooks actually *define* normal using the $ghg^{-1} \in H$ property. Then they define G/H for normal H in the way I did above, using the coset definition

$$(g_1H) \cdot (g_2H) = g_1g_2H.$$

Using purely algebraic manipulations (like I did) this is well-defined, and so now you have this group G/H or something. The underlying homomorphism isn't mentioned at all, or is just mentioned in passing.

I think this is incredibly dumb. The normal condition looks like it gets pulled out of thin air and no one has any clue what's going on, because no one has any clue what a normal subgroup actually should look like.

Other sources like to also write the so-called first isomorphism theorem.[2] It goes like this.

Theorem 3.6.1 (First isomorphism theorem)
Let $\phi\colon G \to H$ be a homomorphism. Then $G/\ker\phi$ is isomorphic to $\phi^{\text{img}}(G)$.

To me, this is just a clumsier way of stating the same idea.

About the only merit this claim has is that if ϕ is injective, then the image $\phi^{\text{img}}(G)$ is an *isomorphic copy* of G inside the group H. (Try to see this directly!) This is a pattern we'll often see in other branches of mathematics: whenever we have an *injective structure-preserving map*, often the image of this map will be some "copy" of G. (Here "structure" refers to the group multiplication, but we'll see some more other examples of "types of objects" later!)

In that sense an injective homomorphism $\phi\colon G \hookrightarrow H$ is an *embedding* of G into H.

§3.7 A few harder problems to think about

Problem 3A (18.701 at MIT)**.** Determine all groups G for which the map $\phi\colon G \to G$ defined by

$$\phi(g) = g^2$$

is a homomorphism.

Problem 3B. Consider the dihedral group $G = D_{10}$.

(a) Is $H = \langle r \rangle$ a normal subgroup of G? If so, compute G/H up to isomorphism.

(b) Is $H = \langle s \rangle$ a normal subgroup of G? If so, compute G/H up to isomorphism.

[2]There is a second and third isomorphism theorem. But four years after learning about them, I *still* don't remember what they are. So I'm guessing they weren't very important.

Problem 3C. Does S_4 have a normal subgroup of order 3?

Problem 3D. Let G and H be finite groups, where $|G| = 1000$ and $|H| = 999$. Show that a homomorphism $G \to H$ must be trivial.

Problem 3E. Let $\mathbb{C}^\times$ denote the nonzero complex numbers under multiplication. Show that there are five homomorphisms $\mathbb{Z}/5\mathbb{Z} \to \mathbb{C}^\times$ but only two homomorphisms $D_{10} \to \mathbb{C}^\times$, even though $\mathbb{Z}/5\mathbb{Z}$ is a subgroup of D_{10}.

Problem 3F. Find a non-abelian group G such that every subgroup of G is normal. (These groups are called **Hamiltonian**.)

Problem 3G (PRIMES entrance exam, 2018)**.** Let G be a group with presentation given by

$$G = \left\langle a, b, c \mid ab = c^2a^4,\ bc = ca^6,\ ac = ca^8,\ c^{2018} = b^{2019} \right\rangle.$$

Determine the order of G.

Problem 3H (Homophony group)**.** The homophony group (of English) is the group with 26 generators a, b, $\ldots$, z and one relation for every pair of English words which sound the same. For example $knight = night$ (and hence $k = 1$). Prove that the group is trivial.

4 Rings and ideals

§4.1 Some motivational metaphors about rings vs groups

In this chapter we'll introduce the notion of a **commutative ring** R. It is a larger structure than a group: it will have two operations addition and multiplication, rather than just one. We will then immediately define a **ring homomorphism** $R \to S$ between pairs of rings.

This time, instead of having normal subgroups $H \trianglelefteq G$, rings will instead have subsets $I \subseteq R$ called **ideals**, which are not themselves rings but satisfy some niceness conditions. We will then show you how to define R/I, in analogy to G/H as before. Finally, like with groups, we will talk a bit about how to generate ideals.

Here is a possibly helpful table of analogies to help you keep track:

	Group	Ring
Notation	G	R
Operations	$\cdot$	$+$, $\times$
Commutativity	only if abelian	for us, always
Sub-structure	subgroup	(not discussed)
Homomorphism	grp hom. $G \to H$	ring hom. $R \to S$
Kernel	normal subgroup	ideal
Quotient	G/H	R/I

§4.2 (Optional) Pedagogical notes on motivation

I wrote most of these examples with a number theoretic eye in mind; thus if you liked elementary number theory, a lot of your intuition will carry over. Basically, we'll try to generalize properties of the ring $\mathbb{Z}$ to any abelian structure in which we can also multiply. That's why, for example, you can talk about "irreducible polynomials in $\mathbb{Q}[x]$" in the same way you can talk about "primes in $\mathbb{Z}$", or about "factoring polynomials modulo p" in the same way we can talk "unique factorization in $\mathbb{Z}$". Even if you only care about $\mathbb{Z}$ (say, you're a number theorist), this has a lot of value: I assure you that trying to solve $x^n + y^n = z^n$ (for $n > 2$) requires going into a ring other than $\mathbb{Z}$!

Thus for all the sections that follow, keep $\mathbb{Z}$ in mind as your prototype.

I mention this here because commutative algebra is *also* closely tied to algebraic geometry. Lots of the ideas in commutative algebra have nice "geometric" interpretations that motivate the definitions, and these connections are explored in the corresponding part later. So, I want to admit outright that this is not the only good way (perhaps not even the most natural one) of motivating what is to follow.

§4.3 Definition and examples of rings

Prototypical example for this section: $\mathbb{Z}$ all the way! Also $R[x]$ and various fields (next section).

Well, I guess I'll define a ring[1].

[1]Or, according to some authors, a "ring with identity"; some authors don't require rings to have multiplicative identity. For us, "ring" always means "ring with 1".

Definition 4.3.1. A **ring** is a triple $(R, +, \times)$, the two operations usually called addition and multiplication, such that

(i) $(R, +)$ is an abelian group, with identity 0_R, or just 0.

(ii) $\times$ is an associative, binary operation on R with some identity, written 1_R or just 1.

(iii) Multiplication distributes over addition.

The ring R is **commutative** if $\times$ is commutative.

Abuse of Notation 4.3.2. As usual, we will abbreviate $(R, +, \times)$ to R.

Abuse of Notation 4.3.3. For simplicity, assume all rings are commutative for the rest of this chapter. We'll run into some noncommutative rings eventually, but for such rings we won't need the full theory of this chapter anyways.

These definitions are just here for completeness. The examples are much more important.

Example 4.3.4 (Typical rings)

(a) The sets $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are all rings with the usual addition and multiplication.

(b) The integers modulo n are also a ring with the usual addition and multiplication. We also denote it by $\mathbb{Z}/n\mathbb{Z}$.

Here is also a trivial example.

Definition 4.3.5. The **zero ring** is the ring R with a single element. We denote the zero ring by 0. A ring is **nontrivial** if it is not the zero ring.

Exercise 4.3.6 (Comedic)**.** Show that a ring is nontrivial if and only if $0_R \neq 1_R$.

Since I've defined this structure, I may as well state the obligatory facts about it.

Fact 4.3.7. For any ring R and $r \in R$, $r \cdot 0_R = 0_R$. Moreover, $r \cdot (-1_R) = -r$.

Here are some more examples of rings.

Example 4.3.8 (Product ring)

Given two rings R and S the **product ring**, denoted $R \times S$, is defined as ordered pairs (r, s) with both operations done component-wise. For example, the Chinese remainder theorem says that

$$\mathbb{Z}/15\mathbb{Z} \cong \mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/5\mathbb{Z}$$

with the isomorphism $n \bmod 15 \mapsto (n \bmod 3, n \bmod 5)$.

Remark 4.3.9 — Equivalently, we can define $R \times S$ as the abelian group $R \oplus S$, and endow it with the multiplication where $r \cdot s = 0$ for $r \in R$, $s \in S$.

Question 4.3.10. Which (r, s) is the identity element of the product ring $R \times S$?

Example 4.3.11 (Polynomial ring)

Given any ring R, the **polynomial ring** $R[x]$ is defined as the set of polynomials with coefficients in R:

$$R[x] = \left\{ a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 \mid a_0, \ldots, a_n \in R \right\}.$$

This is pronounced "R adjoin x". Addition and multiplication are done exactly in the way you would expect.

Remark 4.3.12 (Digression on division) — Happily, polynomial division also does what we expect: if $p \in R[x]$ is a polynomial, and $p(a) = 0$, then $(x - a)q(x) = p(x)$ for some polynomial q. Proof: do polynomial long division.
With that, note the caveat that

$$x^2 - 1 \equiv (x - 1)(x + 1) \pmod 8$$

has *four* roots 1, 3, 5, 7 in $\mathbb{Z}/8\mathbb{Z}$.
The problem is that $2 \cdot 4 = 0$ even though 2 and 4 are not zero; we call 2 and 4 *zero divisors* for that reason. In an *integral domain* (a ring without zero divisors), this pathology goes away, and just about everything you know about polynomials carries over. (I'll say this all again next section.)

Example 4.3.13 (Multi-variable polynomial ring)

We can consider polynomials in n variables with coefficients in R, denoted $R[x_1, \ldots, x_n]$. (We can even adjoin infinitely many x's if we like!)

Example 4.3.14 (Gaussian integers are a ring)

The **Gaussian integers** are the set of complex numbers with integer real and imaginary parts, that is

$$\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}.$$

Abuse of Notation 4.3.15 (Liberal use of adjoinment)**.** Careful readers will detect some abuse in notation here. $\mathbb{Z}[i]$ should officially be "integer-coefficient polynomials in a variable i". However, it is understood from context that $i^2 = -1$; and a polynomial in $i = \sqrt{-1}$ "is" a Gaussian integer.

Example 4.3.16 (Cube root of 2)

As another example (using the same abuse of notation):

$$\mathbb{Z}[\sqrt[3]{2}] = \left\{ a + b\sqrt[3]{2} + c\sqrt[3]{4} \mid a, b, c \in \mathbb{Z} \right\}.$$

§4.4 Fields

Prototypical example for this section: $\mathbb{Q}$ *is a field, but* $\mathbb{Z}$ *is not.*

Although we won't need to know what a field is until next chapter, they're so convenient for examples I will go ahead and introduce them now.

As you might already know, if the multiplication is invertible, then we call the ring a field. To be explicit, let me write the relevant definitions.

Definition 4.4.1. A **unit** of a ring R is an element $u \in R$ which is invertible: for some $x \in R$ we have $ux = 1_R$.

Example 4.4.2 (Examples of units)

(a) The units of $\mathbb{Z}$ are ± 1, because these are the only things which "divide 1" (which is the reason for the name "unit").

(b) On the other hand, in $\mathbb{Q}$ everything is a unit (except 0). For example, $\frac{3}{5}$ is a unit since $\frac{3}{5} \cdot \frac{5}{3} = 1$.

(c) The Gaussian integers $\mathbb{Z}[i]$ have four units: ± 1 and $\pm i$.

Definition 4.4.3. A nontrivial (commutative) ring is a **field** when all its nonzero elements are units.

Colloquially, we say that

A field is a structure where you can add, subtract, multiply, and divide.

Depending on context, they are often denoted either k, K, F.

Example 4.4.4 (First examples of fields)

(a) $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are fields, since the notion $\frac{1}{c}$ makes sense in them.

(b) If p is a prime, then $\mathbb{Z}/p\mathbb{Z}$ is a field, which we usually denote by $\mathbb{F}_p$.

The trivial ring 0 is *not* considered a field, since we require fields to be nontrivial.

§4.5 Homomorphisms

Prototypical example for this section: $\mathbb{Z} \to \mathbb{Z}/5\mathbb{Z}$ *by modding out by* 5.

This section is going to go briskly – it's the obvious generalization of all the stuff we did with quotient groups.[2]

First, we define a homomorphism and isomorphism.

Definition 4.5.1. Let $R = (R, +_R, \times_R)$ and $S = (S, +_S, \times_S)$ be rings. A **ring homomorphism** is a map $\phi \colon R \to S$ such that

[2] I once found an abstract algebra textbook which teaches rings before groups. At the time I didn't understand why, but now I think I get it – modding out by things in commutative rings is far more natural, and you can start talking about all the various flavors of rings and fields. You also have (in my opinion) more vivid first examples for rings than for groups. I actually sympathize a lot with this approach — maybe I'll convert Napkin to follow it one day.

(i) $\phi(x +_R y) = \phi(x) +_S \phi(y)$ for each $x, y \in R$.

(ii) $\phi(x \times_R y) = \phi(x) \times_S \phi(y)$ for each $x, y \in R$.

(iii) $\phi(1_R) = 1_S$.

If ϕ is a bijection then ϕ is an **isomorphism** and we say that rings R and S are **isomorphic**.

Just what you would expect. The only surprise is that we also demand $\phi(1_R)$ to go to 1_S. This condition is not extraneous: consider the map $\mathbb{Z} \to \mathbb{Z}$ called "multiply by zero".

Example 4.5.2 (Examples of homomorphisms)

(a) The identity map, as always.

(b) The map $\mathbb{Z} \to \mathbb{Z}/5\mathbb{Z}$ modding out by 5.

(c) The map $\mathbb{R}[x] \to \mathbb{R}$ by $p(x) \mapsto p(0)$ by taking the constant term.

(d) For any ring R, there is a trivial ring homomorphism $R \to 0$.

Example 4.5.3 (Non-examples of homomorphisms)

Because we require 1_R to 1_S, some maps that you might have thought were homomorphisms will fail.

(a) The map $\mathbb{Z} \to \mathbb{Z}$ by $x \mapsto 2x$ is not a ring homomorphism. Aside from the fact it sends 1 to 2, it also does not preserve multiplication.

(b) If S is a nontrivial ring, the map $R \to S$ by $x \mapsto 0$ is not a ring homomorphism, even though it preserves multiplication.

(c) There is no ring homomorphism $\mathbb{Z}/2016\mathbb{Z} \to \mathbb{Z}$ at all.

In particular, whereas for groups G and H there was always a trivial group homomorphism sending everything in G to 1_H, this is not the case for rings.

§4.6 Ideals

Prototypical example for this section: The multiples of 5 *are an ideal of* $\mathbb{Z}$.

Now, just like we were able to mod out by groups, we'd also like to define quotient rings. So once again,

Definition 4.6.1. The **kernel** of a ring homomorphism $\phi\colon R \to S$, denoted $\ker \phi$, is the set of $r \in R$ such that $\phi(r) = 0$.

In group theory, we were able to characterize the "normal" subgroups by a few obviously necessary conditions (namely, $gHg^{-1} = H$). We can do the same thing for rings, and it's in fact easier because our operations are commutative.

First, note two obvious facts:

- If $\phi(x) = \phi(y) = 0$, then $\phi(x + y) = 0$ as well. So $\ker \phi$ should be closed under addition.

- If $\phi(x) = 0$, then for any $r \in R$ we have $\phi(rx) = \phi(r)\phi(x) = 0$ too. So for $x \in \ker \phi$ and *any* $r \in R$, we have $rx \in \ker \phi$.

A (nonempty) subset $I \subseteq R$ is called an ideal if it satisfies these properties. That is,

Definition 4.6.2. A nonempty subset $I \subseteq R$ is an **ideal** if it is closed under addition, and for each $x \in I$, $rx \in I$ for all $r \in R$. It is **proper** if $I \neq R$.

Note that in the second condition, r need not be in I! So this is stronger than merely saying I is closed under multiplication.

Remark 4.6.3 — If R is not commutative, we also need the condition $xr \in I$. That is, the ideal is *two-sided*: it absorbs multiplication from both the left and the right. But since rings in Napkin are commutative we needn't worry with this distinction.

Example 4.6.4 (Prototypical example of an ideal)
Consider the set $I = 5\mathbb{Z} = \{\dots, -10, -5, 0, 5, 10, \dots\}$ as an ideal in $\mathbb{Z}$. We indeed see I is the kernel of the "take mod 5" homomorphism:

$$\mathbb{Z} \twoheadrightarrow \mathbb{Z}/5\mathbb{Z}.$$

It's clearly closed under addition, but it absorbs multiplication from *all* elements of $\mathbb{Z}$: given $15 \in I$, $999 \in \mathbb{Z}$, we get $15 \cdot 999 \in I$.

Exercise 4.6.5 (Mandatory: fields have two ideals)**.** If K is a field, show that K has exactly two ideals. What are they?

Now we claim that these conditions are sufficient. More explicitly,

Theorem 4.6.6 (Ring analog of normal subgroups)
Let R be a ring and $I \subsetneq R$. Then I is the kernel of some homomorphism if and only if it's an ideal.

Proof. It's quite similar to the proof for the normal subgroup thing, and you might try it yourself as an exercise.

Obviously the conditions are necessary. To see they're sufficient, we *define* a ring by "cosets"

$$S = \{r + I \mid r \in R\}.$$

These are the equivalence classes under $r_1 \sim r_2$ if and only if $r_1 - r_2 \in I$ (think of this as taking "mod I"). To see that these form a ring, we have to check that the addition and multiplication we put on them is well-defined. Specifically, we want to check that if $r_1 \sim s_1$ and $r_2 \sim s_2$, then $r_1 + r_2 \sim s_1 + s_2$ and $r_1 r_2 \sim s_1 s_2$. We actually already did the first part – just think of R and S as abelian groups, forgetting for the moment that we can multiply. The multiplication is more interesting.

Exercise 4.6.7 (Recommended)**.** Show that if $r_1 \sim s_1$ and $r_2 \sim s_2$, then $r_1 r_2 \sim s_1 s_2$. You will need to use the fact that I absorbs multiplication from *any* elements of R, not just those in I.

Anyways, since this addition and multiplication is well-defined there is now a surjective homomorphism $R \to S$ with kernel exactly I. □

Definition 4.6.8. Given an ideal I, we define as above the **quotient ring**

$$R/I \coloneqq \{r + I \mid r \in R\}.$$

It's the ring of these equivalence classes. This ring is pronounced "R mod I".

> **Example 4.6.9** ($\mathbb{Z}/5\mathbb{Z}$)
> The integers modulo 5 formed by "modding out additively by 5" are the $\mathbb{Z}/5\mathbb{Z}$ we have already met.

But here's an important point: just as we don't actually think of $\mathbb{Z}/5\mathbb{Z}$ as consisting of $k + 5\mathbb{Z}$ for $k = 0, \dots, 4$, we also don't really want to think about R/I as elements $r + I$. The better way to think about it is

> **R/I is the result when we declare that elements of I are all zero; that is, we "mod out by elements of I".**

For example, modding out by $5\mathbb{Z}$ means that we consider all elements in $\mathbb{Z}$ divisible by 5 to be zero. This gives you the usual modular arithmetic!

Exercise 4.6.10. Earlier, we wrote $\mathbb{Z}[i]$ for the Gaussian integers, which was a slight abuse of notation. Convince yourself that this ring could instead be written as $\mathbb{Z}[x]/(x^2+1)$, if we wanted to be perfectly formal. (We will stick with $\mathbb{Z}[i]$ though — it's more natural.) Here the shorthand $(x^2+1) \coloneqq (x^2+1)\mathbb{Z}[x] = \{(x^2+1)f \mid f \in \mathbb{Z}[x]\}$ denotes the ideal of multiples of x^2+1 within $\mathbb{Z}[x]$.
Figure out the analogous formalization of $\mathbb{Z}[\sqrt[3]{2}]$.

§4.7 Generating ideals

Prototypical example for this section: In $\mathbb{Z}$, the ideals are all of the form (n).

Let's give you some practice with ideals.

An important piece of intuition is that once an ideal contains a unit, it contains 1, and thus must contain the entire ring. That's why the notion of "proper ideal" is useful language. To expand on that:

> **Proposition 4.7.1** (Proper ideal $\iff$ no units)
> Let R be a ring and $I \subseteq R$ an ideal. Then I is proper (i.e. $I \neq R$) if and only if it contains no units of R.

Proof. Suppose I contains a unit u, i.e. an element u with an inverse u^{-1}. Then it contains $u \cdot u^{-1} = 1$, and thus $I = R$. Conversely, if I contains no units, it is obviously proper. □

As a consequence, if K is a field, then its only ideals are (0) and K (this was Exercise 4.6.5). So for our practice purposes, we'll be working with rings that aren't fields.

First practice: $\mathbb{Z}$.

Exercise 4.7.2. Show that the only ideals of $\mathbb{Z}$ are precisely those sets of the form $n\mathbb{Z}$, where n is a nonnegative integer.

Thus, while ideals of fields are not terribly interesting, ideals of $\mathbb{Z}$ look eerily like elements of $\mathbb{Z}$. Let's make this more precise.

Definition 4.7.3. Let R be a ring. The **ideal generated** by a set of elements $x_1, \dots, x_n \in R$ is denoted by $I = (x_1, x_2, \dots, x_n)$ and given by

$$I = \{r_1 x_1 + \cdots + r_n x_n \mid r_i \in R\}.$$

One can think of this as "the smallest ideal containing all the x_i".

The analogy of putting the $\{x_i\}$ in a sealed box and shaking vigorously kind of works here too.

Remark 4.7.4 (Linear algebra digression) — If you know linear algebra, you can summarize this as: an ideal is an R-module. The ideal $(x_1, \dots, x_n)$ is the submodule spanned by $x_1, \dots, x_n$.

In particular, if $I = (x)$ then I consists of exactly the "multiples of x", i.e. numbers of the form rx for $r \in R$.

Remark 4.7.5 — We can also apply this definition to infinite generating sets, as long as only finitely many of the r_i are not zero (since infinite sums don't make sense in general).

Example 4.7.6 (Examples of generated ideals)

(a) As $(n) = n\mathbb{Z}$ for all $n \in \mathbb{Z}$, every ideal in $\mathbb{Z}$ is of the form (n).

(b) In $\mathbb{Z}[i]$, we have $(5) = \{5a + 5bi \mid a, b \in \mathbb{Z}\}$.

(c) In $\mathbb{Z}[x]$, the ideal (x) consists of polynomials with zero constant terms.

(d) In $\mathbb{Z}[x, y]$, the ideal (x, y) again consists of polynomials with zero constant terms.

(e) In $\mathbb{Z}[x]$, the ideal $(x, 5)$ consists of polynomials whose constant term is divisible by 5.

Question 4.7.7. Please check that the set $I = \{r_1 x_1 + \cdots + r_n x_n \mid r_i \in R\}$ is indeed always an ideal (closed under addition, and absorbs multiplication).

Now suppose $I = (x_1, \dots, x_n)$. What does R/I look like? According to what I said at the end of the last section, it's what happens when we "mod out" by each of the elements x_i. For example...

Example 4.7.8 (Modding out by generated ideals)

(a) Let $R = \mathbb{Z}$ and $I = (5)$. Then R/I is literally $\mathbb{Z}/5\mathbb{Z}$, or the "integers modulo 5": it is the result of declaring $5 = 0$.

(b) Let $R = \mathbb{Z}[x]$ and $I = (x)$. Then R/I means we send x to zero; hence $R/I \cong \mathbb{Z}$ as given any polynomial $p(x) \in R$, we simply get its constant term.

(c) Let $R = \mathbb{Z}[x]$ again and now let $I = (x-3)$. Then R/I should be thought of as the quotient when $x - 3 \equiv 0$, that is, $x \equiv 3$. So given a polynomial $p(x)$ its image after we mod out should be thought of as $p(3)$. Again $R/I \cong \mathbb{Z}$, but in a different way.

(d) Finally, let $I = (x-3, 5)$. Then R/I not only sends x to three, but also 5 to zero. So given $p \in R$, we get $p(3) \pmod 5$. Then $R/I \cong \mathbb{Z}/5\mathbb{Z}$.

Remark 4.7.9 (Mod notation) — By the way, given an ideal I of a ring R, it's totally legit to write

$$x \equiv y \pmod I$$

to mean that $x - y \in I$. Everything you learned about modular arithmetic carries over.

§4.8 Principal ideal domains

Prototypical example for this section: $\mathbb{Z}$ is a PID, $\mathbb{Z}[x]$ is not. $\mathbb{C}[x]$ is a PID, $\mathbb{C}[x,y]$ is not.

What happens if we put multiple generators in an ideal, like $(10, 15) \subseteq \mathbb{Z}$? Well, we have by definition that $(10, 15)$ is given as a set by

$$(10, 15) \coloneqq \{10x + 15y \mid x, y \in \mathbb{Z}\}.$$

If you're good at number theory you'll instantly recognize this as $5\mathbb{Z} = (5)$. Surprise! In $\mathbb{Z}$, the ideal (a, b) is exactly $\gcd(a, b)\mathbb{Z}$. And that's exactly the reason you often see the GCD of two numbers denoted (a, b).

We call such an ideal (one generated by a single element) a **principal ideal**. So, in $\mathbb{Z}$, every ideal is principal. But the same is not true in more general rings.

Example 4.8.1 (A non-principal ideal)

In $\mathbb{Z}[x]$, $I = (x, 2015)$ is *not* a principal ideal.
For if $I = (f)$ for some polynomial $f \in I$ then f divides x and 2015. This can only occur if $f = \pm 1$, but then I contains ± 1, which it does not.

A ring with the property that all its ideals are principal is called a **principal ideal ring**. We like this property because they effectively let us take the "greatest common factor" in a similar way as the GCD in $\mathbb{Z}$.

In practice, we actually usually care about so-called **principal ideal domains (PID's)**. But we haven't defined what a domain is yet. Nonetheless, all the examples below are actually PID's, so we will go ahead and use this word for now, and tell you what the additional condition is in the next chapter.

Example 4.8.2 (Examples of PID's)

To reiterate, for now you should just verify that these are principal ideal rings, even though we are using the word PID.

(a) As we saw, $\mathbb{Z}$ is a PID.

(b) As we also saw, $\mathbb{Z}[x]$ is not a PID, since $I = (x, 2015)$ for example is not principal.

(c) It turns out that for a field k the ring $k[x]$ is always a PID. For example, $\mathbb{Q}[x]$, $\mathbb{R}[x]$, $\mathbb{C}[x]$ are PID's.

If you want to try and prove this, first prove an analog of Bézout's lemma, which implies the result.

(d) $\mathbb{C}[x, y]$ is not a PID, because (x, y) is not principal.

§4.9 Noetherian rings

Prototypical example for this section: $\mathbb{Z}[x_1, x_2, \dots]$ *is not Noetherian, but most reasonable rings are. In particular polynomial rings are. (Equivalently, only weirdos care about non-Noetherian rings).*

If it's too much to ask that an ideal is generated by *one* element, perhaps we can at least ask that our ideals are generated by *finitely many* elements. Unfortunately, in certain weird rings this is also not the case.

Example 4.9.1 (Non-Noetherian ring)

Consider the ring $R = \mathbb{Z}[x_1, x_2, x_3, \dots]$ which has *infinitely* many free variables. Then the ideal $I = (x_1, x_2, \dots) \subseteq R$ cannot be written with a finite generating set.

Nonetheless, most "sane" rings we work in *do* have the property that their ideals are finitely generated. We now name such rings and give two equivalent definitions:

Proposition 4.9.2 (The equivalent definitions of a Noetherian ring)

For a ring R, the following are equivalent:

(a) Every ideal I of R is finitely generated (i.e. can be written with a finite generating set).

(b) There does *not* exist an infinite ascending chain of ideals

$$I_1 \subsetneq I_2 \subsetneq I_3 \subsetneq \dots.$$

The absence of such chains is often called the **ascending chain condition**.

Such rings are called **Noetherian**.

Example 4.9.3 (Non-Noetherian ring breaks ACC)
In the ring $R = \mathbb{Z}[x_1, x_2, x_3, \dots]$ we have an infinite ascending chain

$$(x_1) \subsetneq (x_1, x_2) \subsetneq (x_1, x_2, x_3) \subsetneq \dots.$$

From the example, you can kind of see why the proposition is true: from an infinitely generated ideal you can extract an ascending chain by throwing elements in one at a time. I'll leave the proof to you if you want to do it.[3]

Question 4.9.4. Why are fields Noetherian? Why are PID's (such as $\mathbb{Z}$) Noetherian?

This leaves the question: is our prototypical non-example of a PID, $\mathbb{Z}[x]$, a Noetherian ring? The answer is a glorious yes, according to the celebrated Hilbert basis theorem.

Theorem 4.9.5 (Hilbert basis theorem)
Given a Noetherian ring R, the ring $R[x]$ is also Noetherian. Thus by induction, $R[x_1, x_2, \dots, x_n]$ is Noetherian for any integer n.

The proof of this theorem is really olympiad flavored, so I couldn't possibly spoil it – I've left it as a problem at the end of this chapter.

Noetherian rings really shine in algebraic geometry, and it's a bit hard for me to motivate them right now, other than to say "most rings you'll encounter are Noetherian". Please bear with me!

§4.10 A few harder problems to think about

Problem 4A. The ring $R = \mathbb{R}[x]/(x^2+1)$ is one that you've seen before. What is its name?

Problem 4B. Show that $\mathbb{C}[x]/(x^2 - x) \cong \mathbb{C} \times \mathbb{C}$.

Problem 4C. In the ring $\mathbb{Z}$, let $I = (2016)$ and $J = (30)$. Show that $I \cap J$ is an ideal of $\mathbb{Z}$ and compute its elements.

Problem 4D*. Let R be a ring and I an ideal. Find an inclusion-preserving bijection between

- ideals of R/I, and
- ideals of R which contain I.

Problem 4E. Let R be a ring.

(a) Prove that there is exactly one ring homomorphism $\mathbb{Z} \to R$.

(b) Prove that the number of ring homomorphisms $\mathbb{Z}[x] \to R$ is equal to the number of elements of R.

Problem 4F. Prove the Hilbert basis theorem, Theorem 4.9.5.

[3] On the other hand, every undergraduate class in this topic I've seen makes you do it as homework. Admittedly I haven't gone to that many such classes.

Problem 4G (USA Team Selection Test 2016)**.** Let $\mathbb{F}_p$ denote the integers modulo a fixed prime number p. Define $\Psi \colon \mathbb{F}_p[x] \to \mathbb{F}_p[x]$ by

$$\Psi\left(\sum_{i=0}^{n} a_i x^i\right) = \sum_{i=0}^{n} a_i x^{p^i}.$$

Let S denote the image of Ψ.

(a) Show that S is a ring with addition given by polynomial addition, and multiplication given by *function composition*.

(b) Prove that $\Psi \colon \mathbb{F}_p[x] \to S$ is then a ring isomorphism.

5 Flavors of rings

We continue our exploration of rings by considering some nice-ness properties that rings or ideals can satisfy, which will be valuable later on. As before, number theory is interlaced as motivation. I guess I can tell you at the outset what the completed table is going to look like, so you know what to expect.

Ring noun	Ideal adjective	Relation
PID	principal	R is a PID $\iff$ R is an integral domain, and every I is principal
Noetherian ring	finitely generated	R is Noetherian $\iff$ every I is fin. gen.
field	maximal	R/I is a field $\iff$ I is maximal
integral domain	prime	R/I is an integral domain $\iff$ I is prime

§5.1 Fields

Prototypical example for this section: $\mathbb{Q}$ *is a field, but* $\mathbb{Z}$ *is not.*

We already saw this definition last chapter: a field K is a nontrivial ring for which every nonzero element is a unit.

In particular, there are only two ideals in a field: the ideal (0), which is maximal, and the entire field K.

§5.2 Integral domains

Prototypical example for this section: $\mathbb{Z}$ *is an integral domain.*

In practice, we are often not so lucky that we have a full-fledged field. Now it would be nice if we could still conclude the zero product property: if $ab = 0$ then either $a = 0$ or $b = 0$. If our ring is a field, this is true: if $b \neq 0$, then we can multiply by b^{-1} to get $a = 0$. But many other rings we consider like $\mathbb{Z}$ and $\mathbb{Z}[x]$ also have this property, despite not having division.

Not all rings though: in $\mathbb{Z}/15\mathbb{Z}$,

$$3 \cdot 5 \equiv 0 \pmod{15}.$$

If $a, b \neq 0$ but $ab = 0$ then we say a and b are **zero divisors** of the ring R. So we give a name to such rings.

Definition 5.2.1. A nontrivial ring with no zero divisors is called an **integral domain**.[1]

> **Question 5.2.2.** Show that a field is an integral domain.

> **Exercise 5.2.3** (Cancellation in integral domains)**.** Suppose $ac = bc$ in an integral domain, and $c \neq 0$. Show that $a = b$. (There is no c^{-1} to multiply by, so you have to use the definition.)

[1] Some authors abbreviate this to "domain", notably Artin.

Example 5.2.4 (Examples of integral domains)
Every field is an integral domain, so all the previous examples apply. In addition:

(a) $\mathbb{Z}$ is an integral domain, but it is not a field.

(b) $\mathbb{R}[x]$ is not a field, since there is no polynomial $P(x)$ with $xP(x) = 1$. However, $\mathbb{R}[x]$ is an integral domain, because if $P(x)Q(x) = 0$ then one of P or Q is zero.

(c) $\mathbb{Z}[x]$ is also an example of an integral domain. In fact, $R[x]$ is an integral domain for any integral domain R (why?).

(d) $\mathbb{Z}/n\mathbb{Z}$ is a field (hence integral domain) exactly when n is prime. When n is not prime, it is a ring but not an integral domain.

The trivial ring 0 is *not* considered an integral domain.

At this point, we go ahead and say:

Definition 5.2.5. An integral domain where all ideals are principal is called a **principal ideal domain (PID)**.

Recall that the ideal (a, b) is the ring-analog of the *gcd* operation, so essentially what this definition is saying is that: If any family of elements $\{a_i\}$ is taken, then the ideal generated by all of the a_i is in fact generated by a single element a.

In other words,

In a PID, you can take the *gcd* of any collection of elements.

The ring $\mathbb{Z}/6\mathbb{Z}$ is an example of a ring which is a principal ideal ring, but not an integral domain. As we alluded to earlier, we will never really use "principal ideal ring" in any real way: we typically will want to strengthen it to PID.

§5.3 Prime ideals

Prototypical example for this section: (5) *is a prime ideal of* $\mathbb{Z}$.

We know that every integer can be factored (up to sign) as a unique product of primes; for example $15 = 3 \cdot 5$ and $-10 = -2 \cdot 5$. You might remember the proof involves the so-called Bézout's lemma, which essentially says that $(a, b) = (\gcd(a, b))$; in other words we've carefully used the fact that $\mathbb{Z}$ is a PID.

It turns out that for general rings, the situation is not as nice as factoring elements because most rings are not PID's. The classic example of something going wrong is

$$6 = 2 \cdot 3 = \left(1 - \sqrt{-5}\right)\left(1 + \sqrt{-5}\right)$$

in $\mathbb{Z}[\sqrt{-5}]$. Nonetheless, we can sidestep the issue and talk about factoring *ideals*: somehow the example $10 = 2 \cdot 5$ should be $(10) = (2) \cdot (5)$, which says "every multiple of 10 is the product of a multiple of 2 and a multiple of 5". I'd have to tell you then how to multiply two ideals, which I do in the chapter on unique factorization.

Let's at least figure out what primes are. In $\mathbb{Z}$, we have that $p \neq 1$ is prime if whenever $p \mid xy$, either $p \mid x$ or $p \mid y$. We port over this definition to our world of ideals.

Definition 5.3.1. A *proper* ideal $I \subsetneq R$ is a **prime ideal** if whenever $xy \in I$, either $x \in I$ or $y \in I$.

The condition that I is proper is analogous to the fact that we don't consider 1 to be a prime number.

Example 5.3.2 (Examples and non-examples of prime ideals)

(a) The ideal (7) of $\mathbb{Z}$ is prime.

(b) The ideal (8) of $\mathbb{Z}$ is not prime, since $2 \cdot 4 = 8$.

(c) The ideal (x) of $\mathbb{Z}[x]$ is prime.

(d) The ideal (x^2) of $\mathbb{Z}[x]$ is not prime, since $x \cdot x = x^2$.

(e) The ideal $(3, x)$ of $\mathbb{Z}[x]$ is prime. This is actually easiest to see using Theorem 5.3.5 below.

(f) The ideal $(5) = 5\mathbb{Z} + 5i\mathbb{Z}$ of $\mathbb{Z}[i]$ is not prime, since the elements $3+i$ and $3-i$ have product $10 \in (5)$, yet neither is itself in (5).

Remark 5.3.3 — Ideals have the nice property that they get rid of "sign issues". For example, in $\mathbb{Z}$, do we consider -3 to be a prime? When phrased with ideals, this annoyance goes away: $(-3) = (3)$. More generally, for a ring R, talking about ideals lets us ignore multiplication by a unit. (Note that -1 is a unit in $\mathbb{Z}$.)

Exercise 5.3.4. What do you call a ring R for which the zero ideal (0) is prime?

We also have:

Theorem 5.3.5 (Prime ideal $\iff$ quotient is integral domain)
An ideal I is prime if and only if R/I is an integral domain.

Exercise 5.3.6 (Mandatory)**.** Convince yourself the theorem is true; it is just definition chasing. (A possible start is to consider $R = \mathbb{Z}$ and $I = (15)$.)

I now must regrettably inform you that unique factorization is still not true even with the notion of a "prime" ideal (though again I haven't told you how to multiply two ideals yet). But it will become true with some additional assumptions that will arise in algebraic number theory (relevant buzzword: Dedekind domain).

§5.4 Maximal ideals

Prototypical example for this section: The ideal $(x, 5)$ is maximal in $\mathbb{Z}[x]$, by quotient-ing.

Here's another flavor of an ideal.

Definition 5.4.1. A proper ideal I of a ring R is **maximal** if it is not contained in any other proper ideal.

Example 5.4.2 (Examples of maximal ideals)

(a) The ideal $I = (7)$ of $\mathbb{Z}$ is maximal, because if an ideal J contains 7 and an element n not in I it must contain $\gcd(7, n) = 1$, and hence $J = \mathbb{Z}$.

(b) The ideal (x) is *not* maximal in $\mathbb{Z}[x]$, because it's contained in $(x, 5)$ (among others).

(c) On the other hand, $(x, 5)$ is indeed maximal in $\mathbb{Z}[x]$. This is actually easiest to verify using Theorem 5.4.4 below.

(d) Also, (x) is maximal in $\mathbb{C}[x]$, again appealing to Theorem 5.4.4 below.

Exercise 5.4.3. What do you call a ring R for which the zero ideal (0) is maximal?

There's an analogous theorem to the one for prime ideals.

Theorem 5.4.4 (I maximal $\iff$ R/I field)

An ideal I is maximal if and only if R/I is a field.

Proof. A ring is a field if and only if (0) is the only maximal ideal. So this follows by Problem 4D*. □

Corollary 5.4.5 (Maximal ideals are prime)

If I is a maximal ideal of a ring R, then I is prime.

Proof. If I is maximal, then R/I is a field, hence an integral domain, so I is prime. □

In practice, because modding out by generated ideals is pretty convenient, this is a very efficient way to check whether an ideal is maximal.

Example 5.4.6 (Modding out in $\mathbb{Z}[x]$)

(a) This instantly implies that $(x, 5)$ is a maximal ideal in $\mathbb{Z}[x]$, because if we mod out by x and 5 in $\mathbb{Z}[x]$, we just get $\mathbb{F}_5$, which is a field.

(b) On the other hand, modding out by just x gives $\mathbb{Z}$, which is an integral domain but not a field; that's why (x) is prime but not maximal.

As we saw, any maximal ideal is prime. But now note that $\mathbb{Z}$ has the special property that all of its nonzero prime ideals are also maximal. It's with this condition and a few other minor conditions that you get a so-called *Dedekind domain* where prime factorization of ideals *does* work. More on that later.

§5.5 Field of fractions

Prototypical example for this section: $\operatorname{Frac}(\mathbb{Z}) = \mathbb{Q}$.

As long as we are here, we take the time to introduce a useful construction that turns any integral domain into a field.

Definition 5.5.1. Given an integral domain R, we define its **field of fractions** or **fraction field** $\text{Frac}(R)$ as follows: it consists of elements a/b, where $a, b \in R$ and $b \neq 0$. We set $a/b \sim c/d$ if and only if $bc = ad$. Addition and multiplication is defined by

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$$
$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

In fact everything you know about $\mathbb{Q}$ basically carries over by analogy. You can prove if you want that this indeed a field, but considering how comfortable we are that $\mathbb{Q}$ is well-defined, I wouldn't worry about it...

Definition 5.5.2. Let k be a field. We define $k(x) = \text{Frac}(k[x])$ (read "k of x"), and call it the **field of rational functions**.

Example 5.5.3 (Examples of fraction fields)

(a) By *definition*, $\text{Frac}(\mathbb{Z}) = \mathbb{Q}$.

(b) The field $\mathbb{R}(x)$ consists of rational functions in x:

$$\mathbb{R}(x) = \left\{ \frac{f(x)}{g(x)} \mid f, g \in \mathbb{R}[x] \right\}.$$

For example, $\frac{2x}{x^2-3}$ might be a typical element.

Example 5.5.4 (Gaussian rationals)

Just like we defined $\mathbb{Z}[i]$ by abusing notation, we can also write $\mathbb{Q}(i) = \text{Frac}(\mathbb{Z}[i])$. Officially, it should consist of

$$\mathbb{Q}(i) = \left\{ \frac{f(i)}{g(i)} \mid g(i) \neq 0 \right\}$$

for polynomials f and g with rational coefficients. But since $i^2 = -1$ this just leads to

$$\mathbb{Q}(i) = \left\{ \frac{a+bi}{c+di} \mid a, b, c, d \in \mathbb{Q}, (c,d) \neq (0,0) \right\}.$$

And since $\frac{1}{c+di} = \frac{c-di}{c^2+d^2}$ we end up with

$$\mathbb{Q}(i) = \{a + bi \mid a, b \in \mathbb{Q}\}.$$

§5.6 Unique factorization domains (UFD's)

Prototypical example for this section: $\mathbb{Z}$ and polynomial rings in general.

Here is one stray definition that will be important for those with a number-theoretic inclination. Over the positive integers, we have a fundamental theorem of arithmetic, stating that every integer is uniquely the product of prime numbers.

We can even make an analogous statement in $\mathbb{Z}$ or $\mathbb{Z}[i]$, if we allow representations like $6 = (-2)(-3)$ and so on. The trick is that we only consider everything *up to units*; so $6 = (-2)(-3) = 2 \cdot 3$ are considered the same.

The general definition goes as follows.

Definition 5.6.1. A nonzero non-unit of an integral domain R is **irreducible** if it cannot be written as the product of two non-units.

An integral domain R is a **unique factorization domain** if every nonzero non-unit of R can be written as the product of irreducible elements, which is unique up to multiplication by units.

Question 5.6.2. Verify that $\mathbb{Z}$ is a UFD.

Example 5.6.3 (Examples of UFD's)

(a) Fields are a "degenerate" example of UFD's: every nonzero element is a unit, so there is nothing to check.

(b) $\mathbb{Z}$ is a UFD. The irreducible elements are p and $-p$, for example 5 or -17.

(c) $\mathbb{Q}[x]$ is a UFD: polynomials with rational coefficients can be uniquely factored, up to scaling by constants (as the units of $\mathbb{Q}[x]$ are just the rational numbers).

(d) $\mathbb{Z}[x]$ is a UFD.

(e) The Gaussian integers $\mathbb{Z}[i]$ turns out to be a UFD too (and this will be proved in the chapters on algebraic number theory).

(f) $\mathbb{Z}[\sqrt{-5}]$ is the classic non-example of a UFD: one may write

$$6 = 2 \cdot 3 = \left(1 - \sqrt{-5}\right)\left(1 + \sqrt{-5}\right)$$

but each of 2, 3, $1 \pm \sqrt{-5}$ is irreducible. (It turns out the right way to fix this is by considering prime *ideals* instead, and this is one big motivation for Part XIV.)

(g) Theorem we won't prove: if R is a UFD, so is $R[x]$ (and hence by induction so is $R[x, y]$, $R[x, y, z]$, ...).

We have the following theorem:

Theorem 5.6.4

Let R be a PID. Then R is a UFD.

If we look at the non-example above:

$$6 = 2 \cdot 3 = \left(1 - \sqrt{-5}\right)\left(1 + \sqrt{-5}\right)$$

The failure of $\mathbb{Z}[\sqrt{-5}]$ to be a UFD here is reflected by the fact that we cannot decompose any factor into further irreducible factor, indeed, the ideal

$$\left(2, 1 + \sqrt{-5}\right)$$

is not principal — there is no element that is the gcd of 2 and $1 + \sqrt{-5}$.

In a similar manner, we can prove that a PID is an UFD, assuming a decomposition into irreducible factors exist.

§5.7 Extra: Euclidean domains

This chapter will not be used later on, but it is of historical interest.

Recall that a PID is a ring where you can take the gcd of any family of elements.

We all know that the most popular algorithm to compute the gcd of two elements in $\mathbb{Z}$ is the Euclidean algorithm:

- Start with two integers a and b.
- If either of a and b is 0, we're done. The gcd is the nonzero element.
- Otherwise, assume $|a| \geq |b|$, divide a by b to get some remainder r such that $|r| < |b|$, replace a with r, and continue the algorithm.

This algorithm is very efficient — it can be proven that the algorithm only takes logarithmically many steps in the size of a and b — for instance, if a and b are on the order of 10^{100}, at worst 500 steps are needed.

Naturally, the following questions are raised:

> On which rings can we perform the same algorithm?

We will see that we can in fact do it on several rings! For example, the ring of Gaussian integers $\mathbb{Z}[i]$, the Eisenstein integers $\mathbb{Z}[\omega]$, and so on.

If we look at the algorithm description above, what makes the algorithm work? It's the absolute value $|\cdot|$ which is used to compare the magnitude of two numbers, and this absolute value satisfies two conditions:

- It outputs nonnegative integer values — that way, the algorithm will eventually terminate.
- For any two ring elements a and b, where $b \neq 0$, there exist some q such that $r = a - qb$ has smaller absolute value than b.

So, naturally, for any ring with a similar integer-valued function, we can perform the algorithm. We call a function $\mathrm{N}\colon R \to \mathbb{Z}_{\geq 0}$ that satisfies the two conditions above an **Euclidean norm**, and an integral domain R that has a norm an **Euclidean domain**.

Example 5.7.1 (The ring of Gaussian integers is a Euclidean domain)

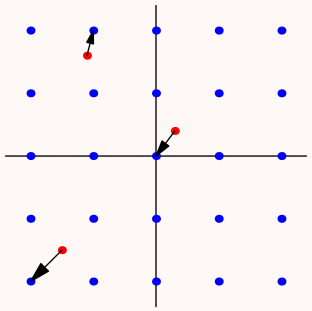
On $\mathbb{Z}[i]$, the usual norm

$$|a + bi| = a^2 + b^2$$

is a Euclidean norm.

Indeed, for any elements a and b with $b \neq 0$, we can compute the remainder r by dividing a by b, let q be the Gaussian integer that is closest to $\frac{a}{b}$ (that is, $|\frac{a}{b} - q|$ is minimized) and let $r = a - bq$, then it can be proven that $|r| < |b|$.

The proof is done by showing $|\frac{a}{b} - q| < 1$ — if we look at the lattice of points contained in $\mathbb{Z}[i]$ embedded in the complex plane, then for any value of $\frac{a}{b} \in \mathbb{C}$, rounding it to the nearest integer will move it by at most $\frac{\sqrt{2}}{2} < 1$.



Example 5.7.2 (The ring of Eisenstein integers is a Euclidean domain)

Similarly, let $\omega = \frac{1+\sqrt{3}i}{2}$ (that is $\omega^3 = -1$), then $\mathbb{Z}[\omega]$ is a Euclidean domain with the usual norm

$$|a+bi| = a^2+b^2$$

or equivalently

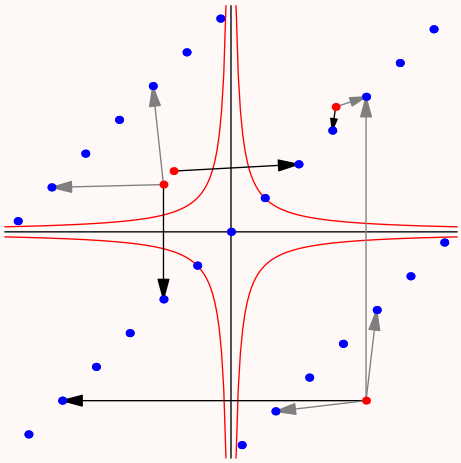
$$|a+b\omega| = a^2+ab+b^2.$$

Example 5.7.3 (The ring $\mathbb{Z}[\sqrt{11}]$ is a Euclidean domain)

As before. This time, the natural norm[a] will be:

$$\mathrm{N}_{\mathbb{Q}(\sqrt{11})/\mathbb{Q}}(a+b\sqrt{11}) = (a+b\sqrt{11})(a-b\sqrt{11}) = a^2-11b^2.$$

Since we need a Euclidean norm, we will take $\mathrm{N}(a+b\sqrt{11}) = |a^2-11b^2|$.
Given two elements a and b in $\mathbb{Z}[\sqrt{11}]$ with $b \neq 0$, we will try to compute r such that $\mathrm{N}(r) < \mathrm{N}(b)$ as $r = a - qb$ as before.
This time around, we cannot draw $\mathbb{Z}[\sqrt{11}]$ as a lattice of points — it is dense in $\mathbb{R}$ — so, each point $a+b\sqrt{11}$ will be drawn at the coordinate $(a+b\sqrt{11}, a-b\sqrt{11})$.
The set of points with norm < 1 will be drawn below.



Instead of a ball (as in imaginary quadratic fields, that is, $\mathbb{Q}(\sqrt{-d})$ for integer d), the set of points with norm 1 forms a hyperbola.
As such, rounding to the nearest point is not always the best way — nevertheless, it can be proven (by exhaustive case checking, similar to the case of $\mathbb{Z}[i]$) that for

every value of $\frac{a}{b} \in \mathbb{Q}(\sqrt{11})$, there is some $q \in \mathbb{Z}[\sqrt{11}]$ such that $\mathrm{N}(\frac{a}{b} - q) < 1$. Thus N is a Euclidean norm.

[a] See Section 54.1 for the explanation why this norm is the natural one.

That having said, sometimes the natural norm of a Euclidean domain need not be Euclidean. $\mathbb{Z}[\frac{1+\sqrt{69}}{2}]$ is the first example.

Example 5.7.4 ($\mathbb{Q}[x]$ is a Euclidean domain)

Similarly, in $\mathbb{Q}[x]$ we can let the norm be the degree of a polynomial — the polynomial division with remainder algorithm will take care of computing the gcd.

Back to the topic of PID. In a Euclidean domain, you can compute the gcd of any two elements. What about an infinite family of elements?

Turns out the situation is very nice:

Proposition 5.7.5

A Euclidean domain is a PID.

Actually, we don't need to provide an explicit algorithm to compute the gcd of an infinite family of elements — of course any such algorithm cannot terminate in a finite amount of time! — but we only need to show the gcd exists, we can cheat our way out.

Note that in the Euclidean algorithm, the norm of the elements *keep decreasing* until one of the elements become 0. So, if we're given an arbitrary family of elements, we take the ideal generated by these elements — certainly the gcd is inside that ideal — and we *take the nonzero element with the smallest norm*. This is the gcd.

With that intuition in mind, we formalize our proof:

Proof. Let I be any ideal. We need to show I is principal.

Let a be a nonzero element with smallest norm in I — such an element exists because $\mathbb{Z}_{\geq 0}$ is well-ordered.

Question 5.7.6. Show that, for every other elements $b \in I$, then $a \mid b$.

Thus, $I = (a)$, we're done. □

Example 5.7.7 (The ring $\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$ is not a Euclidean domain)

Let $R = \mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$.

With the above example in mind, what can we say about this ring?

This is in fact a principal ideal domain (we will not prove it here), but there is in fact no Euclidean norm on this ring.

We will prove the above claim. The general plan is:

- Show that the existence of a Euclidean norm implies the existence of something that we calls an **universal side divisor**.

- Show that R has no universal side divisor.

- Thus, R cannot have a Euclidean norm.

First, look at the examples above of $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$. We see that the units are the elements with the smallest norm.

So far, nothing useful — every ring has the unit 1. Next, we look at the elements with the next-smallest norm:

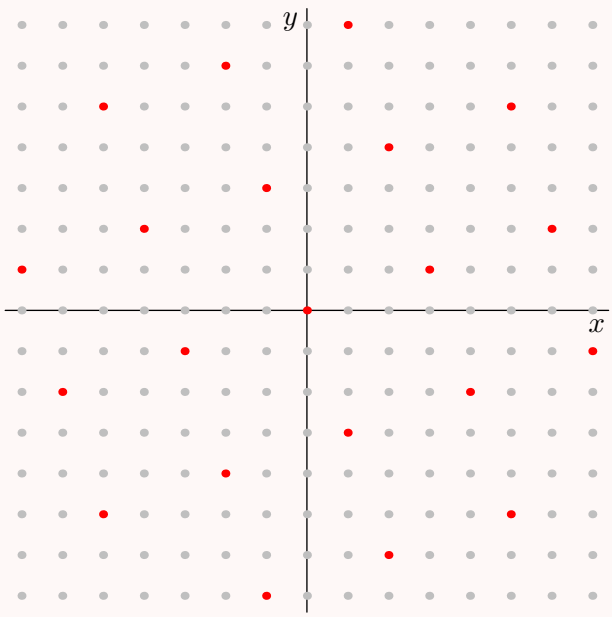
- In $\mathbb{Z}[i]$, the element $1+i$ has norm 2, and is an element with smallest norm that is not 0 or an unit. (The other elements are $\pm 1 \pm i$.)

- In $\mathbb{Z}[\omega]$, the units are $\{\pm 1, \pm\omega, \pm(\omega-1)\}$ with norm 1. An elements with next-smallest norms are $1+\omega$ with norm 3.

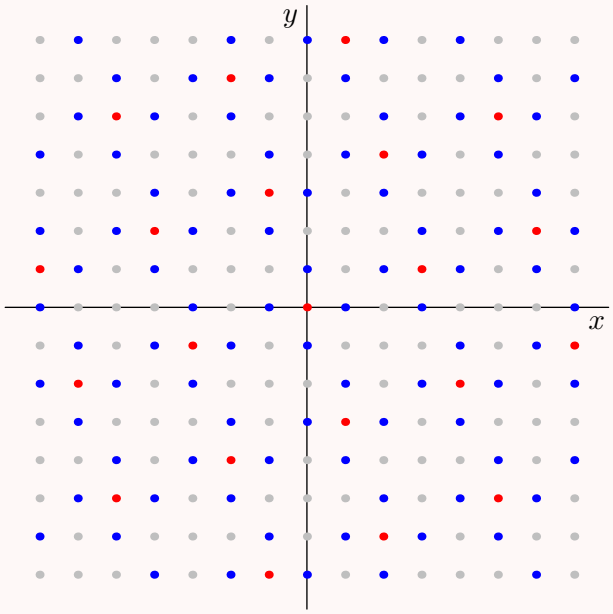
Now, in order to proceed with the proof, we have to define a *side divisor*. Recall that b is a divisor of a if there is some q such that $a = bq$. We say b is a **side divisor** (read: "almost divisor") of a if there is some q such that the remainder $a - bq$ is either 0 or an unit.

Example 5.7.8

In $\mathbb{Z}[i]$, consider $b = 3+i$. The set of numbers a for which $b \mid a$ is drawn in red below.



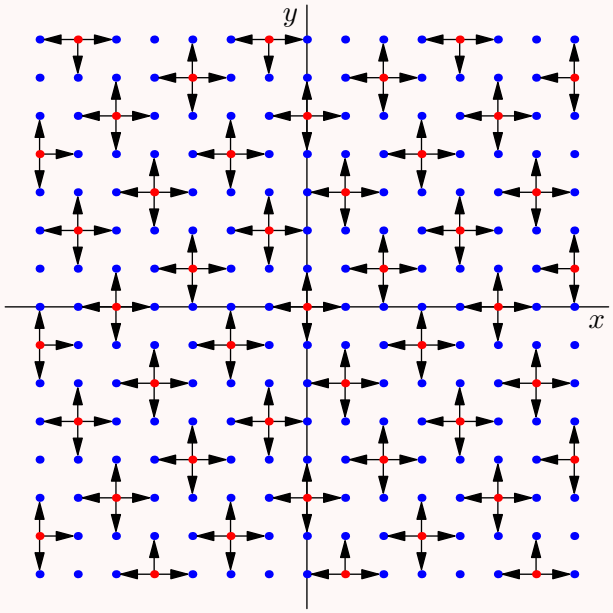
If we add an unit to these values of a, we get the set of numbers a for which b is a side divisor of a, thus give a picture to the concept of "almost divisor".
The points a where b is a side divisor of a is marked in red and blue below.



Finally, we define a **universal side divisor** to be a number b such that b is a side divisor of every element $a \in R$.

Example 5.7.9

Drawing a picture similar to the above, in $\mathbb{Z}[i]$, then $2+i$ and $1+i$ are side divisors.



Now, the connection between the two concepts considered above.

Lemma 5.7.10

In a Euclidean domain, the smallest-norm nonzero element b that is not an unit is a universal side divisor.

Proof. Just run the Euclidean algorithm between any number and b for one step, the remainder must be 0 or an unit. □

And finally,

Proposition 5.7.11

There is no universal side divisor in R.

Proof. All numbers are of the form, $\frac{a}{2} + \frac{b\sqrt{-19}}{2}$ where a and b have the same parity. The absolute value of a complex number, defined as $\frac{a^2}{4} + \frac{19b^2}{4}$, is multiplicative and is greater than 1 for all numbers in $\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$ except for $-1, 0, 1$, which have absolute values of $1, 0, 1$, respectively. Since there are no numbers in the ring with absolute values greater than 0 and less than 1, all numbers in the ring with absolute values greater than 1 do not have multiplicative inverses in the ring. Hence, the only units are 1 and -1.

When we multiply all the elements in the ring by x, the distances between pairs of elements in the complex plane multiplies by $|x|$. Initially, the distances between pairs of elements were at least 1.

- If x is real and $|x| \geq 2$, then xR doesn't contain any elements of the form $a + \frac{\sqrt{-19}}{2}$, so x is not a universal side divisor.
- If x is non-real and $|x| \geq 2$, then an element in xR can only be real if it was previously non-real. This means the elements of xR on the real axis have absolute value of at least $2|\frac{1+\sqrt{-19}}{2}| = 2\sqrt{5} \geq 4$. Hence, x is not a side divisor of 2, so x is not a universal side divisor.

Since R does not have a universal side divisor, R is not an Euclidean domain. □

§5.8 A few harder problems to think about

Not olympiad problems, but again the spirit is very close to what you might see in an olympiad.

Problem 5A. Consider the ring

$$\mathbb{Q}[\sqrt{2}] = \left\{ a + b\sqrt{2} \mid a, b \in \mathbb{Q} \right\}.$$

Is it a field?

Problem 5B (Homomorphisms from fields are injective)**.** Let K be a field and R a nontrivial ring. Prove that any homomorphism $\psi \colon K \to R$ is injective.[2]

Problem 5C* (Pre-image of prime ideals)**.** Suppose $\phi \colon R \to S$ is a ring homomorphism, and $I \subseteq S$ is a prime ideal. Prove that $\phi^{\text{pre}}(I)$ is prime as well.

Problem 5D*. Let R be an integral domain with finitely many elements. Prove that R is a field.

Problem 5E* (Krull's theorem)**.** Let R be a ring and J a proper ideal.

[2]Note that ψ cannot be the zero map for us, since we require $\psi(1_K) = 1_R$. You sometimes find different statements in the literature.

(a) Prove that if R is Noetherian, then J is contained in a maximal ideal I.

(b) Use Zorn's lemma (Chapter 88) to prove the result even if R isn't Noetherian.

Problem 5F (Spec $k[x]$)**.** Describe the prime ideals of $\mathbb{C}[x]$ and $\mathbb{R}[x]$.

Problem 5G†**.** How many prime ideals of $\mathbb{Z}[\sqrt{2017}]$ are *not* maximal ideals?

Problem 5H. Let R denote the set of rational numbers q such that, when q is written in lowest terms, the denominator is not a multiple of 5. Then R is a ring (under the usual addition and multiplication). Classify all the ideals of R. Which of these ideals are prime / maximal?

III

Basic Topology

Part III: Contents

6 Properties of metric spaces

At the end of the last chapter on metric spaces, we introduced two adjectives "open" and "closed". These are important because they'll grow up to be the *definition* for a general topological space, once we graduate from metric spaces.

To move forward, we provide a couple niceness adjectives that applies to *entire metric spaces*, rather than just a set relative to a parent space. They are "(totally) bounded" and "complete". These adjectives are specific to metric spaces, but will grow up to become the notion of *compactness*, which is, in the words of [Pu02], "the single most important concept in real analysis". At the end of the chapter, we will know enough to realize that something is amiss with our definition of homeomorphism, and this will serve as the starting point for the next chapter, when we define fully general topological spaces.

§6.1 Boundedness

Prototypical example for this section: $[0,1]$ *is bounded but* $\mathbb{R}$ *is not.*

Here is one notion of how to prevent a metric space from being a bit too large.

Definition 6.1.1. A metric space M is **bounded** if there is a constant D such that $d(p,q) \le D$ for all $p, q \in M$.

You can change the order of the quantifiers:

Proposition 6.1.2 (Boundedness with radii instead of diameters)
A metric space M is bounded if and only if for every point $p \in M$, there is a radius R (possibly depending on p) such that $d(p,q) \le R$ for all $q \in M$.

Exercise 6.1.3. Use the triangle inequality to show these are equivalent. (The names "radius" and "diameter" are a big hint!)

Example 6.1.4 (Examples of bounded spaces)

(a) Finite intervals like $[0,1]$ and (a,b) are bounded.

(b) The unit square $[0,1]^2$ is bounded.

(c) $\mathbb{R}^n$ is not bounded for any $n \ge 1$.

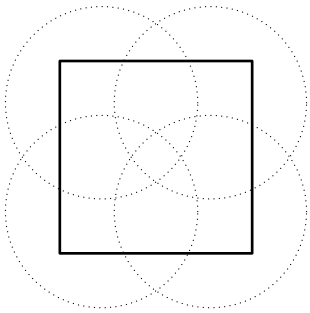
(d) A discrete space on an infinite set is bounded.

(e) $\mathbb{N}$ is not bounded, despite being homeomorphic to the discrete space!

The fact that a discrete space on an infinite set is "bounded" might be upsetting to you, so here is a somewhat stronger condition you can use:

Definition 6.1.5. A metric space is **totally bounded** if for any $\varepsilon > 0$, we can cover M with finitely many ε-neighborhoods.

For example, if $\varepsilon = 1/2$, you can cover $[0,1]^2$ by ε-neighborhoods.



Exercise 6.1.6. Show that "totally bounded" implies "bounded".

Example 6.1.7 (Examples of totally bounded spaces)

(a) A subset of $\mathbb{R}^n$ is bounded if and only if it is totally bounded.

This is for Euclidean geometry reasons: for example in $\mathbb{R}^2$ if I can cover a set by a single disk of radius 2, then I can certainly cover it by finitely many disks of radius $1/2$. (We won't prove this rigorously.)

(b) So for example $[0,1]$ or $[0,2] \times [0,3]$ is totally bounded.

(c) In contrast, a discrete space on an infinite set is not totally bounded.

§6.2 Completeness

Prototypical example for this section: $\mathbb{R}$ *is complete, but* $\mathbb{Q}$ *and* $(0,1)$ *are not.*

So far we can only talk about sequences converging if they have a limit. But consider the sequence

$$x_1 = 1,\ x_2 = 1.4,\ x_3 = 1.41,\ x_4 = 1.414, \ldots.$$

It converges to $\sqrt{2}$ in $\mathbb{R}$, of course. But it fails to converge in $\mathbb{Q}$; there is no *rational* number this sequence converges to. And so somehow, if we didn't know about the existence of $\mathbb{R}$, we would have *no idea* that the sequence (x_n) is "approaching" something.

That seems to be a shame. Let's set up a new definition to describe these sequences whose terms **get close to each other**, even if they don't approach any particular point in the space. Thus, we only want to mention the given points in the definition.

Definition 6.2.1. Let $x_1, x_2, \ldots$ be a sequence which lives in a metric space $M = (M, d_M)$. We say the sequence is **Cauchy** if for any $\varepsilon > 0$, we have

$$d_M(x_m, x_n) < \varepsilon$$

for all sufficiently large m and n.

Question 6.2.2. Show that a sequence which converges is automatically Cauchy. (Draw a picture.)

Now we can define:

Definition 6.2.3. A metric space M is **complete** if every Cauchy sequence converges.

Example 6.2.4 (Examples of complete spaces)

(a) $\mathbb{R}$ is complete. (Depending on your definition of $\mathbb{R}$, this either follows by definition, or requires some work. We won't go through this here.)

(b) The discrete space is complete, as the only Cauchy sequences are eventually constant.

(c) The closed interval $[0,1]$ is complete.

(d) $\mathbb{R}^n$ is complete as well. (You're welcome to prove this by induction on n.)

Example 6.2.5 (Non-examples of complete spaces)

(a) The rationals $\mathbb{Q}$ are not complete.

(b) The open interval $(0,1)$ is not complete, as the sequence 0.9, 0.99, 0.999, 0.9999, ... is Cauchy but does not converge.

So, metric spaces need not be complete, like $\mathbb{Q}$. But we certainly would like them to be complete, and in light of the following theorem this is not unreasonable.

Theorem 6.2.6 (Completion)

Every metric space can be "completed", i.e. made into a complete space by adding in some points.

We won't need this construction at all, so it's left as Problem 6C†.

Example 6.2.7 ($\mathbb{Q}$ completes to $\mathbb{R}$)

The completion of $\mathbb{Q}$ is $\mathbb{R}$.

(In fact, by using a modified definition of completion not depending on the real numbers, other authors often use this as the definition of $\mathbb{R}$.)

§6.3 Let the buyer beware

There is something suspicious about both these notions: neither are preserved under homeomorphism!

Example 6.3.1 (Something fishy is going on here)

Let $M = (0,1)$ and $N = \mathbb{R}$. As we saw much earlier M and N are homeomorphic. However:

- $(0,1)$ is totally bounded, but not complete.
- $\mathbb{R}$ is complete, but not bounded.

This is the first hint of something going awry with the metric. As we progress further into our study of topology, we will see that in fact *open sets and closed sets* (which we motivated by using the metric) are the notion that will really shine later on. I insist on introducing the metric first so that the standard pictures of open sets and closed sets make sense, but eventually it becomes time to remove the training wheels.

§6.4 Subspaces, and (inb4) a confusing linguistic point

Prototypical example for this section: A circle is obtained as a subspace of $\mathbb{R}^2$.

As we've already been doing implicitly in examples, we'll now say:

Definition 6.4.1. Every subset $S \subseteq M$ is a metric space in its own right, by reusing the distance function on M. We say that S is a **subspace** of M.

For example, we saw that the circle S^1 is just a subspace of $\mathbb{R}^2$.
It thus becomes important to distinguish between

(i) **"absolute" adjectives** like "complete" or "bounded", which can be applied to both spaces, and hence even to subsets of spaces (by taking a subspace), and

(ii) **"relative" adjectives** like "open (in M)" and "closed (in M)", which make sense only relative to a space, even though people are often sloppy and omit them.

So "$[0,1]$ is complete" makes sense, as does "$[0,1]$ is a complete subset of $\mathbb{R}$", which we take to mean "$[0,1]$ is a complete subspace of $\mathbb{R}$". This is since "complete" is an absolute adjective.

But here are some examples of ways in which relative adjectives require a little more care:

- Consider the sequence 1, 1.4, 1.41, 1.414, ... Viewed as a sequence in $\mathbb{R}$, it converges to $\sqrt{2}$. But if viewed as a sequence in $\mathbb{Q}$, this sequence does *not* converge! Similarly, the sequence 0.9, 0.99, 0.999, 0.9999 does not converge in the space $(0,1)$, although it does converge in $[0,1]$.

  The fact that these sequences fail to converge even though they "ought to" is weird and bad, and was why we defined complete spaces to begin with.

- In general, it makes no sense to ask a question like "is $[0,1]$ open?". The questions "is $[0,1]$ open in $\mathbb{R}$?" and "is $[0,1]$ open in $[0,1]$?" do make sense, however. The answer to the first question is "no" but the answer to the second question is "yes"; indeed, every space is open in itself. Similarly, $[0, \frac{1}{2})$ is an open set in the space $M = [0,1]$ because it is the ball *in* M of radius $\frac{1}{2}$ centered at 0.

- Dually, it doesn't make sense to ask "is $[0,1]$ closed"? It is closed *in* $\mathbb{R}$ and *in itself* (but every space is closed in itself, anyways).

To make sure you understand the above, here are two exercises to help you practice relative adjectives.

> **Exercise 6.4.2.** Let M be a complete metric space and let $S \subseteq M$. Prove that S is complete if and only if it is closed in M. In particular, $[0,1]$ is complete.

Exercise 6.4.3. Let $M = [0,1] \cup (2,3)$. Show that $[0,1]$ and $(2,3)$ are both open and closed in M.

This illustrates a third point: a nontrivial set can be both open and closed.[1] As we'll see in Chapter 7, this implies the space is disconnected; i.e. the only examples look quite like the one we've given above.

§6.5 A few harder problems to think about

Problem 6A† (Banach fixed point theorem)**.** Let $M = (M, d)$ be a complete metric space. Suppose $T\colon M \to M$ is a continuous map such that for any $p, q \in M$,

$$d(T(p), T(q)) \le 0.999 d(p, q).$$

(We call T a **contraction**.) Show that T has a unique fixed point.

Problem 6B (Henning Makholm, on math.SE)**.** We let M and N denote the metric spaces obtained by equipping $\mathbb{R}$ with the following two metrics:

$$d_M(x,y) = \min\{1, |x-y|\}$$
$$d_N(x,y) = |e^x - e^y|.$$

(a) Fill in the following 2×3 table with "yes" or "no" for each cell.

	Complete?	Bounded?	Totally bounded?
M			
N			

(b) Are M and N homeomorphic?

Problem 6C† (Completion of a metric space)**.** Let M be a metric space. Construct a complete metric space $\overline{M}$ such that M is a subspace of $\overline{M}$, and every nonempty open set of $\overline{M}$ contains a point of M (meaning M is **dense** in $\overline{M}$).

Problem 6D. Show that a metric space is totally bounded if and only if any sequence has a Cauchy subsequence.

Problem 6E. Prove that $\mathbb{Q}$ is not homeomorphic to any complete metric space.

[1] Which always gets made fun of.

7 Topological spaces

In Chapter 2 we introduced the notion of space by describing metrics on them. This gives you a lot of examples, and nice intuition, and tells you how you should draw pictures of open and closed sets.

However, moving forward, it will be useful to begin thinking about topological spaces in terms of just their *open sets*. (One motivation is that our fishy Example 6.3.1 shows that in some ways the notion of homeomorphism really wants to be phrased in terms of open sets, not in terms of the metric.) As we are going to see, the open sets manage to actually retain nearly all the information we need, but are simpler.[1] This will be done in just a few sections, and after that we will start describing more adjectives that we can apply to topological (and hence metric) spaces.

The most important topological notion is missing from this chapter: that of a *compact* space. It is so important that I have dedicated a separate chapter just for it.

Quick note for those who care: the adjectives "Hausdorff", "connected", and later "compact" are all absolute adjectives.

§7.1 Forgetting the metric

Recall Theorem 2.6.11:

> A function $f \colon M \to N$ of metric spaces is continuous if and only if the pre-image of every open set in N is open in M.

Despite us having defined this in the context of metric spaces, this nicely doesn't refer to the metric at all, only the open sets. As alluded to at the start of this chapter, this is a great motivation for how we can forget about the fact that we had a metric to begin with, and rather *start* with the open sets instead.

Definition 7.1.1. A **topological space** is a pair $(X, \mathcal{T})$, where X is a set of points, and $\mathcal{T}$ is the **topology**, which consists of several subsets of X, called the **open sets** of X. The topology must obey the following axioms.

- $\varnothing$ and X are both in $\mathcal{T}$.
- Finite intersections of open sets are also in $\mathcal{T}$.
- Arbitrary unions (possibly infinite) of open sets are also in $\mathcal{T}$.

So this time, the open sets are *given*. Rather than defining a metric and getting open sets from the metric, we instead start from just the open sets.

Abuse of Notation 7.1.2. We abbreviate $(X, \mathcal{T})$ by just X, leaving the topology $\mathcal{T}$ implicit. (Do you see a pattern here?)

[1] The reason I adamantly introduce metric spaces first is because I think otherwise the examples make much less sense.

Example 7.1.3 (Examples of topologies)

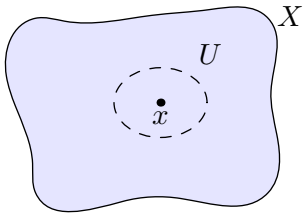
(a) Given a metric space M, we can let $\mathcal{T}$ be the open sets in the metric sense. The point is that the axioms are satisfied.

(b) In particular, **discrete space** is a topological space in which every set is open. (Why?)

(c) Given X, we can let $\mathcal{T} = \{\varnothing, X\}$, the opposite extreme of the discrete space.

Now we can port over our metric definitions.

Definition 7.1.4. An **open neighborhood**[2] of a point $x \in X$ is an open set U which contains x (see figure).



Abuse of Notation 7.1.5. Just to be perfectly clear: by an "open neighborhood" I mean *any* open set containing x. But by an "r-neighborhood" I always mean the points with distance less than r from x, and so I can only use this term if my space is a metric space.

§7.2 Re-definitions

Now that we've defined a topological space, for nearly all of our metric notions we can write down as the definition the one that required only open sets (which will of course agree with our old definitions when we have a metric space).

§7.2.i Continuity

Here was our motivating example, continuity:

Definition 7.2.1. We say function $f \colon X \to Y$ of topological spaces is **continuous** at a point $p \in X$ if the pre-image of any open neighborhood of $f(p)$ is an open neighborhood of p. The function is continuous if it is continuous at every point.

Thus homeomorphism carries over: a bijection which is continuous in both directions.

Definition 7.2.2. A **homeomorphism** of topological spaces (X, τ_X) and (Y, τ_Y) is a bijection $f \colon X \to Y$ which induces a bijection from τ_X to τ_Y: i.e. the bijection preserves open sets.

Question 7.2.3. Show that this is equivalent to f and its inverse both being continuous.

[2] In literature, a "neighborhood" refers to a set which contains some open set around x. We will not use this term, and exclusively refer to "open neighborhoods".

Therefore, any property defined only in terms of open sets is preserved by homeomorphism. Such a property is called a **topological property**. The later adjectives we define ("connected", "Hausdorff", "compact") will all be defined only in terms of open sets, so they will be topological properties.

§7.2.ii Closed sets

We saw last time there were two equivalent definitions for closed sets, but one of them relies only on open sets, and we use it:

Definition 7.2.4. In a general topological space X, we say that $S \subseteq X$ is **closed** in X if the complement $X \setminus S$ is open in X.

If $S \subseteq X$ is any set, the **closure** of S, denoted $\overline{S}$, is defined as the smallest closed set containing S.

Thus for general topological spaces, open and closed sets carry the same information, and it is entirely a matter of taste whether we define everything in terms of open sets or closed sets. In particular, you can translate axioms and properties of open sets to closed ones:

> **Question 7.2.5.** Show that the (possibly infinite) intersection of closed sets is closed while the union of finitely many closed sets is closed. (Look at complements.)

> **Exercise 7.2.6.** Show that a function is continuous if and only if the pre-image of every closed set is closed.

Mathematicians seem to have agreed that they like open sets better.

§7.2.iii Properties that don't carry over

Not everything works:

> **Remark 7.2.7** (Complete and (totally) bounded are metric properties) — The two metric properties we have seen, "complete" and "(totally) bounded", are not topological properties. They rely on a metric, so as written we cannot apply them to topological spaces. One might hope that maybe, there is some alternate definition (like we saw for "continuous function") that is just open-set based. But Example 6.3.1 showing $(0,1) \cong \mathbb{R}$ tells us that it is hopeless.

> **Remark 7.2.8** (Sequences don't work well) — You could also try to port over the notion of sequences and convergent sequences. However, this turns out to break a lot of desirable properties. Therefore I won't bother to do so, and thus if we are discussing sequences you should assume that we are working with a metric space.

§7.3 Hausdorff spaces

Prototypical example for this section: Every space that's not the Zariski topology (defined much later).

As you might have guessed, there exist topological spaces which cannot be realized as metric spaces (in other words, are not **metrizable**). One example is just to take

$X = \{a, b, c\}$ and the topology $\tau_X = \{\varnothing, \{a, b, c\}\}$. This topology is fairly "stupid": it can't tell apart any of the points a, b, c! But any metric space can tell its points apart (because $d(x, y) > 0$ when $x \neq y$).

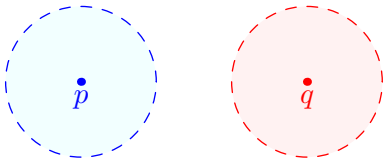
We'll see less trivial examples later, but for now we want to add a little more sanity condition onto our spaces. There is a whole hierarchy of such axioms, labelled T_n for integers n (with $n = 0$ being the weakest and $n = 6$ the strongest); these axioms are called **separation axioms**.

By far the most common hypothesis is the T_2 axiom, which bears a special name.

Definition 7.3.1. A topological space X is **Hausdorff** if for any two distinct points p and q in X, there exists an open neighborhood U of p and an open neighborhood V of q such that

$$U \cap V = \varnothing.$$

In other words, around any two distinct points we should be able to draw disjoint open neighborhoods. Here's a picture to go with above, but not much going on.



Question 7.3.2. Show that all metric spaces are Hausdorff.

I just want to define this here so that I can use this word later. In any case, basically any space we will encounter other than the Zariski topology is Hausdorff.

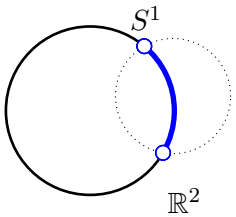
§7.4 Subspaces

Prototypical example for this section: S^1 *is a subspace of* $\mathbb{R}^2$.

One can also take subspaces of general topological spaces.

Definition 7.4.1. Given a topological space X, and a subset $S \subseteq X$, we can make S into a topological space by declaring that the open subsets of S are $U \cap S$ for open $U \subseteq X$. This is called the **subspace topology**.

So for example, if we view S^1 as a subspace of $\mathbb{R}^2$, then any open arc is an open set, because you can view it as the intersection of an open disk with S^1.



Needless to say, for metric spaces it doesn't matter which of these definitions I choose. (Proving this turns out to be surprisingly annoying, so I won't do so.)

§7.5 Connected spaces

Prototypical example for this section: $[0,1] \cup [2,3]$ *is disconnected.*

Even in metric spaces, it is possible for a set to be both open and closed.

Definition 7.5.1. A subset S of a topological space X is **clopen** if it is both closed and open in X. (Equivalently, both S and its complement are open.)

For example $\varnothing$ and the entire space are examples of clopen sets. In fact, the presence of a nontrivial clopen set other than these two leads to a so-called *disconnected* space.

> **Question 7.5.2.** Show that a space X has a nontrivial clopen set (one other than $\varnothing$ and X) if and only if X can be written as a disjoint union of two nonempty open sets.

We say X is **disconnected** if there are nontrivial clopen sets, and **connected** otherwise. To see why this should be a reasonable definition, it might help to solve Problem 7A†.

Example 7.5.3 (Disconnected and connected spaces)

(a) The metric space
$$\{(x,y) \mid x^2 + y^2 \le 1\} \cup \{(x,y) \mid (x-4)^2 + y^2 \le 1\} \subseteq \mathbb{R}^2$$
is disconnected (it consists of two disks).

(b) The space $[0,1] \cup [2,3]$ is disconnected: it consists of two segments, each of which is a clopen set.

(c) A discrete space on more than one point is disconnected, since *every* set is clopen in the discrete space.

(d) Convince yourself that the set
$$\left\{x \in \mathbb{Q} \mid x^2 < 2014\right\}$$
is a clopen subset of $\mathbb{Q}$. Hence $\mathbb{Q}$ is disconnected too – it has *gaps*.

(e) $[0,1]$ is connected.

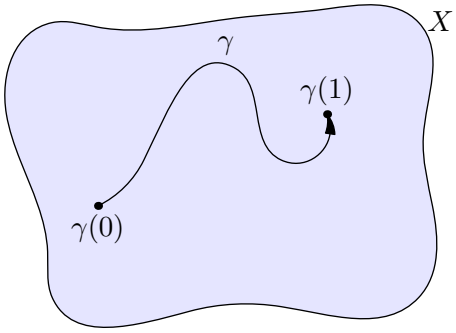
§7.6 Path-connected spaces

Prototypical example for this section: Walking around in $\mathbb{C}$.

A stronger and perhaps more intuitive notion of a connected space is a *path-connected* space. The short description: "walk around in the space".

Definition 7.6.1. A **path** in the space X is a continuous function
$$\gamma \colon [0,1] \to X.$$
Its **endpoints** are the two points $\gamma(0)$ and $\gamma(1)$.

You can think of $[0,1]$ as measuring "time", and so we'll often write $\gamma(t)$ for $t \in [0,1]$ (with t standing for "time"). Here's a picture of a path.



Question 7.6.2. Why does this agree with your intuitive notion of what a "path" is?

Definition 7.6.3. A space X is **path-connected** if any two points in it are connected by some path.

Exercise 7.6.4 (Path-connected implies connected)**.** Let $X = U \sqcup V$ be a disconnected space. Show that there is no path from a point of U to point V. (If $\gamma\colon [0,1] \to X$, then we get $[0,1] = \gamma^{\text{pre}}(U) \sqcup \gamma^{\text{pre}}(V)$, but $[0,1]$ is connected.)

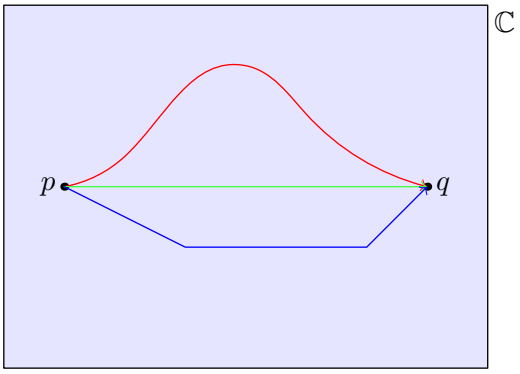
Example 7.6.5 (Examples of path-connected spaces)

- $\mathbb{R}^2$ is path-connected, since we can "connect" any two points with a straight line.
- The unit circle S^1 is path-connected, since we can just draw the major or minor arc to connect two points.

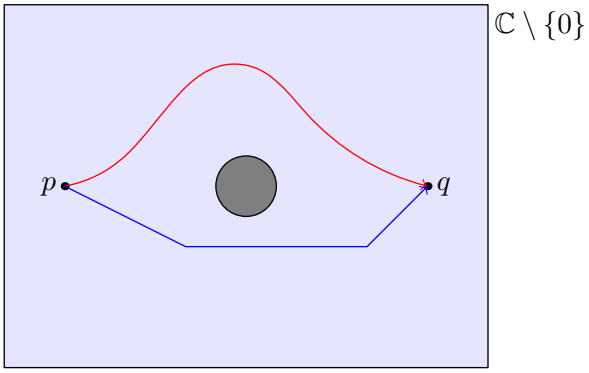
§7.7 Homotopy and simply connected spaces

Prototypical example for this section: $\mathbb{C}$ *and* $\mathbb{C} \setminus \{0\}$.

Now let's motivate the idea of homotopy. Consider the example of the complex plane $\mathbb{C}$ (which you can think of just as $\mathbb{R}^2$) with two points p and q. There's a whole bunch of paths from p to q but somehow they're not very different from one another. If I told you "walk from p to q" you wouldn't have too many questions.



So we're living happily in $\mathbb{C}$ until a meteor strikes the origin, blowing it out of existence. Then suddenly to get from p to q, people might tell you two different things: "go left around the meteor" or "go right around the meteor".



So what's happening? In the first picture, the red, green, and blue paths somehow all looked the same: if you imagine them as pieces of elastic string pinned down at p and q, you can stretch each one to any other one.

But in the second picture, you can't move the red string to match with the blue string: there's a meteor in the way. The paths are actually different.[3]

The formal notion we'll use to capture this is *homotopy equivalence*. We want to write a definition such that in the first picture, the three paths are all *homotopic*, but the two paths in the second picture are somehow not homotopic. And the idea is just continuous deformation.

Definition 7.7.1. Let α and β be paths in X whose endpoints coincide. A (path) **homotopy** from α to β is a continuous function $F \colon [0,1]^2 \to X$, which we'll write $F_s(t)$ for $s, t \in [0,1]$, such that

$$F_0(t) = \alpha(t) \text{ and } F_1(t) = \beta(t) \text{ for all } t \in [0,1]$$

and moreover

$$\alpha(0) = \beta(0) = F_s(0) \text{ and } \alpha(1) = \beta(1) = F_s(1) \text{ for all } s \in [0,1].$$

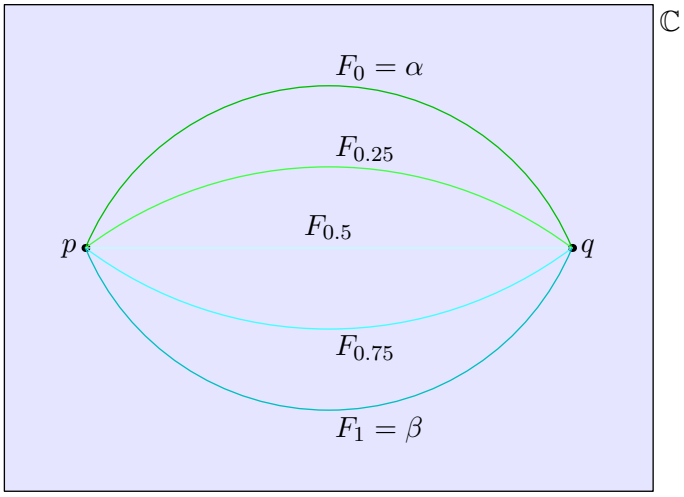
If a path homotopy exists, we say α and β are path **homotopic** and write $\alpha \simeq \beta$.

Abuse of Notation 7.7.2. While I strictly should say "path homotopy" to describe this relation between two paths, I will shorten this to just "homotopy" instead. Similarly I will shorten "path homotopic" to "homotopic".

Animated picture: https://commons.wikimedia.org/wiki/File:HomotopySmall.gif. Needless to say, $\simeq$ is an equivalence relation.

What this definition is doing is taking α and "continuously deforming" it to β, while keeping the endpoints fixed. Note that for each particular s, F_s is itself a function. So s represents time as we deform α to β: it goes from 0 to 1, starting at α and ending at β.

[3] If you know about winding numbers, you might feel this is familiar. We'll talk more about this in the chapter on the fundamental group.



Question 7.7.3. Convince yourself the above definition is right. What goes wrong when the meteor strikes?

So now I can tell you what makes $\mathbb{C}$ special:

Definition 7.7.4. A space X is **simply connected** if it's path-connected and for any points p and q, all paths from p to q are homotopic.

That's why you don't ask questions when walking from p to q in $\mathbb{C}$: there's really only one way to walk. Hence the term "simply" connected.

Question 7.7.5. Convince yourself that $\mathbb{R}^n$ is simply connected for all n.

§7.8 Bases of spaces

Prototypical example for this section: $\mathbb{R}$ has a basis of open intervals, and $\mathbb{R}^2$ has a basis of open disks.

You might have noticed that the open sets of $\mathbb{R}$ are a little annoying to describe: the prototypical example of an open set is $(0,1)$, but there are other open sets like

$$(0,1) \cup \left(1, \frac{3}{2}\right) \cup \left(2, \frac{7}{3}\right) \cup (2014, 2015).$$

Question 7.8.1. Check this is an open set.

But okay, this isn't *that* different. All I've done is taken a bunch of my prototypes and threw a bunch of $\cup$ signs at it. And that's the idea behind a basis.

Definition 7.8.2. A **basis** for a topological space X is a subset $\mathcal{B}$ of the open sets such that every open set in X is a union of some (possibly infinite) number of elements in $\mathcal{B}$.

And all we're doing is saying:

Example 7.8.3 (Basis of $\mathbb{R}$)
The open intervals form a basis of $\mathbb{R}$.

In fact, more generally we have:

Theorem 7.8.4 (Basis of metric spaces)
The r-neighborhoods form a basis of any metric space M.

Proof. Kind of silly – given an open set U, for every point p inside U, draw an r_p-neighborhood U_p contained entirely inside U. Then $\bigcup_p U_p$ is contained in U and covers every point inside it. □

Hence, an open set in $\mathbb{R}^2$ is nothing more than a union of a bunch of open disks, and so on. The point is that in a metric space, the only open sets you really ever have to worry too much about are the r-neighborhoods.

§7.9 A few harder problems to think about

Problem 7A†. Let X be a topological space. Show that there exists a nonconstant continuous function $X \to \{0,1\}$ if and only if X is disconnected (here $\{0,1\}$ is given the discrete topology).

Problem 7B*. Let X and Y be topological spaces and let $f\colon X \to Y$ be a continuous function.

(a) Show that if X is connected then so is $f^{\text{img}}(X)$.

(b) Show that if X is path-connected then so is $f^{\text{img}}(X)$.

Problem 7C (Hausdorff implies T_1 axiom)**.** Let X be a Hausdorff topological space. Prove that for any point $p \in X$ the set $\{p\}$ is closed.

Problem 7D ([Pu02], Exercise 2.56)**.** Let M be a metric space with more than one point but at most countably infinitely many points. Show that M is disconnected.

Problem 7E. Let X be a topological space. The *connected component* of a point $p \in X$ is the union of all subspaces $S \subseteq X$ which are connected and contain p.

(a) Does the connected component of a point have to be itself connected?

(b) Does the connected component of a point have to be an open subset of X?

Problem 7F (Furstenberg)**.** We declare a subset of $\mathbb{Z}$ to be open if it's the union (possibly empty or infinite) of arithmetic sequences $\{a + nd \mid n \in \mathbb{Z}\}$, where a and d are positive integers.

(a) Verify this forms a topology on $\mathbb{Z}$, called the **evenly spaced integer topology**.

(b) Prove there are infinitely many primes by considering $\bigcup_p p\mathbb{Z}$ for primes p.

Problem 7G. Prove that the evenly spaced integer topology on $\mathbb{Z}$ is metrizable. In other words, show that one can impose a metric $d\colon \mathbb{Z}^2 \to \mathbb{R}$ which makes $\mathbb{Z}$ into a metric space whose open sets are those described above.

Problem 7H. We know that any open set $U \subseteq \mathbb{R}$ is a union of open intervals (allowing $\pm\infty$ as endpoints). One can show that it's actually possible to write U as the union of *pairwise disjoint* open intervals.[4] Prove that there exists such a disjoint union with at most *countably many* intervals in it.

[4] You are invited to try and prove this, but I personally found the proof quite boring.

8 Compactness

One of the most important notions of topological spaces is that of *compactness*. It generalizes the notion of "closed and bounded" in Euclidean space to any topological space (e.g. see Problem 8F†).

For metric spaces, there are two equivalent ways of formulating compactness:

- A "natural" definition using *sequences*, called sequential compactness.
- A less natural definition using open covers.

As I alluded to earlier, sequences in metric spaces are super nice, but sequences in general topological spaces *suck* (to the point where I didn't bother to define convergence of general sequences). So it's the second definition that will be used for general spaces.

§8.1 Definition of sequential compactness

Prototypical example for this section: $[0,1]$ *is compact, but* $(0,1)$ *is not.*

To emphasize, compactness is one of the *best* possible properties that a metric space can have.

Definition 8.1.1. A **subsequence** of an infinite sequence $x_1, x_2, \ldots$ is exactly what it sounds like: a sequence $x_{i_1}, x_{i_2}, \ldots$ where $i_1 < i_2 < \cdots$ are positive integers. Note that the sequence is required to be infinite.

Another way to think about this is "selecting infinitely many terms" or "deleting some terms" of the sequence, depending on whether your glass is half empty or half full.

Definition 8.1.2. A metric space M is **sequentially compact** if every sequence has a subsequence which converges.

This time, let me give some non-examples before the examples.

Example 8.1.3 (Non-examples of compact metric spaces)

(a) The space $\mathbb{R}$ is not compact: consider the sequence $1, 2, 3, 4, \ldots$. Any subsequence explodes, hence $\mathbb{R}$ cannot possibly be compact.

(b) More generally, if a space is not bounded it cannot be compact. (You can prove this if you want.)

(c) The open interval $(0,1)$ is bounded but not compact: consider the sequence $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$. No subsequence can converge to a point in $(0,1)$ because the sequence "converges to 0".

(d) More generally, any space which is not complete cannot be compact.

Now for the examples!

Question 8.1.4. Show that a finite set is compact. (Pigeonhole Principle.)

Example 8.1.5 (Examples of compact spaces)
Here are some more examples of compact spaces. I'll prove they're compact in just a moment; for now just convince yourself they are.

(a) $[0,1]$ is compact. Convince yourself of this! Imagine having a large number of dots in the unit interval...

(b) The surface of a sphere, $S^2 = \{(x,y,z) \mid x^2+y^2+z^2 = 1\}$ is compact.

(c) The unit ball $B^2 = \{(x,y) \mid x^2+y^2 \le 1\}$ is compact.

(d) The **Hawaiian earring** living in $\mathbb{R}^2$ is compact: it consists of mutually tangent circles of radius $\frac{1}{n}$ for each n, as in Figure 8.1.

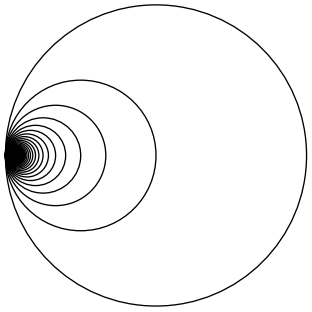


Figure 8.1: Hawaiian Earring.

To aid in generating more examples, we remark:

Proposition 8.1.6 (Closed subsets of compacts)
Closed subsets of sequentially compact sets are compact.

Question 8.1.7. Prove this. (It should follow easily from definitions.)

We need to do a bit more work for these examples, which we do in the next section.

§8.2 Criteria for compactness

Theorem 8.2.1 (Tychonoff's theorem)
If X and Y are compact spaces, then so is $X \times Y$.

Proof. Problem 8E. □

We also have:

Theorem 8.2.2 (The interval is compact)
$[0,1]$ is compact.

Proof. Killed by Problem 8F†; however, here is a sketch of a direct proof. Split $[0,1]$ into $[0,\frac{1}{2}] \cup [\frac{1}{2},1]$. By Pigeonhole, infinitely many terms of the sequence lie in the left half (say); let x_1 be the first one and then keep only the terms in the left half after x_1. Now split $[0,\frac{1}{2}]$ into $[0,\frac{1}{4}] \cup [\frac{1}{4},\frac{1}{2}]$. Again, by Pigeonhole, infinitely many terms fall in some half; pick one of them, call it x_2. Rinse and repeat. In this way we generate a sequence $x_1, x_2, \ldots$ which is Cauchy, implying that it converges since $[0,1]$ is complete. □

Now we can prove the main theorem about Euclidean space: in $\mathbb{R}^n$, compactness is equivalent to being "closed and bounded".

Theorem 8.2.3 (Bolzano-Weierstraß)
A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.

Question 8.2.4. Why does this imply the spaces in our examples are compact?

Proof. Well, look at a closed and bounded $S \subseteq \mathbb{R}^n$. Since it's bounded, it lives inside some box $[a_1,b_1] \times [a_2,b_2] \times \cdots \times [a_n,b_n]$. By Tychonoff's theorem, since each $[a_i,b_i]$ is compact the entire box is. Since S is a closed subset of this compact box, we're done. □

One really has to work in $\mathbb{R}^n$ for this to be true! In other spaces, this criterion can easily fail.

Example 8.2.5 (Closed and bounded but not compact)
Let $S = \{s_1, s_2, \ldots\}$ be any infinite set equipped with the discrete metric. Then S is closed (since all convergent sequences are constant sequences) and S is bounded (all points are a distance 1 from each other) but it's certainly not compact since the sequence $s_1, s_2, \ldots$ doesn't converge.

The Bolzano-Weierstrass theorem, which is Problem 8F†, tells you exactly which sets are compact in metric spaces in a geometric way.

§8.3 Compactness using open covers

Prototypical example for this section: $[0,1]$ *is compact.*

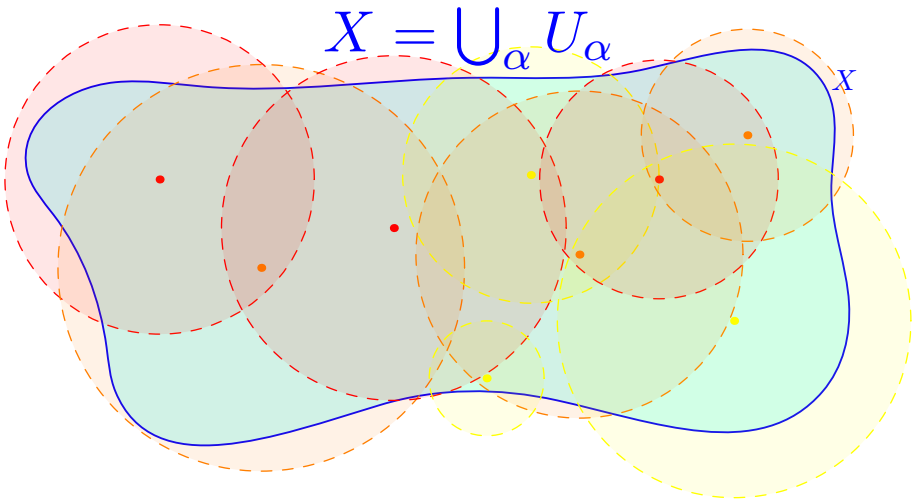
There's a second related notion of compactness which I'll now define. The following definitions might appear very unmotivated, but bear with me.

Definition 8.3.1. An **open cover** of a topological space X is a collection of open sets $\{U_\alpha\}$ (possibly infinite or uncountable) which *cover* it: every point in X lies in at least one of the U_α, so that

$$X = \bigcup U_\alpha.$$

A **subcover** is exactly what it sounds like: it takes only some of the U_α, while ensuring that X remains covered.

Some art:



Definition 8.3.2. A topological space X is **quasicompact** if *every* open cover has a finite subcover. It is **compact** if it is also Hausdorff.

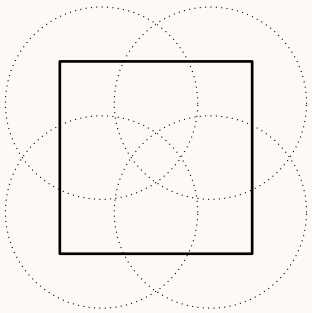
Remark 8.3.3 — The "Hausdorff" hypothesis that I snuck in is a sanity condition which is not worth worrying about unless you're working on the algebraic geometry chapters, since all the spaces you will deal with are Hausdorff. (In fact, some authors don't even bother to include it.) For example all metric spaces are Hausdorff and thus this condition can be safely ignored if you are working with metric spaces.

What does this mean? Here's an example:

Example 8.3.4 (Example of a finite subcover)

Suppose we cover the unit square $M = [0,1]^2$ by putting an open disk of diameter 1 centered at every point (trimming any overflow). This is clearly an open cover because, well, every point lies in *many* of the open sets, and in particular is the center of one.

But this is way overkill – we only need about four of these circles to cover the whole square. That's what is meant by a "finite subcover".



Why do we care? Because of this:

Theorem 8.3.5 (Sequentially compact $\iff$ compact)

A metric space M is sequentially compact if and only if it is compact.

We defer the proof to the last section.

This gives us the motivation we wanted for our definition. Sequential compactness was a condition that made sense. The open-cover definition looked strange, but it turned out to be equivalent. But we now prefer it, because we have seen that whenever possible we want to resort to open-set-only based definitions: so that e.g. they are preserved under homeomorphism.

Example 8.3.6 (An example of non-compactness)
The space $X = [0, 1)$ is not compact in either sense. We can already see it is not sequentially compact, because it is not even complete (look at $x_n = 1 - \frac{1}{n}$). To see it is not compact under the covering definition, consider the sets

$$U_m = \left[0, 1 - \frac{1}{m+1}\right)$$

for $m = 1, 2, \ldots$. Then $X = \bigcup U_i$; hence the U_i are indeed a cover. But no finite collection of the U_i's will cover X.

Question 8.3.7. Convince yourself that $[0, 1]$ *is* compact; this is a little less intuitive than it being sequentially compact.

Abuse of Notation 8.3.8. Thus, we'll never call a metric space "sequentially compact" again — we'll just say "compact". (Indeed, I kind of already did this in the previous few sections.)

§8.4 Applications of compactness

Compactness lets us reduce *infinite* open covers to finite ones. Actually, it lets us do this even if the open covers are *blithely stupid*. Very often one takes an open cover consisting of an open neighborhood of $x \in X$ for every single point x in the space; this is a huge number of open sets, and yet compactness lets us reduce to a finite set.

To give an example of a typical usage:

Proposition 8.4.1 (Compact $\implies$ totally bounded)
Let M be compact. Then M is totally bounded.

Proof using covers. For every point $p \in M$, take an ε-neighborhood of p, say U_p. These cover M for the horrendously stupid reason that each point p is at the very least covered by its open neighborhood U_p. Compactness then lets us take a finite subcover. □

Next, an important result about maps between compact spaces.

Theorem 8.4.2 (Images of compacts are compact)
Let $f \colon X \to Y$ be a continuous function, where X is compact. Then the image

$$f^{\text{img}}(X) \subseteq Y$$

is compact.

Proof using covers. Take any open cover $\{V_\alpha\}$ in Y of $f^{\text{img}}(X)$. By continuity of f, it pulls back to an open cover $\{U_\alpha\}$ of X. Thus some finite subcover of this covers X. The corresponding V's cover $f^{\text{img}}(X)$. □

Question 8.4.3. Give another proof using the sequential definitions of continuity and compactness. (This is even easier.)

Some nice corollaries of this:

Corollary 8.4.4 (Extreme value theorem)

Let X be compact and consider a continuous function $f\colon X \to \mathbb{R}$. Then f achieves a *maximum value* at some point, i.e. there is a point $p \in X$ such that $f(p) \geq f(q)$ for any other $q \in X$.

Corollary 8.4.5 (Intermediate value theorem)

Consider a continuous function $f\colon [0,1] \to \mathbb{R}$. Then the image of f is of the form $[a,b]$ for some real numbers $a \leq b$.

Sketch of Proof. The point is that the image of f is compact in $\mathbb{R}$, and hence closed and bounded. You can convince yourself that the closed sets are just unions of closed intervals. That implies the extreme value theorem.

When $X = [0,1]$, the image is also connected, so there should only be one closed interval in $f^{\text{img}}([0,1])$. Since the image is bounded, we then know it's of the form $[a,b]$. (To give a full proof, you would use the so-called *least upper bound* property, but that's a little involved for a bedtime story; also, I think $\mathbb{R}$ is boring.) □

Example 8.4.6 ($1/x$)

The compactness hypothesis is really important here. Otherwise, consider the function

$$(0,1) \to \mathbb{R} \quad \text{by} \quad x \mapsto \frac{1}{x}.$$

This function (which you plot as a hyperbola) is not bounded; essentially, you can see graphically that the issue is we can't extend it to a function on $[0,1]$ because it explodes near $x = 0$.

One last application: if M is a compact metric space, then continuous functions $f\colon M \to N$ are continuous in an especially "nice" way:

Definition 8.4.7. A function $f\colon M \to N$ of metric spaces is called **uniformly continuous** if for any $\varepsilon > 0$, there exists a $\delta > 0$ (depending only on ε) such that whenever $d_M(x,y) < \delta$ we also have $d_N(f(x), f(y)) < \varepsilon$.

The name means that for $\varepsilon > 0$, we need a δ that works for *every point* of M.

Example 8.4.8 (Uniform continuity)

(a) The functions $\mathbb{R}$ to $\mathbb{R}$ of the form $x \mapsto ax + b$ are all uniformly continuous, since one can always take $\delta = \varepsilon/|a|$ (or $\delta = 1$ if $a = 0$).

(b) Actually, it is true that a differentiable function $\mathbb{R} \to \mathbb{R}$ with a bounded derivative is uniformly continuous. (The converse is false for the reason that uniformly continuous doesn't imply differentiable at all.)

(c) The function $f\colon \mathbb{R} \to \mathbb{R}$ by $x \mapsto x^2$ is *not* uniformly continuous, since for large x, tiny δ changes to x lead to fairly large changes in x^2. (If you like, you can try to prove this formally now.)

Think $f(2017.01) - f(2017) > 40$; even when $\delta = 0.01$, one can still cause large changes in f.

(d) However, when restricted to $(0,1)$ or $[0,1]$ the function $x \mapsto x^2$ becomes uniformly continuous. (For $\varepsilon > 0$ one can now pick for example $\delta = \min\{1, \varepsilon\}/3$.)

(e) The function $(0,1) \to \mathbb{R}$ by $x \mapsto 1/x$ is *not* uniformly continuous (same reason as before).

Now, as promised:

Proposition 8.4.9 (Continuous on compact $\implies$ uniformly continuous)
If M is compact and $f\colon M \to N$ is continuous, then f is uniformly continuous.

Proof using sequences. Fix $\varepsilon > 0$, and assume for contradiction that for every $\delta = 1/k$ there exists points x_k and y_k within δ of each other but with images $\varepsilon > 0$ apart. By compactness, take a convergent subsequence $x_{i_k} \to p$. Then $y_{i_k} \to p$ as well, since the x_k's and y_k's are close to each other. So both sequences $f(x_{i_k})$ and $f(y_{i_k})$ should converge to $f(p)$ by sequential continuity, but this can't be true since the two sequences are always ε apart. □

§8.5 (Optional) Equivalence of formulations of compactness

We will prove that:

Theorem 8.5.1 (Heine-Borel for general metric spaces)
For a metric space M, the following are equivalent:

(i) Every sequence has a convergent subsequence,

(ii) The space M is complete and totally bounded, and

(iii) Every open cover has a finite subcover.

We leave the proof that (i) $\iff$ (ii) as Problem 8F†; the idea of the proof is much in the spirit of Theorem 8.2.2.

Proof that (i) and (ii) $\implies$ (iii). We prove the following lemma, which is interesting in its own right.

Lemma 8.5.2 (Lebesgue number lemma)

Let M be a compact metric space and $\{U_\alpha\}$ an open cover. Then there exists a real number $\delta > 0$, called a **Lebesgue number** for that covering, such that the δ-neighborhood of any point p lies entirely in some U_α.

Proof of lemma. Assume for contradiction that for every $\delta = 1/k$ there is a point $x_k \in M$ such that its $1/k$-neighborhood isn't contained in any U_α. In this way we construct a sequence $x_1, x_2, \ldots$; thus we're allowed to take a subsequence which converges to some x. Then for every $\varepsilon > 0$ we can find an integer n such that $d(x_n, x) + 1/n < \varepsilon$; thus the ε-neighborhood at x isn't contained in any U_α for every $\varepsilon > 0$. This is impossible, because we assumed x was covered by some open set. ■

Now, take a Lebesgue number δ for the covering. Since M is totally bounded, finitely many δ-neighborhoods cover the space, so finitely many U_α do as well. □

Proof that (iii) $\implies$ (ii). One step is immediate:

> **Question 8.5.3.** Show that the covering condition $\implies$ totally bounded.

The tricky part is showing M is complete. Assume for contradiction it isn't and thus that the sequence (x_k) is Cauchy, but it doesn't converge to any particular point.

> **Question 8.5.4.** Show that this implies for each $p \in M$, there is an ε_p-neighborhood U_p which contains at most finitely many of the points of the sequence (x_k). (You will have to use the fact that $x_k \not\to p$ and (x_k) is Cauchy.)

Now if we consider $M = \bigcup_p U_p$ we get a finite subcover of these open neighborhoods; but this finite subcover can only cover finitely many points of the sequence, by contradiction. □

§8.6 A few harder problems to think about

The later problems are pretty hard; some have the flavor of IMO 3/6-style constructions. It's important to draw lots of pictures so one can tell what's happening. Of these Problem 8F† is definitely my favorite.

Problem 8A. Show that the closed interval $[0, 1]$ and open interval $(0, 1)$ are not homeomorphic.

Problem 8B. Let X be a topological space with the discrete topology. Under what conditions is X compact?

Problem 8C (The cofinite topology is quasicompact only)**.** We let X be an infinite set and equip it with the **cofinite topology**: the open sets are the empty set and complements of finite sets. This makes X into a topological space. Show that X is quasicompact but not Hausdorff.

Problem 8D (Cantor's intersection theorem)**.** Let X be a compact topological space, and suppose

$$X = K_0 \supseteq K_1 \supseteq K_2 \supseteq \ldots$$

is an infinite sequence of nested nonempty closed subsets. Show that $\bigcap_{n \geq 0} K_n \neq \varnothing$.

Problem 8E (Tychonoff's theorem)**.** Let X and Y be compact metric spaces. Show that $X \times Y$ is compact. (This is also true for general topological spaces, but the proof is surprisingly hard, and we haven't even defined $X \times Y$ in general yet.)

Problem 8F† (Bolzano-Weierstraß theorem for general metric spaces)**.** Prove that a metric space M is sequentially compact if and only if it is complete and totally bounded.

Problem 8G (Almost Arzelà-Ascoli theorem)**.** Let $f_1, f_2, \ldots : [0,1] \to [-100, 100]$ be an **equicontinuous** sequence of functions, meaning

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall n \, \forall x, y \quad (|x-y| < \delta \implies |f_n(x) - f_n(y)| < \varepsilon)$$

Show that we can extract a subsequence $f_{i_1}, f_{i_2}, \ldots$ of these functions such that for every $x \in [0,1]$, the sequence $f_{i_1}(x)$, $f_{i_2}(x)$, $\ldots$ converges.

Problem 8H. Let $M = (M, d)$ be a bounded metric space. Suppose that whenever d' is another metric on M for which (M, d) and (M, d') are homeomorphic (i.e. have the same open sets), then d' is also bounded. Prove that M is compact.

Problem 8I. In this problem a "circle" refers to the boundary of a disk with *nonzero* radius.

(a) Is it possible to partition the plane $\mathbb{R}^2$ into disjoint circles?

(b) From the plane $\mathbb{R}^2$ we delete two distinct points p and q. Is it possible to partition the remaining points into disjoint circles?

IV

Linear Algebra

Part IV: Contents

9 Vector spaces

This is a pretty light chapter. The point of it is to define what a vector space and a basis are. These are intuitive concepts that you may already know.

§9.1 The definitions of a ring and field

Prototypical example for this section: $\mathbb{Z}$, $\mathbb{R}$, *and* $\mathbb{C}$ *are rings; the latter two are fields.*

I'll very informally define a ring/field here, in case you skipped the earlier chapter.

- A **ring** is a structure with a *commutative* addition and multiplication, as well as subtraction, like $\mathbb{Z}$. It also has an additive identity 0 and multiplicative identity 1.
- If the multiplication is invertible like in $\mathbb{R}$ or $\mathbb{C}$, (meaning $\frac{1}{x}$ makes sense for any $x \neq 0$), then the ring is called a **field**.

In fact, if you replace "field" by "$\mathbb{R}$" everywhere in what follows, you probably won't lose much. It's customary to use the letter R for rings, and k or K for fields.

Finally, in case you skipped the chapter on groups, I should also mention:

- An **additive abelian group** is a structure with a commutative addition, as well as subtraction, plus an additive identity 0. It doesn't have to have multiplication. A good example is $\mathbb{R}^3$ (with addition componentwise).

§9.2 Modules and vector spaces

Prototypical example for this section: Polynomials of degree at most n.

You intuitively know already that $\mathbb{R}^n$ is a "vector space": its elements can be added together, and there's some scaling by real numbers. Let's develop this more generally.

Fix a commutative ring R. Then informally,

An R-module is any structure where you can add two elements and scale by elements of R.

Moreover, a **vector space** is just a module whose commutative ring is actually a field. I'll give you the full definition in a moment, but first, examples...

Example 9.2.1 (Quadratic polynomials, aka my favorite example)

My favorite example of an $\mathbb{R}$-vector space is the set of polynomials of degree at most two, namely

$$\left\{ ax^2 + bx + c \mid a, b, c \in \mathbb{R} \right\}.$$

Indeed, you can add any two quadratics, and multiply by constants. You can't multiply two quadratics to get a quadratic, but that's irrelevant – in a vector space there need not be a notion of multiplying two vectors together.

In a sense we'll define later, this vector space has dimension 3 (as expected!).

Example 9.2.2 (All polynomials)

The set of *all* polynomials with real coefficients is an $\mathbb{R}$-vector space, because you can *add any two polynomials* and *scale by constants*.

Example 9.2.3 (Euclidean space)

(a) The complex numbers
$$\{a + bi \mid a, b \in \mathbb{R}\}$$
form a real vector space. As we'll see later, it has "dimension 2".

(b) The real numbers $\mathbb{R}$ form a real vector space of dimension 1.

(c) The set of 3D vectors
$$\{(x, y, z) \mid x, y, z \in \mathbb{R}\}$$
forms a real vector space, because you can add any two triples component-wise. Again, we'll later explain why it has "dimension 3".

Example 9.2.4 (More examples of vector spaces)

(a) The set
$$\mathbb{Q}[\sqrt{2}] = \left\{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\right\}$$
has a structure of a $\mathbb{Q}$-vector space in the obvious fashion: one can add any two elements, and scale by rational numbers. (It is not an $\mathbb{R}$-vector space — why?)

(b) The set
$$\{(x, y, z) \mid x + y + z = 0 \text{ and } x, y, z \in \mathbb{R}\}$$
is a 2-dimensional real vector space.

(c) The set of all functions $f \colon \mathbb{R} \to \mathbb{R}$ is also a real vector space (since the notions $f + g$ and $c \cdot f$ both make sense for $c \in \mathbb{R}$).

Now let me write the actual rules for how this multiplication behaves.

Definition 9.2.5. Let R be a commutative ring. An R-**module** starts with an additive abelian group $M = (M, +)$ whose identity is denoted $0 = 0_M$. We additionally specify a left multiplication by elements of R. This multiplication must satisfy the following properties for $r, r_1, r_2 \in R$ and $m, m_1, m_2 \in M$:

(i) $r_1 \cdot (r_2 \cdot m) = (r_1 r_2) \cdot m$.

(ii) Multiplication is distributive, meaning
$$(r_1 + r_2) \cdot m = r_1 \cdot m + r_2 \cdot m \text{ and } r \cdot (m_1 + m_2) = r \cdot m_1 + r \cdot m_2.$$

(iii) $1_R \cdot m = m$.

(iv) $0_R \cdot m = 0_M$. (This is actually extraneous; one can deduce it from the first three.)

If R is a field we say M is an R-**vector space**; its elements are called **vectors** and the members of R are called **scalars**.

Abuse of Notation 9.2.6. In the above, we're using the same symbol $+$ for the addition of M and the addition of R. Sorry about that, but it's kind of hard to avoid, and the point of the axioms is that these additions should be related. I'll try to remember to put $r \cdot m$ for the multiplication of the module and $r_1 r_2$ for the multiplication of R.

Question 9.2.7. In Example 9.2.1, I was careful to say "degree at most 2" instead of "degree 2". What's the reason for this? In other words, why is

$$\left\{ax^2 + bx + c \mid a, b, c \in \mathbb{R}, a \neq 0\right\}$$

not an $\mathbb{R}$-vector space?

A couple less intuitive but somewhat important examples...

Example 9.2.8 (Abelian groups are $\mathbb{Z}$-modules)
(Skip this example if you're not comfortable with groups.)

(a) The example of real polynomials

$$\left\{ax^2 + bx + c \mid a, b, c \in \mathbb{R}\right\}$$

is also a $\mathbb{Z}$-module! Indeed, we can add any two such polynomials, and we can scale them by integers.

(b) The set of integers modulo 100, say $\mathbb{Z}/100\mathbb{Z}$, is a $\mathbb{Z}$-module as well. Can you see how?

(c) In fact, *any* abelian group $G = (G, +)$ is a $\mathbb{Z}$-module. The multiplication can be defined by

$$n \cdot g = \underbrace{g + \cdots + g}_{n \text{ times}} \qquad (-n) \cdot g = n \cdot (-g)$$

for $n \geq 0$. (Here $-g$ is the additive inverse of g.)

Example 9.2.9 (Every ring is its own module)

(a) $\mathbb{R}$ can be thought of as an $\mathbb{R}$-vector space over itself. Can you see why?

(b) By the same reasoning, we see that *any* commutative ring R can be thought of as an R-module over itself.

§9.3 Direct sums

Prototypical example for this section: $\{ax^2 + bx + c\} = \mathbb{R} \oplus x\mathbb{R} \oplus x^2\mathbb{R}$, *and* $\mathbb{R}^3$ *is the sum of its axes.*

Let's return to Example 9.2.1, and consider

$$V = \left\{ax^2 + bx + c \mid a, b, c \in \mathbb{R}\right\}.$$

Even though I haven't told you what a dimension is, you can probably see that this vector space "should have" dimension 3. We'll get to that in a moment.

The other thing you may have noticed is that somehow the x^2, x and 1 terms don't "talk to each other". They're totally unrelated. In other words, we can consider the three sets

$$\begin{aligned} x^2\mathbb{R} &:= \left\{ax^2 \mid a \in \mathbb{R}\right\} \\ x\mathbb{R} &:= \{bx \mid b \in \mathbb{R}\} \\ \mathbb{R} &:= \{c \mid c \in \mathbb{R}\}. \end{aligned}$$

In an obvious way, each of these can be thought of as a "copy" of $\mathbb{R}$.

Then V quite literally consists of the "sums of these sets". Specifically, every element of V can be written *uniquely* as the sum of one element from each of these sets. This motivates us to write

$$V = x^2\mathbb{R} \oplus x\mathbb{R} \oplus \mathbb{R}.$$

The notion which captures this formally is the **direct sum**.

Definition 9.3.1. Let M be an R-module. Let M_1 and M_2 be subsets of M which are themselves R-modules. Then we write $M = M_1 \oplus M_2$ and say M is a **direct sum** of M_1 and M_2 if every element from M can be written uniquely as the sum of an element from M_1 and M_2.

Example 9.3.2 (Euclidean plane)
Take the vector space $\mathbb{R}^2 = \{(x, y) \mid x \in \mathbb{R}, y \in \mathbb{R}\}$. We can consider it as a direct sum of its x-axis and y-axis:

$$X = \{(x, 0) \mid x \in \mathbb{R}\} \text{ and } Y = \{(0, y) \mid y \in \mathbb{R}\}.$$

Then $\mathbb{R}^2 = X \oplus Y$.

This gives us a "top-down" way to break down modules into some disconnected components.

By applying this idea in reverse, we can also construct new vector spaces as follows. In a very unfortunate accident, the two names and notations for technically distinct things are exactly the same.

Definition 9.3.3. Let M and N be R-modules. We define the **direct sum** $M \oplus N$ to be the R-module whose elements are pairs $(m, n) \in M \times N$. The operations are given by

$$(m_1, n_1) + (m_2, n_2) = (m_1 + m_2, n_1 + n_2).$$

and

$$r \cdot (m, n) = (r \cdot m, r \cdot n).$$

For example, while we technically wrote $\mathbb{R}^2 = X \oplus Y$, since each of X and Y is a copy of $\mathbb{R}$, we might as well have written $\mathbb{R}^2 \cong \mathbb{R} \oplus \mathbb{R}$.

Abuse of Notation 9.3.4. The above illustrates an abuse of notation in the way we write a direct sum. The symbol $\oplus$ has two meanings.

- If V is a *given* space and W_1 and W_2 are subspaces, then $V = W_1 \oplus W_2$ means that "V *splits* as a direct sum $W_1 \oplus W_2$" in the way we defined above.
- If W_1 and W_2 are two *unrelated* spaces, then $W_1 \oplus W_2$ is *defined* as the vector space whose *elements* are pairs $(w_1, w_2) \in W_1 \times W_2$.

You can see that these definitions "kind of" coincide.

In this way, you can see that V should be isomorphic to $\mathbb{R} \oplus \mathbb{R} \oplus \mathbb{R}$; we had $V = x^2\mathbb{R} \oplus x\mathbb{R} \oplus \mathbb{R}$, but the 1, x, x^2 don't really talk to each other and each of the summands is really just a copy of $\mathbb{R}$ at heart.

Definition 9.3.5. We can also define, for every positive integer n, the module

$$M^{\oplus n} := \underbrace{M \oplus M \oplus \cdots \oplus M}_{n \text{ times}}.$$

§9.4 Linear independence, spans, and basis

Prototypical example for this section: $\{1, x, x^2\}$ *is a basis of* $\{ax^2 + bx + c \mid a, b, c \in \mathbb{R}\}$.

The idea of a basis, the topic of this section, gives us another way to capture the notion that

$$V = \left\{ax^2 + bx + c \mid a, b, c \in \mathbb{R}\right\}$$

is sums of copies of $\{1, x, x^2\}$. This section should be very intuitive, if technical. If you can't see why the theorems here "should" be true, you're doing it wrong.

Let M be an R-module now. We define three very classical notions that you likely are already familiar with. If not, fall upon your notion of Euclidean space or V above.

Definition 9.4.1. A **linear combination** of some vectors $v_1, \ldots, v_n$ is a sum of the form $r_1v_1 + \cdots + r_nv_n$, where $r_1, \ldots, r_n \in R$. The linear combination is called **trivial** if $r_1 = r_2 = \cdots = r_n = 0_R$, and **nontrivial** otherwise.

Definition 9.4.2. Consider a finite set of vectors $v_1, \ldots, v_n$ in a module M.

- It is called **linearly independent** if there is no nontrivial linear combination with value 0_M. (Observe that $0_M = 0 \cdot v_1 + 0 \cdot v_2 + \cdots + 0 \cdot v_n$ is always true – the assertion is that there is no other way to express 0_M in this form.)
- It is called a **generating set** if every $v \in M$ can be written as a linear combination of the $\{v_i\}$. If M is a vector space we say it is **spanning** instead.
- It is called a **basis** (plural **bases**) if every $v \in M$ can be written *uniquely* as a linear combination of the $\{v_i\}$.

The same definitions apply for an infinite set, with the proviso that all sums must be finite.

So by definition, $\{1, x, x^2\}$ is a basis for V. It's not the only one: $\{2, x, x^2\}$ and $\{x + 4, x - 2, x^2 + x\}$ are other examples of bases, though not as natural. However, the set $S = \{3 + x^2, x + 1, 5 + 2x + x^2\}$ is not a basis; it fails for two reasons:

- Note that $0 = (3 + x^2) + 2(x + 1) - (5 + 2x + x^2)$. So the set S is not linearly independent.
- It's not possible to write x^2 as a sum of elements of S. So S fails to be spanning.

With these new terms, we can say a basis is a linearly independent and spanning set.

Example 9.4.3 (More examples of bases)

(a) Regard $\mathbb{Q}[\sqrt{2}] = \left\{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\right\}$ as a $\mathbb{Q}$-vector space. Then $\{1, \sqrt{2}\}$ is a basis.

(b) If V is the set of all real polynomials, there is an infinite basis $\{1, x, x^2, \dots\}$. The condition that we only use finitely many terms just says that the polynomials must have finite degree (which is good).

(c) Let $V = \{(x, y, z) \mid x + y + z = 0 \text{ and } x, y, z \in \mathbb{R}\}$. Then we expect there to be a basis of size 2, but unlike previous examples there is no immediately "obvious" choice. Some working examples include:

- $(1, -1, 0)$ and $(1, 0, -1)$,
- $(0, 1, -1)$ and $(1, 0, -1)$,
- $(5, 3, -8)$ and $(2, -1, -1)$.

Exercise 9.4.4. Show that a set of vectors is a basis if and only if it is linearly independent and spanning. (Think about the polynomial example if you get stuck.)

Now we state a few results which assert that bases in vector spaces behave as nicely as possible.

Theorem 9.4.5 (Maximality and minimality of bases)

Let V be a vector space over some field k and take $e_1, \dots, e_n \in V$. The following are equivalent:

(a) The e_i form a basis.

(b) The e_i are spanning, but no proper subset is spanning.

(c) The e_i are linearly independent, but adding any other element of V makes them not linearly independent.

Remark 9.4.6 — If we replace V by a general module M over a commutative ring R, then (a) $\implies$ (b) and (a) $\implies$ (c) but not conversely.

Proof. Straightforward, do it yourself if you like. The key point to notice is that you need to divide by scalars for the converse direction, hence V is required to be a vector space instead of just a module for the implications (b) $\implies$ (a) and (c) $\implies$ (a). □

Theorem 9.4.7 (Dimension theorem for vector spaces)

If a vector space V has a finite basis, then every other basis has the same number of elements.

Proof. We prove something stronger: Assume $v_1, \dots, v_n$ is a spanning set while $w_1, \dots, w_m$ is linearly independent. We claim that $n \geq m$.

Question 9.4.8. Show that this claim is enough to imply the theorem.

Let $A_0 = \{v_1, \dots, v_n\}$ be the spanning set. Throw in w_1: by the spanning condition, $w_1 = c_1 v_1 + \cdots + c_n v_n$. There's some nonzero coefficient, say c_n. Thus

$$v_n = \frac{1}{c_n} w_1 - \frac{c_1}{c_n} v_1 - \frac{c_2}{c_n} v_2 - \dots.$$

Thus $A_1 = \{v_1, \dots, v_{n-1}, w_1\}$ is spanning. Now do the same thing, throwing in w_2, and deleting some element of the v_i as before to get A_2; the condition that the w_i are linearly independent ensures that some v_i coefficient must always not be zero. Since we can eventually get to A_m, we have $n \geq m$. □

Remark 9.4.9 (Generalizations)

- The theorem is true for an infinite basis as well if we interpret "the number of elements" as "cardinality". This is confusing on a first read through, so we won't elaborate.
- In fact, this is true for modules over any commutative ring. Interestingly, the proof for the general case proceeds by reducing to the case of a vector space.

The dimension theorem, true to its name, lets us define the **dimension** of a vector space as the size of any finite basis, if one exists. When it does exist we say V is **finite-dimensional**. So for example,

$$V = \left\{ ax^2 + bx + c \mid a, b, c \in \mathbb{R} \right\}$$

has dimension three, because $\{1, x, x^2\}$ is a basis. That's not the only basis: we could as well have written

$$\left\{ a(x^2 - 4x) + b(x + 2) + c \mid a, b, c \in \mathbb{R} \right\}$$

and gotten the exact same vector space. But the beauty of the theorem is that no matter how we try to contrive the generating set, we always will get exactly three elements. That's why it makes sense to say V has dimension three.

On the other hand, the set of all polynomials $\mathbb{R}[x]$ is *infinite-dimensional* (which should be intuitively clear).

A basis $e_1, \dots, e_n$ of V is really cool because it means that to specify $v \in V$, I only have to specify $a_1, \dots, a_n \in k$, and then let $v = a_1 e_1 + \cdots + a_n e_n$. You can even think of v as $(a_1, \dots, a_n)$. To put it another way, if V is a k-vector space we always have

$$V = e_1 k \oplus e_2 k \oplus \cdots \oplus e_n k.$$

§9.5 Linear maps

Prototypical example for this section: Evaluation of $\{ax^2 + bx + c\}$ *at* $x = 3$.

We've seen homomorphisms and continuous maps. Now we're about to see linear maps, the structure preserving maps between vector spaces. Can you guess the definition?

Definition 9.5.1. Let V and W be vector spaces over the same field k. A **linear map** is a map $T \colon V \to W$ such that:

(i) We have $T(v_1 + v_2) = T(v_1) + T(v_2)$ for any $v_1, v_2 \in V$.[1]

(ii) For any $a \in k$ and $v \in V$, $T(a \cdot v) = a \cdot T(v)$.

If this map is a bijection (equivalently, if it has an inverse), it is an **isomorphism**. We then say V and W are **isomorphic** vector spaces and write $V \cong W$.

Example 9.5.2 (Examples of linear maps)

(a) For any vector spaces V and W there is a trivial linear map sending everything to $0_W \in W$.

(b) For any vector space V, there is the identity isomorphism $\mathrm{id} \colon V \to V$.

(c) The map $\mathbb{R}^3 \to \mathbb{R}$ by $(a, b, c) \mapsto 4a + 2b + c$ is a linear map.

(d) Let V be the set of real polynomials of degree at most 2. The map $\mathbb{R}^3 \to V$ by $(a, b, c) \mapsto ax^2 + bx + c$ is an *isomorphism*.

(e) Let V be the set of real polynomials of degree at most 2. The map $V \to \mathbb{R}$ by $ax^2 + bx + c \mapsto 9a + 3b + c$ is a linear map, which can be described as "evaluation at 3".

(f) Let W be the set of functions $\mathbb{R} \to \mathbb{R}$. The evaluation map $W \to \mathbb{R}$ by $f \mapsto f(0)$ is a linear map.

(g) There is a map of $\mathbb{Q}$-vector spaces $\mathbb{Q}[\sqrt{2}] \to \mathbb{Q}[\sqrt{2}]$ called "multiply by $\sqrt{2}$"; this map sends $a + b\sqrt{2} \mapsto 2b + a\sqrt{2}$. This map is an isomorphism, because it has an inverse "multiply by $1/\sqrt{2}$".

In the expression $T(a \cdot v) = a \cdot T(v)$, note that the first $\cdot$ is the multiplication of V and the second $\cdot$ is the multiplication of W. Note that this notion of isomorphism really only cares about the size of the basis:

Proposition 9.5.3 (n-dimensional vector spaces are isomorphic)
If V is an n-dimensional vector space, then $V \cong k^{\oplus n}$.

Question 9.5.4. Let $e_1, \ldots, e_n$ be a basis for V. What is the isomorphism? (Your first guess is probably right.)

Remark 9.5.5 — You could technically say that all finite-dimensional vector spaces are just $k^{\oplus n}$ and that no other space is worth caring about. But this seems kind of rude. Spaces often are more than just triples: $ax^2 + bx + c$ is a polynomial, and so it has some "essence" to it that you'd lose if you compressed it into (a, b, c). Moreover, a lot of spaces, like the set of vectors (x, y, z) with $x + y + z = 0$, do not have an obvious choice of basis. Thus to cast such a space into $k^{\oplus n}$ would require you to make arbitrary decisions.

[1] In group language, T is a homomorphism $(V, +) \to (W, +)$.

§9.6 What is a matrix?

Now I get to tell you what a matrix is! This is fun, because now I can finally explain to you how to *derive* the recipes for matrix multiplication, rather than being told.

This section is so important, and also revelatory for so many students, that I'm actually going to do it twice. The first time, I'm going to work in an extremely special case, namely $V = W = \mathbb{R}^2$, using lots of numbers. (This is how I explained this concept when I taught it to first-year undergraduate students that didn't have proof experience.) Then the second time, we'll do it in modern language without all the numbers.

§9.6.i Extended example with $\mathbb{R}^2$, suitable for the general public

Throughout this section, I'll work specifically with $\mathbb{R}^2$, whose elements I will write as $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ rather than (x, y) (you'll see why when I talk about matrix multiplication).

Pop quiz:

- **Question 1**: Suppose that you're given a linear map $T \colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $T\left(\left[\begin{smallmatrix} 3 \\ 4 \end{smallmatrix}\right]\right) = \left[\begin{smallmatrix} \pi \\ 9 \end{smallmatrix}\right]$ and $T\left(\left[\begin{smallmatrix} 100 \\ 100 \end{smallmatrix}\right]\right) = \left[\begin{smallmatrix} 0 \\ 12 \end{smallmatrix}\right]$. What are $T\left(\left[\begin{smallmatrix} 103 \\ 104 \end{smallmatrix}\right]\right)$ and $T\left(\left[\begin{smallmatrix} 203 \\ 204 \end{smallmatrix}\right]\right)$?

  Answer 1: just add them.

$$T\left(\begin{bmatrix} 103 \\ 104 \end{bmatrix}\right) = \begin{bmatrix} \pi \\ 9 \end{bmatrix} + \begin{bmatrix} 0 \\ 12 \end{bmatrix} = \begin{bmatrix} \pi \\ 21 \end{bmatrix}$$
$$T\left(\begin{bmatrix} 203 \\ 204 \end{bmatrix}\right) = \begin{bmatrix} \pi \\ 9 \end{bmatrix} + 2\begin{bmatrix} 0 \\ 12 \end{bmatrix} = \begin{bmatrix} \pi \\ 33 \end{bmatrix}.$$

- **Question 2**: Suppose that you're given a linear map $T \colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $T\left(\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]\right) = \left[\begin{smallmatrix} 1 \\ 3 \end{smallmatrix}\right]$ and $T\left(\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]\right) = \left[\begin{smallmatrix} 2 \\ 4 \end{smallmatrix}\right]$. What is $T\left(\left[\begin{smallmatrix} 50 \\ 70 \end{smallmatrix}\right]\right)$?

  Answer 2:

$$T\left(\begin{bmatrix} 50 \\ 70 \end{bmatrix}\right) = 50\begin{bmatrix} 1 \\ 3 \end{bmatrix} + 70\begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 190 \\ 430 \end{bmatrix}.$$

So what this example illustrates is that the requirements on a linear map $T \colon \mathbb{R}^2 \to \mathbb{R}^2$ are so strong that if you just know $T\left(\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]\right)$ and $T\left(\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]\right)$ then you can *compute* the values of T at any other point. That's true for any two basis vectors (i.e., Question 1 could have been asked for inputs much nastier than the cherry-picked $\left[\begin{smallmatrix} 103 \\ 104 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 203 \\ 204 \end{smallmatrix}\right]$, and it would still be solvable), but of course $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ is an especially convenient choice.

Now we can give the following definition:

Definition 9.6.1. For a linear transform $T \colon \mathbb{R}^2 \to \mathbb{R}^2$, its *matrix* is an encoding of T obtained by gluing the column vectors

$$T\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) \qquad \text{and} \qquad T\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$$

together to get a 2×2 array of numbers.

For example,

$$T\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \text{ and } T\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \iff T \text{ encoded as } \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

Now, what happens if you apply the matrix multiplication rule from high school to the column vector $\left[\begin{smallmatrix}50\\70\end{smallmatrix}\right]$? Well, you get that

$$\begin{bmatrix}1 & 2\\3 & 4\end{bmatrix}\begin{bmatrix}50\\70\end{bmatrix} = \begin{bmatrix}1\cdot 50+2\cdot 70\\3\cdot 50+4\cdot 70\end{bmatrix} = \begin{bmatrix}190\\430\end{bmatrix}$$

...and you can see we're actually just doing the second pop quiz question again. So:

If $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ is encoded as a 2×2 matrix M, then multiplication of M with a (column) vector $v \in \mathbb{R}^2$ is defined to coincide with $T(v)$.

Remark 9.6.2 (The identity matrix deserves its name) — This also gives a more natural reason why the 2×2 identity matrix is $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$ rather than the explanation high school gives (namely, "well, try multiplying by it and notice you get the same thing"). If id is the identity function, then $\mathrm{id}\left(\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]\right) = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$, so that's the first column of the matrix; similarly $\mathrm{id}\left(\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]\right) = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ is the second column.

Now, what happens if we bring two maps S and T into the game, and compose them? We can do the same game with $S\circ T$.

- **Question 3**: Suppose that you're given a linear map $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $T\left(\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]\right) = \left[\begin{smallmatrix}1\\3\end{smallmatrix}\right]$ and $T\left(\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]\right) = \left[\begin{smallmatrix}2\\4\end{smallmatrix}\right]$. Then you're given a second linear map $S\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $S\left(\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]\right) = \left[\begin{smallmatrix}5\\7\end{smallmatrix}\right]$ and $S\left(\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]\right) = \left[\begin{smallmatrix}6\\8\end{smallmatrix}\right]$. What are $S\left(T\left(\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]\right)\right)$ and $S\left(T\left(\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]\right)\right)$?

  Answer 3:

$$S\left(T\left(\begin{bmatrix}1\\0\end{bmatrix}\right)\right) = S\left(\begin{bmatrix}1\\3\end{bmatrix}\right) = 1\begin{bmatrix}5\\7\end{bmatrix} + 3\begin{bmatrix}6\\8\end{bmatrix} = \begin{bmatrix}23\\31\end{bmatrix}.$$
$$S\left(T\left(\begin{bmatrix}0\\1\end{bmatrix}\right)\right) = S\left(\begin{bmatrix}2\\4\end{bmatrix}\right) = 2\begin{bmatrix}5\\7\end{bmatrix} + 4\begin{bmatrix}6\\8\end{bmatrix} = \begin{bmatrix}34\\46\end{bmatrix}.$$

Since $S\circ T$ is itself a linear map, we now know its matrix encoding:

$$S\circ T = \begin{bmatrix}23 & 34\\31 & 46\end{bmatrix}.$$

Now, you might have learned some matrix multiplication rule in school as a definition. If you execute that definition on the matrices for S and T, you should get

$$\underbrace{\begin{bmatrix}5 & 6\\7 & 8\end{bmatrix}}_{\text{encoding of } S}\ \underbrace{\begin{bmatrix}1 & 2\\3 & 4\end{bmatrix}}_{\text{encoding of } T} = \begin{bmatrix}5\cdot 1+6\cdot 3 & 5\cdot 2+6\cdot 4\\7\cdot 1+8\cdot 3 & 7\cdot 2+8\cdot 4\end{bmatrix} = \begin{bmatrix}23 & 34\\31 & 46\end{bmatrix}$$

It's the encoding for $S\circ T$ — indeed, you can see why, because if you trace through the work in Answer 3, it's actually the same arithmetic being carried out.

This shows why our Napkin definition of matrix as the *encoding* of a linear function is better than what many of you have seen. In high school, the recipe for matrix multiplication is provided as an unnatural definition, e.g., in cute pictures like Figure 9.1. However, for us, the recipe in Figure 9.1 is a *theorem*: we can *derive* how to get the encoding of $S\circ T$ given the encodings of S and T.

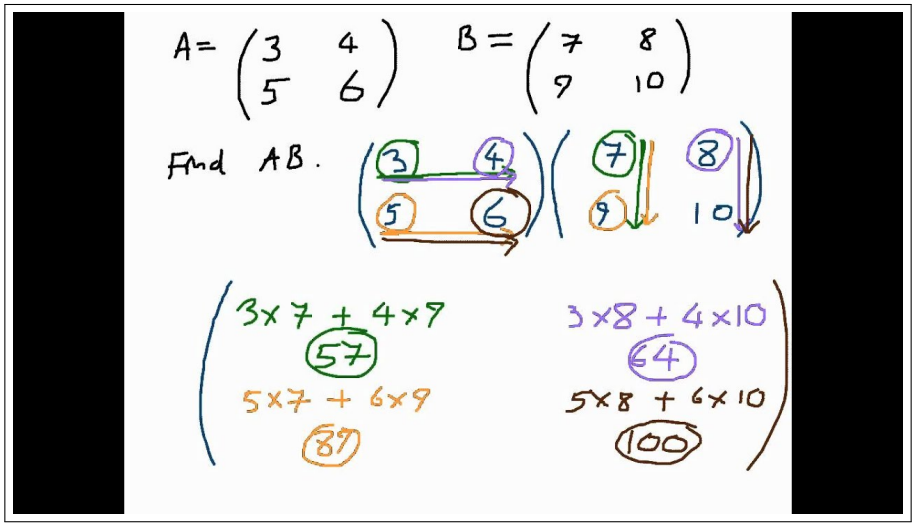


Figure 9.1: Matrix multiplication as taught in American high school: "here's a recipe, trust me bro". Image from [Ma12].

§9.6.ii General discussion, back to Napkin levels of abstraction

Let's go back to modern language, where we work with finite-dimensional spaces over any field, and any basis of the spaces (rather than a fixed basis like in the previous section).

Pick a finite-dimensional vector space V with *some* basis $e_1, \dots, e_m$ and a vector space W with basis $w_1, \dots, w_n$. Suppose I have a map $T\colon V \to W$ and I want to tell you what T is. It would be awfully inconsiderate of me to try and tell you what $T(v)$ is at every point v. But we saw I only have to tell you what $T(e_1)$, …, $T(e_m)$ are, because from there you can work out $T(a_1e_1 + \cdots + a_me_m)$ for yourself:

$$T(a_1e_1 + \cdots + a_me_m) = a_1T(e_1) + \cdots + a_mT(e_m).$$

Since the e_i are a basis, that tells you all you need to know about T.

Example 9.6.3 (Extending linear maps)
Let $V = \{ax^2 + bx + c \mid a, b, c \in \mathbb{R}\}$. Then $T(ax^2+bx+c) = aT(x^2)+bT(x)+cT(1)$.

Now I can even be more concrete. I could tell you what $T(e_1)$ is, but seeing as I have a basis of W, I can actually just tell you what $T(e_1)$ is in terms of this basis. Specifically, there are unique $a_{11}, a_{21}, \dots, a_{n1} \in k$ such that

$$T(e_1) = a_{11}w_1 + a_{21}w_2 + \cdots + a_{n1}w_n.$$

So rather than telling you the value of $T(e_1)$ in some abstract space W, I could just tell you what $a_{11}, a_{21}, \dots, a_{n1}$ were. Then I'd repeat this for $T(e_2)$, $T(e_3)$, all the way up to $T(e_m)$, and that would tell you everything you need to know about T.

That's where the matrix T comes from! It's a concise way of writing down all mn numbers I need to tell you.

To be explicit, the matrix for T is defined as the array

$$T = \underbrace{\begin{bmatrix} | & | & & | \\ T(e_1) & T(e_2) & \dots & T(e_m) \\ | & | & & | \end{bmatrix}}_{m \text{ columns}} \Big\} n \text{ rows}$$
$$= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix}.$$

To drive this point home,

A matrix is the laziest possible way to specify a linear map from V to W.

Example 9.6.4 (An example of a matrix)

Here is a concrete example in terms of a basis. Let $V = \mathbb{R}^3$ with basis e_1, e_2, e_3 and let $W = \mathbb{R}^2$ with basis w_1, w_2. If I have $T \colon V \to W$ then uniquely determined by three values, for example:

$$\begin{aligned} T(e_1) &= 4w_1 + 7w_2 \\ T(e_2) &= 2w_1 + 3w_2 \\ T(e_3) &= w_1 \end{aligned}$$

The columns then correspond to $T(e_1)$, $T(e_2)$, $T(e_3)$:

$$T = \begin{bmatrix} 4 & 2 & 1 \\ 7 & 3 & 0 \end{bmatrix}$$

Example 9.6.5 (An example of a matrix after choosing a basis)

We again let $V = \{ax^2 + bx + c\}$ be the vector space of polynomials of degree at most 2. We fix the basis 1, x, x^2 for it.
Consider the "evaluation at 3" map, a map $V \to \mathbb{R}$. We pick 1 as the basis element of the RHS; then we can write it as a 1×3 matrix

$$\begin{bmatrix} 1 & 3 & 9 \end{bmatrix}$$

with the columns corresponding to $T(1)$, $T(x)$, $T(x^2)$.

From here you can actually work out for yourself what it means to multiply two matrices. Suppose we have picked a basis for three spaces U, V, W. Given maps $T \colon U \to V$ and $S \colon V \to W$, we can consider their composition $S \circ T$, i.e.

$$U \xrightarrow{T} V \xrightarrow{S} W.$$

Matrix multiplication is defined exactly so that the matrix ST is the same thing we get from interpreting the composed function $S \circ T$ as a matrix, as we saw last section.

In particular, since function composition is associative, it follows that matrix multiplication is as well.

This means you can define concepts like the determinant or the trace of a matrix both in terms of an "intrinsic" map $T\colon V \to W$ and in terms of the entries of the matrix. Since the map T itself doesn't refer to any basis, the abstract definition will imply that the numerical definition doesn't depend on the choice of a basis.

§9.7 Subspaces and picking convenient bases

Prototypical example for this section: Any two linearly independent vectors in $\mathbb{R}^3$.

Definition 9.7.1. Let M be a left R-module. A **submodule** N of M is a module N such that every element of N is also an element of M. If M is a vector space then N is called a **subspace**.

Example 9.7.2 (Kernels)

The **kernel** of a map $T\colon V \to W$ (written $\ker T$) is the set of $v \in V$ such that $T(v) = 0_W$. It is a subspace of V, since it's closed under addition and scaling (why?).

Example 9.7.3 (Spans)

Let V be a vector space and $v_1, \dots, v_m$ be any vectors of V. The **span** of these vectors is defined as the set

$$\{a_1 v_1 + \cdots + a_m v_m \mid a_1, \dots, a_m \in k\}.$$

Note that it is a subspace of V as well!

Question 9.7.4. Why is 0_V an element of each of the above examples? In general, why must any subspace contain 0_V?

Subspaces behave nicely with respect to bases.

Theorem 9.7.5 (Basis completion)

Let V be an n-dimensional space, and V' a subspace of V. Then

(a) V' is also finite-dimensional.

(b) If $e_1, \dots, e_m$ is a basis of V', then there exist $e_{m+1}, \dots, e_n$ in V such that $e_1, \dots, e_n$ is a basis of V.

Proof. Omitted, since it is intuitive and the proof is not that enlightening. (However, we will use this result repeatedly later on, so do take the time to internalize it now.) □

A very common use case is picking a convenient basis for a map T.

Theorem 9.7.6 (Picking a basis for linear maps)

Let $T\colon V \to W$ be a map of finite-dimensional vector spaces, with $n = \dim V$, $m = \dim W$. Then there exists a basis $v_1, \dots, v_n$ of V and a basis $w_1, \dots, w_m$ of W, as well as a nonnegative integer k, such that

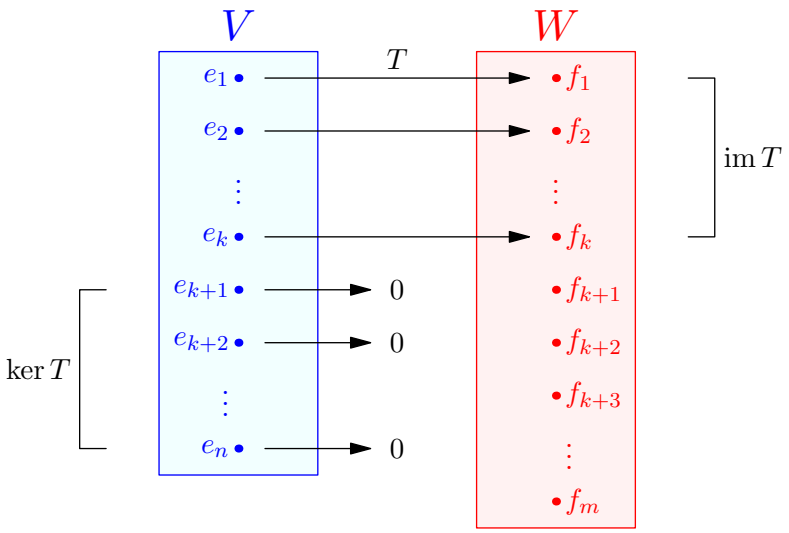
$$T(v_i) = \begin{cases} w_i & \text{if } i \le k \\ 0_W & \text{if } i > k. \end{cases}$$

Moreover $\dim \ker T = n - k$ and $\dim T^{\text{img}}(V) = k$.

Sketch of Proof. You might like to try this one yourself before reading on: it's a repeated application of Theorem 9.7.5.

Let $\ker T$ have dimension $n-k$. We can pick $v_{k+1}, \dots, v_n$ a basis of $\ker T$. Then extend it to a basis $v_1, \dots, v_n$ of V. The map T is injective over the span of $v_1, \dots, v_k$ (since only 0_V is in the kernel) so its images in W are linearly independent. Setting $w_i = T(v_i)$ for each i, we get some linearly independent set in W. Then extend it again to a basis of W. □

This theorem is super important, not only because of applications but also because it will give you the right picture in your head of how a linear map is supposed to look. I'll even draw a cartoon of it to make sure you remember:



In particular, for $T\colon V \to W$, one can write $V = \ker T \oplus V'$, so that T annihilates its kernel while sending V' to an isomorphic copy in W.

A corollary of this (which you should have expected anyways) is the so called rank-nullity theorem, which is the analog of the first isomorphism theorem.

Theorem 9.7.7 (Rank-nullity theorem)

Let V and W be finite-dimensional vector spaces. If $T\colon V \to W$, then

$$\dim V = \dim \ker T + \dim \operatorname{im} T.$$

Question 9.7.8. Conclude the rank-nullity theorem from Theorem 9.7.6.

§9.8 A cute application: Lagrange interpolation

Here's a cute application[2] of linear algebra to a theorem from high school.

Theorem 9.8.1 (Lagrange interpolation)

Let $x_1, \dots, x_{n+1}$ be distinct real numbers and $y_1, \dots, y_{n+1}$ any real numbers. Then there exists a *unique* polynomial P of degree at most n such that

$$P(x_i) = y_i$$

for every i.

When $n = 1$ for example, this loosely says there is a unique line joining two points.

Proof. The idea is to consider the vector space V of polynomials with degree at most n, as well as the vector space $W = \mathbb{R}^{n+1}$.

> **Question 9.8.2.** Check that $\dim V = n + 1 = \dim W$. This is easiest to do if you pick a basis for V, but you can then immediately forget about the basis once you finish this exercise.

Then consider the linear map $T \colon V \to W$ given by

$$P \mapsto (P(x_1), \dots, P(x_{n+1})) .$$

This is indeed a linear map because, well, $T(P+Q) = T(P) + T(Q)$ and $T(cP) = cT(P)$. It also happens to be injective: if $P \in \ker T$, then $P(x_1) = \cdots = P(x_{n+1}) = 0$, but $\deg P \le n$ and so P can only be the zero polynomial.

So T is an injective map between vector spaces of the same dimension. Thus it is actually a bijection, which is exactly what we wanted. □

§9.9 Pedagogical digression: Arrays of numbers are evil

(This whole section is Evan yapping about how to *teach* linear algebra, so it can be safely skipped.)

As I'll stress repeatedly, a matrix represents a *linear map between two vector spaces*. Writing it in the form of an $m \times n$ matrix is merely a very convenient way to see the map concretely. But it obfuscates the fact that this map is, well, a map, not an array of numbers.

If you took high school precalculus, you'll see everything done in terms of matrices. To any typical high school student, a matrix is an array of numbers. No one is sure what exactly these numbers represent, but they're told how to magically multiply these arrays to get more arrays. They're told that the matrix

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

[2] Source: Communicated to me by Joe Harris at the first Harvard-MIT Undergraduate Math Symposium.

is an "identity matrix", because when you multiply by another matrix it doesn't change. Then they're told that the determinant is some magical combination of these numbers formed by this weird multiplication rule. No one knows what this determinant does, other than the fact that $\det(AB) = \det A \det B$, and something about areas and row operations and Cramer's rule.

Then you go into linear algebra in college, and you do more magic with these arrays of numbers. You're told that two matrices T_1 and T_2 are similar if

$$T_2 = ST_1S^{-1}$$

for some invertible matrix S. You're told that the trace of a matrix $\operatorname{Tr} T$ is the sum of the diagonal entries. Somehow this doesn't change if you look at a similar matrix, but you're not sure why. Then you define the characteristic polynomial as

$$p_T(X) = \det(XI - T).$$

Somehow this also doesn't change if you take a similar matrix, but now you really don't know why. And then you have the Cayley-Hamilton theorem in all its black magic: $p_T(T)$ is the zero map. Out of curiosity you Google the proof, and you find some ad-hoc procedure which still leaves you with no idea why it's true.

This is terrible. What's so special about $T_2 = ST_1S^{-1}$? Only if you know that the matrices are linear maps does this make sense: T_2 is just T_1 rewritten with a different choice of basis.

I really want to push the opposite view. Linear algebra is the study of *linear maps*, but it is taught as the study of *arrays of numbers*, and no one knows what these numbers mean. And for a good reason: the numbers are meaningless. They are a highly convenient way of encoding the matrix, but they are not the main objects of study, any more than the dates of events are the main objects of study in history.

The other huge downside is that people get the impression that the only (real) vector space in existence is $\mathbb{R}^{\oplus n}$. As explained in Remark 9.5.5, while you *can* work this way if you're a soulless robot, it's very unnatural for humans to do so.

When I took Math 55a as a freshman at Harvard, I got the exact opposite treatment: we did all of linear algebra without writing down a single matrix. During all this time I was quite confused. What's wrong with a basis? I didn't appreciate until later that this approach was the morally correct way to treat the subject: it made it clear what was happening.

Throughout the Napkin, I've tried to strike a balance between these two approaches, using matrices when appropriate to illustrate the maps and to simplify proofs, but ultimately writing theorems and definitions in their *morally correct* form. I hope that this has both the advantage of giving the "right" definitions while being concrete enough to be digested. But I would like to say for the record that, if I had to pick between the high school approach and the 55a approach, I would pick 55a in a heartbeat.

§9.10 A word on general modules

Prototypical example for this section: $\mathbb{Z}[\sqrt{2}]$ *is a* $\mathbb{Z}$*-module of rank two.*

I focused mostly on vector spaces (aka modules over a field) in this chapter for simplicity, so I want to make a few remarks about modules over a general commutative ring R before concluding.

Firstly, recall that for general modules, we say "generating set" instead of "spanning set". Shrug.

The main issue with rings is that our key theorem Theorem 9.4.5 fails in spectacular ways. For example, consider $\mathbb{Z}$ as a $\mathbb{Z}$-module over itself. Then $\{2\}$ is linearly independent, but it cannot be extended to a basis. Similarly, $\{2, 3\}$ is spanning, but one cannot cut it down to a basis. You can see why defining dimension is going to be difficult.

Nonetheless, there are still analogs of some of the definitions above.

Definition 9.10.1. An R-module M is called **finitely generated** if it has a finite generating set.

Definition 9.10.2. An R-module M is called **free** if it has a basis. As said before, the analogue of the dimension theorem holds, and we use the word **rank** to denote the size of the basis. As before, there's an isomorphism $M \cong R^{\oplus n}$ where n is the rank.

Example 9.10.3 (An example of a $\mathbb{Z}$-module)
The $\mathbb{Z}$-module

$$\mathbb{Z}[\sqrt{2}] = \left\{ a + b\sqrt{2} \mid a, b \in \mathbb{Z} \right\}$$

has a basis $\{1, \sqrt{2}\}$, so we say it is a free $\mathbb{Z}$-module of rank 2.

Abuse of Notation 9.10.4 (Notation for groups)**.** Recall that an abelian group can be viewed a $\mathbb{Z}$-module (and in fact vice-versa!), so we can (and will) apply these words to abelian groups. We'll use the notation $G \oplus H$ for two abelian groups G and H for their Cartesian product, emphasizing the fact that G and H are abelian. This will happen when we study algebraic number theory and homology groups.

§9.11 A few harder problems to think about

General hint: Theorem 9.7.6 will be your best friend for many of these problems.

Problem 9A†. Let V and W be finite-dimensional vector spaces with nonzero dimension, and consider linear maps $T \colon V \to W$. Complete the following table by writing "sometimes", "always", or "never" for each entry.

	T injective	T surjective	T isomorphism
If $\dim V > \dim W \ldots$			
If $\dim V = \dim W \ldots$			
If $\dim V < \dim W \ldots$			

Problem 9B† (Equal dimension vector spaces are usually isomorphisms)**.** Let V and W be finite-dimensional vector spaces with $\dim V = \dim W$. Prove that for a map $T \colon V \to W$, the following are equivalent:

- T is injective,
- T is surjective,
- T is bijective.

Problem 9C. Let's say a *magic square* is a 3×3 matrix of real numbers where the sum of all diagonals, columns, and rows is equal, such as $\left[\begin{smallmatrix} 8 & 1 & 6 \\ 3 & 5 & 7 \\ 4 & 9 & 2 \end{smallmatrix}\right]$. Find the dimension of the set of magic squares, as a real vector space under addition.

Problem 9D (Multiplication by $\sqrt{5}$)**.** Let $V = \mathbb{Q}[\sqrt{5}] = \left\{a + b\sqrt{5}\right\}$ be a two-dimensional $\mathbb{Q}$-vector space, and fix the basis $\{1, \sqrt{5}\}$ for it. Write down the 2×2 matrix with rational coefficients that corresponds to multiplication by $\sqrt{5}$.

Problem 9E (Multivariable Lagrange interpolation)**.** Let $S \subset \mathbb{Z}^2$ be a set of n lattice points. Prove that there exists a nonzero two-variable polynomial p with real coefficients, of degree at most $\sqrt{2n}$, such that $p(x, y) = 0$ for every $(x, y) \in S$.

Problem 9F (Putnam 2003)**.** Do there exist polynomials $a(x)$, $b(x)$, $c(y)$, $d(y)$ such that

$$1 + xy + (xy)^2 = a(x)c(y) + b(x)d(y)$$

holds identically?

Problem 9G (TSTST 2014)**.** Let $P(x)$ and $Q(x)$ be arbitrary polynomials with real coefficients, and let d be the degree of $P(x)$. Assume that $P(x)$ is not the zero polynomial. Prove that there exist polynomials $A(x)$ and $B(x)$ such that

(i) Both A and B have degree at most $d/2$,

(ii) At most one of A and B is the zero polynomial,

(iii) P divides $A + Q \cdot B$.

Problem 9H* (Idempotents are projection maps)**.** Let $P \colon V \to V$ be a linear map, where V is a vector space (not necessarily finite-dimensional). Suppose P is **idempotent**, meaning $P(P(v)) = P(v)$ for each $v \in V$, or equivalently P is the identity on its image. Prove that

$$V = \ker P \oplus \operatorname{im} P.$$

Thus we can think of P as *projection* onto the subspace $\operatorname{im} P$.

Problem 9I*. Let V be a finite dimensional vector space. Let $T \colon V \to V$ be a linear map, and let $T^n \colon V \to V$ denote T applied n times. Prove that there exists an integer N such that

$$V = \ker T^N \oplus \operatorname{im} T^N.$$

10 Eigen-things

This chapter will develop the theory of eigenvalues and eigenvectors, the so-called "Jordan canonical form". (Later on we will use it to define the characteristic polynomial.)

§10.1 Why you should care

We know that a square matrix T is really just a linear map from V to V. What's the simplest type of linear map? It would just be multiplication by some scalar λ, which would have associated matrix (in any basis!)

$$T = \begin{bmatrix} \lambda & 0 & \dots & 0 \\ 0 & \lambda & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda \end{bmatrix}.$$

That's perhaps *too* simple, though. If we had a fixed basis $e_1, \dots, e_n$ then another very "simple" operation would just be scaling each basis element e_i by λ_i, i.e. a **diagonal matrix** of the form

$$T = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}.$$

These maps are more general. Indeed, you can, for example, compute T^{100} in a heartbeat: the map sends $e_i \to \lambda_i^{100} e_i$. (Try doing that with an arbitrary $n \times n$ matrix.)

Of course, most linear maps are probably not that nice. Or are they?

Example 10.1.1 (Getting lucky)

Let V be some two-dimensional vector space with e_1 and e_2 as basis elements. Let's consider a map $T \colon V \to V$ by $e_1 \mapsto 2e_1$ and $e_2 \mapsto e_1 + 3e_2$, which you can even write concretely as

$$T = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix} \quad \text{in basis } e_1, e_2.$$

This doesn't look anywhere as nice until we realize we can rewrite it as

$$\begin{aligned} e_1 &\mapsto 2e_1 \\ e_1 + e_2 &\mapsto 3(e_1 + e_2). \end{aligned}$$

So suppose we change to the basis e_1 and $e_1 + e_2$. Thus in the new basis,

$$T = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \quad \text{in basis } e_1, e_1 + e_2.$$

So our completely random-looking map, under a suitable change of basis, looks like the very nice maps we described before!

In this chapter, we will be *making* our luck, and we will see that our better understanding of matrices gives us the right way to think about this.

§10.2 Warning on assumptions

Most theorems in this chapter only work for

- finite-dimensional vector spaces V,
- over a field k which is *algebraically closed*.

On the other hand, the definitions work fine without these assumptions.

§10.3 Eigenvectors and eigenvalues

Let k be a field and V a vector space over it. In the above example, we saw that there were two very nice vectors, e_1 and $e_1 + e_2$, for which V did something very simple. Naturally, these vectors have a name.

Definition 10.3.1. Let $T \colon V \to V$ and $v \in V$ a *nonzero* vector. We say that v is an **eigenvector** if $T(v) = \lambda v$ for some $\lambda \in k$ (possibly zero, but remember $v \neq 0$). The value λ is called an **eigenvalue** of T.

We will sometimes abbreviate "v is an eigenvector with eigenvalue λ" to just "v is a λ-eigenvector".

Of course, no mention to a basis anywhere.

Example 10.3.2 (An example of an eigenvector and eigenvalue)

Consider the example earlier with $T = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}$.

(a) Note that e_1 and $e_1 + e_2$ are 2-eigenvectors and 3-eigenvectors.

(b) Of course, $5e_1$ is also an 2-eigenvector.

(c) And, $7e_1 + 7e_2$ is also a 3-eigenvector.

So you can quickly see the following observation.

Question 10.3.3. Show that the λ-eigenvectors, together with $\{0\}$ form a subspace.

Definition 10.3.4. For any λ, we define the λ-**eigenspace** as the set of λ-eigenvectors together with 0.

This lets us state succinctly that "2 is an eigenvalue of T with one-dimensional eigenspace spanned by e_1".

Unfortunately, it's not exactly true that eigenvalues always exist.

Example 10.3.5 (Eigenvalues need not exist)

Let $V = \mathbb{R}^2$ and let T be the map which rotates a vector by 90° around the origin.

Then $T(v)$ is not a multiple of v for any $v \in V$, other than the trivial $v = 0$.

However, it is true if we replace k with an algebraically closed field.[1]

Theorem 10.3.6 (Eigenvalues always exist over algebraically closed fields)
Suppose k is an *algebraically closed* field. Let V be a finite dimensional k-vector space. Then if $T\colon V \to V$ is a linear map, there exists an eigenvalue $\lambda \in k$.

Proof. (From [Ax97]) The idea behind this proof is to consider "polynomials" in T. For example, $2T^2 - 4T + 5$ would be shorthand for $2T(T(v)) - 4T(v) + 5v$. In this way we can consider "polynomials" $P(T)$; this lets us tie in the "algebraically closed" condition. These polynomials behave nicely:

Question 10.3.7. Show that $P(T) + Q(T) = (P+Q)(T)$ and $P(T) \circ Q(T) = (P \cdot Q)(T)$.

Let $n = \dim V < \infty$ and fix any nonzero vector $v \in V$, and consider vectors v, $T(v)$, $\dots$, $T^n(v)$. There are $n+1$ of them, so they can't be linearly independent for dimension reasons; thus there is a nonzero polynomial P such that $P(T)$ is zero when applied to v. WLOG suppose P is a monic polynomial, and thus $P(z) = (z - r_1)\dots(z - r_m)$ say. Then we get

$$0 = (T - r_1\mathrm{id}) \circ (T - r_2\mathrm{id}) \circ \cdots \circ (T - r_m\mathrm{id})(v)$$

(where id is the identity matrix). This means at least one of $T - r_i\mathrm{id}$ is not injective, i.e. has a nontrivial kernel, which is the same as an eigenvector. □

So in general we like to consider algebraically closed fields. This is not a big loss: any real matrix can be interpreted as a complex matrix whose entries just happen to be real, for example.

§10.4 The Jordan form

So that you know exactly where I'm going, here's the main theorem.

Definition 10.4.1. A **Jordan block** is an $n \times n$ matrix of the following shape:

$$\begin{bmatrix} \lambda & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \lambda & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & \lambda & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & \lambda \end{bmatrix}.$$

In other words, it has λ on the diagonal, and 1 above it. We allow $n = 1$, so $\begin{bmatrix}\lambda\end{bmatrix}$ is a Jordan block.

[1] A field is **algebraically closed** if all its polynomials have roots, the archetypal example being $\mathbb{C}$.

Theorem 10.4.2 (Jordan canonical form)

Let $T\colon V \to V$ be a linear map of finite-dimensional vector spaces over an algebraically closed field k. Then we can choose a basis of V such that the matrix T is "block-diagonal" with each block being a Jordan block.

Such a matrix is said to be in **Jordan form**. This form is unique up to rearranging the order of the blocks.

As an example, this means the matrix should look something like:

$$\begin{bmatrix}
\lambda_1 & 1 & & & & & & & \\
0 & \lambda_1 & & & & & & & \\
& & \lambda_2 & & & & & & \\
& & & \lambda_3 & 1 & 0 & & & \\
& & & 0 & \lambda_3 & 1 & & & \\
& & & 0 & 0 & \lambda_3 & & & \\
& & & & & & \ddots & & \\
& & & & & & & \lambda_m & 1 \\
& & & & & & & 0 & \lambda_m
\end{bmatrix}$$

Question 10.4.3. Check that diagonal matrices are the special case when each block is 1×1.

What does this mean? Basically, it means *our dream is almost true*. What happens is that V can get broken down as a direct sum

$$V = J_1 \oplus J_2 \oplus \cdots \oplus J_m$$

and T acts on each of these subspaces independently. These subspaces correspond to the blocks in the matrix above. In the simplest case, $\dim J_i = 1$, so J_i has a basis element e for which $T(e) = \lambda_i e$; in other words, we just have a simple eigenvalue. But on occasion, the situation is not quite so simple, and we have a block of size greater than 1; this leads to 1's just above the diagonals.

I'll explain later how to interpret the 1's, when I make up the word *descending staircase*. For now, you should note that even if $\dim J_i \geq 2$, we still have a basis element which is an eigenvector with eigenvalue λ_i.

Example 10.4.4 (A concrete example of Jordan form)

Let $T\colon k^6 \to k^6$ and suppose T is given by the matrix

$$T = \begin{bmatrix}
5 & 0 & 0 & 0 & 0 & 0 \\
0 & 2 & 1 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 \\
0 & 0 & 0 & 7 & 0 & 0 \\
0 & 0 & 0 & 0 & 3 & 0 \\
0 & 0 & 0 & 0 & 0 & 3
\end{bmatrix}.$$

Reading the matrix, we can compute all the eigenvectors and eigenvalues: for any

constants $a, b \in k$ we have

$$\begin{aligned} T(a \cdot e_1) &= 5a \cdot e_1 \\ T(a \cdot e_2) &= 2a \cdot e_2 \\ T(a \cdot e_4) &= 7a \cdot e_4 \\ T(a \cdot e_5 + b \cdot e_6) &= 3\left[a \cdot e_5 + b \cdot e_6\right]. \end{aligned}$$

The element e_3 on the other hand, is not an eigenvector since $T(e_3) = e_2 + 2e_3$.

§10.5 Nilpotent maps

Bear with me for a moment. First, define:

Definition 10.5.1. A map $T \colon V \to V$ is **nilpotent** if T^m is the zero map for some integer m. (Here T^m means "T applied m times".)

What's an example of a nilpotent map?

Example 10.5.2 (The "descending staircase")
Let $V = k^{\oplus 3}$ have basis e_1, e_2, e_3. Then the map T which sends

$$e_3 \mapsto e_2 \mapsto e_1 \mapsto 0$$

is nilpotent, since $T(e_1) = T^2(e_2) = T^3(e_3) = 0$, and hence $T^3(v) = 0$ for all $v \in V$.

The 3×3 descending staircase has matrix representation

$$T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

You'll notice this is a Jordan block.

Exercise 10.5.3. Show that the descending staircase above has 0 as its only eigenvalue.

That's a pretty nice example. As another example, we can have multiple such staircases.

Example 10.5.4 (Double staircase)
Let $V = k^{\oplus 5}$ have basis e_1, e_2, e_3, e_4, e_5. Then the map

$$e_3 \mapsto e_2 \mapsto e_1 \mapsto 0 \text{ and } e_5 \mapsto e_4 \mapsto 0$$

is nilpotent.

Picture, with some zeros omitted for emphasis:

$$T = \begin{bmatrix} 0 & 1 & 0 & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & & \\ & & & 0 & 1 \\ & & & 0 & 0 \end{bmatrix}$$

You can see this isn't really that different from the previous example; it's just the same idea repeated multiple times. And in fact we now claim that *all* nilpotent maps have essentially that form.

Theorem 10.5.5 (Nilpotent Jordan)

Let V be a finite-dimensional vector space over an algebraically closed field k. Let $T\colon V \to V$ be a nilpotent map. Then we can write $V = \bigoplus_{i=1}^{m} V_i$ where each V_i has a basis of the form v_i, $T(v_i)$, $\dots$, $T^{\dim V_i - 1}(v_i)$ for some $v_i \in V_i$, and such that $T^{\dim V_i}(v_i) = 0$.

Hence:

Every nilpotent map can be viewed as independent staircases.

Each chain v_i, $T(v_i)$, $T(T(v_i))$, $\dots$ is just one staircase. The proof is given later, but first let me point out where this is going.

Here's the punch line. Let's take the double staircase again. Expressing it as a matrix gives, say

$$S = \begin{bmatrix} 0 & 1 & 0 & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & & \\ & & & 0 & 1 \\ & & & 0 & 0 \end{bmatrix}.$$

Then we can compute

$$S + \lambda\text{id} = \begin{bmatrix} \lambda & 1 & 0 & & \\ 0 & \lambda & 1 & & \\ 0 & 0 & \lambda & & \\ & & & \lambda & 1 \\ & & & 0 & \lambda \end{bmatrix}.$$

It's a bunch of λ Jordan blocks! This gives us a plan to proceed: we need to break V into a bunch of subspaces such that $T - \lambda\text{id}$ is nilpotent over each subspace. Then Nilpotent Jordan will finish the job.

§10.6 Reducing to the nilpotent case

Definition 10.6.1. Let $T\colon V \to V$. A subspace $W \subseteq V$ is called T-**invariant** if $T(w) \in W$ for any $w \in W$. In this way, T can be thought of as a map $W \to W$.

In this way, the Jordan form is a decomposition of V into invariant subspaces.

Now I'm going to be cheap, and define:

Definition 10.6.2. A map $T\colon V \to V$ is called **indecomposable** if it's impossible to write $V = W_1 \oplus W_2$ where both W_1 and W_2 are nontrivial T-invariant spaces.

Picture of a *decomposable* map:

$$\left[\begin{array}{cc|ccc} \multicolumn{2}{c|}{\multirow{2}{*}{W_1}} & 0 & 0 & 0 \\ & & 0 & 0 & 0 \\ \hline 0 & 0 & \multicolumn{3}{c}{\multirow{3}{*}{W_2}} \\ 0 & 0 & \multicolumn{3}{c}{} \\ 0 & 0 & \multicolumn{3}{c}{} \end{array}\right]$$

As you might expect, we can break a space apart into "indecomposable" parts.

Proposition 10.6.3 (Invariant subspace decomposition)
Let V be a finite-dimensional vector space. Given any map $T\colon V \to V$, we can write
$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_m$$
where each V_i is T-invariant, and for any i the map $T\colon V_i \to V_i$ is indecomposable.

Proof. Same as the proof that every integer is the product of primes. If V is not decomposable, we are done. Otherwise, by definition write $V = W_1 \oplus W_2$ and then repeat on each of W_1 and W_2. □

Incredibly, with just that we're almost done! Consider a decomposition as above, so that $T\colon V_1 \to V_1$ is an indecomposable map. Then T has an eigenvalue λ_1, so let $S = T - \lambda_1 \mathrm{id}$; hence $\ker S \neq \{0\}$.

Question 10.6.4. Show that V_1 is also S-invariant, so we can consider $S\colon V_1 \to V_1$.

By Problem 9I*, we have
$$V_1 = \ker S^N \oplus \operatorname{im} S^N$$
for some N. But we assumed T was indecomposable, so this can only happen if $\operatorname{im} S^N = \{0\}$ and $\ker S^N = V_1$ (since $\ker S^N$ contains our eigenvector). Hence S is nilpotent, so it's a collection of staircases. In fact, since T is indecomposable, there is only one staircase. Hence V_1 is a Jordan block, as desired.

§10.7 (Optional) Proof of nilpotent Jordan

The proof is just induction on $\dim V$. Assume $\dim V \geq 1$, and let $W = T^{\text{img}}(V)$ be the image of V. Since T is nilpotent, we must have $W \subsetneq V$. Moreover, if $W = \{0\}$ (i.e. T is the zero map) then we're already done. So assume $\{0\} \subsetneq W \subsetneq V$.

By the inductive hypothesis, we can select a good basis of W:
$$\begin{aligned}\mathcal{B}' = \Big\{ & T(v_1), T(T(v_1)), T(T(T(v_1))), \dots \\ & T(v_2), T(T(v_2)), T(T(T(v_2))), \dots \\ & \dots, \\ & T(v_\ell), T(T(v_\ell)), T(T(T(v_\ell))), \dots \Big\}\end{aligned}$$
for some $T(v_i) \in W$ (here we have taken advantage of the fact that each element of W is itself of the form $T(v)$ for some v).

Also, note that there are exactly ℓ elements of $\mathcal{B}'$ which are in $\ker T$ (namely the last element of each of the ℓ staircases). We can thus complete it to a basis $v_{\ell+1}, \dots, v_m$ (where $m = \dim \ker T$). (In other words, the last element of each staircase plus the $m - \ell$ new ones are a basis for $\ker T$.)

Now consider

$$\begin{aligned}\mathcal{B} = \Big\{ & v_1, T(v_1), T(T(v_1)), T(T(T(v_1))), \dots \\ & v_2, T(v_2), T(T(v_2)), T(T(T(v_2))), \dots \\ & \dots, \\ & v_\ell, T(v_\ell), T(T(v_\ell)), T(T(T(v_\ell))), \dots \\ & v_{\ell+1}, v_{\ell+2}, \dots, v_m \Big\}.\end{aligned}$$

Question 10.7.1. Check that there are exactly $\ell + \dim W + (\dim \ker T - \ell) = \dim V$ elements.

Exercise 10.7.2. Show that all the elements are linearly independent. (Assume for contradiction there is some linear dependence, then take T of both sides.)

Hence $\mathcal{B}$ is a basis of the desired form.

§10.8 Algebraic and geometric multiplicity

Prototypical example for this section: The matrix T below.

This is some convenient notation: let's consider the matrix in Jordan form

$$T = \begin{bmatrix} 7 & 1 & & & & & \\ 0 & 7 & & & & & \\ & & 9 & & & & \\ & & & 7 & 1 & 0 \\ & & & 0 & 7 & 1 \\ & & & 0 & 0 & 7 \end{bmatrix}.$$

We focus on the eigenvalue 7, which appears multiple times, so it is certainly "repeated". However, there are two different senses in which you could say it is repeated.

- *Algebraic*: You could say it is repeated five times, because it appears five times on the diagonal.
- *Geometric*: You could say it really only appears two times: because there are only two eigen*vectors* with eigenvalue 7, namely e_1 and e_4.

  Indeed, the vector e_2 for example has $T(e_2) = 7e_2 + e_1$, so it's not really an eigenvector! If you apply $T - 7\text{id}$ to e_2 twice though, you do get zero.

Question 10.8.1. In this example, how many times do you need to apply $T - 7\text{id}$ to e_6 to get zero?

Both these notions are valid, so we will name both. To preserve generality, we first state the "intrinsic" definition.

Definition 10.8.2. Let $T \colon V \to V$ be a linear map and λ a scalar.

- The **geometric multiplicity** of λ is the dimension $\dim V_\lambda$ of the λ-eigenspace.

- Define the **generalized eigenspace** V^λ to be the subspace of V for which $(T - \lambda\text{id})^n(v) = 0$ for some $n \geq 1$. The **algebraic multiplicity** of λ is the dimension $\dim V^\lambda$.

(Silly edge case: we allow "multiplicity zero" if λ is not an eigenvalue at all.)

However in practice you should just count the Jordan blocks.

Example 10.8.3 (An example of eigenspaces via Jordan form)
Retain the matrix T mentioned earlier and let $\lambda = 7$.

- The eigenspace V_λ has basis e_1 and e_4, so the geometric multiplicity is 2.
- The generalized eigenspace V^λ has basis e_1, e_2, e_4, e_5, e_6 so the algebraic multiplicity is 5.

To be completely explicit, here is how you think of these in practice:

Proposition 10.8.4 (Geometric and algebraic multiplicity vs Jordan blocks)
Assume $T \colon V \to V$ is a linear map of finite-dimensional vector spaces, written in Jordan form. Let λ be a scalar. Then

- The geometric multiplicity of λ is the number of Jordan blocks with eigenvalue λ; the eigenspace has one basis element per Jordan block.
- The algebraic multiplicity of λ is the sum of the dimensions of the Jordan blocks with eigenvalue λ; the eigenspace is the direct sum of the subspaces corresponding to those blocks.

Proof. Definition 10.8.2 was essentially chosen to be a basis-free rephrasing of this proposition. □

Question 10.8.5. Show that the geometric multiplicity is always less than or equal to the algebraic multiplicity.

This actually gives us a tentative definition:

- The trace is the sum of the eigenvalues, counted with algebraic multiplicity.
- The determinant is the product of the eigenvalues, counted with algebraic multiplicity.

This definition is okay, but it has the disadvantage of requiring the ground field to be algebraically closed. It is also not the definition that is easiest to work with computationally. The next two chapters will give us a better definition.

§10.9 A few harder problems to think about

Problem 10A (Sum of algebraic multiplicities)**.** Given a 2018-dimensional complex vector space V and a map $T \colon V \to V$, what is the sum of the algebraic multiplicities of all eigenvalues of T?

Problem 10B (The word "diagonalizable"). A linear map $T\colon V \to V$ (where $\dim V$ is finite) is said to be **diagonalizable** if it has a basis $e_1, \dots, e_n$ such that each e_i is an eigenvector.

(a) Explain the name "diagonalizable".

(b) Suppose we are working over an algebraically closed field. Then show that T is diagonalizable if and only if for any λ, the geometric multiplicity of λ equals the algebraic multiplicity of λ.

Problem 10C (Switcharoo). Let V be the $\mathbb{C}$-vector space with basis e_1 and e_2. The map $T\colon V \to V$ sends $T(e_1) = e_2$ and $T(e_2) = e_1$. Determine the eigenspaces of T.

Problem 10D. Suppose $T\colon \mathbb{C}^{\oplus 2} \to \mathbb{C}^{\oplus 2}$ is a linear map of $\mathbb{C}$-vector spaces such that $T^{2011} = \mathrm{id}$. Must T be diagonalizable?

Problem 10E (Writing a polynomial backwards). Define the complex vector space V of polynomials with degree at most 2, say $V = \{ax^2 + bx + c \mid a, b, c \in \mathbb{C}\}$. Define $T\colon V \to V$ by

$$T(ax^2 + bx + c) = cx^2 + bx + a.$$

Determine the eigenspaces of T.

Problem 10F (Differentiation of polynomials). Let $V = \mathbb{R}[x]$ be the infinite-dimensional real vector space of all polynomials with real coefficients. Note that $\frac{d}{dx}\colon V \to V$ is a linear map (for example it sends x^3 to $3x^2$). Which real numbers are eigenvalues of this map?

Problem 10G (Differentiation of functions). Let V be the infinite-dimensional real vector space of all infinitely differentiable functions $\mathbb{R} \to \mathbb{R}$. Note that $\frac{d}{dx}\colon V \to V$ is a linear map (for example it sends $\cos x$ to $-\sin x$). Which real numbers are eigenvalues of this map?

11 Dual space and trace

You may have learned in high school that given a matrix

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

the trace is the sum along the diagonals $a+d$ and the determinant is $ad-bc$. But we know that a matrix is somehow just encoding a linear map using a choice of basis. Why would these random formulas somehow not depend on the choice of a basis?

In this chapter, we are going to give an intrinsic definition of $\operatorname{Tr} T$, where $T\colon V \to V$ and $\dim V < \infty$. This will give a coordinate-free definition which will in particular imply the trace $a+d$ doesn't change if we take a different basis.

In doing so, we will introduce two new constructions: the *tensor product* $V \otimes W$ (which is a sort of product of two spaces, with dimension $\dim V \cdot \dim W$) and the *dual space* $V^\vee$, which is the set of linear maps $V \to k$ (a k-vector space). Later on, when we upgrade from a vector space V to an inner product space, we will see that the dual space $V^\vee$ gives a nice interpretation of the "transpose" of a matrix. You'll already see some of that come through here.

The trace is only defined for finite-dimensional vector spaces, so if you want you can restrict your attention to finite-dimensional vector spaces for this chapter. (On the other hand we do not need the ground field to be algebraically closed.)

The next chapter will then do the same for the determinant.

§11.1 Tensor product

Prototypical example for this section: $\mathbb{R}[x] \otimes \mathbb{R}[y] = \mathbb{R}[x,y]$.

We know that $\dim(V \oplus W) = \dim V + \dim W$, even though as sets $V \oplus W$ looks like $V \times W$. What if we wanted a real "product" of spaces, with multiplication of dimensions?

For example, let's pull out my favorite example of a real vector space, namely

$$V = \left\{ ax^2 + bx + c \mid a,b,c \in \mathbb{R} \right\}.$$

Here's another space, a little smaller:

$$W = \left\{ dy + e \mid d,e \in \mathbb{R} \right\}.$$

If we take the direct sum, then we would get some rather unnatural vector space of dimension five (whose elements can be thought of as pairs $(ax^2+bx+c, dy+e)$). But suppose we want a vector space whose elements are *products* of polynomials in V and W; it would contain elements like $4x^2y + 5xy + y + 3$. In particular, the basis would be

$$\left\{ x^2y, x^2, xy, x, y, 1 \right\}$$

and thus have dimension six.

For this we resort to the *tensor product*. It does exactly this, except that the "multiplication" is done by a scary[1] symbol $\otimes$: think of it as a "wall" that separates the elements between the two vector spaces. For example, the above example might be written as

$$4x^2 \otimes y + 5x \otimes y + 1 \otimes y + 3 \otimes 1.$$

[1]Seriously, $\otimes$ looks *terrifying* to non-mathematicians, and even to many math undergraduates.

(This should be read as $(4x^2 \otimes y) + (5x \otimes y) + \dots$; addition comes after $\otimes$.) Of course there should be no distinction between writing $4x^2 \otimes y$ and $x^2 \otimes 4y$ or even $2x^2 \otimes 2y$. While we want to keep the x and y separate, the scalars should be free to float around.

Of course, there's no need to do everything in terms of just the monomials. We are free to write

$$(x+1) \otimes (y+1).$$

If you like, you can expand this as

$$x \otimes y + 1 \otimes y + x \otimes 1 + 1 \otimes 1.$$

Same thing. The point is that we can take any two of our polynomials and artificially "tensor" them together.

The definition of the tensor product does exactly this, and nothing else.[2]

Definition 11.1.1. Let V and W be vector spaces over the same field k. The **tensor product** $V \otimes_k W$ is the abelian group generated by elements of the form $v \otimes w$, subject to relations

$$\begin{aligned}(v_1 + v_2) \otimes w &= v_1 \otimes w + v_2 \otimes w \\ v \otimes (w_1 + w_2) &= v \otimes w_1 + v \otimes w_2 \\ (c \cdot v) \otimes w &= v \otimes (c \cdot w).\end{aligned}$$

As a vector space, its action is given by $c \cdot (v \otimes w) = (c \cdot v) \otimes w = v \otimes (c \cdot w)$.

Here's another way to phrase the same idea. We define a **pure tensor** as an element of the form $v \otimes w$ for $v \in V$ and $w \in W$. But we let the $\otimes$ wall be "permeable" in the sense that

$$(c \cdot v) \otimes w = v \otimes (c \cdot w) = c \cdot (v \otimes w)$$

and we let multiplication and addition distribute as we expect. Then $V \otimes W$ consists of sums of pure tensors.

Example 11.1.2 (Infinite-dimensional example of tensor product: two-variable polynomials)

Although it's not relevant to this chapter, this definition works equally well with infinite-dimensional vector spaces. The best example might be

$$\mathbb{R}[x] \otimes_{\mathbb{R}} \mathbb{R}[y] = \mathbb{R}[x,y].$$

That is, the tensor product of polynomials in x with real polynomials in y turns out to just be two-variable polynomials $\mathbb{R}[x,y]$.

Remark 11.1.3 (Warning on sums of pure tensors) — Remember the elements of $V \otimes_k W$ really are *sums* of these pure tensors! If you liked the previous example, this fact has a nice interpretation — not every polynomial in $\mathbb{R}[x,y] = \mathbb{R}[x] \otimes_{\mathbb{R}} \mathbb{R}[y]$ factors as a polynomial in x times a polynomial in y (i.e. as pure tensors $f(x) \otimes g(y)$). But they all can be written as sums of pure tensors $x^a \otimes y^b$.

[2] I'll only define this for vector spaces for simplicity. The definition for modules over a commutative ring R is exactly the same.

As the example we gave suggested, the basis of $V \otimes_k W$ is literally the "product" of the bases of V and W. In particular, this fulfills our desire that $\dim(V \otimes_k W) = \dim V \cdot \dim W$.

Proposition 11.1.4 (Basis of $V \otimes W$)
Let V and W be finite-dimensional k-vector spaces. If $e_1, \dots, e_m$ is a basis of V and $f_1, \dots, f_n$ is a basis of W, then the basis of $V \otimes_k W$ is precisely $e_i \otimes f_j$, where $i = 1, \dots, m$ and $j = 1, \dots, n$.

Proof. Omitted; it's easy at least to see that this basis is spanning. □

Example 11.1.5 (Explicit computation)
Let V have basis e_1, e_2 and W have basis f_1, f_2. Let $v = 3e_1 + 4e_2 \in V$ and $w = 5f_1 + 6f_2 \in W$. Let's write $v \otimes w$ in this basis for $V \otimes_k W$:

$$\begin{aligned} v \otimes w &= (3e_1 + 4e_2) \otimes (5f_1 + 6f_2) \\ &= (3e_1) \otimes (5f_1) + (4e_2) \otimes (5f_1) + (3e_1) \otimes (6f_2) + (4e_2) \otimes (6f_2) \\ &= 15(e_1 \otimes f_1) + 20(e_2 \otimes f_1) + 18(e_1 \otimes f_2) + 24(e_2 \otimes f_2). \end{aligned}$$

So you can see why tensor products are a nice "product" to consider if we're really interested in $V \times W$ in a way that's more intimate than just a direct sum.

Abuse of Notation 11.1.6. Moving forward, we'll almost always abbreviate $\otimes_k$ to just $\otimes$, since k is usually clear.

Remark 11.1.7 — Observe that to define a linear map $V \otimes W \to X$, I only have to say what happens to each pure tensor $v \otimes w$, since the pure tensors *generate* $V \otimes W$. But again, keep in mind that $V \otimes W$ consists of *sums* of these pure tensors! In other words, $V \otimes W$ is generated by pure tensors.

Remark 11.1.8 — Much like the Cartesian product $A \times B$ of sets, you can tensor together any two vector spaces V and W over the same field k; the relationship between V and W is completely irrelevant. One can think of the $\otimes$ as a "wall" through which one can pass scalars in k, but otherwise keeps the elements of V and W separated. Thus, $\otimes$ is **content-agnostic**.
This also means that even if V and W have some relation to each other, the tensor product doesn't remember this. So for example $v \otimes 1 \neq 1 \otimes v$, just like $(g, 1_G) \neq (1_G, g)$ in the group $G \times G$.

§11.2 Dual space

Prototypical example for this section: Rotate a column matrix by 90 degrees.

Consider the following vector space:

Example 11.2.1 (Functions from $\mathbb{R}^3 \to \mathbb{R}$)
The set of real functions $f(x, y, z)$ is an infinite-dimensional real vector space.

Indeed, we can add two functions to get $f+g$, and we can think of functions like $2f$.

This is a terrifyingly large vector space, but you can do some reasonable reductions. For example, you can restrict your attention to just the *linear maps* from $\mathbb{R}^3$ to $\mathbb{R}$.

That's exactly what we're about to do. This definition might seem strange at first, but bear with me.

Definition 11.2.2. Let V be a k-vector space. Then $V^\vee$, the **dual space** of V, is defined as the vector space whose elements are *linear maps from V to k*.

The addition and multiplication are pointwise: it's the same notation we use when we write $cf+g$ to mean $c \cdot f(x) + g(x)$. The dual space itself is less easy to think about.

Let's try to find a basis for $V^\vee$. First, here is a very concrete interpretation of the vector space. Suppose for example $V = \mathbb{R}^3$. We can think of elements of V as column matrices, like

$$v = \begin{bmatrix} 2 \\ 5 \\ 9 \end{bmatrix} \in V.$$

Then a linear map $f \colon V \to k$ can be interpreted as a *row matrix*:

$$f = \begin{bmatrix} 3 & 4 & 5 \end{bmatrix} \in V^\vee.$$

Then

$$f(v) = \begin{bmatrix} 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \\ 9 \end{bmatrix} = 71.$$

More precisely: **to specify a linear map $V \to k$, I only have to tell you where each basis element of V goes**. In the above example, f sends e_1 to 3, e_2 to 4, and e_3 to 5. So f sends

$$2e_1 + 5e_2 + 9e_3 \mapsto 2 \cdot 3 + 5 \cdot 4 + 9 \cdot 5 = 71.$$

Let's make all this precise.

Proposition 11.2.3 (The dual basis for $V^\vee$)

Let V be a finite-dimensional vector space with basis $e_1, \dots, e_n$. For each i consider the function $e_i^\vee \colon V \to k$ defined by

$$e_i^\vee(e_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

In more humane terms, $e_i^\vee(v)$ gives the coefficient of e_i in v.
Then $e_1^\vee, e_2^\vee, \dots, e_n^\vee$ is a basis of $V^\vee$.

Example 11.2.4 (Explicit example of element in $V^\vee$)

In this notation, $f = 3e_1^\vee + 4e_2^\vee + 5e_3^\vee$. Do you see why the "sum" notation works

as expected here? Indeed

$$\begin{aligned} f(e_1) &= (3e_1^\vee + 4e_2^\vee + 5e_3^\vee)(e_1) \\ &= 3e_1^\vee(e_1) + 4e_2^\vee(e_1) + 5e_3^\vee(e_1) \\ &= 3 \cdot 1 + 4 \cdot 0 + 5 \cdot 0 = 3. \end{aligned}$$

That's exactly what we wanted.

You might be inclined to point out that $V \cong V^\vee$ at this point, with an isomorphism given by $e_i \mapsto e_i^\vee$. You might call it "rotating the column matrix by 90°".

This statement is technically true, but for a generic vector space V without any extra information, you can just think of this as an artifact of the $\dim V = \dim V^\vee$ (as *any* two vector spaces of equal dimension are isomorphic). Most importantly, the isomorphism given above depends on what basis you picked.

Remark 11.2.5 (Explicit example showing that the isomorphism $V \to V^\vee$ given above is unnatural) Alice or Bob are looking at the same two-dimensional real vector space

$$V = \{(x, y, z) \mid x + y + z = 0\}.$$

Also, let $v_{\text{example}} = (3, 5, -8)$ be an example of an arbitrary element of V for concreteness.
Suppose Alice chooses the following basis vectors for V.

$$\begin{aligned} e_1 &= (1, 0, -1) \\ e_2 &= (0, 1, -1). \end{aligned}$$

Alice uses this to construct an isomorphism $A \colon V \to V^\vee$ as described above, and considers $e_1^\vee = A(e_1)$. The element $e_1^\vee \in V^\vee$ is a function $e_1^\vee \colon V \to \mathbb{R}$, meaning Alice can plug any vector in V into it. As an example, for v_{example}

$$e_1^\vee(v_{\text{example}}) = e_1^\vee((3, 5, -8)) = e_1^\vee(3e_1 + 5e_2) = 3.$$

Meanwhile, Bob chooses the different basis vectors

$$\begin{aligned} f_1 &= (1, 0, -1) \\ f_2 &= (1, -1, 0). \end{aligned}$$

This gives Bob an isomorphism $B \colon V \to V^\vee$, and a corresponding $f_1^\vee = B(f_1)$. Bob can also evaluate it anywhere, e.g.

$$f_1^\vee(v_{\text{example}}) = f_1^\vee((3, 5, -8)) = f_1^\vee(8f_1 - 5f_2) = 8.$$

It follows that $e_1^\vee = A((1, 0, -1))$ and $f_1^\vee = B((1, 0, -1))$ are different elements of $V^\vee$. In other words Alice and Bob got different isomorphisms because they picked different bases.

§11.3 $V^\vee \otimes W$ gives matrices from V to W

Goal of this section:

If V and W are finite-dimensional k-vector spaces then $V^\vee \otimes W$ represents linear maps $V \to W$.

Here's the intuition. If V is three-dimensional and W is five-dimensional, then we can think of the maps $V \to W$ as a 5×3 array of numbers. We want to think of these maps as a vector space: (since one can add or scale matrices). So it had better be a vector space with dimension 15, but just saying "$k^{\oplus 15}$" is not really that satisfying (what is the basis?).

To do better, we consider the tensor product

$$V^\vee \otimes W$$

which somehow is a product of maps out of V and the target space W. We claim that this is in fact the space we want: i.e. **there is a natural bijection between elements of $V^\vee \otimes W$ and linear maps from V to W.**

First, how do we interpret an element of $V^\vee \otimes W$ as a map $V \to W$? For concreteness, suppose V has a basis e_1, e_2, e_3, and W has a basis f_1, f_2, f_3, f_4, f_5. Consider an element of $V^\vee \otimes W$, say

$$e_1^\vee \otimes (f_2 + 2f_4) + 4e_2^\vee \otimes f_5.$$

We want to interpret this element as a function $V \to W$: so given a $v \in V$, we want to output an element of W. There's really only one way to do this: feed in $v \in V$ into the $V^\vee$ guys on the left. That is, take the map

$$v \mapsto e_1^\vee(v) \cdot (f_2 + 2f_4) + 4e_2^\vee(v) \cdot f_5 \in W.$$

So, there's a natural way to interpret any element $\xi_1 \otimes w_1 + \cdots + \xi_m \otimes w_m \in V^\vee \otimes W$ as a linear map $V \to W$. The claim is that in fact, every linear map $V \to W$ has such an interpretation.

First, for notational convenience,

Definition 11.3.1. Let $\operatorname{Hom}(V, W)$ denote the set of linear maps from V to W (which one can interpret as matrices which send V to W), viewed as a vector space over k. (The "Hom" stands for homomorphism.)

Question 11.3.2. Identify $\operatorname{Hom}(V, k)$ by name.

We can now write down something that's more true generally.

Theorem 11.3.3 ($V^\vee \otimes W \iff$ linear maps $V \to W$)
Let V and W be finite-dimensional vector spaces. We described a map

$$\Psi \colon V^\vee \otimes W \to \operatorname{Hom}(V, W)$$

by sending $\xi_1 \otimes w_1 + \cdots + \xi_m \otimes w_m$ to the linear map

$$v \mapsto \xi_1(v)w_1 + \cdots + \xi_m(v)w_m.$$

Then Ψ is an isomorphism of vector spaces, i.e. every linear map $V \to W$ can be uniquely represented as an element of $V^\vee \otimes W$ in this way.

The above is perhaps a bit dense, so here is a concrete example.

Example 11.3.4 (Explicit example)
Let $V = \mathbb{R}^2$ and take a basis e_1, e_2 of V. Then define $T\colon V \to V$ by

$$T = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

Then we have

$$\Psi(e_1^\vee \otimes e_1 + 2e_2^\vee \otimes e_1 + 3e_1^\vee \otimes e_2 + 4e_2^\vee \otimes e_2) = T.$$

The beauty is that the Ψ definition is basis-free; thus even if we change the basis, although the above expression will look completely different, the *actual element* in $V^\vee \otimes V$ doesn't change.

Despite this, we'll indulge ourselves in using coordinates for the proof.

Proof of Theorem 11.3.3. This looks intimidating, but it's actually not difficult. We proceed in two steps:

1. First, we check that Ψ is *surjective*; every linear map has at least one representation in $V^\vee \otimes W$. To see this, take any $T\colon V \to W$. Suppose V has basis e_1, e_2, e_3 and that $T(e_1) = w_1$, $T(e_2) = w_2$ and $T(e_3) = w_3$. Then the element

$$e_1^\vee \otimes w_1 + e_2^\vee \otimes w_2 + e_3^\vee \otimes w_3$$

works, as it is contrived to agree with T on the basis elements e_i.

2. So it suffices to check now that $\dim V^\vee \otimes W = \dim \operatorname{Hom}(V, W)$. Certainly, $V^\vee \otimes W$ has dimension $\dim V \cdot \dim W$. But by viewing $\operatorname{Hom}(V, W)$ as $\dim V \cdot \dim W$ matrices, we see that it too has dimension $\dim V \cdot \dim W$. □

So there is a **natural isomorphism** $V^\vee \otimes W \cong \operatorname{Hom}(V, W)$. While we did use a basis liberally in the *proof that it works*, this doesn't change the fact that the isomorphism is "God-given", depending only on the spirit of V and W itself and not which basis we choose to express the vector spaces in.

§11.4 The trace

We are now ready to give the definition of a trace. Recall that a square matrix T can be thought of as a map $T\colon V \to V$. According to the above theorem,

$$\operatorname{Hom}(V, V) \cong V^\vee \otimes V$$

so every map $V \to V$ can be thought of as an element of $V^\vee \otimes V$. But we can also define an *evaluation map* $\operatorname{ev}\colon V^\vee \otimes V \to k$ by "collapsing" each pure tensor: $f \otimes v \mapsto f(v)$. So this gives us a composed map

$$\operatorname{Hom}(V, V) \xrightarrow{\cong} V^\vee \otimes V \xrightarrow{\operatorname{ev}} k.$$

This result is called the **trace** of a matrix T.

Example 11.4.1 (Example of a trace)
Continuing the previous example,

$$\operatorname{Tr} T = e_1^\vee(e_1) + 2e_2^\vee(e_1) + 3e_1^\vee(e_2) + 4e_2^\vee(e_2) = 1 + 0 + 0 + 4 = 5.$$

And that is why the trace is the sum of the diagonal entries.

§11.5 A few harder problems to think about

Problem 11A (Trace is sum of eigenvalues)**.** Let V be an n-dimensional vector space over an algebraically closed field k. Let $T\colon V \to V$ be a linear map with eigenvalues λ_1, λ_2, $\ldots$, λ_n (counted with algebraic multiplicity). Show that $\operatorname{Tr} T = \lambda_1 + \cdots + \lambda_n$.

Problem 11B† (Product of traces)**.** Let $T\colon V \to V$ and $S\colon W \to W$ be linear maps of finite-dimensional vector spaces V and W. Define $T \otimes S\colon V \otimes W \to V \otimes W$ by $v \otimes w \mapsto T(v) \otimes S(w)$. Prove that

$$\operatorname{Tr}(T \otimes S) = \operatorname{Tr}(T)\operatorname{Tr}(S).$$

Problem 11C† (Traces kind of commute)**.** Let $T\colon V \to W$ and $S\colon W \to V$ be linear maps between finite-dimensional vector spaces V and W. Show that

$$\operatorname{Tr}(T \circ S) = \operatorname{Tr}(S \circ T).$$

Problem 11D (Putnam 1988)**.** Let V be an n-dimensional vector space. Let $T\colon V \to V$ be a linear map and suppose there exists $n+1$ eigenvectors, any n of which are linearly independent. Does it follow that T is a scalar multiple of the identity?

12 Determinant

The goal of this chapter is to give the basis-free definition of the determinant: that is, we're going to define $\det T$ for $T\colon V \to V$ without making reference to the encoding for T. This will make it obvious that the determinant of a matrix does not depend on the choice of basis, and that several properties are vacuously true (e.g. that the determinant is multiplicative).

The determinant is only defined for finite-dimensional vector spaces, so if you want you can restrict your attention to finite-dimensional vector spaces for this chapter. On the other hand we do not need the ground field to be algebraically closed.

§12.1 Wedge product

Prototypical example for this section: $\bigwedge^2(\mathbb{R}^2)$ *gives parallelograms.*

We're now going to define something called the wedge product. It will look at first like the tensor product $V \otimes V$, but we'll have one extra relation.

For simplicity, I'll first define the wedge product $\bigwedge^2(V)$. But we will later replace 2 with any n.

Definition 12.1.1. Let V be a k-vector space. The 2-wedge product $\bigwedge^2(V)$ is the abelian group generated by elements of the form $v \wedge w$ (where $v, w \in V$), subject to the same relations

$$\begin{aligned}(v_1 + v_2) \wedge w &= v_1 \wedge w + v_2 \wedge w \\ v \wedge (w_1 + w_2) &= v \wedge w_1 + v \wedge w_2 \\ (c \cdot v) \wedge w &= v \wedge (c \cdot w)\end{aligned}$$

plus two additional relations:

$$v \wedge v = 0 \quad \text{and} \quad v \wedge w = -w \wedge v.$$

As a vector space, its action is given by $c \cdot (v \wedge w) = (c \cdot v) \wedge w = v \wedge (c \cdot w)$.

> **Exercise 12.1.2.** Show that the condition $v \wedge w = -(w \wedge v)$ is actually extraneous: you can derive it from the fact that $v \wedge v = 0$. (Hint: expand $(v + w) \wedge (v + w) = 0$.)

This looks almost exactly the same as the definition for a tensor product, with two subtle differences. The first is that we only have V now, rather than V and W as with the tensor product.[1] Secondly, there is a new *mysterious* relation

$$v \wedge v = 0 \implies v \wedge w = -(w \wedge v).$$

What's that doing there? It seems kind of weird.

I'll give you a hint.

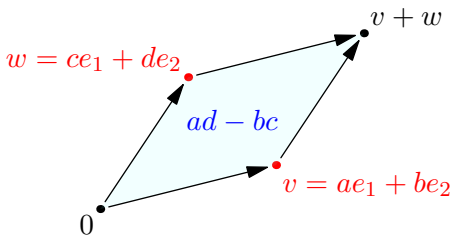
[1] So maybe the wedge product might be more accurately called the "wedge power"!

Example 12.1.3 (Wedge product explicit computation)
Let $V = \mathbb{R}^2$, and let $v = ae_1 + be_2$, $w = ce_1 + de_2$. Now let's compute $v \wedge w$ in $\bigwedge^2(V)$.

$$\begin{aligned} v \wedge w &= (ae_1 + be_2) \wedge (ce_1 + de_2) \\ &= ac(e_1 \wedge e_1) + bd(e_2 \wedge e_2) + ad(e_1 \wedge e_2) + bc(e_2 \wedge e_1) \\ &= ad(e_1 \wedge e_2) + bc(e_2 \wedge e_1) \\ &= (ad - bc)(e_1 \wedge e_2). \end{aligned}$$

What is $ad - bc$? You might already recognize it:

- You might know that the area of the parallelogram formed by v and w is $ad - bc$.
- You might recognize it as the determinant of $\begin{bmatrix} a & c \\ b & d \end{bmatrix}$. In fact, you might even know that the determinant is meant to interpret hypervolumes.



This is absolutely no coincidence. The wedge product is designed to interpret signed areas. That is, $v \wedge w$ is meant to interpret the area of the parallelogram formed by v and w. You can see why the condition $(cv) \wedge w = v \wedge (cw)$ would make sense now. And now of course you know why $v \wedge v$ ought to be zero: it's an area zero parallelogram!

The **miracle of wedge products** is that the only additional condition we need to add to the tensor product axioms is that $v \wedge v = 0$. Then suddenly, the wedge will do all our work of interpreting volumes for us.

Remark 12.1.4 (Side digression on definitions in mathematics) This "property-based" philosophy is a common trope in modern mathematics. You have some intuition about an object you wish to define, and then you write down a wishlist of properties that "should" follow. But then it turns out the properties are sufficient to work with, and so for the definition, you just define an abstract object satisfying all the properties on your wishlist. Thereafter the intuition plays no "official" role; it serves only as cheerleading motivation for the wishlist.
For wedge products, the wishlist has only the single property $v \wedge v = 0$.

In analog to earlier:

Proposition 12.1.5 (Basis of $\bigwedge^2(V)$)
Let V be a vector space with basis $e_1, \dots, e_n$. Then a basis of $\bigwedge^2(V)$ is

$$e_i \wedge e_j$$

where $i < j$. Hence $\bigwedge^2(V)$ has dimension $\binom{n}{2}$.

Proof. Surprisingly slippery, and also omitted. (You can derive it from the corresponding theorem on tensor products.) □

Now I have the courage to define a multi-dimensional wedge product. It's just the same thing with more wedges.

Definition 12.1.6. Let V be a vector space and m a positive integer. The space $\bigwedge^m(V)$ is generated by wedges of the form

$$v_1 \wedge v_2 \wedge \cdots \wedge v_m$$

subject to relations

$$\begin{aligned}
\cdots \wedge (v_1 + v_2) \wedge \ldots &= (\cdots \wedge v_1 \wedge \ldots) + (\cdots \wedge v_2 \wedge \ldots) \\
\cdots \wedge (cv_1) \wedge v_2 \wedge \ldots &= \cdots \wedge v_1 \wedge (cv_2) \wedge \ldots \\
\cdots \wedge v \wedge v \wedge \ldots &= 0 \\
\cdots \wedge v \wedge w \wedge \ldots &= -(\cdots \wedge w \wedge v \wedge \ldots)
\end{aligned}$$

As a vector space

$$c \cdot (v_1 \wedge v_2 \wedge \cdots \wedge v_m) = (cv_1) \wedge v_2 \wedge \cdots \wedge v_m = v_1 \wedge (cv_2) \wedge \cdots \wedge v_m = \dots.$$

This definition is pretty wordy, but in English the three conditions say

- We should be able to add products like before,
- You can put constants onto any of the m components (as is directly pointed out in the "vector space" action), and
- Switching any two *adjacent* wedges negates the whole wedge.

So this is the natural generalization of $\bigwedge^2(V)$. You can convince yourself that any element of the form

$$\cdots \wedge v \wedge \cdots \wedge v \wedge \ldots$$

should still be zero.

Just like $e_1 \wedge e_2$ was a basis earlier, we can find the basis for general m and n.

Proposition 12.1.7 (Basis of the wedge product)
Let V be a vector space with basis $e_1, \dots, e_n$. A basis for $\bigwedge^m(V)$ consists of the elements

$$e_{i_1} \wedge e_{i_2} \wedge \cdots \wedge e_{i_m}$$

where

$$1 \le i_1 < i_2 < \cdots < i_m \le n.$$

Hence $\bigwedge^m(V)$ has dimension $\binom{n}{m}$.

Sketch of proof. We knew earlier that $e_{i_1} \otimes \cdots \otimes e_{i_m}$ was a basis for the tensor product. Here we have the additional property that (a) if two basis elements re-appear then the whole thing becomes zero, thus we should assume the i's are all distinct; and (b) we can shuffle around elements, and so we arbitrarily decide to put the basis elements in increasing order. □

§12.2 The determinant

Prototypical example for this section: $(ae_1 + be_2) \wedge (ce_1 + de_2) = (ad - bc)(e_1 \wedge e_2)$.

Now we're ready to define the determinant. Suppose $T \colon V \to V$ is a square matrix. We claim that the map $\bigwedge^m(V) \to \bigwedge^m(V)$ given on wedges by

$$v_1 \wedge v_2 \wedge \cdots \wedge v_m \mapsto T(v_1) \wedge T(v_2) \wedge \cdots \wedge T(v_m).$$

and extending linearly to all of $\bigwedge^m(V)$ is a well-defined linear map (Here "well-defined" means that equivalent elements of the domain get mapped to equivalent elements of the codomain. This, and linearity, both follow from T being a linear map.) We call that map $\bigwedge^m(T)$.

Example 12.2.1 (Example of $\bigwedge^m(T)$)
In $V = \mathbb{R}^4$ with standard basis e_1, e_2, e_3, e_4, let $T(e_1) = e_2$, $T(e_2) = 2e_3$, $T(e_3) = e_3$ and $T(e_4) = 2e_2 + e_3$. Then, for example, $\bigwedge^2(T)$ sends

$$\begin{aligned}(e_1 \wedge e_2) + (e_3 \wedge e_4) &\mapsto T(e_1) \wedge T(e_2) + T(e_3) \wedge T(e_4) \\ &= e_2 \wedge 2e_3 + e_3 \wedge (2e_2 + e_3) \\ &= 2(e_2 \wedge e_3 + e_3 \wedge e_2) \\ &= 0.\end{aligned}$$

Now here's something interesting. Suppose V has dimension n, and let $m = n$. Then $\bigwedge^n(V)$ has dimension $\binom{n}{n} = 1$ — it's a one dimensional space! Hence $\bigwedge^n(V) \cong k$.

So $\bigwedge^n(T)$ can be thought of as a linear map from k to k. But we know that *a linear map from k to k is just multiplication by a constant.* Hence $\bigwedge^n(T)$ is multiplication by some constant.

Definition 12.2.2. Let $T \colon V \to V$, where V is an n-dimensional vector space. Then $\bigwedge^n(T)$ is multiplication by a constant c; we define the **determinant** of T as $c = \det T$.

Example 12.2.3 (The determinant of a 2×2 matrix)
Let $V = \mathbb{R}^2$ again with basis e_1 and e_2. Let

$$T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}.$$

In other words, $T(e_1) = ae_1 + be_2$ and $T(e_2) = ce_1 + de_2$.
Now let's consider $\bigwedge^2(V)$. It has a basis $e_1 \wedge e_2$. Now $\bigwedge^2(T)$ sends it to

$$e_1 \wedge e_2 \xmapsto{\bigwedge^2(T)} T(e_1) \wedge T(e_2) = (ae_1 + be_2) \wedge (ce_1 + de_2) = (ad - bc)(e_1 \wedge e_2).$$

So $\bigwedge^2(T) \colon \bigwedge^2(V) \to \bigwedge^2(V)$ is multiplication by $\det T = ad - bc$, because it sent

$e_1 \wedge e_2$ to $(ad-bc)(e_1 \wedge e_2)$.

And that is the definition of a determinant. Once again, since we defined it in terms of $\bigwedge^n(T)$, this definition is totally independent of the choice of basis. In other words, the determinant can be defined based on $T\colon V \to V$ alone without any reference to matrices.

Question 12.2.4. Why does $\bigwedge^n(S \circ T) = \bigwedge^n(S) \circ \bigwedge^n(T)$?

In this way, we also get

$$\det(S \circ T) = \det(S)\det(T)$$

for free.

More generally if we replace 2 by n, an write out the result of expanding

$$(a_{11}e_1 + a_{21}e_2 + \cdots) \wedge \cdots \wedge (a_{1n}e_1 + a_{2n}e_2 + \cdots + a_{nn}e_n)$$

then you will get the formula

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) a_{1,\sigma(1)} a_{2,\sigma(2)} \dots a_{n,\sigma(n)}$$

called the **Leibniz formula** for determinants. American high school students will recognize it; this is (unfortunately) taught as the definition of the determinant, rather than a corollary of the better definition using wedge products.

Exercise 12.2.5. Verify that expanding the wedge product yields the Leibniz formula for $n = 3$.

§12.3 Characteristic polynomials, and Cayley-Hamilton

Let's connect with the theory of eigenvalues. Take a map $T\colon V \to V$, where V is n-dimensional over an algebraically closed field, and suppose its eigenvalues are λ_1, λ_2, $\dots$, λ_n (with repetition). Then the **characteristic polynomial** is given by

$$p_T(X) = (X-\lambda_1)(X-\lambda_2)\dots(X-\lambda_n).$$

Note that if we've written T in Jordan form, that is,

$$T = \begin{bmatrix} \lambda_1 & * & 0 & \dots & 0 \\ 0 & \lambda_2 & * & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

(here each $*$ is either 0 or 1), then we can hack together the definition

$$p_T(X) \coloneqq \det\left(X \cdot \mathrm{id}_n - T\right) = \det \begin{bmatrix} X-\lambda_1 & * & 0 & \dots & 0 \\ 0 & X-\lambda_2 & * & \dots & 0 \\ 0 & 0 & X-\lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & X-\lambda_n \end{bmatrix}.$$

The latter definition is what you'll see in most linear algebra books because it lets you define the characteristic polynomial without mentioning the word "eigenvalue" (i.e. entirely in terms of arrays of numbers). I'll admit it does have the merit that it means that given any matrix, it's easy to compute the characteristic polynomial and hence compute the eigenvalues; but I still think the definition should be done in terms of eigenvalues to begin with. For instance the determinant definition obscures the following theorem, which is actually a complete triviality.

Theorem 12.3.1 (Cayley-Hamilton)

Let $T\colon V \to V$ be a map of finite-dimensional vector spaces over an algebraically closed field. Then for any $T\colon V \to V$, the map $p_T(T)$ is the zero map.

Here, by $p_T(T)$ we mean that if

$$p_T(X) = X^n + c_{n-1}X^{n-1} + \cdots + c_0$$

then

$$p_T(T) = T^n + c_{n-1}T^{n-1} + \cdots + c_1 T + c_0 I$$

is the zero map, where T^k denotes T applied k times. We saw this concept already when we proved that T had at least one nonzero eigenvector.

Example 12.3.2 (Example of Cayley-Hamilton using determinant definition)

Suppose $T = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$. Using the determinant definition of characteristic polynomial, we find that $p_T(X) = (X-1)(X-4) - (-2)(-3) = X^2 - 5X - 2$. Indeed, you can verify that

$$T^2 - 5T - 2 = \begin{bmatrix} 7 & 10 \\ 15 & 22 \end{bmatrix} - 5 \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - 2 \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

If you define p_T without the word eigenvalue, and adopt the evil view that matrices are arrays of numbers, then this looks like a complete miracle. (Indeed, just look at the terrible proofs on Wikipedia.)

But if you use the abstract viewpoint of T as a linear map, then the theorem is almost obvious:

Proof of Cayley-Hamilton. Suppose we write V in Jordan normal form as

$$V = J_1 \oplus \cdots \oplus J_m$$

where J_i has eigenvalue λ_i and dimension d_i. By definition,

$$p_T(T) = (T - \lambda_1)^{d_1}(T - \lambda_2)^{d_2} \dots (T - \lambda_m)^{d_m}.$$

By definition, $(T - \lambda_1)^{d_1}$ is the zero map on J_1. So $p_T(T)$ is zero on J_1. Similarly it's zero on each of the other J_i's — end of story. □

Remark 12.3.3 (Tensoring up) — The Cayley-Hamilton theorem holds without the hypothesis that k is algebraically closed: because for example any real matrix can be regarded as a matrix with complex coefficients (a trick we've mentioned before). I'll briefly hint at how you can use tensor products to formalize this idea.
Let's take the space $V = \mathbb{R}^3$, with basis e_1, e_2, e_3. Thus objects in V are of the form $r_1e_1 + r_2e_2 + r_3e_3$ where r_1, r_2, r_3 are real numbers. We want to consider essentially the same vector space, but with complex coefficients z_i rather than real coefficients r_i.
So here's what we do: view $\mathbb{C}$ as a $\mathbb{R}$-vector space (with basis $\{1, i\}$, say) and consider the **complexification**

$$V_{\mathbb{C}} \coloneqq \mathbb{C} \otimes_{\mathbb{R}} V.$$

Then you can check that our elements are actually of the form

$$z_1 \otimes e_1 + z_2 \otimes e_2 + z_3 \otimes e_3.$$

Here, the tensor product is over $\mathbb{R}$, so we have $z \otimes re_i = (zr) \otimes e_i$ for $r \in \mathbb{R}$. Then $V_{\mathbb{C}}$ can be thought as a three-dimensional vector space over $\mathbb{C}$, with basis $1 \otimes e_i$ for $i \in \{1, 2, 3\}$. In this way, the tensor product lets us formalize the idea that we "fuse on" complex coefficients.
If $T \colon V \to W$ is a map, then $T_{\mathbb{C}} \colon V_{\mathbb{C}} \to W_{\mathbb{C}}$ is just the map $z \otimes v \mapsto z \otimes T(v)$. You'll see this written sometimes as $T_{\mathbb{C}} = \mathrm{id} \otimes T$. One can then apply theorems to $T_{\mathbb{C}}$ and try to deduce the corresponding results on T.

§12.4 A few harder problems to think about

Problem 12A (Column operations)**.** Show that for any real numbers x_{ij} (here $1 \le i, j \le n$) we have

$$\det \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{bmatrix} = \det \begin{bmatrix} x_{11} + cx_{12} & x_{12} & \dots & x_{1n} \\ x_{21} + cx_{22} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} + cx_{n2} & x_{n2} & \dots & x_{nn} \end{bmatrix}.$$

Problem 12B (Determinant is product of eigenvalues)**.** Let V be an n-dimensional vector space over an algebraically closed field k. Let $T \colon V \to V$ be a linear map with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ (counted with algebraic multiplicity). Show that $\det T = \lambda_1 \dots \lambda_n$.

Problem 12C (Exponential matrix)**.** Let X be an $n \times n$ matrix with complex coefficients. We define the exponential map by

$$\exp(X) = 1 + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots$$

(take it for granted that this converges to some $n \times n$ matrix). Prove that

$$\det(\exp(X)) = e^{\operatorname{Tr} X}.$$

Problem 12D (Extension to Problem 9B†)**.** Let $T \colon V \to V$ be a map of finite-dimensional vector spaces. Prove that T is an isomorphism if and only if $\det T \neq 0$.

Problem 12E (Based on Sweden 2010)**.** A herd of 1000 cows of nonzero weight is given. Prove that we can remove one cow such that the remaining 999 cows cannot be partitioned into two sets with equal sum of weights.

Problem 12F (Putnam 2015)**.** Define S to be the set of real matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ such that a, b, c, d form an arithmetic progression in that order. Find all $M \in S$ such that for some integer $k > 1$, $M^k \in S$.

Problem 12G. Let V be a finite-dimensional vector space over k and $T \colon V \to V$. Show that

$$\det(a \cdot \mathrm{id}_V - T) = \sum_{n=0}^{\dim V} a^{\dim V - n} \cdot (-1)^n \operatorname{Tr}\left(\bigwedge^n (T)\right)$$

where the trace is taken by viewing $\bigwedge^n(T) \colon \bigwedge^n(V) \to \bigwedge^n(V)$.

13 Inner product spaces

It will often turn out that our vector spaces which look more like $\mathbb{R}^n$ not only have the notion of addition, but also a notion of *orthogonality* and the notion of *distance*. All this is achieved by endowing the vector space with a so-called **inner form**, which you likely already know as the "dot product" for $\mathbb{R}^n$. Indeed, in $\mathbb{R}^n$ you already know that

- $v \cdot w = 0$ if and only if v and w are perpendicular, and
- $|v|^2 = v \cdot v$.

The purpose is to quickly set up this structure in full generality. Some highlights of the chapter:

- We'll see that the high school "dot product" formulation is actually very natural: it falls out from the two axioms we listed above. If you ever wondered why $\sum a_i b_i$ behaves as nicely as it does, now you'll know.
- We show how the inner form can be used to make V into a *metric space*, giving it more geometric structure.
- A few chapters later, we'll identify $V \cong V^\vee$ in a way that wasn't possible before, and as a corollary deduce the nice result that symmetric matrices with real entries always have real eigenvalues.

Throughout this chapter, *all vector spaces are over* $\mathbb{C}$ *or* $\mathbb{R}$, unless otherwise specified. We'll generally prefer working over $\mathbb{C}$ instead of $\mathbb{R}$ since $\mathbb{C}$ is algebraically closed (so, e.g. we have Jordan forms). Every real matrix can be thought of as a matrix with complex entries anyways.

§13.1 The inner product

Prototypical example for this section: Dot product in $\mathbb{R}^n$.

§13.1.i For real numbers: bilinear forms

First, let's define the inner form for real spaces. Rather than the notation $v \cdot w$ it is most customary to use $\langle v, w \rangle$ for general vector spaces.

Definition 13.1.1. Let V be a real vector space. A **real inner form**[1] is a function

$$\langle \bullet, \bullet \rangle : V \times V \to \mathbb{R}$$

which satisfies the following properties:

- The form is **symmetric**: for any $v, w \in V$ we have

$$\langle v, w \rangle = \langle w, v \rangle .$$

  Of course, one would expect this property from a product.

[1] Other names include "inner product", "dot product", "positive definite nondegenerate symmetric bilinear form", ...

- The form is **bilinear**, or **linear in both arguments**, meaning that $\langle -, v\rangle$ and $\langle v, -\rangle$ are linear functions for any fixed v. Spelled explicitly this means that
$$\begin{aligned} \langle cx, v\rangle &= c\langle x, v\rangle \\ \langle x+y, v\rangle &= \langle x, v\rangle + \langle y, v\rangle . \end{aligned}$$
and similarly if v was on the left. This is often summarized by the single equation $\langle cx+y, z\rangle = c\langle x, z\rangle + \langle y, z\rangle$.
- The form is **positive definite**, meaning $\langle v, v\rangle \geq 0$ is a nonnegative real number, and equality takes place only if $v = 0_V$.

Exercise 13.1.2. Show that linearity in the first argument plus symmetry already gives you linearity in the second argument, so we could edit the above definition by only requiring $\langle -, v\rangle$ to be linear.

Example 13.1.3 ($\mathbb{R}^n$)
As we already know, one can define the inner form on $\mathbb{R}^n$ as follows. Let $e_1 = (1, 0, \ldots, 0)$, $e_2 = (0, 1, \ldots, 0)$, $\ldots$, $e_n = (0, \ldots, 0, 1)$ be the usual basis. Then we let
$$\langle a_1e_1 + \cdots + a_ne_n, b_1e_1 + \cdots + b_ne_n\rangle \coloneqq a_1b_1 + \cdots + a_nb_n.$$
It's easy to see this is bilinear (symmetric and linear in both arguments). To see it is positive definite, note that if $a_i = b_i$ then the dot product is $a_1^2 + \cdots + a_n^2$, which is zero exactly when all a_i are zero.

§13.1.ii For complex numbers: sesquilinear forms

The definition for a complex product space is similar, but has one difference: rather than symmetry we instead have *conjugate symmetry* meaning $\langle v, w\rangle = \overline{\langle w, v\rangle}$. Thus, while we still have linearity in the first argument, we actually have a different linearity for the second argument. To be explicit:

Definition 13.1.4. Let V be a complex vector space. A **complex inner product** is a function
$$\langle \bullet, \bullet\rangle : V \times V \to \mathbb{C}$$
which satisfies the following properties:

- The form has **conjugate symmetry**, which means that for any $v, w \in V$ we have
$$\langle v, w\rangle = \overline{\langle w, v\rangle}.$$
- The form is **sesquilinear** (the name means "one-and-a-half linear"). This means that:
  - The form is **linear in the first argument**, so again we have
$$\begin{aligned} \langle x+y, v\rangle &= \langle x, v\rangle + \langle y, v\rangle \\ \langle cx, v\rangle &= c\langle x, v\rangle . \end{aligned}$$
Again this is often abbreviated to the single line $\langle cx+y, v\rangle = c\langle x, v\rangle + \langle y, v\rangle$ in the literature.

– However, it is now **anti-linear in the second argument**: for any complex number c and vectors x and y we have

$$\langle v, x+y\rangle = \langle v, x\rangle + \langle v, y\rangle$$
$$\langle v, cx\rangle = \overline{c}\,\langle v, x\rangle\,.$$

Note the appearance of the complex conjugate $\overline{c}$, which is new! Again, we can abbreviate this to just $\langle v, cx+y\rangle = \overline{c}\,\langle v, x\rangle + \langle v, y\rangle$ if we only want to write one equation.

- The form is **positive definite**, meaning $\langle v, v\rangle$ is a nonnegative real number, and equals zero exactly when $v = 0_V$.

Exercise 13.1.5. Show that anti-linearity follows from conjugate symmetry plus linearity in the first argument.

Example 13.1.6 ($\mathbb{C}^n$)
The dot product in $\mathbb{C}^n$ is defined as follows: let $\mathbf{e}_1$, $\mathbf{e}_2$, …, $\mathbf{e}_n$ be the standard basis. For complex numbers w_i, z_i we set

$$\langle w_1\mathbf{e}_1 + \cdots + w_n\mathbf{e}_n, z_1\mathbf{e}_1 + \cdots + z_n\mathbf{e}_n\rangle \coloneqq w_1\overline{z_1} + \cdots + w_n\overline{z_n}.$$

Question 13.1.7. Check that the above is in fact a complex inner form.

§13.1.iii Inner product space

It'll be useful to treat both types of spaces simultaneously:

Definition 13.1.8. An **inner product space** is either a real vector space equipped with a real inner form, or a complex vector space equipped with a complex inner form.

A linear map between inner product spaces is a map between the underlying vector spaces (we do *not* require any compatibility with the inner form).

Remark 13.1.9 (Why sesquilinear?) — The above example explains one reason why we want to satisfy conjugate symmetry rather than just symmetry. If we had tried to define the dot product as $\sum w_i z_i$, then we would have lost the condition of being positive definite, because there is no guarantee that $\langle v, v\rangle = \sum z_i^2$ will even be a real number at all. On the other hand, with conjugate symmetry we actually enforce $\langle v, v\rangle = \overline{\langle v, v\rangle}$, i.e. $\langle v, v\rangle \in \mathbb{R}$ for every v.
Let's make this point a bit more forcefully. Suppose we tried to put a bilinear form $\langle -, -\rangle$, on a *complex* vector space V. Let e be any vector with $\langle e, e\rangle = 1$ (a unit vector). Then we would instead get $\langle ie, ie\rangle = -\langle e, e\rangle = -1$; this is a vector with length $\sqrt{-1}$, which is not okay! That's why it is important that, when we have a complex inner product space, our form is sesquilinear, not bilinear.

Now that we have a dot product, we can talk both about the norm and orthogonality.

§13.2 Norms

Prototypical example for this section: $\mathbb{R}^n$ *becomes its usual Euclidean space with the vector norm.*

The inner form equips our vector space with a notion of distance, which we call the norm.

Definition 13.2.1. Let V be an inner product space. The **norm** of $v \in V$ is defined by

$$\|v\| = \sqrt{\langle v, v \rangle}.$$

This definition makes sense because we assumed our form to be positive definite, so $\langle v, v \rangle$ is a nonnegative real number.

Example 13.2.2 ($\mathbb{R}^n$ and $\mathbb{C}^n$ are normed vector spaces)
When $V = \mathbb{R}^n$ or $V = \mathbb{C}^n$ with the standard dot product norm, then the norm of v corresponds to the absolute value that we are used to.

Our goal now is to prove that

With the metric $d(v, w) = \|v - w\|$, V becomes a metric space.

Question 13.2.3. Verify that $d(v, w) = 0$ if and only if $v = w$.

So we just have to establish the triangle inequality. Let's now prove something we all know and love, which will be a stepping stone later:

Lemma 13.2.4 (Cauchy-Schwarz)
Let V be an inner product space. For any $v, w \in V$ we have

$$|\langle v, w \rangle| \le \|v\| \, \|w\|$$

with equality if and only if v and w are linearly dependent.

Proof. The theorem is immediate if $\langle v, w \rangle = 0$. It is also immediate if $\|v\| \, \|w\| = 0$, since then one of v or w is the zero vector. So henceforth we assume all these quantities are nonzero (as we need to divide by them later).

The key to the proof is to think about the equality case: we'll use the inequality $\langle cv - w, cv - w \rangle \ge 0$. Deferring the choice of c until later, we compute

$$\begin{aligned}
0 &\le \langle cv - w, cv - w \rangle \\
&= \langle cv, cv \rangle - \langle cv, w \rangle - \langle w, cv \rangle + \langle w, w \rangle \\
&= |c|^2 \langle v, v \rangle - c \langle v, w \rangle - \overline{c} \langle w, v \rangle + \langle w, w \rangle \\
&= |c|^2 \|v\|^2 + \|w\|^2 - c \langle v, w \rangle - \overline{c \langle v, w \rangle} \\
2 \operatorname{Re} [c \langle v, w \rangle] &\le |c|^2 \|v\|^2 + \|w\|^2
\end{aligned}$$

At this point, a good choice of c is

$$c = \frac{\|w\|}{\|v\|} \cdot \frac{|\langle v, w\rangle|}{\langle v, w\rangle}$$

since then

$$c\langle v, w\rangle = \frac{\|w\|}{\|v\|} |\langle v, w\rangle| \in \mathbb{R}$$
$$|c| = \frac{\|w\|}{\|v\|}$$

whence the inequality becomes

$$2\frac{\|w\|}{\|v\|} |\langle v, w\rangle| \leq 2\|w\|^2$$
$$|\langle v, w\rangle| \leq \|v\| \|w\|.$$

□

Thus:

Theorem 13.2.5 (Triangle inequality)
We always have

$$\|v\| + \|w\| \geq \|v + w\|$$

with equality if and only if v and w are linearly dependent and point in the same direction.

Exercise 13.2.6. Prove this by squaring both sides, and applying Cauchy-Schwarz.

In this way, our vector space now has a topological structure of a metric space.

§13.3 Orthogonality

Prototypical example for this section: Still $\mathbb{R}^n$!

Our next goal is to give the geometric notion of "perpendicular". The definition is easy enough:

Definition 13.3.1. Two nonzero vectors v and w in an inner product space are **orthogonal** if $\langle v, w\rangle = 0$.

As we expect from our geometric intuition in $\mathbb{R}^n$, this implies independence:

Lemma 13.3.2 (Orthogonal vectors are independent)
Any set of pairwise orthogonal vectors $v_1, v_2, \ldots, v_n$, with $\|v_i\| \neq 0$ for each i, is linearly independent.

Proof. Consider a dependence

$$a_1 v_1 + \cdots + a_n v_n = 0$$

for a_i in $\mathbb{R}$ or $\mathbb{C}$. Then

$$0 = \left\langle v_1, \sum a_i v_i \right\rangle = \overline{a_1} \|v_1\|^2.$$

Hence $a_1 = 0$, since we assumed $\|v_1\| \neq 0$. Similarly $a_2 = \cdots = a_m = 0$. □

In light of this, we can now consider a stronger condition on our bases:

Definition 13.3.3. An **orthonormal** basis of a *finite-dimensional* inner product space V is a basis $e_1, \dots, e_n$ such that $\|e_i\| = 1$ for every i and $\langle e_i, e_j \rangle = 0$ for any $i \neq j$.

Example 13.3.4 ($\mathbb{R}^n$ and $\mathbb{C}^n$ have standard bases)
In $\mathbb{R}^n$ and $\mathbb{C}^n$ equipped with the standard dot product, the standard basis $\mathbf{e}_1, \dots, \mathbf{e}_n$ is also orthonormal.

This is no loss of generality:

Theorem 13.3.5 (Gram-Schmidt)
Let V be a finite-dimensional inner product space. Then it has an orthonormal basis.

Sketch of Proof. One constructs the orthonormal basis explicitly from any basis $e_1, \dots, e_n$ of V. Define $\text{proj}_u(v) = \frac{\langle v,u \rangle}{\langle u,u \rangle} u$. Then recursively define

$$\begin{aligned}
u_1 &= e_1 \\
u_2 &= e_2 - \text{proj}_{u_1}(e_2) \\
u_3 &= e_3 - \text{proj}_{u_1}(e_3) - \text{proj}_{u_2}(e_3) \\
&\vdots \\
u_n &= e_n - \text{proj}_{u_1}(e_n) - \cdots - \text{proj}_{u_{n-1}}(e_n).
\end{aligned}$$

One can show the u_i are pairwise orthogonal and not zero. □

Thus, we can generally assume our bases are orthonormal.
Worth remarking:

Example 13.3.6 (The dot product is the "only" inner form)
Let V be a finite-dimensional inner product space, and consider *any* orthonormal basis $e_1, \dots, e_n$. Then we have that

$$\langle a_1 e_1 + \cdots + a_n e_n, b_1 e_1 + \cdots + b_n e_n \rangle = \sum_{i,j=1}^{n} a_i \overline{b_j} \langle e_i, e_j \rangle = \sum_{i=1}^{n} a_i \overline{b_i}$$

owing to the fact that the $\{e_i\}$ are orthonormal.

And now you know why the dot product expression is so ubiquitous.

§13.4 Hilbert spaces

In algebra we are usually scared of infinity, and so when we defined a basis of a vanilla vector space many chapters ago, we only allowed finite linear combinations. However, if we have an inner product space, then it is a metric space and we *can* sometimes actually talk about convergence.

Here is how it goes:

Definition 13.4.1. A **Hilbert space** is an inner product space V, such that the corresponding metric space is complete.

In that case, it will now often make sense to take infinite linear combinations, because we can look at the sequence of partial sums and let it converge. Here is how we might do it. Let's suppose we have e_1, e_2, ... an infinite sequence of vectors with norm 1 and which are pairwise orthogonal. Suppose c_1, c_2, ..., is a sequence of real or complex numbers. Then consider the sequence

$$\begin{aligned} v_1 &= c_1 e_1 \\ v_2 &= c_1 e_1 + c_2 e_2 \\ v_3 &= c_1 e_1 + c_2 e_2 + c_3 e_3 \\ &\vdots \end{aligned}$$

Proposition 13.4.2 (Convergence criteria in a Hilbert space)
The sequence (v_i) defined above converges if and only if $\sum |c_i|^2 < \infty$.

Proof. This will make more sense if you read Chapter 26, so you could skip this proof if you haven't read the chapter. The sequence v_i converges if and only if it is Cauchy, meaning that when $i < j$,

$$\|v_j - v_i\|^2 = |c_{i+1}|^2 + \cdots + |c_j|^2$$

tends to zero as i and j get large. This is equivalent to the sequence $s_n = |c_1|^2 + \cdots + |c_n|^2$ being Cauchy.

Since $\mathbb{R}$ is complete, s_n is Cauchy if and only if it converges. Since s_n consists of nonnegative real numbers, converges holds if and only if s_n is bounded, or equivalently if $\sum |c_i|^2 < \infty$. □

Thus, when we have a Hilbert space, we change our definition slightly:

Definition 13.4.3. An **orthonormal basis** for a Hilbert space V is a (possibly infinite) sequence e_1, e_2, ..., of vectors such that

- $\langle e_i, e_i \rangle = 1$ for all i,
- $\langle e_i, e_j \rangle = 0$ for $i \neq j$, i.e. the vectors are pairwise orthogonal
- every element of V can be expressed uniquely as an infinite linear combination

$$\sum_i c_i e_i$$

where $\sum_i |c_i|^2 < \infty$, as described above.

That's the official definition, anyways. (Note that if $\dim V < \infty$, this agrees with our usual definition, since then there are only finitely many e_i.) But for our purposes you can mostly not worry about it and instead think:

A Hilbert space is an inner product space whose basis requires infinite linear combinations, not just finite ones.

The technical condition $\sum |c_i|^2 < \infty$ is exactly the one which ensures the infinite sum makes sense.

§13.5 A few harder problems to think about

Problem 13A (Pythagorean theorem)**.** Show that if $\langle v, w \rangle = 0$ in an inner product space, then $\|v\|^2 + \|w\|^2 = \|v+w\|^2$.

Problem 13B* (Finite-dimensional $\implies$ Hilbert)**.** Show that a finite-dimensional inner product space is a Hilbert space.

Problem 13C (Taiwan IMO camp)**.** In a town there are n people and k clubs. Each club has an odd number of members, and any two clubs have an even number of common members. Prove that $k \le n$.

Problem 13D* (Inner product structure of tensors)**.** Let V and W be finite-dimensional inner product spaces over k, where k is either $\mathbb{R}$ or $\mathbb{C}$.

(a) Find a canonical way to make $V \otimes_k W$ into an inner product space too.

(b) Let $e_1, \dots, e_n$ be an orthonormal basis of V and $f_1, \dots, f_m$ be an orthonormal basis of W. What's an orthonormal basis of $V \otimes W$?

Problem 13E (Putnam 2014)**.** Let n be a positive integer. What is the largest k for which there exist $n \times n$ matrices $M_1, \dots, M_k$ and $N_1, \dots, N_k$ with real entries such that for all i and j, the matrix product $M_i N_j$ has a zero entry somewhere on its diagonal if and only if $i \neq j$?

Problem 13F (Sequence space)**.** Consider the space ℓ^2 of infinite sequences of real numbers $a = (a_1, a_2, \dots)$ satisfying $\sum_i a_i^2 < \infty$. We equip it with the dot product

$$\langle a, b \rangle = \sum_i a_i b_i.$$

Is this a Hilbert space? If so, identify a Hilbert basis.

Problem 13G (Kuratowski embedding)**.** A **Banach space** is a normed vector space V, such that the corresponding metric space is complete. (So a Hilbert space is a special case of a Banach space.)

Let (M, d) be any metric space. Prove that there exists a Banach space X and an injective function $f \colon M \hookrightarrow X$ such that $d(x,y) = \|f(x) - f(y)\|$ for any x and y.

14 Bonus: Fourier analysis

Now that we've worked hard to define abstract inner product spaces, I want to give an (optional) application: how to set up Fourier analysis correctly, using this language.

For fun, I also prove a form of Arrow's Impossibility Theorem using binary Fourier analysis.

In what follows, we let $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ denote the "circle group", thought of as the additive group of "real numbers modulo 1". There is a canonical map $e \colon \mathbb{T} \to \mathbb{C}$ sending $\mathbb{T}$ to the complex unit circle, given by

$$e(\theta) = \exp(2\pi i\theta).$$

§14.1 Synopsis

Suppose we have a domain Z and are interested in functions $f \colon Z \to \mathbb{C}$. Naturally, the set of such functions form a complex vector space. We like to equip the set of such functions with a positive definite *inner product*.

The idea of Fourier analysis is to then select an *orthonormal basis* for this set of functions, say $(e_\xi)_\xi$, which we call the **characters**; the indexing ξ are called **frequencies**. In that case, since we have a basis, every function $f \colon Z \to \mathbb{C}$ becomes a sum

$$f(x) = \sum_\xi \widehat{f}(\xi) e_\xi$$

where $\widehat{f}(\xi)$ are complex coefficients of the basis; appropriately we call $\widehat{f}$ the **Fourier coefficients**. The variable $x \in Z$ is referred to as the **physical** variable. This is generally good because the characters are deliberately chosen to be nice "symmetric" functions, like sine or cosine waves or other periodic functions. Thus we decompose an arbitrarily complicated function into a sum of nice ones.

§14.2 A reminder on Hilbert spaces

For convenience, we record a few facts about orthonormal bases.

Proposition 14.2.1 (Facts about orthonormal bases)

Let V be a complex Hilbert space with inner form $\langle -, - \rangle$ and suppose $x = \sum_\xi a_\xi e_\xi$ and $y = \sum_\xi b_\xi e_\xi$ where e_ξ are an orthonormal basis. Then

$$\begin{aligned} \langle x, x \rangle &= \sum_\xi |a_\xi|^2 \\ a_\xi &= \langle x, e_\xi \rangle \\ \langle x, y \rangle &= \sum_\xi a_\xi \overline{b_\xi}. \end{aligned}$$

Exercise 14.2.2. Prove all of these. (You don't need any of the preceding section, it's only there to motivate the notation with lots of scary ξ's.)

In what follows, most of the examples will be of finite-dimensional inner product spaces (which are thus Hilbert spaces), but the example of "square-integrable functions" will actually be an infinite dimensional example. Fortunately, as I alluded to earlier, this is no cause for alarm and you can mostly close your eyes and not worry about infinity.

§14.3 Common examples

§14.3.i Binary Fourier analysis on $\{\pm 1\}^n$

Let $Z = \{\pm 1\}^n$ for some positive integer n, so we are considering functions $f(x_1, \dots, x_n)$ accepting binary values. Then the functions $Z \to \mathbb{C}$ form a 2^n-dimensional vector space $\mathbb{C}^Z$, and we endow it with the inner form

$$\langle f, g \rangle = \frac{1}{2^n} \sum_{x \in Z} f(x)\overline{g(x)}.$$

In particular,

$$\langle f, f \rangle = \frac{1}{2^n} \sum_{x \in Z} |f(x)|^2$$

is the average of the squares; this establishes also that $\langle -, - \rangle$ is positive definite.

In that case, the **multilinear polynomials** form a basis of $\mathbb{C}^Z$, that is the polynomials

$$\chi_S(x_1, \dots, x_n) = \prod_{s \in S} x_s.$$

Exercise 14.3.1. Show that they're actually orthonormal under $\langle -, - \rangle$. This proves they form a basis, since there are 2^n of them.

Thus our frequency set is actually the subsets $S \subseteq \{1, \dots, n\}$. Thus, we have a decomposition

$$f = \sum_{S \subseteq \{1, \dots, n\}} \widehat{f}(S)\chi_S.$$

Example 14.3.2 (An example of binary Fourier analysis)
Let $n = 2$. Then binary functions $\{\pm 1\}^2 \to \mathbb{C}$ have a basis given by the four polynomials

$$1, \quad x_1, \quad x_2, \quad x_1x_2.$$

For example, consider the function f which is 1 at $(1, 1)$ and 0 elsewhere. Then we can put

$$f(x_1, x_2) = \frac{x_1 + 1}{2} \cdot \frac{x_2 + 1}{2} = \frac{1}{4}\left(1 + x_1 + x_2 + x_1x_2\right).$$

So the Fourier coefficients are $\widehat{f}(S) = \frac{1}{4}$ for each of the four S's.

This notion is useful in particular for binary functions $f \colon \{\pm 1\}^n \to \{\pm 1\}$; for these functions (and products thereof), we always have $\langle f, f \rangle = 1$.

It is worth noting that the frequency $\varnothing$ plays a special role:

Exercise 14.3.3. Show that
$$\widehat{f}(\varnothing) = \frac{1}{|Z|} \sum_{x \in Z} f(x).$$

§14.3.ii Fourier analysis on finite groups Z

This time, suppose we have a finite abelian group Z, and consider functions $Z \to \mathbb{C}$; this is a $|Z|$-dimensional vector space. The inner product is the same as before:
$$\langle f, g \rangle = \frac{1}{|Z|} \sum_{x \in Z} f(x)\overline{g(x)}.$$

To proceed, we'll need to be able to multiply two elements of Z. This is a bit of a nuisance since it actually won't really matter what map I pick, so I'll move briskly; feel free to skip most or all of the remaining paragraph.

Definition 14.3.4. We select a *symmetric non-degenerate bilinear form*
$$\cdot \colon Z \times Z \to \mathbb{T}$$
satisfying the following properties:

- $\xi \cdot (x_1 + x_2) = \xi \cdot x_1 + \xi \cdot x_2$ and $(\xi_1 + \xi_2) \cdot x = \xi_1 \cdot x + \xi_2 \cdot x$ (this is the word "bilinear")
- $\cdot$ is symmetric,
- For any $\xi \neq 0$, there is an x with $\xi \cdot x \neq 0$ (this is the word "nondegenerate").

Example 14.3.5 (The form on $\mathbb{Z}/n\mathbb{Z}$)
If $Z = \mathbb{Z}/n\mathbb{Z}$ then $\xi \cdot x = (\xi x)/n$ satisfies the above.

In general, it turns out finite abelian groups decompose as the sum of cyclic groups (see Section 18.1), which makes it relatively easy to find such a $\cdot$; but as I said the choice won't matter, so let's move on.

Now for the fun part: defining the characters.

Proposition 14.3.6 (e_ξ are orthonormal)
For each $\xi \in Z$ we define the character
$$e_\xi(x) = e(\xi \cdot x).$$
The $|Z|$ characters form an orthonormal basis of the space of functions $Z \to \mathbb{C}$.

Proof. I recommend skipping this one, but it is:
$$\begin{aligned}
\langle e_\xi, e_{\xi'} \rangle &= \frac{1}{|Z|} \sum_{x \in Z} e(\xi \cdot x)\overline{e(\xi' \cdot x)} \\
&= \frac{1}{|Z|} \sum_{x \in Z} e(\xi \cdot x) e(-\xi' \cdot x) \\
&= \frac{1}{|Z|} \sum_{x \in Z} e\left((\xi - \xi') \cdot x\right).
\end{aligned}$$

□

In this way, the set of frequencies is also Z, but the $\xi \in Z$ play very different roles from the "physical" $x \in Z$. Here is an example which might be enlightening.

Example 14.3.7 (Cube roots of unity filter)
Suppose $Z = \mathbb{Z}/3\mathbb{Z}$, with the inner form given by $\xi \cdot x = (\xi x)/3$. Let $\omega = \exp(\frac{2}{3}\pi i)$ be a primitive cube root of unity. Note that

$$e_\xi(x) = \begin{cases} 1 & \xi = 0 \\ \omega^x & \xi = 1 \\ \omega^{2x} & \xi = 2. \end{cases}$$

Then given $f \colon Z \to \mathbb{C}$ with $f(0) = a$, $f(1) = b$, $f(2) = c$, we obtain

$$f(x) = \frac{a+b+c}{3} \cdot 1 + \frac{a+\omega^2 b+\omega c}{3} \cdot \omega^x + \frac{a+\omega b+\omega^2 c}{3} \cdot \omega^{2x}.$$

In this way we derive that the transforms are

$$\begin{aligned} \widehat{f}(0) &= \frac{a+b+c}{3} \\ \widehat{f}(1) &= \frac{a+\omega^2 b+\omega c}{3} \\ \widehat{f}(2) &= \frac{a+\omega b+\omega^2 c}{3}. \end{aligned}$$

Exercise 14.3.8. Show that in analogy to $\widehat{f}(\varnothing)$ for binary Fourier analysis, we now have

$$\widehat{f}(0) = \frac{1}{|Z|} \sum_{x \in Z} f(x).$$

Olympiad contestants may recognize the previous example as a "roots of unity filter", which is exactly the point. For concreteness, suppose one wants to compute

$$\binom{1000}{0} + \binom{1000}{3} + \cdots + \binom{1000}{999}.$$

In that case, we can consider the function

$$w \colon \mathbb{Z}/3 \to \mathbb{C}.$$

such that $w(0) = 1$ but $w(1) = w(2) = 0$. By abuse of notation we will also think of w as a function $w \colon \mathbb{Z} \twoheadrightarrow \mathbb{Z}/3 \to \mathbb{C}$. Then the sum in question is

$$\begin{aligned} \sum_n \binom{1000}{n} w(n) &= \sum_n \binom{1000}{n} \sum_{k=0,1,2} \widehat{w}(k) \omega^{kn} \\ &= \sum_{k=0,1,2} \widehat{w}(k) \sum_n \binom{1000}{n} \omega^{kn} \\ &= \sum_{k=0,1,2} \widehat{w}(k) (1+\omega^k)^{1000}. \end{aligned}$$

In our situation, we have $\widehat{w}(0) = \widehat{w}(1) = \widehat{w}(2) = \frac{1}{3}$, and we have evaluated the desired sum. More generally, we can take any periodic weight w and use Fourier analysis in order to interchange the order of summation.

Example 14.3.9 (Binary Fourier analysis)

Suppose $Z = \{\pm 1\}^n$, viewed as an abelian group under pointwise multiplication hence isomorphic to $(\mathbb{Z}/2\mathbb{Z})^{\oplus n}$. Assume we pick the dot product defined by

$$\xi \cdot x \coloneqq \frac{1}{2} \sum_i \frac{\xi_i - 1}{2} \cdot \frac{x_i - 1}{2}$$

where $\xi = (\xi_1, \dots, \xi_n)$ and $x = (x_1, \dots, x_n)$.
We claim this coincides with the first example we gave. Indeed, let $S \subseteq \{1, \dots, n\}$ and let $\xi \in \{\pm 1\}^n$ which is -1 at positions in S, and $+1$ at positions not in S. Then the character χ_S from the previous example coincides with the character e_ξ in the new notation. In particular, $\widehat{f}(S) = \widehat{f}(\xi)$.
Thus Fourier analysis on a finite group Z subsumes binary Fourier analysis.

§14.3.iii Fourier series for functions $L^2([-\pi, \pi])$

This is the most famous one, and hence the one you've heard of.

Definition 14.3.10. The space $L^2([-\pi, \pi])$ consists of all functions $f \colon [-\pi, \pi] \to \mathbb{C}$ such that the integral $\int_{[-\pi,\pi]} |f(x)|^2 \, dx$ exists and is finite, modulo the relation that a function which is zero "almost everywhere" is considered to equal zero.[1]

It is made into an inner product space according to

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{[-\pi,\pi]} f(x) \overline{g(x)} \, dx.$$

It turns out (we won't prove) that this is an (infinite-dimensional) Hilbert space!
Now, the beauty of Fourier analysis is that **this space has a great basis**:

Theorem 14.3.11 (The classical Fourier basis)

For each integer n, define

$$e_n(x) = \exp(inx).$$

Then e_n form an orthonormal basis of the Hilbert space $L^2([-\pi, \pi])$.

Thus this time the frequency set $\mathbb{Z}$ is infinite, and we have

$$f(x) = \sum_n \widehat{f}(n) \exp(inx) \quad \text{almost everywhere}$$

for coefficients $\widehat{f}(n)$ with $\sum_n \left|\widehat{f}(n)\right|^2 < \infty$. Since the frequency set is indexed by $\mathbb{Z}$, we call this a **Fourier series** to reflect the fact that the index is $n \in \mathbb{Z}$.

[1] We won't define this, yet, as it won't matter to us for now. But we will elaborate more on this in the parts on measure theory.

There is one point at which this is relevant. Often we require that the function f satisfies $f(-\pi) = f(\pi)$, so that f becomes a periodic function, and we can think of it as $f \colon \mathbb{T} \to \mathbb{C}$. This makes no essential difference since we merely change the value at one point.

Exercise 14.3.12. Show once again

$$\widehat{f}(0) = \frac{1}{2\pi} \int_{[-\pi,\pi]} f(x)\,dx.$$

§14.4 Summary, and another teaser

We summarize our various flavors of Fourier analysis in the following table.

Type	Physical var	Frequency var	Basis functions
Binary	$\{\pm 1\}^n$	Subsets $S \subseteq \{1,\dots,n\}$	$\prod_{s\in S} x_s$
Finite group	Z	$\xi \in Z$, choice of $\cdot$	$e(\xi \cdot x)$
Fourier series	$\mathbb{T}$ or $[-\pi,\pi]$	$n \in \mathbb{Z}$	$\exp(inx)$
Discrete	$\mathbb{Z}/n\mathbb{Z}$	$\xi \in \mathbb{Z}/n\mathbb{Z}$	$e(\xi x/n)$

I snuck in a fourth row with $Z = \mathbb{Z}/n\mathbb{Z}$, but it's a special case of the second row, so no cause for alarm.

Alluding to the future, I want to hint at how Chapter 39 starts. Each one of these is really a statement about how functions from $G \to \mathbb{C}$ can be expressed in terms of functions $\widehat{G} \to \mathbb{C}$, for some "dual" $\widehat{G}$. In that sense, we could rewrite the above table as:

Name	Domain G	Dual $\widehat{G}$	Characters
Binary	$\{\pm 1\}^n$	$S \subseteq \{1,\dots,n\}$	$\prod_{s\in S} x_s$
Finite group	Z	$\xi \in \widehat{Z} \cong Z$	$e(i\xi \cdot x)$
Fourier series	$\mathbb{T} \cong [-\pi,\pi]$	$n \in \mathbb{Z}$	$\exp(inx)$
Discrete	$\mathbb{Z}/n\mathbb{Z}$	$\xi \in \mathbb{Z}/n\mathbb{Z}$	$e(\xi x/n)$

It will turn out that in general we can say something about many different domains G, once we know what it means to integrate a measure. This is the so-called *Pontryagin duality*; and it is discussed as a follow-up bonus in Chapter 39.

§14.5 Parseval and friends

Here is a fun section in which you get to learn a lot of big names quickly. Basically, we can take each of the three results from Proposition 14.2.1, translate it into the context of our Fourier analysis (for which we have an orthonormal basis of the Hilbert space), and get a big-name result.

Corollary 14.5.1 (Parseval theorem)

Let $f\colon Z \to \mathbb{C}$, where Z is a finite abelian group. Then

$$\sum_{\xi} |\widehat{f}(\xi)|^2 = \frac{1}{|Z|} \sum_{x\in Z} |f(x)|^2.$$

Similarly, if $f\colon [-\pi,\pi] \to \mathbb{C}$ is square-integrable then its Fourier series satisfies

$$\sum_{n} |\widehat{f}(n)|^2 = \frac{1}{2\pi} \int_{[-\pi,\pi]} |f(x)|^2\,dx.$$

Proof. Recall that $\langle f, f\rangle$ is equal to the square sum of the coefficients. □

Corollary 14.5.2 (Fourier inversion formula)
Let $f\colon Z \to \mathbb{C}$, where Z is a finite abelian group. Then

$$\widehat{f}(\xi) = \frac{1}{|Z|} \sum_{x \in Z} f(x)\overline{e_\xi(x)}.$$

Similarly, if $f\colon [-\pi, \pi] \to \mathbb{C}$ is square-integrable then its Fourier series is given by

$$\widehat{f}(n) = \frac{1}{2\pi} \int_{[-\pi,\pi]} f(x) \exp(-inx)\, dx.$$

Proof. Recall that in an orthonormal basis $(e_\xi)_\xi$, the coefficient of e_ξ in f is $\langle f, e_\xi \rangle$. □

Question 14.5.3. What happens when $\xi = 0$ above?

Corollary 14.5.4 (Plancherel theorem)
Let $f\colon Z \to \mathbb{C}$, where Z is a finite abelian group. Then

$$\langle f, g \rangle = \sum_{\xi \in Z} \widehat{f}(\xi)\overline{\widehat{g}(\xi)}.$$

Similarly, if $f\colon [-\pi, \pi] \to \mathbb{C}$ is square-integrable then

$$\langle f, g \rangle = \sum_{n} \widehat{f}(n)\overline{\widehat{g}(n)}.$$

Question 14.5.5. Prove this one in one line (like before).

§14.6 Application: Basel problem

One cute application about Fourier analysis on $L^2([-\pi, \pi])$ is that you can get some otherwise hard-to-compute sums, as long as you are willing to use a little calculus.

Here is the classical one:

Theorem 14.6.1 (Basel problem)
We have

$$\sum_{n \geq 1} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

The proof is to consider the identity function $f(x) = x$, which is certainly square-integrable. Then by Parseval, we have

$$\sum_{n \in \mathbb{Z}} \left|\widehat{f}(n)\right|^2 = \langle f, f \rangle = \frac{1}{2\pi} \int_{[-\pi,\pi]} |f(x)|^2 \, dx.$$

A calculus computation gives

$$\frac{1}{2\pi} \int_{[-\pi,\pi]} x^2 \, dx = \frac{\pi^2}{3}.$$

On the other hand, we will now compute all Fourier coefficients. We have already that

$$\widehat{f}(0) = \frac{1}{2\pi}\int_{[-\pi,\pi]} f(x)\,dx = \frac{1}{2\pi}\int_{[-\pi,\pi]} x\,dx = 0.$$

For $n \neq 0$, we have by definition (or "Fourier inversion formula", if you want to use big words) the formula

$$\begin{aligned}\widehat{f}(n) &= \langle f, \exp(inx)\rangle \\ &= \frac{1}{2\pi}\int_{[-\pi,\pi]} x \cdot \overline{\exp(inx)}\,dx \\ &= \frac{1}{2\pi}\int_{[-\pi,\pi]} x\exp(-inx)\,dx.\end{aligned}$$

The anti-derivative is equal to $\frac{1}{n^2}\exp(-inx)(1+inx)$, which thus with some more calculation gives that

$$\widehat{f}(n) = \frac{(-1)^n}{n}i.$$

So

$$\sum_n \left|\widehat{f}(n)\right|^2 = 2\sum_{n\geq 1}\frac{1}{n^2}$$

implying the result.

§14.7 Application: Arrow's Impossibility Theorem

As an application of binary Fourier analysis, we now prove a form of Arrow's theorem.

Consider n voters voting among 3 candidates A, B, C. Each voter specifies a tuple $v_i = (x_i, y_i, z_i) \in \{\pm 1\}^3$ as follows:

- $x_i = 1$ if person i ranks A ahead of B, and $x_i = -1$ otherwise.
- $y_i = 1$ if person i ranks B ahead of C, and $y_i = -1$ otherwise.
- $z_i = 1$ if person i ranks C ahead of A, and $z_i = -1$ otherwise.

Tacitly, we only consider $3! = 6$ possibilities for v_i: we forbid "paradoxical" votes of the form $x_i = y_i = z_i$ by assuming that people's votes are consistent (meaning the preferences are transitive).

For brevity, let $x_\bullet = (x_1, \dots, x_n)$ and define $y_\bullet$ and $z_\bullet$ similarly. Then, we can consider a voting mechanism

$$\begin{aligned} f&\colon \{\pm 1\}^n \to \{\pm 1\} \\ g&\colon \{\pm 1\}^n \to \{\pm 1\} \\ h&\colon \{\pm 1\}^n \to \{\pm 1\}\end{aligned}$$

such that

- $f(x_\bullet)$ is the global preference of A vs. B,
- $g(y_\bullet)$ is the global preference of B vs. C,
- and $h(z_\bullet)$ is the global preference of C vs. A.

We'd like to avoid situations where the global preference $(f(x_\bullet), g(y_\bullet), h(z_\bullet))$ is itself paradoxical.

Let $\mathbb{E}f$ denote the average value of f across all 2^n inputs. Define $\mathbb{E}g$ and $\mathbb{E}h$ similarly. We'll add an assumption that $\mathbb{E}f = \mathbb{E}g = \mathbb{E}h = 0$, which provides symmetry (and e.g. excludes the possibility that f, g, h are constant functions which ignore voter input). With that we will prove the following result:

Theorem 14.7.1 (Arrow Impossibility Theorem)

Assume that (f, g, h) always avoids paradoxical outcomes, and assume $\mathbb{E}f = \mathbb{E}g = \mathbb{E}h = 0$. Then (f, g, h) is either a dictatorship or anti-dictatorship: there exists a "dictator" k such that

$$f(x_\bullet) = \pm x_k, \qquad g(y_\bullet) = \pm y_k, \qquad h(z_\bullet) = \pm z_k$$

where all three signs coincide.

Unlike the usual Arrow theorem, we do *not* assume that $f(+1, \dots, +1) = +1$ (hence possibility of anti-dictatorship).

Proof. Suppose the voters each randomly select one of the $3! = 6$ possible consistent votes. In Problem 14B it is shown that the exact probability of a paradoxical outcome for any functions f, g, h is given exactly by

$$\frac{1}{4} + \frac{1}{4} \sum_{S \subseteq \{1,\dots,n\}} \left(-\frac{1}{3}\right)^{|S|} \left(\widehat{f}(S)\widehat{g}(S) + \widehat{g}(S)\widehat{h}(S) + \widehat{h}(S)\widehat{f}(S)\right).$$

Assume that this probability (of a paradoxical outcome) equals 0. Then, we derive

$$1 = \sum_{S \subseteq \{1,\dots,n\}} -\left(-\frac{1}{3}\right)^{|S|} \left(\widehat{f}(S)\widehat{g}(S) + \widehat{g}(S)\widehat{h}(S) + \widehat{h}(S)\widehat{f}(S)\right).$$

But now we can just use weak inequalities. We have $\widehat{f}(\varnothing) = \mathbb{E}f = 0$ and similarly for $\widehat{g}$ and $\widehat{h}$, so we restrict attention to $|S| \geq 1$. We then combine the famous inequality $|ab + bc + ca| \leq a^2 + b^2 + c^2$ (which is true across all real numbers) to deduce that

$$\begin{aligned}
1 &= \sum_{S \subseteq \{1,\dots,n\}} -\left(-\frac{1}{3}\right)^{|S|} \left(\widehat{f}(S)\widehat{g}(S) + \widehat{g}(S)\widehat{h}(S) + \widehat{h}(S)\widehat{f}(S)\right) \\
&\leq \sum_{S \subseteq \{1,\dots,n\}} \left(\frac{1}{3}\right)^{|S|} \left(\widehat{f}(S)^2 + \widehat{g}(S)^2 + \widehat{h}(S)^2\right) \\
&\leq \sum_{S \subseteq \{1,\dots,n\}} \left(\frac{1}{3}\right)^{1} \left(\widehat{f}(S)^2 + \widehat{g}(S)^2 + \widehat{h}(S)^2\right) \\
&= \frac{1}{3}(1 + 1 + 1) = 1.
\end{aligned}$$

with the last step by Parseval. So all inequalities must be sharp, and in particular $\widehat{f}$, $\widehat{g}$, $\widehat{h}$ are supported on one-element sets, i.e. they are linear in inputs. As f, g, h are ± 1 valued, each f, g, h is itself either a dictator or anti-dictator function. Since (f, g, h) is always consistent, this implies the final result. □

§14.8 A few harder problems to think about

Problem 14A (For calculus fans)**.** Prove that

$$\sum_{n \geq 1} \frac{1}{n^4} = \frac{\pi^4}{90}.$$

Problem 14B. Let $f, g, h \colon \{\pm 1\}^n \to \{\pm 1\}$ be any three functions. For each i, we randomly select $(x_i, y_i, z_i) \in \{\pm 1\}^3$ subject to the constraint that not all are equal (hence, choosing among $2^3 - 2 = 6$ possibilities). Prove that the probability that

$$f(x_1, \dots, x_n) = g(y_1, \dots, y_n) = h(z_1, \dots, z_n)$$

is given by the formula

$$\frac{1}{4} + \frac{1}{4} \sum_{S \subseteq \{1,\dots,n\}} \left(-\frac{1}{3}\right)^{|S|} \left(\widehat{f}(S)\widehat{g}(S) + \widehat{g}(S)\widehat{h}(S) + \widehat{h}(S)\widehat{f}(S)\right)$$

15 Duals, adjoint, and transposes

This chapter is dedicated to the basis-free interpretation of the transpose and conjugate transpose of a matrix.

Poster corollary: we will see that symmetric matrices with real coefficients are diagonalizable and have real eigenvalues.

§15.1 Dual of a map

Prototypical example for this section: The example below.

We go ahead and now define a notion that will grow up to be the transpose of a matrix.

Definition 15.1.1. Let V and W be vector spaces. Suppose $T\colon V \to W$ is a linear map. Then we actually get a map

$$\begin{aligned} T^\vee\colon W^\vee &\to V^\vee \\ f &\mapsto f \circ T. \end{aligned}$$

This map is called the **dual map**.

Example 15.1.2 (Example of a dual map)

Work over $\mathbb{R}$. Let's consider V with basis e_1, e_2, e_3 and W with basis f_1, f_2. Suppose that

$$\begin{aligned} T(e_1) &= f_1 + 2f_2 \\ T(e_2) &= 3f_1 + 4f_2 \\ T(e_3) &= 5f_1 + 6f_2. \end{aligned}$$

Now consider $V^\vee$ with its dual basis $e_1^\vee$, $e_2^\vee$, $e_3^\vee$ and $W^\vee$ with its dual basis $f_1^\vee$, $f_2^\vee$. Let's compute $T^\vee(f_1^\vee) = f_1^\vee \circ T$: it is given by

$$\begin{aligned} f_1^\vee\left(T(ae_1 + be_2 + ce_3)\right) &= f_1^\vee\left((a+3b+5c)f_1 + (2a+4b+6c)f_2\right) \\ &= a + 3b + 5c. \end{aligned}$$

So accordingly we can write

$$T^\vee(f_1^\vee) = e_1^\vee + 3e_2^\vee + 5e_3^\vee$$

Similarly,

$$T^\vee(f_2^\vee) = 2e_1^\vee + 4e_2^\vee + 6e_3^\vee.$$

This determines $T^\vee$ completely.

If we write the matrices for T and $T^\vee$ in terms of our basis, we now see that

$$T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix} \quad \text{and} \quad T^\vee = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}.$$

So in our selected basis, we find that the matrices are **transposes**: mirror images of each other over the diagonal.

Of course, this should work in general.

Theorem 15.1.3 (Transpose interpretation of $T^\vee$)
Let V and W be finite-dimensional k-vector spaces. Then, for any $T\colon V \to W$, the following two matrices are transposes:

- The matrix for $T\colon V \to W$ expressed in the basis (e_i), (f_j).
- The matrix for $T^\vee\colon W^\vee \to V^\vee$ expressed in the basis $(f_j^\vee)$, $(e_i^\vee)$.

Proof. The (i,j)th entry of the matrix T corresponds to the coefficient of f_i in $T(e_j)$, which corresponds to the coefficient of $e_j^\vee$ in $f_i^\vee \circ T$. □

The nice part of this is that the definition of $T^\vee$ is basis-free. So it means that if we start with any linear map T, and then pick whichever basis we feel like, then T and $T^\vee$ will still be transposes.

§15.2 Identifying with the dual space

For the rest of this chapter, though, we'll now bring inner products into the picture.

Earlier I complained that there was no natural isomorphism $V \cong V^\vee$. But in fact, given an inner form we can actually make such an identification: that is we can naturally associate every linear map $\xi\colon V \to k$ with a vector $v \in V$.

To see how we might do this, suppose $V = \mathbb{R}^3$ for now with an orthonormal basis e_1, e_2, e_3. How might we use the inner product to represent a map from $V \to \mathbb{R}$? For example, take $\xi \in V^\vee$ by $\xi(e_1) = 3$, $\xi(e_2) = 4$ and $\xi(e_3) = 5$. Actually, I claim that

$$\xi(v) = \langle v, 3e_1 + 4e_2 + 5e_3 \rangle$$

for every v.

Question 15.2.1. Check this.

And this works beautifully in the real case.

Theorem 15.2.2 ($V \cong V^\vee$ for real inner forms)
Let V be a finite-dimensional *real* inner product space and $V^\vee$ its dual. Then the map $V \to V^\vee$ by

$$v \mapsto \langle -, v \rangle \in V^\vee$$

is an isomorphism of real vector spaces.

Proof. It suffices to show that the map is injective and surjective.

- Injective: suppose $\langle v, v_1 \rangle = \langle v, v_2 \rangle$ for every vector $v \in V$. This means $\langle v, v_1 - v_2 \rangle = 0$ for every vector $v \in V$. This can only happen if $v_1 - v_2 = 0$; for example, take $v = v_1 - v_2$ and use positive definiteness.

- Surjective: take an orthonormal basis $e_1, \dots e_n$ and let $e_1^\vee, \dots, e_n^\vee$ be the dual basis on $V^\vee$. Then e_1 maps to $e_1^\vee$, et cetera. □

Actually, since we already know $\dim V = \dim V^\vee$ we only had to prove one of the above. As a matter of personal taste, I find the proof of injectivity more elegant, and the proof of surjectivity more enlightening, so I included both. Thus

If a real inner product space V is given an inner form, then V and $V^\vee$ are canonically isomorphic.

Unfortunately, things go awry if V is complex. Here is the result:

Theorem 15.2.3 (V versus $V^\vee$ for complex inner forms)
Let V be a finite-dimensional *complex* inner product space and $V^\vee$ its dual. Then the map $V \to V^\vee$ by

$$v \mapsto \langle -, v \rangle \in V^\vee$$

is a bijection of sets.

Wait, what? Well, the proof above shows that it is both injective and surjective, but why is it not an isomorphism? The answer is that it is not a linear map: since the form is sesquilinear we have for example

$$iv \mapsto \langle -, iv \rangle = -i \langle -, v \rangle$$

which has introduced a minus sign! In fact, it is an *anti-linear* map, in the sense we defined before.

Eager readers might try to fix this by defining the isomorphism $v \mapsto \langle v, - \rangle$ instead. However, this also fails, because the right-hand side is not even an element of $V^\vee$: it is an "anti-linear", not linear.

And so we are stuck. Fortunately, we will only need the "bijection" result for what follows, so we can continue on anyways. (If you want to fix this, Problem 15D gives a way to do so.)

§15.3 The adjoint (conjugate transpose)

We will see that, as a result of the flipping above, the *conjugate transpose* is actually the better concept for inner product spaces: since it can be defined using only the inner product without making mention to dual spaces at all.

Definition 15.3.1. Let V and W be finite-dimensional inner product spaces, and let $T \colon V \to W$. The **adjoint** (or **conjugate transpose**) of T, denoted $T^\dagger \colon W \to V$, is defined as follows: for every vector $w \in W$, we let $T^\dagger(w) \in V$ be the unique vector with

$$\left\langle v, T^\dagger(w) \right\rangle_V = \langle T(v), w \rangle_W$$

for every $v \in V$.

Some immediate remarks about this definition:

- Our $T^\dagger$ is well-defined, because $v \mapsto \langle T(v), w \rangle_W$ is some function in $V^\vee$, and hence by the bijection earlier it should be uniquely of the form $\langle -, v \rangle$ for some $v \in V$.

- This map $T^\dagger$ is indeed a linear map (why?).
- The niceness of this definition is that it doesn't make reference to any basis or even $V^\vee$, so it is the "right" definition for a inner product space.

Example 15.3.2 (Example of an adjoint map)

We'll work over $\mathbb{C}$, so the conjugates are more visible. Let's consider V with orthonormal basis e_1, e_2, e_3 and W with orthonormal basis f_1, f_2. We put

$$\begin{aligned} T(e_1) &= if_1 + 2f_2 \\ T(e_2) &= 3f_1 + 4f_2 \\ T(e_3) &= 5f_1 + 6if_2. \end{aligned}$$

We compute $T^\dagger(f_1)$. It is the unique vector $x \in V$ such that

$$\langle v, x \rangle_V = \langle T(v), f_1 \rangle_W$$

for any $v \in V$. If we expand $v = ae_1 + be_2 + ce_3$ the above equality becomes

$$\begin{aligned} \langle ae_1 + be_2 + ce_3, x \rangle_V &= \langle T(ae_1 + be_2 + ce_3), f_1 \rangle_W \\ &= ia + 3b + 5c. \end{aligned}$$

However, since x is in the second argument, this means we actually want to take

$$T^\dagger(f_1) = -ie_1 + 3e_2 + 5e_3$$

so that the sesquilinearity will conjugate the i.

The pattern continues, though we remind the reader that we need the basis to be orthonormal to proceed.

Theorem 15.3.3 (Adjoints are conjugate transposes)

Fix an *orthonormal* basis of a finite-dimensional inner product space V. Let $T \colon V \to V$ be a linear map. If we write T as a matrix in this basis, then the matrix $T^\dagger$ (in the same basis) is the *conjugate transpose* of the matrix of T; that is, the (i,j)th entry of $T^\vee$ is the complex conjugate of the (j,i)th entry of T.

Proof. One-line version: take v and w to be basis elements, and this falls right out.

Full proof: let

$$T = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$$

in this basis $e_1, \dots, e_n$. Then, letting $w = e_i$ and $v = e_j$ we deduce that

$$\left\langle e_i, T^\dagger(e_j) \right\rangle = \langle T(e_i), e_j \rangle = a_{ji} \implies \left\langle T^\dagger(e_j), e_i \right\rangle = \overline{a_{ji}}$$

for any i, which is enough to deduce the result. □

§15.4 Eigenvalues of normal maps

We now come to the advertised theorem. Restrict to the situation where $T\colon V \to V$. You see, the world would be a very beautiful place if it turned out that we could pick a basis of eigenvectors that was also *orthonormal*. This is of course far too much to hope for; even without the orthonormal condition, we saw that Jordan form could still have 1's off the diagonal.

However, it turns out that there is a complete characterization of exactly when our overzealous dream is true.

Definition 15.4.1. We say a linear map T (from a finite-dimensional inner product space to itself) is **normal** if $TT^\dagger = T^\dagger T$.

We say a complex T is **self-adjoint** or **Hermitian** if $T = T^\dagger$; i.e. as a matrix in any orthonormal basis, T is its own conjugate transpose. For real T we say "self-adjoint", "Hermitian" or **symmetric**.

Theorem 15.4.2 (Normal $\iff$ diagonalizable with orthonormal basis)
Let V be a finite-dimensional complex inner product space. A linear map $T\colon V \to V$ is normal if and only if one can pick an orthonormal basis of eigenvectors.

Exercise 15.4.3. Show that if there exists such an orthonormal basis then $T\colon V \to V$ is normal, by writing T as a diagonal matrix in that basis.

Proof. This is long, and maybe should be omitted on a first reading. If T has an orthonormal basis of eigenvectors, this result is immediate.

Now assume T is normal. We first prove T is diagonalizable; this is the hard part.

Claim 15.4.4. If T is normal, then $\ker T = \ker T^r = \ker T^\dagger$ for $r \geq 1$. (Here T^r is T applied r times.)

Proof of Claim. Let $S = T^\dagger \circ T$, which is self-adjoint. We first note that S is Hermitian and $\ker S = \ker T$. To see it's Hermitian, note $\langle Sv, w\rangle = \langle Tv, Tw\rangle = \langle v, Sw\rangle$. Taking $v = w$ also implies $\ker S \subseteq \ker T$ (and hence equality since obviously $\ker T \subseteq \ker S$).

First, since we have $\langle S^r(v), S^{r-2}(v)\rangle = \langle S^{r-1}(v), S^{r-1}(v)\rangle$, an induction shows that $\ker S = \ker S^r$ for $r \geq 1$. Now, since T is normal, we have $S^r = (T^\dagger)^r \circ T^r$, and thus we have the inclusion

$$\ker T \subseteq \ker T^r \subseteq \ker S^r = \ker S = \ker T$$

where the last equality follows from the first claim. Thus in fact $\ker T = \ker T^r$.

Finally, to show equality with $\ker T^\dagger$ we

$$\begin{aligned} \langle Tv, Tv\rangle &= \left\langle v, T^\dagger T v\right\rangle \\ &= \left\langle v, TT^\dagger v\right\rangle \\ &= \left\langle T^\dagger v, T^\dagger v\right\rangle. \end{aligned}$$

■

Now consider the given T, and any λ.

Question 15.4.5. Show that $(T - \lambda\mathrm{id})^\dagger = T^\dagger - \overline{\lambda}\mathrm{id}$. Thus if T is normal, so is $T - \lambda\mathrm{id}$.

In particular, for any eigenvalue λ of T, we find that $\ker(T - \lambda\text{id}) = \ker(T - \lambda\text{id})^r$. This implies that all the Jordan blocks of T have size 1; i.e. that T is in fact diagonalizable. Finally, we conclude that the eigenvectors of T and $T^\dagger$ match, and the eigenvalues are complex conjugates.

So, diagonalize T. We just need to show that if v and w are eigenvectors of T with distinct eigenvalues, then they are orthogonal. (We can use Gram-Schmidt on any eigenvalue that appears multiple times.) To do this, suppose $T(v) = \lambda v$ and $T(w) = \mu w$ (thus $T^\dagger(w) = \overline{\mu} w$). Then

$$\lambda \langle v, w \rangle = \langle \lambda v, w \rangle = \langle Tv, w \rangle = \left\langle v, T^\dagger(w) \right\rangle = \langle v, \overline{\mu} w \rangle = \mu \langle v, w \rangle .$$

Since $\lambda \neq \mu$, we conclude $\langle v, w \rangle = 0$. □

This means that not only can we write

$$T = \begin{bmatrix} \lambda_1 & \dots & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

but moreover that the basis associated with this matrix happens to be orthonormal vectors.

As a corollary:

Theorem 15.4.6 (Hermitian matrices have real eigenvalues)
A Hermitian matrix T is diagonalizable, and all its eigenvalues are real.

Proof. Obviously Hermitian $\implies$ normal, so write it in the orthonormal basis of eigenvectors. To see that the eigenvalues are real, note that $T = T^\dagger$ means $\lambda_i = \overline{\lambda_i}$ for every i. □

§15.5 A few harder problems to think about

Problem 15A* (Double dual)**.** Let V be a finite-dimensional vector space. Prove that

$$\begin{aligned} V &\to (V^\vee)^\vee \\ v &\mapsto (\xi \mapsto \xi(v)) \end{aligned}$$

gives an isomorphism. (This is significant because the isomorphism is *canonical*, and in particular does not depend on the choice of basis. So this is more impressive.)

Problem 15B (Fundamental theorem of linear algebra)**.** Let $T \colon V \to W$ be a map of finite-dimensional k-vector spaces. Prove that

$$\dim \operatorname{im} T = \dim \operatorname{im} T^\vee = \dim V - \dim \ker T = \dim W - \dim \ker T^\vee.$$

Problem 15C† (Row rank is column rank)**.** A $m \times n$ matrix M of real numbers is given. The *column rank* of M is the dimension of the span in $\mathbb{R}^m$ of its n column vectors. The *row rank* of M is the dimension of the span in $\mathbb{R}^n$ of its m row vectors. Prove that the row rank and column rank are equal.

Problem 15D (The complex conjugate spaces)**.** Let $V = (V, +, \cdot)$ be a complex vector space. Define the **complex conjugate vector space**, denoted $\overline{V} = (V, +, *)$ by changing just the multiplication:

$$c * v = \overline{c} \cdot v.$$

Show that for any sesquilinear form on V, if V is finite-dimensional, then

$$\begin{aligned} \overline{V} &\to V^\vee \\ v &\mapsto \langle -, v \rangle \end{aligned}$$

is an isomorphism of complex vector spaces.

Problem 15E ($T^\dagger$ vs $T^\vee$)**.** Let V and W be real inner product spaces and let $T \colon V \to W$ be a linear map. Show that the following diagram commutes:

$$\begin{array}{ccc} W & \xrightarrow{T^\dagger} & V \\ \cong\downarrow & & \downarrow\cong \\ W^\vee & \xrightarrow[T^\vee]{} & V^\vee \end{array}$$

Here the isomorphisms are $v \mapsto \langle -, v \rangle$. Thus, for real inner product spaces, $T^\dagger$ is just $T^\vee$ with the duals eliminated (by Theorem 15.2.2).

Problem 15F (Polynomial criteria for normality)**.** Let V be a complex inner product space and let $T \colon V \to V$ be a linear map. Show that T is normal if and only if there is a polynomial[1] $p \in \mathbb{C}[t]$ such that

$$T^\dagger = p(T).$$

Problem 15G (Kronecker product of matrices)**.** Find an equivalence between the following two definitions of the Kronecker product, the former from a mathematician and the latter from a computer scientist:

- Suppose $A \colon V_1 \to W_1$ and $B \colon V_2 \to W_2$ are linear maps of finite-dimensional vector spaces over $\mathbb{R}$. Then we define $A \otimes B \colon V_1 \otimes V_2 \to W_1 \otimes W_2$ on simple tensors by $v_1 \otimes v_2 \mapsto A(v_1) \otimes B(v_2)$.
- Suppose A is an $m \times n$ matrix and B is a $p \times q$ matrix. Then $A \otimes B$ is an operator which takes a $q \times n$ matrix X and returns the $p \times m$ matrix $BXA^\top$.

[1] Here, we mean $p(T)$ in the same composition sense as in Cayley-Hamilton.

V

More on Groups

Part V: Contents

16 Group actions overkill AIME problems

Consider this problem from the 1996 AIME:

> (AIME 1996) Two of the squares of a 7×7 checkerboard are painted yellow, and the rest are painted green. Two color schemes are equivalent if one can be obtained from the other by applying a rotation in the plane of the board. How many inequivalent color schemes are possible?

What's happening here? Let X be the set of the $\binom{49}{2}$ possible colorings of the board. What's the natural interpretation of "rotation"? Answer: the group $\mathbb{Z}/4\mathbb{Z} = \langle r \mid r^4 = 1 \rangle$ somehow "acts" on this set X by sending one state $x \in X$ to another state $r \cdot x$, which is just x rotated by 90°. Intuitively we're just saying that two configurations are the same if they can be reached from one another by this "action".

We can make all of this precise using the idea of a group action.

§16.1 Definition of a group action

Prototypical example for this section: The AIME problem.

Definition 16.1.1. Let X be a set and G a group. A **group action** is a binary operation $\cdot : G \times X \to X$ which lets a $g \in G$ send an $x \in X$ to $g \cdot x$. It satisfies the axioms

- $(g_1 g_2) \cdot x = g_1 \cdot (g_2 \cdot x)$ for any $g_1, g_2 \in G$ for all $x \in X$.
- $1_G \cdot x = x$ for any $x \in X$.

Example 16.1.2 (Examples of group actions)
Let $G = (G, \star)$ be a group.

(a) The group $\mathbb{Z}/4\mathbb{Z}$ can act on the set of ways to color a 7×7 board either yellow or green.

(b) The group $\mathbb{Z}/4\mathbb{Z} = \langle r \mid r^4 = 1 \rangle$ acts on the xy-plane $\mathbb{R}^2$ as follows: $r \cdot (x, y) = (y, -x)$. In other words, it's a rotation by 90°.

(c) The dihedral group D_{2n} acts on the set of ways to color the vertices of an n-gon.

(d) The group S_n acts on $X = \{1, 2, \ldots, n\}$ by applying the permutation σ: $\sigma \cdot x := \sigma(x)$.

(e) The group G can act on itself (i.e. $X = G$) by left multiplication: put $g \cdot g' := g \star g'$.

Exercise 16.1.3. Show that a group action can equivalently be described as a group homomorphism from G to S_X, where S_X is the symmetric group of permutations on X.

§16.2 Stabilizers and orbits

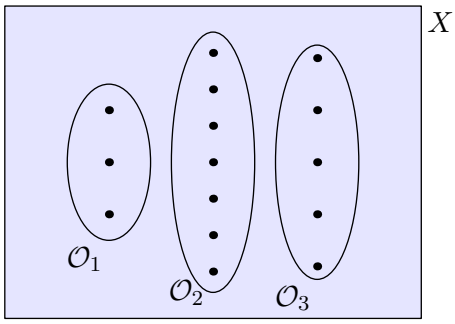
Prototypical example for this section: Again the AIME problem.

Given a group action G on X, we can define an equivalence relation $\sim$ on X as follows: $x \sim y$ if $x = g \cdot y$ for some $g \in G$. For example, in the AIME problem, $\sim$ means "one can be obtained from the other by a rotation".

Question 16.2.1. Why is this an equivalence relation?

In that case, the AIME problem wants the number of equivalence classes under $\sim$. So let's give these equivalence classes a name: **orbits**. We usually denote orbits by $\mathcal{O}$.

As usual, orbits carve out X into equivalence classes.



It turns out that a very closely related concept is:

Definition 16.2.2. The **stabilizer** of a point $x \in X$, denoted $\text{Stab}_G(x)$, is the set of $g \in G$ which fix x; in other words

$$\text{Stab}_G(x) \coloneqq \{g \in G \mid g \cdot x = x\}.$$

Example 16.2.3

Consider the AIME problem again, with X the possible set of states (again $G = \mathbb{Z}/4\mathbb{Z}$). Let x be the configuration where two opposite corners are colored yellow. Evidently 1_G fixes x, but so does the 180° rotation r^2. But r and r^3 do not preserve x, so $\text{Stab}_G(x) = \{1, r^2\} \cong \mathbb{Z}/2\mathbb{Z}$.

Question 16.2.4. Why is $\text{Stab}_G(x)$ a subgroup of G?

Once we realize the stabilizer is a group, this leads us to what I privately call the "fundamental theorem of how big an orbit is".

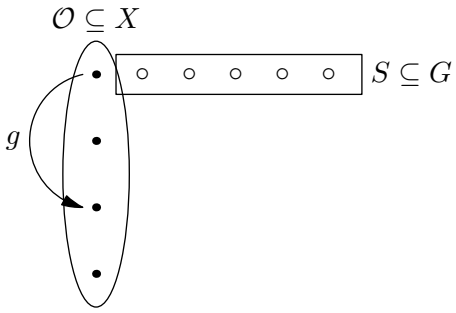
Theorem 16.2.5 (Orbit-stabilizer theorem)

Let $\mathcal{O}$ be an orbit, and pick any $x \in \mathcal{O}$. Let $S = \text{Stab}_G(x)$ be a subgroup of G. There is a natural bijection between $\mathcal{O}$ and left cosets. In particular,

$$|\mathcal{O}| \, |S| = |G|.$$

In particular, the stabilizers of each $x \in \mathcal{O}$ have the same size.

Proof. The point is that every coset gS just specifies an element of $\mathcal{O}$, namely $g \cdot x$. The fact that S is a stabilizer implies that it is irrelevant which representative we pick.



Since the $|\mathcal{O}|$ cosets partition G, each of size $|S|$, we obtain the second result. □

§16.3 Burnside's lemma

Now for the crux of this chapter: a way to count the number of orbits.

Theorem 16.3.1 (Burnside's lemma)

Let G act on a set X. The number of orbits of the action is equal to

$$\frac{1}{|G|} \sum_{g \in G} |\operatorname{FixPt} g|$$

where $\operatorname{FixPt} g$ is the set of points $x \in X$ such that $g \cdot x = x$.

The proof is deferred as a bonus problem, since it has a very olympiad-flavored solution. As usual, this lemma was not actually proven by Burnside; Cauchy got there first, and thus it is sometimes called *the lemma that is not Burnside's*. Example application:

Example 16.3.2 (AIME 1996)

Two of the squares of a 7×7 checkerboard are painted yellow, and the rest are painted green. Two color schemes are equivalent if one can be obtained from the other by applying a rotation in the plane of the board. How many inequivalent color schemes are possible?

We know that $G = \mathbb{Z}/4\mathbb{Z}$ acts on the set X of $\binom{49}{2}$ possible coloring schemes. Now we can compute $\operatorname{FixPt} g$ explicitly for each $g \in \mathbb{Z}/4\mathbb{Z}$.

- If $g = 1_G$, then every coloring is fixed, for a count of $\binom{49}{2} = 1176$.
- If $g = r^2$ there are exactly 24 coloring schemes fixed by g: this occurs when the two squares are reflections across the center, which means they are preserved under a 180° rotation.
- If $g = r$ or $g = r^3$, then there are no fixed coloring schemes.

As $|G| = 4$, the average is

$$\frac{1176 + 24 + 0 + 0}{4} = 300.$$

Exercise 16.3.3 (MathCounts Chapter Target Round)**.** A circular spinner has seven sections of equal size, each of which is colored either red or blue. Two colorings are considered the same if one can be rotated to yield the other. In how many ways can the spinner be colored? (Answer: 20)

Consult [Ma13b] for some more examples of "hands-on" applications.

§16.4 Conjugation of elements

Prototypical example for this section: In S_n, conjugacy classes are "cycle types".

A particularly common type of action is the so-called **conjugation**. We let G act on itself as follows:

$$g \colon h \mapsto ghg^{-1}.$$

You might think this definition is a little artificial. Who cares about the element ghg^{-1}? Let me try to convince you this definition is not so unnatural.

Example 16.4.1 (Conjugacy in S_n)

Let $G = S_5$, and fix a $\pi \in S_5$. Here's the question: is $\pi\sigma\pi^{-1}$ related to σ? To illustrate this, I'll write out a completely random example of a permutation $\sigma \in S_5$.

$$\text{If } \sigma = \begin{array}{ccc} 1 & \mapsto & 3 \\ 2 & \mapsto & 1 \\ 3 & \mapsto & 5 \\ 4 & \mapsto & 2 \\ 5 & \mapsto & 4 \end{array} \quad \text{then} \quad \pi\sigma\pi^{-1} = \begin{array}{ccc} \pi(1) & \mapsto & \pi(3) \\ \pi(2) & \mapsto & \pi(1) \\ \pi(3) & \mapsto & \pi(5) \\ \pi(4) & \mapsto & \pi(2) \\ \pi(5) & \mapsto & \pi(4) \end{array}$$

Thus our fixed π doesn't really change the structure of σ at all: it just "renames" each of the elements 1, 2, 3, 4, 5 to $\pi(1)$, $\pi(2)$, $\pi(3)$, $\pi(4)$, $\pi(5)$.

But wait, you say. That's just a very particular type of group behaving nicely under conjugation. Why does this mean anything more generally? All I have to say is: remember Cayley's theorem! (This was Problem 1F†.)

In any case, we may now define:

Definition 16.4.2. The **conjugacy classes** of a group G are the orbits of G under the conjugacy action.

Let's see what the conjugacy classes of S_n are, for example.

Example 16.4.3 (Conjugacy classes of S_n correspond to cycle types)

Intuitively, the discussion above says that two elements of S_n should be conjugate if they have the same "shape", regardless of what the elements are named. The right way to make the notion of "shape" rigorous is cycle notation. For example, consider the permutation

$$\sigma_1 = (1\ 3\ 5)(2\ 4)$$

in cycle notation, meaning $1 \mapsto 3 \mapsto 5 \mapsto 1$ and $2 \mapsto 4 \mapsto 2$. It is conjugate to the permutation

$$\sigma_2 = (1\ 2\ 3)(4\ 5)$$

or any other way of relabeling the elements. So, we could think of σ as having conjugacy class

$$(-\ -\ -)(-\ -).$$

More generally, you can show that two elements of S_n are conjugate if and only if they have the same "shape" under cycle decomposition.

Question 16.4.4. Show that the number of conjugacy classes of S_n equals the number of *partitions* of n.

As long as I've put the above picture, I may as well also define:

Definition 16.4.5. Let G be a group. The **center** of G, denoted $Z(G)$, is the set of elements $x \in G$ such that $xg = gx$ for every $g \in G$. More succinctly,

$$Z(G) \coloneqq \{x \in G \mid gx = xg \ \forall g \in G\}.$$

You can check this is indeed a subgroup of G.

Question 16.4.6. Why is $Z(G)$ normal in G?

Question 16.4.7. What are the conjugacy classes of elements in the center?

A trivial result that gets used enough that I should explicitly call it out:

Corollary 16.4.8 (Conjugacy in abelian groups is trivial)
If G is abelian, then the conjugacy classes all have size one.

§16.5 A few harder problems to think about

Problem 16A (PUMaC 2009 C8)**.** Taotao wants to buy a bracelet consisting of seven beads, each of which is orange, white or black. (The bracelet can be rotated and reflected in space.) Find the number of possible bracelets.

Problem 16B. Show that two elements in the same conjugacy class have the same order.

Problem 16C. Prove Burnside's lemma.

Problem 16D* (The "class equation")**.** Let G be a finite group. We define the **centralizer** $C_G(g) = \{x \in G \mid xg = gx\}$ for each $g \in G$. Show that

$$|G| = |Z(G)| + \sum_{s \in S} \frac{|G|}{|C_G(s)|}$$

where $S \subseteq G$ is defined as follows: for each conjugacy class $C \subseteq G$ with $|C| > 1$, we pick a representative of C and add it to S.

Problem 16E† (Classical)**.** Assume G is a finite group of order $n \geq 2$ and p is the smallest prime dividing n. Let H be a subgroup of G with $|G| / |H| = p$. Show that H is normal in G.

17 Find all groups

The following problem will hopefully never be proposed at the IMO.

> Let n be a positive integer and let $S = \{1, \dots, n\}$. Find all functions $f : S \times S \to S$ such that
>
> (a) $f(x, 1) = f(1, x) = x$ for all $x \in S$.
>
> (b) $f(f(x, y), z) = f(x, f(y, z))$ for all $x, y, z \in S$.
>
> (c) For every $x \in S$ there exists a $y \in S$ such that $f(x, y) = f(y, x) = 1$.

Nonetheless, it's remarkable how much progress we've made on this "problem". In this chapter I'll try to talk about some things we have accomplished.

§17.1 Sylow theorems

Here we present the famous Sylow theorems, some of the most general results we have about finite groups.

Theorem 17.1.1 (The Sylow theorems)

Let G be a group of order $p^n m$, where $\gcd(p, m) = 1$ and p is a prime. A **Sylow p-subgroup** is a subgroup of order p^n. Let n_p be the number of Sylow p-subgroups of G. Then

(a) $n_p \equiv 1 \pmod{p}$. In particular, $n_p \neq 0$ and a Sylow p-subgroup exists.

(b) n_p divides m.

(c) Any two Sylow p-subgroups are conjugate subgroups (hence isomorphic).

Sylow's theorem is really huge for classifying groups; in particular, the conditions $n_p \equiv 1 \pmod{p}$ and $n_p \mid m$ can often pin down the value of n_p to just a few values. Here are some results which follow from the Sylow theorems.

- A Sylow p-subgroup is normal if and only if $n_p = 1$.
- Any group G of order pq, where $p < q$ are primes, must have $n_q = 1$, since $n_q \equiv 1 \pmod{q}$ yet $n_q \mid p$. Thus G has a normal subgroup of order q.
- Since any abelian group has all subgroups normal, it follows that any abelian group has exactly one Sylow p-subgroup for every p dividing its order.
- If $p \neq q$, the intersection of a Sylow p-subgroup and a Sylow q-subgroup is just $\{1_G\}$. That's because the intersection of any two subgroups is also a subgroup, and Lagrange's theorem tells us that its order must divide both a power of p and a power of q; this can only happen if the subgroup is trivial.

Here's an example of another "practical" application.

Proposition 17.1.2 (Triple product of primes)

If $|G| = pqr$ is the product of distinct primes, then G must have a normal Sylow subgroup.

Proof. WLOG, assume $p < q < r$. Notice that $n_p \equiv 1 \pmod{p}$, $n_p | qr$ and cyclically, and assume for contradiction that $n_p, n_q, n_r > 1$.

Since $n_r | pq$, we have $n_r = pq$ since n_r divides neither p nor q as $n_r \geq 1 + r > p, q$. Also, $n_p \geq 1 + p$ and $n_q \geq 1 + q$. So we must have at least $1 + p$ Sylow p-subgroups, at least $1 + q$ Sylow q-subgroups, and at least pq Sylow r-subgroups.

But these groups are pretty exclusive.

Question 17.1.3. Take the $n_p + n_q + n_r$ Sylow subgroups and consider two of them, say H_1 and H_2. Show that $|H_1 \cap H_2| = 1$ as follows: check that $H_1 \cap H_2$ is a subgroup of both H_1 and H_2, and then use Lagrange's theorem.

We claim that there are too many elements now. Indeed, if we count the non-identity elements contributed by these subgroups, we get

$$n_p(p-1) + n_q(q-1) + n_r(r-1) \geq (1+p)(p-1) + (1+q)(q-1) + pq(r-1) > pqr$$

which is more elements than G has! □

§17.2 (Optional) Proving Sylow's theorem

The proof of Sylow's theorem is somewhat involved, and in fact many proofs exist. I'll present one below here. It makes extensive use of group actions, so I want to recall a few facts first. If G acts on X, then

- The orbits of the action form a partition of X.
- if $\mathcal{O}$ is any orbit, then the orbit-stabilizer theorem says that
$$|\mathcal{O}| = |G| / |\text{Stab}_G(x)|$$
for any $x \in \mathcal{O}$.
- In particular: suppose in the above that G is a **p-group**, meaning $|G| = p^t$ for some t. Then either $|\mathcal{O}| = 1$ or p divides $|\mathcal{O}|$. In the case $\mathcal{O} = \{x\}$, then by definition, x is a **fixed point** of every element of G: we have $g \cdot x = x$ for every g.

Note that when I say x is a fixed point, I mean it is fixed by **every** element of the group, i.e. the orbit really has size one. Hence that's a really strong condition.

§17.2.i Definitions

Prototypical example for this section: Conjugacy in S_n.

I've defined conjugacy of elements previously, but I now need to define it for groups:

Definition 17.2.1. Let G be a group, and let X denote the set of subgroups of G. Then **conjugation** is the action of G on X that sends

$$H \mapsto gHg^{-1} = \left\{ ghg^{-1} \mid h \in H \right\}.$$

If H and K are subgroups of G such that $H = gKg^{-1}$ for some $g \in G$ (in other words, they are in the same orbit under this action), then we say they are **conjugate** subgroups.

Because we somehow don't think of conjugate elements as "that different" (for example, in permutation groups), the following shouldn't be surprising:

> **Question 17.2.2.** Show that for any subgroup H of a group G, the map $H \to gHg^{-1}$ by $h \mapsto ghg^{-1}$ is in fact an isomorphism. This implies that any two conjugate subgroups are isomorphic.

Definition 17.2.3. For any subgroup H of G the **normalizer** of H is defined as

$$N_G(H) \coloneqq \left\{ g \in G \mid gHg^{-1} = H \right\}.$$

In other words, it is the stabilizer of H under the conjugation action.

We are now ready to present the proof.

§17.2.ii Step 1: Prove that a Sylow p-subgroup exists

What follows is something like the probabilistic method. By considering the set X of ALL subsets of size p^n at once, we can exploit the "deep number theoretic fact" that

$$|X| = \binom{p^n m}{p^n} \not\equiv 0 \pmod{p}.$$

(It's not actually deep: use Lucas' theorem.)

Here is the proof.

- Let G act on X by $g \cdot S \coloneqq \{gs \mid s \in S\}$.
- Take an orbit $\mathcal{O}$ with size not divisible by p. (This is possible because of our deep number theoretic fact. Since $|X|$ is nonzero mod p and the orbits partition X, the claimed orbit must exist.)
- Let $S \in \mathcal{O}$, $H = \mathrm{Stab}_G(S)$. Then p^n divides $|H|$, by the orbit-stabilizer theorem.
- Consider a second action: let H act on S by $h \cdot s \coloneqq hs$ (we know $hs \in S$ since $H = \mathrm{Stab}_G(S)$).
- Observe that $\mathrm{Stab}_H(s) = \{1_H\}$. Then all orbits of the second action must have size $|H|$. Thus $|H|$ divides $|S| = p^n$.
- This implies $|H| = p^n$, and we're done.

§17.2.iii Step 2: Any two Sylow p-subgroups are conjugate

If P is a Sylow p-subgroup and Q is a p-group, we prove $Q \subseteq gPg^{-1}$. Note that if Q is also a Sylow p-subgroup, then $Q = gPg^{-1}$ for size reasons; this implies that any two Sylow subgroups are indeed conjugate.

Let Q act on the set of left cosets of P by left multiplication. Note that

- Q is a p-group, so any orbit has size divisible by p unless it's 1.
- But the number of left cosets is m, which isn't divisible by p.

Hence some coset gP is a fixed point for every q, meaning $qgP = gP$ for all q. Equivalently, $qg \in gP$ for all $q \in Q$, so $Q \subseteq gPg^{-1}$ as desired.

§17.2.iv Step 3: Showing $n_p \equiv 1 \pmod p$

Let $\mathcal{S}$ denote the set of all the Sylow p-subgroups. By our first step, there exists some $P \in \mathcal{S}$.

> **Question 17.2.4.** Why does $|\mathcal{S}|$ equal n_p? (In other words, are you awake?)

Now we can proceed with the proof. Let P act on $\mathcal{S}$ by conjugation. Then:

- Because P is a p-group, $n_p \pmod p$ is the number of fixed points of this action. Now we claim P is the only fixed point of this action.
- Let Q be any other fixed point, meaning $xQx^{-1} = Q$ for any $x \in P$.
- Define the normalizer $N_G(Q) = \{g \in G \mid gQg^{-1} = Q\}$. It contains both P and Q.
- Now for the crazy part: apply Step 2 to $N_G(Q)$. Since P and Q are Sylow p-subgroups of it, they must be conjugate.
- Hence $P = Q$, as desired.

§17.2.v Step 4: n_p divides m

Since $n_p \equiv 1 \pmod p$, it suffices to show n_p divides $|G|$. Let G act on the set of all Sylow p-groups by conjugation. Step 2 says this action has only one orbit, so the orbit-stabilizer theorem implies n_p divides $|G|$.

§17.3 (Optional) Simple groups and Jordan-Hölder

Prototypical example for this section: Decomposition of $\mathbb{Z}/12\mathbb{Z}$ *is* $1 \trianglelefteq \mathbb{Z}/2\mathbb{Z} \trianglelefteq \mathbb{Z}/4\mathbb{Z} \trianglelefteq \mathbb{Z}/12\mathbb{Z}$.

Just like every integer breaks down as the product of primes, we can try to break every group down as a product of "basic" groups. Armed with our idea of quotient groups, the right notion is this.

Definition 17.3.1. A **simple group** is a group with no normal subgroups other than itself and the trivial group.

> **Question 17.3.2.** For which n is $\mathbb{Z}/n\mathbb{Z}$ simple? (Hint: remember that $\mathbb{Z}/n\mathbb{Z}$ is abelian.)

Then we can try to define what it means to "break down a group".

Definition 17.3.3. A **composition series** of a group G is a sequence of subgroups H_0, H_1, …, H_n such that

$$\{1\} = H_0 \trianglelefteq H_1 \trianglelefteq H_2 \trianglelefteq \ldots \trianglelefteq H_n = G$$

of maximal length (i.e. n is as large as possible, but all H_i are of course distinct). The **composition factors** are the groups H_1/H_0, H_2/H_1, …, H_n/H_{n-1}.

You can show that the "maximality" condition implies that the composition factors are all simple groups.

Let's say two composition series are equivalent if they have the same composition factors (up to permutation); in particular they have the same length. Then it turns out that the following theorem *is* true.

Theorem 17.3.4 (Jordan-Hölder)
Every finite group G admits a unique composition series up to equivalence.

Example 17.3.5 (Fundamental theorem of arithmetic when $n = 12$)
Let's consider the group $\mathbb{Z}/12\mathbb{Z}$. It's not hard to check that the possible composition series are

$$\{1\} \trianglelefteq \mathbb{Z}/2\mathbb{Z} \trianglelefteq \mathbb{Z}/4\mathbb{Z} \trianglelefteq \mathbb{Z}/12\mathbb{Z} \text{ with factors } \mathbb{Z}/2\mathbb{Z},\ \mathbb{Z}/2\mathbb{Z},\ \mathbb{Z}/3\mathbb{Z}$$
$$\{1\} \trianglelefteq \mathbb{Z}/2\mathbb{Z} \trianglelefteq \mathbb{Z}/6\mathbb{Z} \trianglelefteq \mathbb{Z}/12\mathbb{Z} \text{ with factors } \mathbb{Z}/2\mathbb{Z},\ \mathbb{Z}/3\mathbb{Z},\ \mathbb{Z}/2\mathbb{Z}$$
$$\{1\} \trianglelefteq \mathbb{Z}/3\mathbb{Z} \trianglelefteq \mathbb{Z}/6\mathbb{Z} \trianglelefteq \mathbb{Z}/12\mathbb{Z} \text{ with factors } \mathbb{Z}/3\mathbb{Z},\ \mathbb{Z}/2\mathbb{Z},\ \mathbb{Z}/2\mathbb{Z}.$$

These correspond to the factorization $12 = 2^2 \cdot 3$.

This suggests that classifying all finite simple groups would be great progress, since every finite group is somehow a "product" of simple groups; the only issue is that there are multiple ways of building a group from constituents.

Amazingly, we actually *have* a full list of simple groups, but the list is really bizarre. Every finite simple group falls in one of the following categories:

- $\mathbb{Z}/p\mathbb{Z}$ for p a prime,
- For $n \geq 5$, the subgroup of S_n consisting of "even" permutations.
- A simple group of Lie type (which I won't explain), and
- Twenty-six "sporadic" groups which do not fit into any nice family.

The two largest of the sporadic groups have cute names. The **baby monster group** has order

$$2^{41} \cdot 3^{13} \cdot 5^6 \cdot 7^2 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 \cdot 31 \cdot 47 \approx 4 \cdot 10^{33}$$

and the **monster group** (also "**friendly giant**") has order

$$2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71 \approx 8 \cdot 10^{53}.$$

It contains twenty of the sporadic groups as subquotients (including itself), and these twenty groups are called the "**happy family**".

Math is weird.

Question 17.3.6. Show that "finite simple group of order 2" is redundant in the sense that any group of order 2 is both finite and simple.

§17.4 A few harder problems to think about

Problem 17A* (Cauchy's theorem)**.** Let G be a group and let p be a prime dividing $|G|$. Prove[1] that G has an element of order p.

[1] Cauchy's theorem can be proved without the Sylow theorems, and in fact can often be used to give alternate proofs of Sylow.

Problem 17B. Let G be a finite simple group. Show that $|G| \neq 56$.

Problem 17C (Engel's PSS?)**.** Consider the set of all words consisting of the letters a and b. Given such a word, we can change the word either by inserting a word of the form www, where w is a word, anywhere in the given word, or by deleting such a sequence from the word. Can we turn the word ab into the word ba?

Problem 17D. Let p be a prime and suppose G is a simple group whose order is a power of p. Show that $G \cong \mathbb{Z}/p\mathbb{Z}$.

Problem 17E (Athemath Community-Building Event #1, Fall 2022)**.** A group action $\cdot$ of a group G on set X is said to be

- **transitive** if for all $x_1, x_2 \in X$, there exists a g such that $g \cdot x_1 = x_2$;
- **faithful** if the only element $g \in G$ such that $g \cdot x = x$ for every $x \in X$ is $g = 1_G$. In other words, the only element which acts trivially on the entire set X is the identity element of G.

Does there exist a faithful transitive action of S_5 on a six-element set?

18 The PID structure theorem

The main point of this chapter is to discuss a classification theorem for finitely generated abelian groups. This won't take long to do, and if you like, you can read just the first section and then move on.

However, since I'm here, I will go ahead and state the result as a special case of the much more general *structure theorem*. Its corollaries include

- All finite-dimensional vector spaces are $k^{\oplus n}$.
- The classification theorem for finitely generated abelian groups,
- The Jordan decomposition of a matrix from before,
- Another canonical form for a matrix: "Frobenius normal form".

§18.1 Finitely generated abelian groups

Remark 18.1.1 — We talk about abelian groups in what follows, but really the morally correct way to think about these structures is as $\mathbb{Z}$-modules.

Definition 18.1.2. An abelian group $G = (G, +)$ is **finitely generated** if it is finitely generated as a $\mathbb{Z}$-module. (That is, there exists a finite collection $b_1, \dots, b_m \in G$, such that every $x \in G$ can be written in the form $c_1 b_1 + \cdots + c_m b_m$ for some $c_1, \dots, c_m \in \mathbb{Z}$.)

Example 18.1.3 (Examples of finitely generated abelian groups)

(a) $\mathbb{Z}$ is finitely generated (by 1).

(b) $\mathbb{Z}/n\mathbb{Z}$ is finitely generated (by 1).

(c) $\mathbb{Z}^{\oplus 2}$ is finitely generated (by two elements $(1,0)$ and $(0,1)$).

(d) $\mathbb{Z}^{\oplus 3} \oplus \mathbb{Z}/9\mathbb{Z} \oplus \mathbb{Z}/2016\mathbb{Z}$ is finitely generated by five elements.

(e) $\mathbb{Z}/3\mathbb{Z} \oplus \mathbb{Z}/5\mathbb{Z}$ is finitely generated by two elements.

Exercise 18.1.4. In fact $\mathbb{Z}/3\mathbb{Z} \oplus \mathbb{Z}/5\mathbb{Z}$ is generated by *one* element. What is it?

You might notice that these examples are not very diverse. That's because they are actually the only examples:

Theorem 18.1.5 (Fundamental theorem of finitely generated abelian groups)
Let G be a finitely generated abelian group. Then there exists an integer r, prime powers $q_1, \dots, q_m$ (not necessarily distinct) such that

$$G \cong \mathbb{Z}^{\oplus r} \oplus \mathbb{Z}/q_1\mathbb{Z} \oplus \mathbb{Z}/q_2\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/q_m\mathbb{Z}.$$

This decomposition is unique up to permutation of the $\mathbb{Z}/q_i\mathbb{Z}$.

Definition 18.1.6. The **rank** of a finitely generated abelian group G is the integer r above.

Now, we could prove this theorem, but it is more interesting to go for the gold and state and prove the entire structure theorem.

§18.2 Some ring theory prerequisites

Prototypical example for this section: $R = \mathbb{Z}$.

Before I can state the main theorem, I need to define a few terms for UFD's, which behave much like $\mathbb{Z}$:

Our intuition from the case $R = \mathbb{Z}$ basically carries over verbatim.

We don't even need to deal with prime ideals and can factor elements instead.

Definition 18.2.1. If R is a UFD, then $p \in R$ is a **prime element** if (p) is a prime ideal and $p \neq 0$. For UFD's this is equivalent to: if $p = xy$ then either x or y is a unit.

So for example in $\mathbb{Z}$ the set of prime elements is $\{\pm 2, \pm 3, \pm 5, \dots\}$. Now, since R is a UFD, every element r factors into a product of prime elements

$$r = up_1^{e_1}p_2^{e_2}\dots p_m^{e_m}$$

Definition 18.2.2. We say r **divides** s if $s = r'r$ for some $r' \in R$. This is written $r \mid s$.

Example 18.2.3 (Divisibility in $\mathbb{Z}$)
The number 0 is divisible by every element of $\mathbb{Z}$. All other divisibility as expected.

Question 18.2.4. Show that $r \mid s$ if and only if the exponent of each prime in r is less than or equal to the corresponding exponent in s.

Now, the case of interest is the even stronger case when R is a PID:

Proposition 18.2.5 (PID's are Noetherian UFD's)
If R is a PID, then it is Noetherian and also a UFD.

Proof. The fact that R is Noetherian is obvious. For R to be a UFD we essentially repeat the proof for $\mathbb{Z}$, using the fact that (a, b) is principal in order to extract $\gcd(a, b)$. □

In this case, we have a Chinese remainder theorem for elements.

Theorem 18.2.6 (Chinese remainder theorem for rings)
Let m and n be relatively prime elements, meaning $(m) + (n) = (1)$. Then

$$R/(mn) \cong R/(m) \times R/(n).$$

Here the ring product is as defined in Example 4.3.8.

Proof. This is the same as the proof of the usual Chinese remainder theorem. First, since $(m, n) = (1)$ we have $am + bn = 1$ for some a and b. Then we have a map

$$R/(m) \times R/(n) \to R/(mn) \quad \text{by} \quad (r, s) \mapsto r \cdot bn + s \cdot am.$$

One can check that this map is well-defined and an isomorphism of rings. (Diligent readers invited to do so.) □

Finally, we need to introduce the concept of a Noetherian R-module.

Definition 18.2.7. An R-module M is **Noetherian** if it satisfies one of the two equivalent conditions:

- Its submodules obey the ascending chain condition: there is no infinite sequence of modules $M_1 \subsetneq M_2 \subsetneq \dots$.
- All submodules of M (including M itself) are finitely generated.

This generalizes the notion of a Noetherian ring: a Noetherian ring R is one for which R is Noetherian as an R-module.

Question 18.2.8. Check these two conditions are equivalent. (Copy the proof for rings.)

§18.3 The structure theorem

Our structure theorem takes two forms:

Theorem 18.3.1 (Structure theorem, invariant form)
Let R be a PID and let M be any finitely generated R-module. Then

$$M \cong \bigoplus_{i=1}^{m} R/(s_i)$$

for some s_i (possibly zero) satisfying $s_1 \mid s_2 \mid \cdots \mid s_m$.

Corollary 18.3.2 (Structure theorem, primary form)
Let R be a PID and let M be any finitely generated R-module. Then

$$M \cong R^{\oplus r} \oplus R/(q_1) \oplus R/(q_2) \oplus \cdots \oplus R/(q_m)$$

where $q_i = p_i^{e_i}$ for some prime element p_i and integer $e_i \geq 1$.

Proof of corollary. Factor each s_i into prime factors (since R is a UFD), then use the Chinese remainder theorem. □

Remark 18.3.3 — In both theorems the decomposition is unique up to permutations of the summands.

§18.4 Reduction to maps of free R-modules

Definition 18.4.1. A **free R-module** is a module of the form $R^{\oplus n}$ (or more generally, $\bigoplus_I R$ for some indexing set I, just to allow an infinite basis).

The proof of the structure theorem proceeds in two main steps. First, we reduce the problem to a *linear algebra* problem involving free R-modules $R^{\oplus d}$. Once that's done, we just have to play with matrices; this is done in the next section.

Suppose M is finitely generated by d elements. Then there is a surjective map of R-modules

$$R^{\oplus d} \twoheadrightarrow M$$

whose image on the basis of $R^{\oplus d}$ are the generators of M. Let K denote the kernel.

We claim that K is finitely generated as well. To this end we prove that

Lemma 18.4.2 (Direct sum of Noetherian modules is Noetherian)
Let M and N be two Noetherian R-modules. Then the direct sum $M \oplus N$ is also a Noetherian R-module.

Proof. It suffices to show that if $L \subseteq M \oplus N$, then L is finitely generated. One guess is that $L = P \oplus Q$, where P and Q are the projections of L onto M and N. Unfortunately this is false (take $M = N = \mathbb{Z}$ and $L = \{(n, n) \mid n \in \mathbb{Z}\}$) so we will have to be more careful.

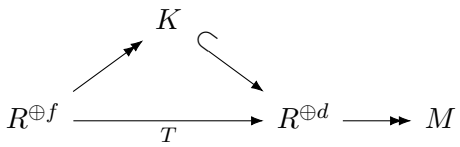
Consider the submodules

$$A = \{x \in M \mid (x, 0) \in L\} \subseteq M$$
$$B = \{y \in N \mid \exists x \in M : (x, y) \in L\} \subseteq N.$$

(Note the asymmetry for A and B: the proof doesn't work otherwise.) Then A is finitely generated by $a_1, \dots, a_k$, and B is finitely generated by $b_1, \dots, b_\ell$. Let $x_i = (a_i, 0)$ and let $y_i = (*, b_i)$ be elements of L (where the $*$'s are arbitrary things we don't care about). Then x_i and y_i together generate L. □

Question 18.4.3. Deduce that for R a PID, $R^{\oplus d}$ is Noetherian.

Hence $K \subseteq R^{\oplus d}$ is finitely generated as claimed. So we can find another surjective map $R^{\oplus f} \twoheadrightarrow K$. Consequently, we have a composition



Observe that M is the *cokernel* of the linear map T, i.e. we have that

$$M \cong R^{\oplus d} / \operatorname{im}(T).$$

So it suffices to understand the map T well.

§18.5 Uniqueness of primary form

In this section, we will prove that if $M \cong R^{\oplus r} \oplus R/(q_1) \oplus R/(q_2) \oplus \cdots \oplus R/(q_m)$, then the integer r and the prime powers q_i are unique, up to permutations.

First, we consider the case where M is free.

Theorem 18.5.1 (Uniqueness of free module's rank)

For a commutative integral domain R, if a free module M has a finite basis, then every other basis has the same number of elements.

It was mentioned once in Theorem 9.4.7 that the strategy of the proof is to pass to the field case. Indeed, we're going to pass to the field F being the fraction field of R, then directly apply the dimension theorem for vector spaces.

Proof. As before, but we prove by contradiction this time. Assume $v_1, \dots, v_n$ is a basis for the free module M of rank n, while $w_1, \dots, w_m$ are any elements of M such that $m > n$.

Let F be the fraction field of R, and embed the R-module $M \cong R^n$ into the F-vector space $V \cong F^n$.

Then, because $m > n$, as elements of V, the elements $w_1, \dots, w_m$ are linearly dependent, which means there are some elements $f_1, \dots, f_m \in F$ not all zero, such that $f_1 w_1 + \cdots + f_m w_m = 0$.

By clearing denominators, we can obtain ring elements $r_1, \dots, r_m \in R$ not all zero such that $r_1 w_1 + \cdots + r_m w_m = 0$. This means $w_1, \dots, w_m$ cannot be a basis for M. □

Next, we prove the case where the rank r is 0. This case needs a different strategy, but it still boils down to applying the dimension theorem for appropriately constructed vector spaces.

Theorem 18.5.2

Let R be a PID, let p be a prime element of R, and let $M \cong R/(p^{e_1}) \oplus R/(p^{e_2}) \oplus \cdots \oplus R/(p^{e_m})$ for positive integers $e_1, \dots, e_m$. Then the e_i are unique, up to permutations.

Intuitively, what the following proof is trying to do is:

If we can compute the exponents e_i from intrinsic properties of M, then the exponents must be unique.

Let us consider a simple case — consider $R = \mathbb{Z}$ and $M = \mathbb{Z}/4\mathbb{Z}$. This module has 4 elements, but it's not the same as $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$. In this case, the difference between the two modules can be detected by the fact that in M, the element 1 (mod 4) is not zero when multiplied by 2, on the other hand, multiplying by 2 makes every element in $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ zero.

For notational convenience,

Definition 18.5.3. For $r \in R$ and a R-module M, define $rM = \{rm \mid m \in M\}$. (Check that this is still a R-module.)

Then, what the paragraph above says is that $M \ncong \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ because $|2M| = 2 \neq 1 = |2(\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z})|$. In other words, in this case, counting the number of elements of $2M$ suffices to distinguish the two modules.

For modules of the form $M \cong R/(p^{e_1}) \oplus R/(p^{e_2}) \oplus \cdots \oplus R/(p^{e_m})$ where $R = \mathbb{Z}$, this almost works in general — except we also need to consider the number of elements in pM, p^2M, p^3M, etc.

Equivalently, we may also consider the number of elements in the successive quotients: M/pM, pM/p^2M, p^2M/p^3M, etc.

Example 18.5.4

Let $R = \mathbb{Z}$, $p = 3$, and $M = \mathbb{Z}/3^2\mathbb{Z} \oplus \mathbb{Z}/3^5\mathbb{Z}$. Then:

- $|M/3M| = 9$,
- $|3M/3^2M| = 9$,
- $|3^2M/3^3M| = 3$,
- $|3^3M/3^4M| = 3$,
- $|3^4M/3^5M| = 3$,
- $|3^eM/3^{e+1}M| = 1$ for all integer $e \geq 5$.

You can already see where this is going — each decrement of the size of the quotient corresponds to a prime power p^{e_i}.

When the quotient is infinite however, we can no longer do this. However, note that:

Lemma 18.5.5

For each integer $e \geq 0$, then $p^eM/p^{e+1}M$ is a $R/(p)$-vector space.

Thus, instead of counting the number of elements in $p^eM/p^{e+1}M$, we count the *dimension* of the $p^eM/p^{e+1}M$ as a $R/(p)$-vector space — by Theorem 9.4.7, this is indeed intrinsic to the module M.

Proof of Theorem 18.5.2. Note that, since $M \cong R/(p^{e_1}) \oplus R/(p^{e_2}) \oplus \cdots \oplus R/(p^{e_m})$, we have

$$\pi(M) \cong \pi(R/(p^{e_1})) \oplus \pi(R/(p^{e_2})) \oplus \cdots \oplus \pi(R/(p^{e_m}))$$

where $\pi(M) = p^eM/p^{e+1}M$ for any integer $e \geq 0$.

This means, as $R/(p)$-vector space,

$$\dim \pi(M) = \dim \pi(R/(p^{e_1})) + \dim \pi(R/(p^{e_2})) + \cdots + \dim \pi(R/(p^{e_m})).$$

Note that, for each term $R/(p^{e_1})$, then

$$\dim p^e(R/(p^{e_i}))/p^{e+1}(R/(p^{e_1})) = \begin{cases} 1 & e < e_i \\ 0 & \text{otherwise.} \end{cases}$$

With some arithmetic, you can see that the values e_i are indeed uniquely determined by $\dim p^eM/p^{e+1}M$, up to permutation. □

Note that this can be easily generalized to the case where the primes in the denominator may be different – because for different primes p and q of R, then $p^e(R/(q))/p^{e+1}(R/(q))$ is a 0-dimensional $R/(p)$-vector space.

Finally, we handle the general case.

Theorem 18.5.6

If $M \cong R^{\oplus r} \oplus R/(q_1) \oplus R/(q_2) \oplus \cdots \oplus R/(q_m)$, then the integer r and the prime powers q_i are unique, up to permutations.

Proof. From the two theorems above, it suffices if we can prove that the $R^{\oplus r}$ part and the $R/(q_1) \oplus R/(q_2) \oplus \cdots \oplus R/(q_m)$ part are uniquely determined from M.

For notational convenience, we call an element $a \in M$ a **torsion element** if there is $r \in R$, $r \neq 0$ such that $ra = 0$.

Then,

- If an element $a \in M$ has the $R^{\oplus r}$ component zero, then $q_1 q_2 \cdots q_m \cdot a = 0$, thus a is a torsion element.
- If an element $a \in M$ has the $R^{\oplus r}$ component nonzero, then a is not a torsion element.

In other words, the submodule consisting of all torsion elements is identical to the submodule of the elements with $R^{\oplus r}$ component zero, thus is isomorphic to $R/(q_1) \oplus R/(q_2) \oplus \cdots \oplus R/(q_m)$.

For notation convenience, let $\operatorname{Tor}(M)$ be the submodule of M consisting of all torsion elements. Then $\operatorname{Tor}(M) \cong R/(q_1) \oplus R/(q_2) \oplus \cdots \oplus R/(q_m)$ and $M/\operatorname{Tor}(M) \cong R^{\oplus r}$, in other words, the $R^{\oplus r}$ part and the $R/(q_1) \oplus R/(q_2) \oplus \cdots \oplus R/(q_m)$ part are uniquely determined from M, so we're done. □

§18.6 Smith normal form

The idea is now that we have reduced our problem to studying linear maps $T\colon R^{\oplus m} \to R^{\oplus n}$, which can be thought of as a generic matrix

$$T = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix}$$

for a basis $e_1, \dots, e_m$ of $R^{\oplus m}$ and $f_1, \dots, f_n$ of $R^{\oplus n}$.

Of course, as you might expect it ought to be possible to change the given basis of T such that T has a nicer matrix form. We already saw this in *Jordan form*, where we had a map $T\colon V \to V$ and changed the basis so that T was "almost diagonal". This time, we have *two* sets of bases we can change, so we would hope to get a diagonal basis, or even better.

Before proceeding let's think about how we might edit the matrix: what operations are permitted? Here are some examples:

- Swapping rows and columns, which just corresponds to re-ordering the basis.
- Adding a multiple of a column to another column. For example, if we add 3 times the first column to the second column, this is equivalent to replacing the basis

$$(e_1, e_2, e_3, \dots, e_m) \mapsto (e_1, e_2 + 3e_1, e_3, \dots, e_m).$$

- Adding a multiple of a row to another row. One can see that adding 3 times the first row to the second row is equivalent to replacing the basis

$$(f_1, f_2, f_3, \dots, f_n) \mapsto (f_1 - 3f_2, f_2, f_3, \dots, f_n).$$

More generally,

If A is an invertible $n \times n$ matrix we can replace T with AT.

This corresponds to replacing

$$(f_1, \dots, f_n) \mapsto ((FA^{-1})_1, \dots, (FA^{-1})_n)$$

(the "invertible" condition just guarantees the latter is a basis). Here, F is the $n \times n$ matrix with columns being $f_1, \dots, f_n$, and $(FA^{-1})_1$ denotes the first column of FA^{-1}.

Of course similarly we can replace T with TB where B is an invertible $m \times m$ matrix; this corresponds to

$$(e_1, \dots, e_m) \mapsto ((EB)_1, \dots, (EB)_m)$$

where E is the $m \times m$ matrix with columns being $e_1, \dots, e_m$.

Armed with this knowledge, we can now approach:

Theorem 18.6.1 (Smith normal form)
Let R be a PID. Let $M = R^{\oplus m}$ and $N = R^{\oplus n}$ be free R-modules and let $T \colon M \to N$ be a linear map. Set $k = \min\{m, n\}$.
Then we can select a pair of new bases for M and N such that T has only diagonal entries s_1, s_2, …, s_k and $s_1 \mid s_2 \mid \cdots \mid s_k$.

So if $m > n$, the matrix should take the form

$$\begin{bmatrix} s_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & s_2 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & s_n & \dots & 0 \end{bmatrix}.$$

and similarly when $m \le n$.

Question 18.6.2. Show that Smith normal form implies the structure theorem.

Remark 18.6.3 — Note that this is not a generalization of Jordan form.

- In Jordan form we consider maps $T \colon V \to V$; note that the source and target space are the *same*, and we are considering one basis for the space V.
- In Smith form the maps $T \colon M \to N$ are between *different* modules, and we pick *two* sets of bases (one for M and one for N).

Example 18.6.4 (Example of Smith normal form)

To give a flavor of the idea of the proof, let's work through a concrete example with the $\mathbb{Z}$-matrix

$$\begin{bmatrix} 18 & 38 & 48 \\ 14 & 30 & 32 \end{bmatrix}.$$

The GCD of all the entries is 2, and so motivated by this, we perform the **Euclidean algorithm on the left column**: subtract the second row from the first row, then three times the first row from the second:

$$\begin{bmatrix} 18 & 38 & 48 \\ 14 & 30 & 32 \end{bmatrix} \mapsto \begin{bmatrix} 4 & 8 & 16 \\ 14 & 30 & 32 \end{bmatrix} \mapsto \begin{bmatrix} 4 & 8 & 16 \\ 2 & 6 & -16 \end{bmatrix}.$$

Now that the GCD of 2 is present, we move it to the upper-left by switching the two rows, and then kill off all the entries in the same row/column; since 2 was the GCD all along, we isolate 2 completely:

$$\begin{bmatrix} 4 & 8 & 16 \\ 2 & 6 & -16 \end{bmatrix} \mapsto \begin{bmatrix} 2 & 6 & -16 \\ 4 & 8 & 16 \end{bmatrix} \mapsto \begin{bmatrix} 2 & 6 & -16 \\ 0 & -4 & 48 \end{bmatrix} \mapsto \begin{bmatrix} 2 & 0 & 0 \\ 0 & -4 & 48 \end{bmatrix}.$$

This reduces the problem to a 1×2 matrix. So we just apply the Euclidean algorithm again there:

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \end{bmatrix} \mapsto \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \end{bmatrix}.$$

Now all we have to do is generalize this proof to work with any PID. It's intuitively clear how to do this: the PID condition more or less lets you perform a Euclidean algorithm.

Proof of Smith normal form. Begin with a generic matrix

$$T = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix}$$

We want to show, by a series of operations (gradually changing the given basis) that we can rearrange the matrix into Smith normal form.

Define $\gcd(x, y)$ to be any generator of the principal ideal (x, y).

Claim 18.6.5 ("Euclidean algorithm")**.** If a and b are entries in the same row or column, we can change bases to replace a with $\gcd(a, b)$ and b with something else.

Proof. We do just the case of columns. By hypothesis, $\gcd(a, b) = xa + yb$ for some $x, y \in R$. We must have $(x, y) = (1)$ now (we're in a UFD). So there are u and v such that $xu + yv = 1$. Then

$$\begin{bmatrix} x & y \\ -v & u \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \gcd(a, b) \\ \text{something} \end{bmatrix}$$

and the first matrix is invertible (check this!), as desired. ∎

Let $s_1 = (a_{ij})_{i,j}$ be the GCD of all entries. Now by repeatedly applying this algorithm, we can cause s to appear in the upper left hand corner. Then, we use it to kill off all the

entries in the first row and the first column, thus arriving at a matrix

$$\begin{bmatrix} s_1 & 0 & 0 & \dots & 0 \\ 0 & a'_{22} & a'_{23} & \dots & a'_{2n} \\ 0 & a'_{32} & a'_{33} & \dots & a'_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a'_{m2} & a'_{m3} & \dots & a'_{mn} \end{bmatrix}.$$

Now we repeat the same procedure with this lower-right $(m-1) \times (n-1)$ matrix, and so on. This gives the Smith normal form. □

With the Smith normal form, we have in the original situation that

$$M \cong R^{\oplus d} / \operatorname{im} T$$

and applying the theorem to T completes the proof of the structure theorem.

§18.7 A few harder problems to think about

Now, we can apply our structure theorem!

Problem 18A† (Finite-dimensional vector spaces are all isomorphic)**.** A vector space V over a field k has a finite spanning set of vectors. Show that $V \cong k^{\oplus n}$ for some n.

Problem 18B† (Frobenius normal form)**.** Let $T \colon V \to V$ where V is a finite-dimensional vector space over an arbitrary field k (not necessarily algebraically closed). Show that one can write T as a block-diagonal matrix whose blocks are all of the form

$$\begin{bmatrix} 0 & 0 & 0 & \dots & 0 & * \\ 1 & 0 & 0 & \dots & 0 & * \\ 0 & 1 & 0 & \dots & 0 & * \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & * \end{bmatrix}.$$

(View V as a $k[x]$-module with action $x \cdot v = T(v)$.)

Problem 18C† (Jordan normal form)**.** Let $T \colon V \to V$ where V is a finite-dimensional vector space over an arbitrary field k which is algebraically closed. Prove that T can be written in Jordan form.

Problem 18D. Find two abelian groups G and H which are not isomorphic, but for which there are injective homomorphisms $G \hookrightarrow H$ and $H \hookrightarrow G$.

Problem 18E. Let $A \subseteq B \subseteq C$ be rings. Suppose C is a finitely generated A-module. Does it follow that B is a finitely generated A-module?

VI

Representation Theory

Part VI: Contents

19 Representations of algebras

In the 19th century, the word "group" hadn't been invented yet; all work was done with subsets of $\mathrm{GL}(n)$ or S_n. Only much later was the abstract definition of a group was given, an abstract set G which was an object in its own right.

While this abstraction is good for some reasons, it is often also useful to work with concrete representations. This is the subject of representation theory. Linear algebra is easier than abstract algebra, so if we can take a group G and represent it concretely as a set of matrices in $\mathrm{GL}(n)$, this makes them easier to study. This is the *representation theory of groups*: how can we take a group and represent its elements as matrices?

§19.1 Algebras

Prototypical example for this section: $k[x_1, \dots, x_n]$ *and* $k[G]$.

Rather than working directly with groups from the beginning, it will be more convenient to deal with so-called k-algebras. This setting is more natural and general than that of groups, so once we develop the theory of algebras well enough, it will be fairly painless to specialize to the case of groups.

Colloquially,

An associative k-algebra is a possibly noncommutative ring with a copy of k inside it. It is thus a k-vector space.

I'll present examples before the definition:

Example 19.1.1 (Examples of k-algebras)

Let k be any field. The following are examples of k-algebras:

(a) The field k itself.

(b) The polynomial ring $k[x_1, \dots, x_n]$.

(c) The set of $n \times n$ matrices with entries in k, which we denote by $\mathrm{Mat}_n(k)$. Note the multiplication here is not commutative.

(d) The set $\mathrm{Mat}(V)$ of linear maps $T \colon V \to V$, with multiplication given by the composition of operators. (Here V is some vector space over k.) This is really the same as the previous example.

Definition 19.1.2. Let k be a field. A **k-algebra** A is a *possibly noncommutative* ring, equipped with a ring homomorphism $k \hookrightarrow A$, whose image is the "copy of k". (In particular, $1_k \mapsto 1_A$.)

Thus we can consider k as a subset of A, and we then additionally require $\lambda \cdot a = a \cdot \lambda$ for each $\lambda \in k$ and $a \in A$.

If the multiplication operation is also commutative, then we say A is a **commutative algebra**.

Definition 19.1.3. Equivalently, a **k-algebra** A is a *k-vector space* which also has an associative, bilinear multiplication operation (with an identity 1_A). The "copy of k" is obtained by considering elements $\lambda 1_A$ for each $\lambda \in k$ (i.e. scaling the identity by the elements of k, taking advantage of the vector space structure).

Abuse of Notation 19.1.4. Some other authors don't require A to be associative or to have an identity, so to them what we have just defined is an "associative algebra with 1". However, this is needlessly wordy for our purposes.

Example 19.1.5 (Group algebra)
The **group algebra** $k[G]$ is the k-vector space whose *basis elements* are the elements of a group G, and where the product of two basis elements is the group multiplication. For example, suppose $G = \mathbb{Z}/2\mathbb{Z} = \{1_G, x\}$. Then

$$k[G] = \{a1_G + bx \mid a, b \in k\}$$

with multiplication given by

$$(a1_G + bx)(c1_G + dx) = (ac + bd)1_G + (bc + ad)x.$$

Question 19.1.6. When is $k[G]$ commutative?

The example $k[G]$ is very important, because (as we will soon see) a representation of the algebra $k[G]$ amounts to a representation of the group G itself.

It is worth mentioning at this point that:

Definition 19.1.7. A **homomorphism** of k-algebras A, B is a linear map $T \colon A \to B$ which respects multiplication (i.e. $T(xy) = T(x)T(y)$) and which sends 1_A to 1_B. In other words, T is both a homomorphism as a ring and as a vector space.

We will also need to recall the "product ring" from Example 4.3.8, but for algebras, we will prefer a different name and notation.

Definition 19.1.8. Given k-algebras A and B, the **direct sum** $A \oplus B$ is defined as pairs $a + b$, where addition is done in the obvious way, but we declare $ab = 0$ for any $a \in A$ and $b \in B$.

Question 19.1.9. Show that $1_A + 1_B$ is the multiplicative identity of $A \oplus B$.

Equivalently, similar to Definition 9.3.1 and Example 4.3.8, you can define the direct sum $A \oplus B$ to be the set of pairs (a, b), where multiplication is defined by $(a, b)(a', b') = (aa', bb')$. In this notation, $(1_A, 1_B)$ would be the multiplicative identity of $A \oplus B$.

§19.2 Representations

Prototypical example for this section: $k[S_3]$ acting on $k^{\oplus 3}$ is my favorite.

Definition 19.2.1. A **representation** of a k-algebra A (also a **left A-module**) is:

(i) A k-vector space V, and

(ii) An *action* $\cdot$ of A on V: thus, for every $a \in A$ we can take $v \in V$ and act on it to get $a \cdot v$. This satisfies the usual axioms:
- $(a+b) \cdot v = a \cdot v + b \cdot v$, $a \cdot (v+w) = a \cdot v + a \cdot w$, and $(ab) \cdot v = a \cdot (b \cdot v)$.
- $\lambda \cdot v = \lambda v$ for $\lambda \in k$. In particular, $1_A \cdot v = v$.

Definition 19.2.2. The action of A can be more succinctly described as saying that there is a *k-algebra homomorphism* $\rho \colon A \to \operatorname{Mat}(V)$. (So $a \cdot v = \rho(a)(v)$.) Thus we can also define a **representation** of A as a pair

$$(V, \rho \colon A \to \operatorname{Mat}(V)) .$$

This is completely analogous to how a group action G on a set X with n elements just amounts to a group homomorphism $G \to S_n$. From this perspective, what we are really trying to do is:

If A is an algebra, we are trying to *represent* the elements of A as matrices.

Abuse of Notation 19.2.3. While a representation is a pair (V, ρ) of *both* the vector space V and the action ρ, we frequently will just abbreviate it to "V". This is probably one of the worst abuses I will commit, but everyone else does it and I fear the mob.

Abuse of Notation 19.2.4. Rather than $\rho(a)(v)$ we will just write $\rho(a)v$.

Example 19.2.5 (Representations of $\operatorname{Mat}(V)$)

(a) Let $A = \operatorname{Mat}_2(\mathbb{R})$. Then there is a representation $(\mathbb{R}^{\oplus 2}, \rho)$ where a matrix $a \in A$ just acts by $a \cdot v = \rho(a)(v) = a(v)$.

(b) More generally, given a vector space V over any field k, there is an obvious representation of $A = \operatorname{Mat}(V)$ by $a \cdot v = \rho(a)(v) = a(v)$ (since $a \in \operatorname{Mat}(V)$).

From the matrix perspective: if $A = \operatorname{Mat}(V)$, then we can just represent A as matrices over V.

(c) There are other representations of $A = \operatorname{Mat}_2(\mathbb{R})$. A silly example is the representation $(\mathbb{R}^{\oplus 4}, \rho)$ given by

$$\rho : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mapsto \begin{bmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & c & d \end{bmatrix} .$$

More abstractly, viewing $\mathbb{R}^{\oplus 4}$ as $(\mathbb{R}^{\oplus 2}) \oplus (\mathbb{R}^{\oplus 2})$, this is $a \cdot (v_1, v_2) = (a \cdot v_1, a \cdot v_2)$.

Example 19.2.6 (Representations of polynomial algebras)

(a) Let $A = k$. Then a representation of k is just any k-vector space V.

(b) If $A = k[x]$, then a representation (V, ρ) of A amounts to a vector space V plus the choice of a linear map $T \in \operatorname{Mat}(V)$ (by $T = \rho(x)$).

(c) If $A = k[x]/(x^2)$ then a representation (V, ρ) of A amounts to a vector space V

plus the choice of a linear map $T \in \mathrm{Mat}(V)$ satisfying $T^2 = 0$.

(d) We can create arbitrary "functional equations" with this pattern. For example, if $A = k[x,y]/(x^2 - x + y, y^4)$ then representing A by V amounts to finding commuting operators $S, T \in \mathrm{Mat}(V)$ satisfying $S^2 = S - T$ and $T^4 = 0$.

Example 19.2.7 (Representations of groups)

(a) Let $A = \mathbb{R}[S_3]$. Then let

$$V = \mathbb{R}^{\oplus 3} = \{(x,y,z) \mid x,y,z \in \mathbb{R}\}.$$

We can let A act on V as follows: given a permutation $\pi \in S_3$, we permute the corresponding coordinates in V. So for example, if

$$\text{If } \pi = (1\ 2) \text{ then } \pi \cdot (x,y,z) = (y,x,z).$$

This extends linearly to let A act on V, by permuting the coordinates.

From the matrix perspective, what we are doing is representing the permutations in S_3 as permutation matrices on $k^{\oplus 3}$, like

$$(1\ 2) \mapsto \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

(b) More generally, let $A = k[G]$. Then a representation (V, ρ) of A amounts to a group homomorphism $\psi \colon G \to \mathrm{GL}(V)$. (In particular, $\rho(1_G) = \mathrm{id}_V$.) We call this a **group representation** of G.

Example 19.2.8 (Regular representation)

Any k-algebra A is a representation (A, ρ) over itself, with $a \cdot b = \rho(a)(b) = ab$ (i.e. multiplication given by A). This is called the **regular representation**, denoted $\mathrm{Reg}(A)$.

§19.3 Direct sums

Prototypical example for this section: The example with $\mathbb{R}[S_3]$ seems best.

Definition 19.3.1. Let A be k-algebra and let $V = (V, \rho_V)$ and $W = (W, \rho_W)$ be two representations of A. Then $V \oplus W$ is a representation, with action ρ given by

$$a \cdot (v,w) = (a \cdot v, a \cdot w).$$

This representation is called the **direct sum** of V and W.

Example 19.3.2

Earlier we let $\mathrm{Mat}_2(\mathbb{R})$ act on $\mathbb{R}^{\oplus 4}$ by

$$\rho : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mapsto \begin{bmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & c & d \end{bmatrix}.$$

So this is just a direct sum of two two-dimensional representations.
You can also view the vectors of $\mathbb{R}^{\oplus 4}$ as two vectors in $\mathbb{R}^{\oplus 2}$ "stacked horizontally" as $\begin{pmatrix} e & f \\ g & h \end{pmatrix}$, so the action would be given by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}.$$

Remark 19.3.3 — Perhaps this is the reason why people tend to write V as the representation without the accompanied ρ_V, as long as it's possible to embed the k-algebra A into a subalgebra of $\mathrm{Mat}_d(k)$, then V can be isomorphically embedded as a subrepresentation of $(k^{\oplus d})^{\oplus m}$, being m copies of the obvious $k^{\oplus d}$ representation stacked horizontally.

More generally, given representations (V, ρ_V) and (W, ρ_W) the representation ρ of $V \oplus W$ looks like

$$\rho(a) = \begin{bmatrix} \rho_V(a) & 0 \\ 0 & \rho_W(a) \end{bmatrix}.$$

Example 19.3.4 (Representation of S_n decomposes)
Let $A = \mathbb{R}[S_3]$ again, acting via permutation of coordinates on

$$V = \mathbb{R}^{\oplus 3} = \{(x,y,z) \mid x,y,z \in \mathbb{R}\}.$$

Consider the two subspaces

$$W_1 = \{(t,t,t) \mid t \in \mathbb{R}\}$$
$$W_2 = \{(x,y,z) \mid x+y+z = 0\}.$$

Note $V = W_1 \oplus W_2$ as vector spaces. But each of W_1 and W_2 is a subrepresentation (since the action of A keeps each W_i in place), so $V = W_1 \oplus W_2$ as representations too.

Direct sums also come up when we play with algebras.

Proposition 19.3.5 (Representations of $A \oplus B$ are $V_A \oplus V_B$)
Let A and B be k-algebras. Then every representation of $A \oplus B$ is of the form

$$V_A \oplus V_B$$

where V_A and V_B are representations of A and B, respectively.

Example 19.3.6
Take $A = B = \mathrm{Mat}_2(\mathbb{R})$. There are two obvious representations of the k-algebra $A \oplus B$, V_A and V_B, corresponds to the action of A and B respectively.
Each of V_A and V_B are isomorphic to $\mathbb{R}^{\oplus 2}$ as $\mathbb{R}$-vector spaces.
What this proposition says is that, you cannot "mix" the action of A and B in order to get some representation $V \cong \mathbb{R}^2$ of $A \oplus B$, such as by $(a+b) \cdot v = a \cdot v + 2b \cdot v$ for $a \in A$ and $b \in B$.

Sketch of Proof. Let (V, ρ) be a representation of $A \oplus B$. For any $v \in V$, $\rho(1_A + 1_B)v = \rho(1_A)v + \rho(1_B)v$. One can then set $V_A = \{\rho(1_A)v \mid v \in V\}$ and $V_B = \{\rho(1_B)v \mid v \in V\}$. These are disjoint, since if $\rho(1_A)v = \rho(1_B)v'$, we have $\rho(1_A)v = \rho(1_A 1_A)v = \rho(1_A 1_B)v' = 0_V$, and similarly for the other side. □

In the example above, if you see the representation $V_A \oplus V_B$ as $\mathbb{R}^4$, then any element in A acting on an element in $V_A \oplus V_B$ would zero out the V_B-component of the vector.
So, the key idea of the proof is:

The A and B component of $A \oplus B$ is used to act on V, in order to project the vector space V into the components V_A and V_B to separate out the subrepresentations.

§19.4 Irreducible and indecomposable representations

Prototypical example for this section: $k[S_3]$ *decomposes as the sum of two spaces.*

One of the goals of representation theory will be to classify all possible representations of an algebra A. If we want to have a hope of doing this, then we want to discard "silly" representations such as

$$\rho : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mapsto \begin{bmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & c & d \end{bmatrix}$$

and focus our attention instead on "irreducible" representations. This motivates:

Definition 19.4.1. Let V be a representation of A. A **subrepresentation** $W \subseteq V$ is a subspace W with the property that for any $a \in A$ and $w \in W$, $a \cdot w \in W$. In other words, this subspace is invariant under actions by A.

Thus for example if $V = W_1 \oplus W_2$ for representations W_1, W_2 then W_1 and W_2 are subrepresentations of V.

Definition 19.4.2. If V has no proper nonzero subrepresentations then it is **irreducible**. If there is no pair of proper subrepresentations W_1, W_2 such that $V = W_1 \oplus W_2$, then we say V is **indecomposable**.

Definition 19.4.3. For brevity, an **irrep** of an algebra/group is a *finite-dimensional* irreducible representation.

Example 19.4.4 (Representation of S_n decomposes)
Let $A = \mathbb{R}[S_3]$ again, acting via permutation of coordinates on

$$V = \mathbb{R}^{\oplus 3} = \{(x,y,z) \mid x,y,z \in \mathbb{R}\}.$$

Consider again the two subspaces

$$\begin{aligned} W_1 &= \{(t,t,t) \mid t \in \mathbb{R}\} \\ W_2 &= \{(x,y,z) \mid x+y+z = 0\}. \end{aligned}$$

As we've seen, $V = W_1 \oplus W_2$, and thus V is not irreducible. But one can show that W_1 and W_2 are irreducible (and hence indecomposable) as follows.

- For W_1 it's obvious, since W_1 is one-dimensional.
- For W_2, consider any vector $w = (a,b,c)$ with $a+b+c = 0$ and not all zero. Then WLOG we can assume $a \neq b$ (since not all three coordinates are equal). In that case, $(1\ 2)$ sends w to $w' = (b,a,c)$. Then w and w' span W_2.

Thus V breaks down completely into irreps.

Unfortunately, if W is a subrepresentation of V, then it is not necessarily the case that we can find a supplementary vector space W' such that $V = W \oplus W'$. Put another way, if V is reducible, we know that it has a subrepresentation, but a decomposition requires *two* subrepresentations. Here is a standard counterexample:

Exercise 19.4.5. Let $A = \mathbb{R}[x]$, and $V = \mathbb{R}^{\oplus 2}$ be the representation with action

$$\rho(x) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Show that the only subrepresentation is $W = \{(t,0) \mid t \in \mathbb{R}\}$. So V is not irreducible, but it is indecomposable.

Here is a slightly more optimistic example, and the "prototypical example" that you should keep in mind.

Exercise 19.4.6. Let $A = \mathrm{Mat}_d(k)$ and consider the obvious representation $k^{\oplus d}$ of A that we described earlier. Show that it is irreducible. (This is obvious if you understand the definitions well enough.)

§19.5 Morphisms of representations

We now proceed to define the morphisms between representations.

Definition 19.5.1. Let (V, ρ_V) and (W, ρ_W) be representations of A. An **intertwining operator**, or **morphism**, is a linear map $T\colon V \to W$ such that

$$T(a \cdot v) = a \cdot T(v)$$

for any $a \in A$, $v \in V$. (Note that the first $\cdot$ is the action of ρ_V and the second $\cdot$ is the action of ρ_W.) This is exactly what you expect if you think that V and W are "left A-modules". If T is invertible, then it is an **isomorphism** of representations and we say $V \cong W$.

Remark 19.5.2 (For commutative diagram lovers) — The condition $T(a\cdot v) = a\cdot T(v)$ can be read as saying that

$$\begin{array}{ccc} V & \xrightarrow{\rho_1(a)} & V \\ {\scriptstyle T}\downarrow & & \downarrow{\scriptstyle T} \\ W & \xrightarrow[\rho_2(a)]{} & W \end{array}$$

commutes for any $a \in A$.

Remark 19.5.3 (For category lovers) — A representation is just a "bilinear" functor from an abelian one-object category $\{*\}$ (so $\operatorname{Hom}(*,*) \cong A$) to the abelian category Vect_k. Then an intertwining operator is just a *natural transformation*.

Here are some examples of intertwining operators.

Example 19.5.4 (Intertwining operators)

(a) For any $\lambda \in k$, the scalar map $T(v) = \lambda v$ is intertwining.

(b) If $W \subseteq V$ is a subrepresentation, then the inclusion $W \hookrightarrow V$ is an intertwining operator.

(c) The projection map $V_1 \oplus V_2 \twoheadrightarrow V_1$ is an intertwining operator.

(d) Let $V = \mathbb{R}^{\oplus 2}$ and represent $A = k[x]$ by (V, ρ) where

$$\rho(x) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Thus $\rho(x)$ is rotation by 90° around the origin. Let T be rotation by 30°. Then $T\colon V \to V$ is intertwining (the rotations commute).

Example 19.5.5 (A non-example: Representation of $\operatorname{Mat}(V)$)

Let $A = \operatorname{Mat}_2(\mathbb{R}) \oplus \operatorname{Mat}_2(\mathbb{R})$. Then A can be viewed as a subset of $\operatorname{Mat}_4(\mathbb{R})$ of the matrices of the form

$$\begin{bmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & e & f \\ 0 & 0 & g & h \end{bmatrix}.$$

There are two obvious irreps of A, given by V_1 consisting of the vectors in $\mathbb{R}^4$ of the form $(m, n, 0, 0)$, and V_2 consisting of the vectors in $\mathbb{R}^4$ of the form $(0, 0, p, q)$. In this case, even though V_1 and V_2 are isomorphic as $\mathbb{R}$-vector spaces, they're not isomorphic as representations of A – so any intertwining operator from V_1 to V_2 must be identically zero.

Exercise 19.5.6 (Kernel and image are subrepresentations)**.** Let $T\colon V \to W$ be an intertwining operator.

(a) Show that $\ker T \subseteq V$ is a subrepresentation of V.

(b) Show that $\operatorname{im} T \subseteq W$ is a subrepresentation of W.

The previous exercise gives us the famous Schur's lemma.

Theorem 19.5.7 (Schur's lemma)

Let V and W be representations of a k-algebra A. Let $T\colon V \to W$ be a *nonzero* intertwining operator. Then

(a) If V is irreducible, then T is injective.

(b) If W is irreducible, then T is surjective.

In particular if both V and W are irreducible then T is an isomorphism.

An important special case is if k is algebraically closed: then the only intertwining operators $T\colon V \to V$ are multiplication by a constant.

Theorem 19.5.8 (Schur's lemma for algebraically closed fields)

Let k be an algebraically closed field. Let V be an irrep of a k-algebra A. Then any intertwining operator $T\colon V \to V$ is multiplication by a scalar.

Exercise 19.5.9. Use the fact that T has an eigenvalue λ to deduce this from Schur's lemma. (Consider $T - \lambda \cdot \mathrm{id}_V$, and use Schur to deduce it's zero.)

We have already seen the counterexample of rotation by 90° for $k = \mathbb{R}$; this was the same counterexample we gave to the assertion that all linear maps have eigenvalues.

§19.6 The representations of $\mathrm{Mat}_d(k)$

To give an example of the kind of progress already possible, we prove:

Theorem 19.6.1 (Representations of $\mathrm{Mat}_d(k)$)
Let k be any field, d be a positive integer and let $W = k^{\oplus d}$ be the obvious representation of $A = \mathrm{Mat}_d(k)$. Then the only finite-dimensional representations of $\mathrm{Mat}_d(k)$ are $W^{\oplus n}$ for some positive integer n (up to isomorphism). In particular, it is irreducible if and only if $n = 1$.

For concreteness, I'll just sketch the case $d = 2$, since the same proof applies verbatim to other situations. This shows that the examples of representations of $\mathrm{Mat}_2(\mathbb{R})$ we gave earlier are the only ones.

As we've said this is essentially a functional equation. The algebra $A = \mathrm{Mat}_2(k)$ has basis given by four matrices

$$E_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad E_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \qquad E_3 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad E_4 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

satisfying relations like $E_1 + E_2 = \mathrm{id}_A$, $E_i^2 = E_i$, $E_1E_2 = 0$, etc. So let V be a representation of A, and let $M_i = \rho(E_i)$ for each i; we want to classify the possible matrices M_i on V satisfying the same functional equations. This is because, for example,

$$\mathrm{id}_V = \rho(\mathrm{id}_A) = \rho(E_1 + E_2) = M_1 + M_2.$$

By the same token $M_1M_3 = M_3$. Proceeding in a similar way, we can obtain the following multiplication table:

$\times$	M_1	M_2	M_3	M_4
M_1	M_1	0	M_3	0
M_2	0	M_2	0	M_4
M_3	0	M_3	0	M_1
M_4	M_4	0	M_2	0

and $M_1 + M_2 = \mathrm{id}_V$

Note that each M_i is a linear map $V \to V$; for all we know, it could have hundreds of entries. Nonetheless, given the multiplication table of the basis E_i we get the corresponding table for the M_i.

So, in short, the problem is as follows:

Find all vector spaces V and quadruples of matrices M_i satisfying the multiplication table above.

Let $W_1 = M_1^{\mathrm{img}}(V)$ and $W_2 = M_2^{\mathrm{img}}(V)$ be the images of M_1 and M_2.

Claim 19.6.2. $V = W_1 \oplus W_2$.

Proof. First, note that for any $v \in V$ we have

$$v = \rho(\mathrm{id})(v) = (M_1 + M_2)v = M_1v + M_2v.$$

Moreover, we have that $W_1 \cap W_2 = \{0\}$, because if $M_1v_1 = M_2v_2$ then $M_1v_1 = M_1(M_1v_1) = M_1(M_2v_2) = 0$. □

Claim 19.6.3. $W_1 \cong W_2$.

Proof. Check that the maps

$$W_1 \xrightarrow{\times M_4} W_2 \quad \text{and} \quad W_2 \xrightarrow{\times M_3} W_1$$

are well-defined and mutually inverse. □

Now, let $e_1, \dots, e_n$ be basis elements of W_1; thus $M_4e_1, \dots, M_4e_n$ are basis elements of W_2. However, each $\{e_j, M_4e_j\}$ forms a basis of a subrepresentation isomorphic to $W = k^{\oplus 2}$ (what's the isomorphism?).

This finally implies that all representations of A are of the form $W^{\oplus n}$. In particular, W is irreducible because there are no representations of smaller dimension at all!

§19.7 A few harder problems to think about

Problem 19A†. Suppose we have *one-dimensional* representations $V_1 = (V_1, \rho_1)$ and $V_2 = (V_2, \rho_2)$ of A. Show that $V_1 \cong V_2$ if and only if $\rho_1(a)$ and $\rho_2(a)$ are multiplication by the same constant for every $a \in A$.

Problem 19B† (Schur's lemma for commutative algebras)**.** Let A be a *commutative* algebra over an algebraically closed field k. Prove that any irrep of A is one-dimensional.

Problem 19C*. Let (V, ρ) be a representation of A. Then $\mathrm{Mat}(V)$ is a representation of A with action given by

$$a \cdot T = \rho(a) \circ T$$

for $T \in \mathrm{Mat}(V)$.

(a) Show that $\rho\colon \mathrm{Reg}(A) \to \mathrm{Mat}(V)$ is an intertwining operator.

(b) If V is d-dimensional, show that $\mathrm{Mat}(V) \cong V^{\oplus d}$ as representations of A.

Problem 19D*. Fix an algebra A. Find all intertwining operators

$$T\colon \mathrm{Reg}(A) \to \mathrm{Reg}(A).$$

Problem 19E. Let (V, ρ) be an *indecomposable* (not irreducible) representation of an algebra A. Prove that any intertwining operator $T\colon V \to V$ is either nilpotent or an isomorphism.

(Note that Theorem 19.5.8 doesn't apply, since the field k may not be algebraically closed.)

20 Semisimple algebras

In what follows, **assume the field k is algebraically closed**.

Fix an algebra A and suppose you want to study its representations. We have a "direct sum" operation already. So, much like we pay special attention to prime numbers, we're motivated to study irreducible representations and then build all the representations of A from there.

Unfortunately, we have seen (Exercise 19.4.5) that there exists a representation which is not irreducible, and yet cannot be broken down as a direct sum (indecomposable). This is *weird and bad*, so we want to give a name to representations which are more well-behaved. We say that a representation is **completely reducible** if it doesn't exhibit this bad behavior.

Even better, we say a finite-dimensional algebra A is **semisimple** if all its finite-dimensional representations are completely reducible. So when we study finite-dimensional representations of semisimple algebras A, we just have to figure out what the irreps are, and then piecing them together will give all the representations of A.

In fact, semisimple algebras A have even nicer properties. The culminating point of the chapter is when we prove that A is semisimple if and only if $A \cong \bigoplus_i \operatorname{Mat}(V_i)$, where the V_i are the irreps of A (yes, there are only finitely many!).

In the end, we will see that the group algebras $k[G]$ of a finite group G are all semisimple (at least when k has characteristic 0), thus we're justified in focusing on studying the semisimple algebras.

> **Remark 20.0.1** (Digression) — The converse does not hold, however — if k has characteristic 0, not every finite-dimensional semisimple k-algebra is isomorphic to some group algebra. Classifying exactly when a k-algebra is isomorphic to a group algebra turns out to be a hard question, see `https://mathoverflow.net/q/314502`.

§20.1 Schur's lemma continued

Prototypical example for this section: For V irreducible, $\operatorname{Hom}_{\text{rep}}(V^{\oplus 2}, V^{\oplus 2}) \cong k^{\oplus 4}$.

Definition 20.1.1. For an algebra A and representations V and W, we let $\operatorname{Hom}_{\text{rep}}(V, W)$ be the set of intertwining operators between them. (It is also a k-algebra.)

By Schur's lemma (since k is algebraically closed, which again, we are taking as a standing assumption), we already know that if V and W are irreps, then

$$\operatorname{Hom}_{\text{rep}}(V, W) \cong \begin{cases} k & \text{if } V \cong W \\ 0 & \text{if } V \ncong W. \end{cases}$$

Can we say anything more? For example, it also tells us that

$$\operatorname{Hom}_{\text{rep}}(V, V^{\oplus 2}) = k^{\oplus 2}.$$

The possible maps are $v \mapsto (c_1 v_1, c_2 v_2)$ for some choice of $c_1, c_2 \in k$.

More generally, suppose V is an irrep and consider $\operatorname{Hom}_{\text{rep}}(V^{\oplus m}, V^{\oplus n})$. Intertwining operators $T \colon V^{\oplus m} \to V^{\oplus n}$ are determined completely by the mn choices of compositions

$$V \hookrightarrow V^{\oplus m} \xrightarrow{T} V^{\oplus n} \twoheadrightarrow V$$

where the first arrow is inclusion to the ith component of $V^{\oplus m}$ (for $1 \le i \le m$) and the second arrow is inclusion to the jth component of $V^{\oplus n}$ (for $1 \le j \le n$). However, by Schur's lemma on each of these compositions, we know they must be constant.

Thus, $\text{Hom}_{\text{rep}}(V^{\oplus n}, V^{\oplus m})$ consist of $n \times m$ "matrices" of constants, and the map is provided by

$$\begin{bmatrix} c_{11} & c_{12} & \dots & c_{1(n-1)} & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2(n-1)} & c_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{m1} & c_{m2} & \dots & c_{m(n-1)} & c_{mn} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \in V^{\oplus m}$$

where the $c_{ij} \in k$ but $v_i \in V$; note the type mismatch! This is *not* just a k-linear map $V^{\oplus n} \to V^{\oplus m}$; rather, the outputs are m *linear combinations* of the inputs.

More generally, we have:

Theorem 20.1.2 (Schur's lemma for completely reducible representations)
Let V and W be completely reducible representations, and set $V = \bigoplus V_i^{\oplus n_i}$, $W = \bigoplus V_i^{\oplus m_i}$ for integers $n_i, m_i \ge 0$, where each V_i is an irrep. Then

$$\text{Hom}_{\text{rep}}(V, W) \cong \bigoplus_i \text{Mat}_{m_i \times n_i}(k)$$

meaning that an intertwining operator $T \colon V \to W$ amounts to, for each i, an $m_i \times n_i$ matrix of constants which gives a map $V_i^{\oplus n_i} \to V_i^{\oplus m_i}$.

Corollary 20.1.3 (Subrepresentations of completely reducible representations)
Let $V = \bigoplus V_i^{\oplus n_i}$ be completely reducible. Then any subrepresentation W of V is isomorphic to $\bigoplus V_i^{\oplus m_i}$ where $m_i \le n_i$ for each i, and the inclusion $W \hookrightarrow V$ is given by the direct sum of inclusion $V_i^{\oplus m_i} \hookrightarrow V_i^{\oplus n_i}$, which are $n_i \times m_i$ matrices.

Proof. Apply Schur's lemma to the inclusion $W \hookrightarrow V$. □

Recall from Section 9.5 that a linear maps from a n-dimensional vector space to a m-dimensional vector space can be written as a $n \times m$ matrix. Here the situation is similar, however the matrices are made for each irrep independently, and the non-isomorphic irreps, in some sense, "doesn't talk to each other".

Remark 20.1.4 — The representation $V^{\oplus n}$ can also be viewed as n vectors of V "stacked horizontally", as we did in Example 19.3.2:

$$\begin{pmatrix} \vdots & \vdots & & \vdots \\ v_1 & v_2 & \cdots & v_n \\ \vdots & \vdots & & \vdots \end{pmatrix} \in V^{\oplus n}.$$

That way, the action is given by

$$\begin{pmatrix} \vdots & \vdots & & \vdots \\ v_1 & v_2 & \cdots & v_n \\ \vdots & \vdots & & \vdots \end{pmatrix} \begin{bmatrix} c_{11} & c_{21} & \cdots & c_{(m-1)1} & c_{m1} \\ c_{12} & c_{22} & \cdots & c_{(m-1)1} & c_{m2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{1n} & c_{2n} & \cdots & c_{(m-1)n} & c_{mn} \end{bmatrix} \in V^{\oplus m}.$$

It may be clearer this way to see the type mismatch happening. And this also gives a natural explanation why the intertwining operators in Problem 19D* corresponds to right matrix multiplication.

§20.2 Density theorem

We are going to take advantage of the previous result to prove that finite-dimensional algebras have finitely many irreps.

Theorem 20.2.1 (Jacobson density theorem)

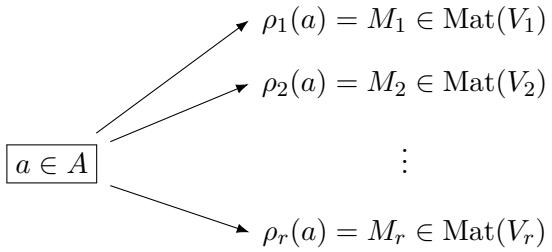
Let $(V_1, \rho_1), \ldots, (V_r, \rho_r)$ be pairwise nonisomorphic irreps of A. Then there is a surjective map of vector spaces

$$\bigoplus_{i=1}^{r} \rho_i \colon A \twoheadrightarrow \bigoplus_{i=1}^{r} \operatorname{Mat}(V_i).$$

The right way to think about this theorem is that

Density is the "Chinese remainder theorem" for irreps of A.

Recall that in number theory, the Chinese remainder theorem tells us that given lots of "unrelated" congruences, we can find a single N which simultaneously satisfies them all. Similarly, given lots of different nonisomorphic irreps of A, this means that we can select a single $a \in A$ which induces any tuple $(\rho_1(a), \ldots, \rho_r(a))$ of actions we want — a surprising result, since even the $r = 1$ case is not obvious at all!



This also gives us the non-obvious corollary:

Corollary 20.2.2 (Finiteness of number of representations)

Any finite-dimensional algebra A has at most $\dim A$ irreps.

Proof. If V_i are such irreps then $A \twoheadrightarrow \bigoplus_i V_i^{\oplus \dim V_i}$, hence we have the inequality $\sum(\dim V_i)^2 \le \dim A$. □

Proof of density theorem. Let $V = V_1 \oplus \cdots \oplus V_r$, so A acts on $V = (V, \rho)$ by $\rho = \bigoplus_i \rho_i$. Thus by Problem 19C*, we can instead consider ρ as an *intertwining operator*

$$\rho\colon \operatorname{Reg}(A) \to \bigoplus_{i=1}^{r} \operatorname{Mat}(V_i) \cong \bigoplus_{i=1}^{r} V_i^{\oplus d_i}.$$

We will use this instead as it will be easier to work with.

First, we handle the case $r = 1$. Fix a basis $e_1, \dots, e_n$ of $V = V_1$. Assume for contradiction that the map is not surjective. Then there is a map of representations (by ρ and the isomorphism) $\operatorname{Reg}(A) \to V^{\oplus n}$ given by $a \mapsto (a \cdot e_1, \dots, a \cdot e_n)$. By hypothesis, it is not surjective: its image is a *proper* subrepresentation of $V^{\oplus n}$. Assume its image is isomorphic to $V^{\oplus m}$ for $m < n$, so by Theorem 20.1.2 there is a matrix of constants X with

$$\begin{array}{ccccc} \operatorname{Reg}(A) & \longrightarrow & V^{\oplus n} & \xleftarrow{X\cdot -} & V^{\oplus m} \\ a & \longmapsto & (a\cdot e_1, \dots, a\cdot e_n) & & \\ 1_A & \longmapsto & (e_1, \dots, e_n) & \longmapsfrom & (v_1, \dots, v_m) \end{array}$$

where the two arrows in the top row have the same image; hence the pre-image $(v_1, \dots, v_m)$ of $(e_1, \dots, e_n)$ can be found. But since $m < n$ we can find constants $c_1, \dots, c_n$ not all zero such that X applied to the column vector $(c_1, \dots, c_n)$ is zero:

$$\sum_{i=1}^{n} c_i e_i = \begin{bmatrix} c_1 & \dots & c_n \end{bmatrix} \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} c_1 & \dots & c_n \end{bmatrix} X \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} = 0$$

contradicting the fact that e_i are linearly independent. Hence we conclude the theorem for $r = 1$.

As for $r \ge 2$, the image $\rho^{\text{img}}(A)$ is necessarily of the form $\bigoplus_i V_i^{\oplus d_i}$ (by Corollary 20.1.3) and by the above $d_i = \dim V_i$ for each i. □

Example 20.2.3 (Applying the proof of density theorem on an explicit example)

We can run through the argument on an explicit example to better understand how it works — in order to do this, we need V to be an irrep, otherwise the image of $\operatorname{Reg}(A)$ would not be isomorphic to $V^{\oplus m}$, and we will not be able to run to the end of the argument.

Let $A = \operatorname{Mat}_2(k)$, and $V \cong k^{\oplus 2}$ with the obvious action. As we know, this is an irrep.

The density theorem claims that $\rho\colon A \to \operatorname{Mat}(V)$ is surjective, which means for any $e_1, e_2 \in V$ independent, and any $w_1, w_2 \in V$, we can find $a \in A$ such that $a \cdot (e_1, e_2) = (w_1, w_2)$.

Because we're working through a counterexample, pick $e_1 = (1, 0)$, $e_2 = (2, 0)$ instead. Then, for some $w_1, w_2 \in V$, there may be no a that sends e_1 to w_1 to e_2 to w_2.

Consider the representation morphism $\operatorname{Reg}(A) \to V^{\oplus 2}$ by $a \mapsto (a \cdot e_1, a \cdot e_2)$; its image is thus $\{(v, 2v) \mid v \in V\}$, which is a subrepresentation of $V^{\oplus 2}$, isomorphic as

a representation to $V^{\oplus 1} \cong V$ by

$$v \mapsto (v, 2v) = v \begin{bmatrix} 1 & 2 \end{bmatrix}.$$

Then, we can find $v = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in V^{\oplus 1}$, for which

$$\begin{pmatrix} e_1 & e_2 \end{pmatrix} = v \begin{bmatrix} 1 & 2 \end{bmatrix}.$$

Now, with the explicit array of numbers $\begin{bmatrix} 1 & 2 \end{bmatrix}$, it is easy to find a linear dependence on e_1 and e_2.

§20.3 Semisimple algebras

Definition 20.3.1. A finite-dimensional algebra A is **semisimple** if every finite-dimensional representation of A is completely reducible.

Theorem 20.3.2 (Semisimple algebras)
Let A be a finite-dimensional algebra. Then the following are equivalent:

(i) $A \cong \bigoplus_i \operatorname{Mat}_{d_i}(k)$ for some d_i.

(ii) A is semisimple.

(iii) $\operatorname{Reg}(A)$ is completely reducible.

Proof. (i) $\implies$ (ii) follows from using Proposition 19.3.5 to breaks any finite-dimensional representation of A into a direct sum of representations of $\operatorname{Mat}_{d_i}(k)$, then Theorem 19.6.1 shows any such representations are completely reducible. (ii) $\implies$ (iii) is tautological.

To see (iii) $\implies$ (i), we use the following clever trick. Consider

$$\operatorname{Hom}_{\text{rep}}(\operatorname{Reg}(A), \operatorname{Reg}(A)).$$

On one hand, by Problem 19D*, it is isomorphic to A^{op} (A with opposite multiplication), because the only intertwining operators $\operatorname{Reg}(A) \to \operatorname{Reg}(A)$ are those of the form $- \cdot a$. On the other hand, suppose that we have set $\operatorname{Reg}(A) = \bigoplus_i V_i^{\oplus n_i}$. By Theorem 20.1.2, we have

$$A^{\text{op}} \cong \operatorname{Hom}_{\text{rep}}(\operatorname{Reg}(A), \operatorname{Reg}(A)) = \bigoplus_i \operatorname{Mat}_{n_i \times n_i}(k).$$

But $\operatorname{Mat}_n(k)^{\text{op}} \cong \operatorname{Mat}_n(k)$ (just by transposing), so we recover the desired conclusion. □

Remark 20.3.3 — The trick of the proof above resembles Cayley's theorem (Problem 1F†), in that we make the object act on itself to get an explicit representation.

Remark 20.3.4 — We can compare this to Corollary 18.3.2. Here, any finite-dimensional representation of A is a finite-dimensional left A-module, and from the theorem above, we know that if A is semisimple, any such module can be broken down into a direct sum of irreps $V_i \cong k^{\oplus d_i}$.

Note that unlike the case where A is a PID, $k^{\oplus d_i}$ is not isomorphic to a quotient of the ring $\mathrm{Mat}_{d_i}(k)$.

In fact, if we combine the above result with the density theorem (and Corollary 20.2.2), we obtain:

Theorem 20.3.5 (Sum of squares formula)
For a finite-dimensional algebra A we have

$$\sum_i \dim(V_i)^2 \le \dim A$$

where the V_i are the irreps of A; equality holds exactly when A is semisimple, in which case

$$\mathrm{Reg}(A) \cong \bigoplus_i \mathrm{Mat}(V_i) \cong \bigoplus_I V_i^{\oplus \dim V_i}.$$

Proof. The inequality was already mentioned in Corollary 20.2.2. It is equality if and only if the map $\rho\colon A \to \bigoplus_i \mathrm{Mat}(V_i)$ is an isomorphism; this means all V_i are present. □

Remark 20.3.6 (Digression) — For any finite-dimensional A, the kernel of the map $\rho\colon A \to \bigoplus_i \mathrm{Mat}(V_i)$ is denoted $\mathrm{Rad}(A)$ and is the so-called **Jacobson radical** of A; it's the set of all $a \in A$ which act by zero in all irreps of A. The usual definition of "semisimple" given in books is that this Jacobson radical is trivial.

§20.4 Maschke's theorem

We now prove that the representation theory of groups is as nice as possible.

Theorem 20.4.1 (Maschke's theorem)
Let G be a finite group, and k an algebraically closed field whose characteristic does not divide $|G|$. Then $k[G]$ is semisimple.

This tells us that when studying representations of groups, all representations are completely reducible.

Proof. Consider any finite-dimensional representation (V, ρ) of $k[G]$. Given a proper subrepresentation $W \subseteq V$, our goal is to construct a supplementary G-invariant subspace W' which satisfies

$$V = W \oplus W'.$$

This will show that indecomposable $\iff$ irreducible, which is enough to show $k[G]$ is semisimple.

Let $\pi\colon V \to W$ be any projection of V onto W, meaning $\pi(v) = v \iff v \in W$. We consider the *averaging* map $P\colon V \to V$ by

$$P = \frac{1}{|G|} \sum_{g \in G} \rho(g^{-1}) \circ \pi \circ \rho(g).$$

We'll use the following properties of the map:

Exercise 20.4.2. Show that the map P satisfies:

- For any $w \in W$, $P(w) = w$.
- For any $v \in V$, $P(v) \in W$.
- The map $P \colon V \to V$ is an intertwining operator.

Thus P is idempotent (it is the identity on its image W), so by Problem 9H* we have $V = \ker P \oplus \operatorname{im} P$, but both $\ker P$ and $\operatorname{im} P$ are subrepresentations as desired. □

Remark 20.4.3 — In the case where $k = \mathbb{C}$, there is a shorter proof. Suppose $B \colon V \times V \to \mathbb{C}$ is an arbitrary bilinear form. Then we can "average" it to obtain a new bilinear form

$$\langle v, w \rangle \coloneqq \frac{1}{|G|} \sum_{g \in G} B(g \cdot v, g \cdot w).$$

The averaged form $\langle -, - \rangle$ is G-invariant, in the sense that $\langle v, w \rangle = \langle g \cdot v, g \cdot w \rangle$. Then, one sees that if $W \subseteq V$ is a subrepresentation, so is its orthogonal complement $W^\perp$. This implies the result.

§20.5 Example: the representations of $\mathbb{C}[S_3]$

We compute all irreps of $\mathbb{C}[S_3]$. I'll take for granted right now there are exactly three such representations (which will be immediate by the first theorem in the next chapter: we'll in fact see that the number of representations of G is exactly equal to the number of conjugacy classes of G).

Given that, if the three representations of $\mathbb{C}[S_3]$ have dimension d_1, d_2, d_3 , then we ought to have

$$d_1^2 + d_2^2 + d_3^2 = |G| = 6.$$

From this, combined with some deep arithmetic, we deduce that we should have $d_1 = d_2 = 1$ and $d_3 = 2$ or some permutation.

In fact, we can describe these representations explicitly. First, we define:

Definition 20.5.1. Let G be a group. The complex **trivial group representation** of a group G is the one-dimensional representation $\mathbb{C}_{\text{triv}} = (\mathbb{C}, \rho)$ where $g \cdot v = v$ for all $g \in G$ and $v \in \mathbb{C}$ (i.e. $\rho(g) = \text{id}$ for all $g \in G$).

Remark 20.5.2 (Warning) **—** The trivial representation of an *algebra* A doesn't make sense for us: we might want to set $a \cdot v = v$ but this isn't linear in A. (You *could* try to force it to work by deleting the condition $1_A \cdot v = v$ from our definition; then one can just set $a \cdot v = 0$. But even then $\mathbb{C}_{\text{triv}}$ would not be the trivial representation of $k[G]$.)
Another way to see this is that the trivial representation depends on how the k-algebra is written as a group algebra: $k[\mathbb{Z}/2\mathbb{Z}]$ has a k-algebra automorphism given by $g \mapsto -g$, where g is the generator of the group $\mathbb{Z}/2\mathbb{Z}$; however the corresponding trivial representations are different.

Then the representations are:

- The one-dimensional $\mathbb{C}_{\text{triv}}$; each $\sigma \in S_3$ acts by the identity.

- There is a nontrivial one-dimensional representation $\mathbb{C}_{\text{sign}}$ where the map $S_3 \to \mathbb{C}^\times$ is given by sending σ to the sign of σ. Thus in $\mathbb{C}_{\text{sign}}$ every $\sigma \in S_3$ acts as ± 1. Of course, $\mathbb{C}_{\text{triv}}$ and $\mathbb{C}_{\text{sign}}$ are not isomorphic (as one-dimensional representations are never isomorphic unless the constants they act on coincide for all a, as we saw in Problem 19A†).
- Finally, we have already seen the two-dimensional representation, but now we give it a name. Define refl_0 to be the representation whose vector space is $\{(x,y,z) \mid x+y+z=0\}$, and whose action of S_3 on it is permutation of coordinates.

  > **Exercise 20.5.3.** Show that refl_0 is irreducible, for example by showing directly that no subspace is invariant under the action of S_3.

  Thus V is also not isomorphic to the previous two representations.

This implies that these are all the irreps of S_3. Note that, if we take the representation V of S_3 on $k^{\oplus 3}$, we just get that $V = \text{refl}_0 \oplus \mathbb{C}_{\text{triv}}$.

§20.6 A few harder problems to think about

Problem 20A. Find all the irreps of $\mathbb{C}[\mathbb{Z}/n\mathbb{Z}]$.

Problem 20B (Maschke requires $|G|$ finite)**.** Consider the representation of the group $\mathbb{R}$ on $\mathbb{C}^{\oplus 2}$ under addition by a homomorphism

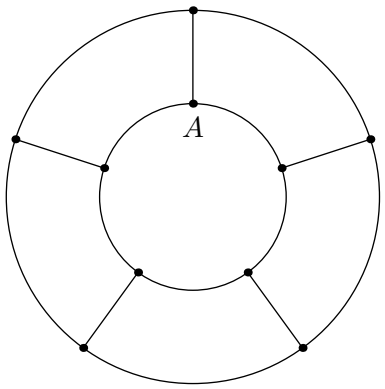
$$\mathbb{R} \to \text{Mat}_2(\mathbb{C}) \quad \text{by} \quad t \mapsto \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}.$$

Show that this representation is not irreducible, but it is indecomposable.

Problem 20C. Prove that all irreducible representations of a finite group are finite-dimensional.

Problem 20D. Determine all the complex irreps of D_{10}.

Problem 20E (AIME 2018)**.** The wheel shown below consists of two circles and five spokes, with a label where a spoke meets a circle. A bug walks along the wheel, starting from A. The bug takes 15 steps. At each step, the bug moves to an adjacent label such that it only walks counterclockwise along the inner circle and clockwise along the outer circle. In how many ways can the bug move to end up at A after all steps?



21 Characters

Characters are basically the best thing ever. To every representation V of A we will attach a so-called character $\chi_V \colon A \to k$. It will turn out that the characters of irreps of V will determine the representation V completely. Thus an irrep is just specified by a set of $\dim A$ numbers.

§21.1 Definitions

Definition 21.1.1. Let $V = (V, \rho)$ be a finite-dimensional representation of A. The **character** $\chi_V \colon A \to k$ attached to A is defined by $\chi_V = \operatorname{Tr} \circ \rho$, i.e.

$$\chi_V(a) \coloneqq \operatorname{Tr}(\rho(a) \colon V \to V).$$

Since Tr and ρ are additive, this is a k-linear map (but it is not multiplicative). Note also that $\chi_{V \oplus W} = \chi_V + \chi_W$ for any representations V and W.

We are especially interested in the case $A = k[G]$, of course. As usual, we just have to specify $\chi_V(g)$ for each $g \in G$ to get the whole map $k[G] \to k$. Thus we often think of χ_V as a function $G \to k$, called a character of the group G. Here is the case $G = S_3$:

Example 21.1.2 (Character table of S_3)

Let's consider the three irreps of $G = S_3$ from before. For $\mathbb{C}_{\text{triv}}$ all traces are 1; for $\mathbb{C}_{\text{sign}}$ the traces are ± 1 depending on sign (obviously, for one-dimensional maps $k \to k$ the trace "is" just the map itself). For refl_0 we take a basis $(1, 0, -1)$ and $(0, 1, -1)$, say, and compute the traces directly in this basis.

$\chi_V(g)$	id	(1 2)	(2 3)	(3 1)	(1 2 3)	(3 2 1)
$\mathbb{C}_{\text{triv}}$	1	1	1	1	1	1
$\mathbb{C}_{\text{sign}}$	1	-1	-1	-1	1	1
refl_0	2	0	0	0	-1	-1

The above table is called the **character table** of the group G. The table above has certain mysterious properties, which we will prove as the chapter progresses.

(I) The value of $\chi_V(g)$ only depends on the conjugacy class of g.

(II) The number of rows equals the number of conjugacy classes.

(III) The sum of the squares of any row is 6 again!

(IV) The "dot product" of any two rows is zero.

Abuse of Notation 21.1.3. The name "character" for $\chi_V \colon G \to k$ is a bit of a misnomer. This χ_V is not multiplicative in any way, as the above example shows: one can almost think of it as an element of $k^{\oplus |G|}$.

Question 21.1.4. Show that $\chi_V(1_A) = \dim V$, so one can read the dimensions of the representations from the leftmost column of a character table.

§21.2 The dual space modulo the commutator

For any algebra, we first observe that since $\operatorname{Tr}(TS) = \operatorname{Tr}(ST)$, we have for any V that

$$\chi_V(ab) = \chi_V(ba).$$

This explains observation (I) from earlier:

> **Question 21.2.1.** Deduce that if g and h are in the same conjugacy class of a group G, and V is a representation of $k[G]$, then $\chi(g) = \chi(h)$.

Now, given our algebra A we define the **commutator** $[A, A]$ to be the k-vector subspace spanned by $xy - yx$ for $x, y \in A$. Thus $[A, A]$ is contained in the kernel of each χ_V.

Definition 21.2.2. The space $A^{\text{ab}} \coloneqq A/[A, A]$ is called the **abelianization** of A. Each character of A can be viewed as a map $A^{\text{ab}} \to k$, i.e. an element of $(A^{\text{ab}})^\vee$.

Example 21.2.3 (Examples of abelianizations)

(a) If A is commutative, then $[A, A] = \{0\}$ and $A^{\text{ab}} = A$.

(b) If $A = \operatorname{Mat}_k(d)$, then $[A, A]$ consists exactly of the $d \times d$ matrices of trace zero. (Proof: harmless exercise.) Consequently, A^{ab} is one-dimensional.

(c) Suppose $A = k[G]$. Then in A^{ab}, we identify gh and hg for each $g, h \in G$; equivalently $ghg^{-1} = h$. So in other words, A^{ab} is isomorphic to the space of k-linear combinations of the *conjugacy classes* of G.

Remark 21.2.4 (Warning) — For a group G, the abelianization of G is defined to be $G/[G, G]$, where $[G, G]$ is the subgroup generated by all the commutators.
When $A = k[G]$, the space A^{ab} is not isomorphic to the group algebra $k[G/[G, G]]$! This is because, in the abelianization of the group G, we identify $ghg^{-1}h^{-1} = 1$ for all $g, h \in G$, which is not the same as $gh - hg$.
In fact, in the general case, A^{ab} does not even inherit the structure of a k-algebra from A, it can only get a k-vector space structure.

Theorem 21.2.5 (Character of representations of algebras)

Let A be an algebra over an algebraically closed field. Then

(a) Characters of pairwise non-isomorphic irreps are linearly independent in $(A^{\text{ab}})^\vee$.

(b) If A is finite-dimensional and semisimple, then the characters attached to irreps form a basis of $(A^{\text{ab}})^\vee$.

In particular, in (b) the number of irreps of A equals $\dim A^{\text{ab}}$.

Proof. Part (a) is more or less obvious by the density theorem: suppose there is a linear dependence, so that for every a we have

$$c_1\chi_{V_1}(a) + c_2\chi_{V_2}(a) + \cdots + c_r\chi_{V_r}(a) = 0$$

for some integer r.

Question 21.2.6. Deduce that $c_1 = \cdots = c_r = 0$ from the density theorem.

For part (b), assume there are r irreps. We may assume that

$$A = \bigoplus_{i=1}^{r} \operatorname{Mat}(V_i)$$

where $V_1, \ldots, V_r$ are the irreps of A. Since we have already showed the characters are linearly independent we need only show that $\dim(A/[A,A]) = r$, which follows from the observation earlier that each $\operatorname{Mat}(V_i)$ has a one-dimensional abelianization. □

Since G has $\dim k[G]^{\text{ab}}$ conjugacy classes, this completes the proof of (II).

§21.3 Orthogonality of characters

Now we specialize to the case of finite groups G, represented over $\mathbb{C}$.

Definition 21.3.1. Let $\text{Classes}(G)$ denote the set of conjugacy classes of G.

If G has r conjugacy classes, then it has r irreps. Each (finite-dimensional) representation V, irreducible or not, gives a character χ_V.

Abuse of Notation 21.3.2. From now on, we will often regard χ_V as a function $G \to \mathbb{C}$ or as a function $\text{Classes}(G) \to \mathbb{C}$. So for example, we will write both $\chi_V(g)$ (for $g \in G$) and $\chi_V(C)$ (for a conjugacy class C); the latter just means $\chi_V(g_C)$ for any representative $g_C \in C$.

Definition 21.3.3. Let $\text{Fun}_{\text{class}}(G)$ denote the set of functions $\text{Classes}(G) \to \mathbb{C}$ viewed as a vector space over $\mathbb{C}$. We endow it with the inner form

$$\langle f_1, f_2 \rangle = \frac{1}{|G|} \sum_{g \in G} f_1(g)\overline{f_2(g)}.$$

This is the same "dot product" that we mentioned at the beginning, when we looked at the character table of S_3. We now aim to prove the following orthogonality theorem, which will imply (III) and (IV) from earlier.

Theorem 21.3.4 (Orthogonality)
For any finite-dimensional complex representations V and W of G we have

$$\langle \chi_V, \chi_W \rangle = \dim \text{Hom}_{\text{rep}}(W, V).$$

In particular, if V and W are irreps then

$$\langle \chi_V, \chi_W \rangle = \begin{cases} 1 & V \cong W \\ 0 & \text{otherwise.} \end{cases}$$

Corollary 21.3.5 (Irreps give an orthonormal basis)
The characters associated to irreps form an *orthonormal* basis of $\text{Fun}_{\text{class}}(G)$.

In order to prove this theorem, we have to define the dual representation and the tensor representation, which give a natural way to deal with the quantity $\chi_V(g)\overline{\chi_W(g)}$.

Definition 21.3.6. Let $V = (V, \rho)$ be a representation of G. The **dual representation** $V^\vee$ is the representation on $V^\vee$ with the action of G given as follows: for each $\xi \in V^\vee$, the action of g gives a $g \cdot \xi \in V^\vee$ specified by

$$v \xmapsto{g\cdot\xi} \xi\left(\rho(g^{-1})(v)\right).$$

Definition 21.3.7. Let $V = (V, \rho_V)$ and $W = (W, \rho_W)$ be *group* representations of G. The **tensor product** of V and W is the group representation on $V \otimes W$ with the action of G given on pure tensors by

$$g \cdot (v \otimes w) = (\rho_V(g)(v)) \otimes (\rho_W(g)(w))$$

which extends linearly to define the action of G on all of $V \otimes W$.

Remark 21.3.8 — Warning: the definition for tensors does *not* extend to algebras. We might hope that $a \cdot (v \otimes w) = (a \cdot v) \otimes (a \cdot w)$ would work, but this is not even linear in $a \in A$ (what happens if we take $a = 2$, for example?).

Theorem 21.3.9 (Character traces)

If V and W are finite-dimensional representations of G, then for any $g \in G$.

(a) $\chi_{V\oplus W}(g) = \chi_V(g) + \chi_W(g)$.

(b) $\chi_{V\otimes W}(g) = \chi_V(g) \cdot \chi_W(g)$.

(c) $\chi_{V^\vee}(g) = \overline{\chi_V(g)}$.

Proof. Parts (a) and (b) follow from the identities $\operatorname{Tr}(S \oplus T) = \operatorname{Tr}(S) + \operatorname{Tr}(T)$ and $\operatorname{Tr}(S \otimes T) = \operatorname{Tr}(S)\operatorname{Tr}(T)$. However, part (c) is trickier. As $(\rho(g))^{|G|} = \rho(g^{|G|}) = \rho(1_G) = \mathrm{id}_V$ by Lagrange's theorem, we can diagonalize $\rho(g)$, say with eigenvalues $\lambda_1, \dots, \lambda_n$ which are $|G|$th roots of unity, corresponding to eigenvectors $e_1, \dots, e_n$. Then we see that in the basis $e_1^\vee, \dots, e_n^\vee$, the action of g on $V^\vee$ has eigenvalues $\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_n^{-1}$. So

$$\chi_V(g) = \sum_{i=1}^n \lambda_i \quad \text{and} \quad \chi_{V^\vee}(g) = \sum_{i=1}^n \lambda_i^{-1} = \sum_{i=1}^n \overline{\lambda_i}$$

where the last step follows from the identity $|z| = 1 \iff z^{-1} = \overline{z}$. □

Remark 21.3.10 (Warning) — The identities (b) and (c) do not extend linearly to $\mathbb{C}[G]$, i.e. it is not true for example that $\chi_{V^\vee}(a) = \overline{\chi_V(a)}$ if we think of χ_V as a map $\mathbb{C}[G] \to \mathbb{C}$.

Proof of orthogonality relation. The key point is that we can now reduce the sums of products to just a single character by

$$\chi_V(g)\overline{\chi_W(g)} = \chi_{V\otimes W^\vee}(g).$$

So we can rewrite the sum in question as just

$$\langle \chi_V, \chi_W \rangle = \frac{1}{|G|}\sum_{g\in G} \chi_{V\otimes W^\vee}(g) = \chi_{V\otimes W^\vee}\left(\frac{1}{|G|}\sum_{g\in G} g\right).$$

Let $P\colon V \otimes W^\vee \to V \otimes W^\vee$ be the action of $\frac{1}{|G|}\sum_{g\in G} g$, so we wish to find $\operatorname{Tr} P$.

Exercise 21.3.11. Show that P is idempotent. (Compute $P \circ P$ directly.)

Hence $V \otimes W^\vee = \ker P \oplus \operatorname{im} P$ (by Problem 9H*) and $\operatorname{im} P$ is the subspace of elements which are fixed under G. From this we deduce that

$$\operatorname{Tr} P = \dim \operatorname{im} P = \dim \left\{ x \in V \otimes W^\vee \mid g \cdot x = x \ \forall g \in G \right\}.$$

Now, consider the natural isomorphism $V \otimes W^\vee \to \operatorname{Hom}(W, V)$.

Exercise 21.3.12. Let $g \in G$. Show that under this isomorphism, $T \in \operatorname{Hom}(W, V)$ satisfies $g \cdot T = T$ if and only if $T(g \cdot w) = g \cdot T(w)$ for each $w \in W$. (This is just unwinding three or four definitions.)

Consequently, $\chi_{V \otimes W^\vee}(P) = \operatorname{Tr} P = \dim \operatorname{Hom}_{\text{rep}}(W, V)$ as desired. □

The orthogonality relation gives us a fast and mechanical way to check whether a finite-dimensional representation V is irreducible. Namely, compute the traces $\chi_V(g)$ for each $g \in G$, and then check whether $\langle \chi_V, \chi_V \rangle = 1$. So, for example, we could have seen the three representations of S_3 that we found were irreps directly from the character table. Thus, we can now efficiently verify any time we have a complete set of irreps.

§21.4 Examples of character tables

Example 21.4.1 (Dihedral group on 10 elements)
Let $D_{10} = \langle r, s \mid r^5 = s^2 = 1, rs = sr^{-1} \rangle$. Let $\omega = \exp(\frac{2\pi i}{5})$. We write four representations of D_{10}:

- $\mathbb{C}_{\text{triv}}$, all elements of D_{10} act as the identity.
- $\mathbb{C}_{\text{sign}}$, r acts as the identity while s acts by negation.
- V_1, which is two-dimensional and given by $r \mapsto \begin{bmatrix} \omega & 0 \\ 0 & \omega^4 \end{bmatrix}$ and $s \mapsto \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.
- V_2, which is two-dimensional and given by $r \mapsto \begin{bmatrix} \omega^2 & 0 \\ 0 & \omega^3 \end{bmatrix}$ and $s \mapsto \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

We claim that these four representations are irreducible and pairwise non-isomorphic. We do so by writing the character table:

D_{10}	1	r, r^4	r^2, r^3	sr^k
$\mathbb{C}_{\text{triv}}$	1	1	1	1
$\mathbb{C}_{\text{sign}}$	1	1	1	-1
V_1	2	$\omega + \omega^4$	$\omega^2 + \omega^3$	0
V_2	2	$\omega^2 + \omega^3$	$\omega + \omega^4$	0

Then a direct computation shows the orthogonality relations, hence we indeed have an orthonormal basis. For example, $\langle \mathbb{C}_{\text{triv}}, \mathbb{C}_{\text{sign}} \rangle = 1 + 2 \cdot 1 + 2 \cdot 1 + 5 \cdot (-1) = 0$.

Example 21.4.2 (Character table of S_4)

We now have enough machinery to compute the character table of S_4, which has five conjugacy classes (corresponding to cycle types id, 2, 3, 4 and $2+2$). First of all, we note that it has two one-dimensional representations, $\mathbb{C}_{\text{triv}}$ and $\mathbb{C}_{\text{sign}}$, and these are the only ones (because there are only two homomorphisms $S_4 \to \mathbb{C}^\times$). So thus far we have the table

S_4	1	$(\bullet\,\bullet)$	$(\bullet\,\bullet\,\bullet)$	$(\bullet\,\bullet\,\bullet\,\bullet)$	$(\bullet\,\bullet)(\bullet\,\bullet)$
$\mathbb{C}_{\text{triv}}$	1	1	1	1	1
$\mathbb{C}_{\text{sign}}$	1	-1	1	-1	1
$\vdots$			$\vdots$		

Note the columns represent $1+6+8+6+3 = 24$ elements.
Now, the remaining three representations have dimensions d_1, d_2, d_3 with

$$d_1^2 + d_2^2 + d_3^2 = 4! - 2 = 22$$

which has only $(d_1, d_2, d_3) = (2,3,3)$ and permutations. Now, we can take the refl_0 representation

$$\{(w,x,y,z) \mid w+x+y+z = 0\}$$

with basis $(1,0,0,-1)$, $(0,1,0,-1)$ and $(0,0,1,-1)$. This can be geometrically checked to be irreducible, but we can also do this numerically by computing the character directly (this is tedious): it comes out to have 3, 1, 0, -1, -1 which indeed gives norm

$$\langle \chi_{\text{refl}_0}, \chi_{\text{refl}_0} \rangle = \frac{1}{4!}\left(\underbrace{3^2}_{\text{id}} + \underbrace{6\cdot(1)^2}_{(\bullet\,\bullet)} + \underbrace{8\cdot(0)^2}_{(\bullet\,\bullet\,\bullet)} + \underbrace{6\cdot(-1)^2}_{(\bullet\,\bullet\,\bullet\,\bullet)} + \underbrace{3\cdot(-1)^2}_{(\bullet\,\bullet)(\bullet\,\bullet)} \right) = 1.$$

Note that we can also tensor this with the sign representation, to get another irreducible representation (since $\mathbb{C}_{\text{sign}}$ has all traces ± 1, the norm doesn't change). Finally, we recover the final row using orthogonality (which we name $\mathbb{C}^2$, for lack of a better name); hence the completed table is as follows.

S_4	1	$(\bullet\,\bullet)$	$(\bullet\,\bullet\,\bullet)$	$(\bullet\,\bullet\,\bullet\,\bullet)$	$(\bullet\,\bullet)(\bullet\,\bullet)$
$\mathbb{C}_{\text{triv}}$	1	1	1	1	1
$\mathbb{C}_{\text{sign}}$	1	-1	1	-1	1
$\mathbb{C}^2$	2	0	-1	0	2
refl_0	3	1	0	-1	-1
$\text{refl}_0 \otimes \mathbb{C}_{\text{sign}}$	3	-1	0	1	-1

§21.5 A few harder problems to think about

Problem 21A† (Reading decompositions from characters)**.** Let W be a complex representation of a finite group G. Let $V_1, \dots, V_r$ be the complex irreps of G and set $n_i = \langle \chi_W, \chi_{V_i} \rangle$. Prove that each n_i is a non-negative integer and

$$W = \bigoplus_{i=1}^{r} V_i^{\oplus n_i}.$$

Problem 21B. Consider complex representations of $G = S_4$. The representation $\text{refl}_0 \otimes \text{refl}_0$ is 9-dimensional, so it is clearly reducible. Compute its decomposition in terms of the five irreducible representations.

Problem 21C (Tensoring by one-dimensional irreps)**.** Let V and W be irreps of G, with $\dim W = 1$. Show that $V \otimes W$ is irreducible.

Problem 21D (Quaternions)**.** Compute the character table of the quaternion group Q_8.

Problem 21E* (Second orthogonality formula)**.** Let g and h be elements of a finite group G, and let $V_1, \ldots, V_r$ be the irreps of G. Prove that

$$\sum_{i=1}^{r} \chi_{V_i}(g)\overline{\chi_{V_i}(h)} = \begin{cases} |C_G(g)| & \text{if } g \text{ and } h \text{ are conjugates} \\ 0 & \text{otherwise.} \end{cases}$$

Here, $C_G(g) = \{x \in G : xg = gx\}$ is the centralizer of g.

22 Some applications

With all this setup, we now take the time to develop some nice results which are of independent interest.

§22.1 Frobenius divisibility

Theorem 22.1.1 (Frobenius divisibility)
Let V be a complex irrep of a finite group G. Then $\dim V$ divides $|G|$.

The proof of this will require algebraic integers (developed in the algebraic number theory chapter). Recall that an *algebraic integer* is a complex number which is the root of a monic polynomial with integer coefficients, and that these algebraic integers form a ring $\overline{\mathbb{Z}}$ under addition and multiplication, and that $\overline{\mathbb{Z}} \cap \mathbb{Q} = \mathbb{Z}$.

First, we prove:

Lemma 22.1.2 (Elements of $\mathbb{Z}[G]$ are integral)
Let $\alpha \in \mathbb{Z}[G]$. Then there exists a monic polynomial P with integer coefficients such that $P(\alpha) = 0$.

Proof. Let A_k be the $\mathbb{Z}$-span of $1, \alpha^1, \dots, \alpha^k$. Since $\mathbb{Z}[G]$ is Noetherian, the inclusions $A_0 \subseteq A_1 \subseteq A_2 \subseteq \dots$ cannot all be strict, hence $A_k = A_{k+1}$ for some k, which means α^{k+1} can be expressed in terms of lower powers of α. □

Proof of Frobenius divisibility. Let $C_1, \dots, C_m$ denote the conjugacy classes of G. Then consider the rational number

$$\frac{|G|}{\dim V};$$

we will show it is an algebraic integer, which will prove the theorem. Observe that we can rewrite it as

$$\frac{|G|}{\dim V} = \frac{|G| \langle \chi_V, \chi_V \rangle}{\dim V} = \sum_{g \in G} \frac{\chi_V(g)\overline{\chi_V(g)}}{\dim V}.$$

We split the sum over conjugacy classes, so

$$\frac{|G|}{\dim V} = \sum_{i=1}^{m} \overline{\chi_V(C_i)} \cdot \frac{|C_i| \chi_V(C_i)}{\dim V}.$$

We claim that for every i,

$$\frac{|C_i| \chi_V(C_i)}{\dim V} = \frac{1}{\dim V} \operatorname{Tr} T_i$$

is an algebraic integer, where

$$T_i \coloneqq \rho \left(\sum_{h \in C_i} h \right).$$

To see this, note that T_i commutes with elements of G, and hence is an intertwining operator $T_i \colon V \to V$. Thus by Schur's lemma, $T_i = \lambda_i \cdot \mathrm{id}_V$ and $\operatorname{Tr} T = \lambda_i \dim V$. By Lemma 22.1.2, $\lambda_i \in \overline{\mathbb{Z}}$, as desired.

Now we are done, since $\overline{\chi_V(C_i)} \in \overline{\mathbb{Z}}$ too (it is the sum of conjugates of roots of unity), so $\frac{|G|}{\dim V}$ is the sum of products of algebraic integers, hence itself an algebraic integer. □

§22.2 Burnside's theorem

We now prove a group-theoretic result. This is the famous poster child for representation theory (in the same way that RSA is the poster child of number theory) because the result is purely group theoretic.

Recall that a group is **simple** if it has no normal subgroups. In fact, we will prove:

Theorem 22.2.1 (Burnside)
Let G be a nonabelian group of order $p^a q^b$ (where p, q are distinct primes and $a, b \geq 0$). Then G is not simple.

In what follows p and q will always denote prime numbers.

Lemma 22.2.2 (On $\gcd(|C|, \dim V) = 1$)
Let $V = (V, \rho)$ be an complex irrep of G. Assume C is a conjugacy class of G with $\gcd(|C|, \dim V) = 1$. Then for any $g \in C$, either

- $\rho(g)$ is multiplication by a scalar, or
- $\chi_V(g) = \operatorname{Tr} \rho(g) = 0$.

Proof. If ε_i are the n eigenvalues of $\rho(g)$ (which are roots of unity), then from the proof of Frobenius divisibility we know $\frac{|C|}{n}\chi_V(g) \in \overline{\mathbb{Z}}$, thus from $\gcd(|C|, n) = 1$ we get

$$\frac{1}{n}\chi_V(g) = \frac{1}{n}(\varepsilon_1 + \cdots + \varepsilon_n) \in \overline{\mathbb{Z}}.$$

So this follows readily from a fact from algebraic number theory, namely Problem 53C⋆: either $\varepsilon_1 = \cdots = \varepsilon_n$ (first case) or $\varepsilon_1 + \cdots + \varepsilon_n = 0$ (second case). □

Lemma 22.2.3 (Simple groups don't have prime power conjugacy classes)
Let G be a finite simple group. Then G cannot have a conjugacy class of order p^k (where $k > 0$).

Proof. By contradiction. Assume C is such a conjugacy class, and fix any $g \in C$. By the second orthogonality formula (Problem 21E⋆) applied g and 1_G (which are not conjugate since $g \neq 1_G$) we have

$$\sum_{i=1}^{r} \dim V_i \chi_{V_i}(g) = 0$$

where V_i are as usual all irreps of G.

Exercise 22.2.4. Show that there exists a nontrivial irrep V such that $p \nmid \dim V$ and $\chi_V(g) \neq 0$. (Proceed by contradiction to show that $-\frac{1}{p} \in \overline{\mathbb{Z}}$ if not.)

Let $V = (V, \rho)$ be the irrep mentioned. By the previous lemma, we now know that $\rho(g)$ acts as a scalar in V.

Now consider the subgroup

$$H = \left\langle ab^{-1} \mid a, b \in C \right\rangle \subseteq G.$$

We claim this is a nontrivial normal subgroup of G. It is easy to check H is normal, and since $|C| > 1$ we have that H is nontrivial. As represented by V each element of H acts trivially in G, so since V is nontrivial and irreducible, $H \neq G$. This contradicts the assumption that G was simple. □

With this lemma, Burnside's theorem follows by partitioning the $|G|$ elements of our group into conjugacy classes. Assume for contradiction G is simple. Each conjugacy class must have order either 1 (of which there are $|Z(G)|$ by Problem 16D*) or divisible by pq (by the previous lemma), but on the other hand the sum equals $|G| = p^a q^b$. Consequently, we must have $|Z(G)| > 1$. But G is not abelian, hence $Z(G) \neq G$, thus the center $Z(G)$ is a nontrivial normal subgroup, contradicting the assumption that G was simple.

§22.3 Frobenius determinant

We finish with the following result, the problem that started the branch of representation theory. Given a finite group G, we create n variables $\{x_g\}_{g \in G}$, and an $n \times n$ matrix M_G whose (g, h)th entry is x_{gh}.

Example 22.3.1 (Frobenius determinants)

(a) If $G = \mathbb{Z}/2\mathbb{Z} = \langle T \mid T^2 = 1 \rangle$ then the matrix would be

$$M_G = \begin{bmatrix} x_{\text{id}} & x_T \\ x_T & x_{\text{id}} \end{bmatrix}.$$

Then $\det M_G = (x_{\text{id}} - x_T)(x_{\text{id}} + x_T)$.

(b) If $G = S_3$, a long computation gives the irreducible factorization of $\det M_G$ is

$$\left(\sum_{\sigma \in S_3} x_\sigma \right) \left(\sum_{\sigma \in S_3} \text{sign}(\sigma) x_\sigma \right) \left(F\left(x_{\text{id}}, x_{(123)}, x_{(321)}\right) - F\left(x_{(12)}, x_{(23)}, x_{(31)}\right) \right)^2$$

where $F(a, b, c) = a^2 + b^2 + c^2 - ab - bc - ca$; the latter factor is irreducible.

Theorem 22.3.2 (Frobenius determinant)

The polynomial $\det M_G$ (in $|G|$ variables) factors into a product of irreducible polynomials such that

(i) The number of polynomials equals the number of conjugacy classes of G, and

(ii) The multiplicity of each polynomial equals its degree.

You may already be able to guess how the "sum of squares" result is related! (Indeed, look at $\deg \det M_G$.)

Legend has it that Dedekind observed this behavior first in 1896. He didn't know how to prove it in general, so he sent it in a letter to Frobenius, who created representation theory to solve the problem.

With all the tools we've built, it is now fairly straightforward to prove the result.

Proof. Let $V = (V, \rho) = \operatorname{Reg}(\mathbb{C}[G])$ and let $V_1, \dots, V_r$ be the irreps of G. Let's consider the map $T \colon \mathbb{C}[G] \to \mathbb{C}[G]$ which has matrix M_G in the usual basis of $\mathbb{C}[G]$, namely

$$T \colon T(\{x_g\}_{g \in G}) = \sum_{g \in G} x_g \rho(g) \in \operatorname{Mat}(V).$$

Thus we want to examine $\det T$.

But we know that $V = \bigoplus_{i=1}^{r} V_i^{\oplus \dim V_i}$ as before, and so breaking down T over its subspaces we know

$$\det T = \prod_{i=1}^{r} \left(\det(T \restriction_{V_i}) \right)^{\dim V_i}.$$

So we only have to show two things: the polynomials $\det T_{V_i}$ are irreducible, and they are pairwise different for different i.

Let $V_i = (V_i, \rho)$, and pick $k = \dim V_i$.

- *Irreducible*: By the density theorem, for any $M \in \operatorname{Mat}(V_i)$ there exists a *particular* choice of complex numbers $x_g \in G$ such that

  $$M = \sum_{g \in G} x_g \cdot \rho_i(g) = (T \restriction_{V_i})(\{x_g\}).$$

  View $\rho_i(g)$ as a $k \times k$ matrix with complex coefficients. Thus the "generic" $(T \restriction_{V_i})(\{x_g\})$, viewed as a matrix with polynomial entries, must have linearly independent entries (or there would be some matrix in $\operatorname{Mat}(V_i)$ that we can't achieve).

  Then, the assertion follows (by a linear variable change) from the simple fact that the polynomial $\det(y_{ij})_{1 \le i,j \le m}$ in m^2 variables is always irreducible.

- *Pairwise distinct*: We show that from $\det T|_{V_i}(\{x_g\})$ we can read off the character χ_{V_i}, which proves the claim. In fact

  > **Exercise 22.3.3.** Pick *any* basis for V_i. If $\dim V_i = k$, and $1_G \neq g \in G$, then
  >
  > $$\chi_{V_i}(g) \text{ is the coefficient of } x_g x_{1_G}^{k-1}.$$

  Thus, we are done. □

VII

Quantum Algorithms

Part VII: Contents

23 Quantum states and measurements

In this chapter we'll explain how to set up quantum states using linear algebra. This will allow me to talk about quantum *circuits* in the next chapter, which will set the stage for Shor's algorithm.

I won't do very much physics (read: none at all). That is, I'll only state what the physical reality is in terms of linear algebras, and defer the philosophy of why this is true to your neighborhood "Philosophy of Quantum Mechanics" class (which is a "social science" class at MIT!).

§23.1 Bra-ket notation

Physicists have their own notation for vectors: whereas I previously used something like v, e_1, and so on, in this chapter you'll see the infamous **bra-ket** notation: a vector will be denoted by $|\bullet\rangle$, where $\bullet$ is some variable name: unlike in math or Python, this can include numbers, symbols, Unicode characters, whatever you like. This is called a "ket". To pay a homage to physicists everywhere, we'll use this notation for this chapter too.

Abuse of Notation 23.1.1 (For this part, $\dim H < \infty$)**.** In this part on quantum computation, we'll use the word "Hilbert space" as defined earlier, but in fact all our Hilbert spaces will be finite-dimensional.

If $\dim H = n$, then its orthonormal basis elements are often denoted

$$|0\rangle, |1\rangle, \dots, |n-1\rangle$$

(instead of e_i) and a generic element of H denoted by

$$|\psi\rangle, |\phi\rangle, \dots$$

and various other Greek letters.

Now for any $|\psi\rangle \in H$, we can consider the canonical dual element in $H^\vee$ (since H has an inner form), which we denote by $\langle\psi|$ (a "bra"). For example, if $\dim H = 2$ then we can write

$$|\psi\rangle = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

in an orthonormal basis, in which case

$$\langle\psi| = \begin{bmatrix} \overline{\alpha} & \overline{\beta} \end{bmatrix}.$$

We even can write dot products succinctly in this notation: if $|\phi\rangle = \begin{bmatrix} \gamma \\ \delta \end{bmatrix}$, then the dot product of $\langle\psi|$ and $|\phi\rangle$ is given by

$$\langle\psi|\phi\rangle = \begin{bmatrix} \overline{\alpha} & \overline{\beta} \end{bmatrix} \begin{bmatrix} \gamma \\ \delta \end{bmatrix} = \overline{\alpha}\gamma + \overline{\beta}\delta.$$

So we will use the notation $\langle\psi|\phi\rangle$ instead of the more mathematical $\langle|\psi\rangle, |\phi\rangle\rangle$. In particular, the squared norm of $|\psi\rangle$ is just $\langle\psi|\psi\rangle$. Concretely, for $\dim H = 2$ we have $\langle\psi|\psi\rangle = |\alpha|^2 + |\beta|^2$.

§23.2 The state space

If you think that's weird, well, it gets worse.

In classical computation, a bit is either 0 or 1. More generally, we can think of a classical space of n possible states $0, \dots, n-1$. Thus in the classical situation, the space of possible states is just a discrete set with n elements.

In quantum computation, a **qubit** is instead any *complex linear combination* of 0 and 1. To be precise, consider the normed complex vector space

$$H = \mathbb{C}^{\oplus 2}$$

and denote the orthonormal basis elements by $|0\rangle$ and $|1\rangle$. Then a *qubit* is a nonzero element $|\psi\rangle \in H$, so that it can be written in the form

$$|\psi\rangle = \alpha |0\rangle + \beta |1\rangle$$

where α and β are not both zero. Typically, we normalize so that $|\psi\rangle$ has norm 1:

$$\langle \psi | \psi \rangle = 1 \iff |\alpha|^2 + |\beta|^2 = 1.$$

In particular, we can recover the "classical" situation with $|0\rangle \in H$ and $|1\rangle \in H$, but now we have some "intermediate" states, such as

$$\frac{1}{\sqrt{2}} \left(|0\rangle + |1\rangle \right).$$

Philosophically, what has happened is that:

> **Instead of allowing just the states $|0\rangle$ and $|1\rangle$, we allow any complex linear combination of them.**

More generally, if $\dim H = n$, then the possible states are nonzero elements

$$c_0 |0\rangle + c_1 |1\rangle + \cdots + c_{n-1} |n-1\rangle$$

which we usually normalize so that $|c_0|^2 + |c_1|^2 + \cdots + |c_{n-1}|^2 = 1$.

§23.3 Observations

Prototypical example for this section: id *corresponds to not making a measurement since all its eigenvalues are equal, but any operator with distinct eigenvalues will cause collapse.*

If you think that's weird, well, it gets worse. First, some linear algebra review (Definition 15.4.1):

Definition 23.3.1. Let V be a finite-dimensional inner product space. For a map $T \colon V \to V$, the following conditions are equivalent:

- $\langle Tx, y \rangle = \langle x, Ty \rangle$ for any $x, y \in V$.
- $T = T^\dagger$.

A map T satisfying these conditions is called **Hermitian**.

Question 23.3.2. Show that T is normal.

Thus, we know that T is diagonalizable with respect to the inner form, so for a suitable basis we can write it in an orthonormal basis as

$$T = \begin{bmatrix} \lambda_0 & 0 & \dots & 0 \\ 0 & \lambda_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_{n-1} \end{bmatrix}.$$

As we've said, this is fantastic: not only do we have a basis of eigenvectors, but the eigenvectors are pairwise orthogonal, and so they form an orthonormal basis of V.

Question 23.3.3. Show that all eigenvalues of T are real. ($T = T^\dagger$.)

Back to quantum computation. Suppose we have a state $|\psi\rangle \in H$, where $\dim H = 2$; we haven't distinguished a particular basis yet, so we just have a nonzero vector. Then the way observations work (and this is physics, so you'll have to take my word for it) is as follows:

Pick a Hermitian operator $T\colon H \to H$; then observations of T return eigenvalues of T.

To be precise:

- Pick a Hermitian operator $T\colon H \to H$, which is called the **observable**.
- Consider its eigenvalues $\lambda_0, \dots, \lambda_{n-1}$ and corresponding eigenvectors $|0\rangle_T, \dots, |n-1\rangle_T$. Tacitly we may assume that $|0\rangle_T, \dots, |n-1\rangle_T$ form an orthonormal basis of H. (The subscript T is here to distinguish the eigenvectors of T from the basis elements of H.)
- Write $|\psi\rangle$ in the orthonormal basis as

$$c_0 |0\rangle_T + c_1 |1\rangle_T + \cdots + c_{n-1} |n-1\rangle_T .$$

- Then the probability of observing λ_i is

$$\frac{|c_i|^2}{|c_0|^2 + \cdots + |c_{n-1}|^2}.$$

  This is called making an **observation along T**.

Note that in particular, for any nonzero constant c, $|\psi\rangle$ and $c|\psi\rangle$ are indistinguishable, which is why we like to normalize $|\psi\rangle$. But the queerest thing of all is what happens to $|\psi\rangle$: by measuring it, we actually destroy information. This behavior is called **quantum collapse**.

- Suppose for simplicity that we observe $|\psi\rangle$ with T and obtain an eigenvalue λ, and that $|i\rangle_T$ is the only eigenvector with this eigenvalue. Then, the state $|\psi\rangle$ *collapses* to just the state $c_i |i\rangle_T$: all the other information is destroyed. (In fact, we may as well say it collapses to $|i\rangle_T$, since again constant factors are not relevant.)

- More generally, if we observe λ, consider the generalized eigenspace H_λ (i.e. the span of eigenvectors with the same eigenvalue). Then the physical state $|\psi\rangle$ has been changed as well: it has now been projected onto the eigenspace H_λ. In still other words, after observation, the state collapses to
$$\sum_{\substack{0 \le i \le n \\ \lambda_i = \lambda}} c_i |i\rangle_T .$$

In other words,

When we make a measurement, the coefficients from different eigenspaces are destroyed.

Why does this happen? Beats me... physics (and hence real life) is weird. But anyways, an example.

Example 23.3.4 (Quantum measurement of a state $|\psi\rangle$)
Let $H = \mathbb{C}^{\oplus 2}$ with orthonormal basis $|0\rangle$ and $|1\rangle$ and consider the state
$$|\psi\rangle = \frac{i}{\sqrt{5}}|0\rangle + \frac{2}{\sqrt{5}}|1\rangle = \begin{bmatrix} i/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix} \in H.$$

(a) Let
$$T = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$
This has eigenvectors $|0\rangle = |0\rangle_T$ and $|1\rangle = |1\rangle_T$, with eigenvalues $+1$ and -1. So if we measure $|\psi\rangle$ to T, we get $+1$ with probability $1/5$ and -1 with probability $4/5$. After this measurement, the original state collapses to $|0\rangle$ if we measured $+1$, and $|1\rangle$ if we measured -1. So we never learn the original probabilities.

(b) Now consider $T = \mathrm{id}$, and arbitrarily pick two orthonormal eigenvectors $|0\rangle_T$, $|1\rangle_T$; thus $\psi = c_0|0\rangle_T + c_1|1\rangle_T$. Since all eigenvalues of T are $+1$, our measurement will always be $+1$ no matter what we do. But there is also no collapsing, because none of the coefficients get destroyed.

(c) Now consider
$$T = \begin{bmatrix} 0 & 7 \\ 7 & 0 \end{bmatrix}.$$
The two normalized eigenvectors are
$$|0\rangle_T = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad |1\rangle_T = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$
with eigenvalues $+7$ and -7 respectively. In this basis, we have
$$|\psi\rangle = \frac{2+i}{\sqrt{10}}|0\rangle_T + \frac{-2+i}{\sqrt{10}}|1\rangle_T .$$
So we get $+7$ with probability $\frac{1}{2}$ and -7 with probability $\frac{1}{2}$, and after the measurement, $|\psi\rangle$ collapses to one of $|0\rangle_T$ and $|1\rangle_T$.

Question 23.3.5. Suppose we measure $|\psi\rangle$ with T and get λ. What happens if we measure with T again?

For $H = \mathbb{C}^{\oplus 2}$ we can come up with more classes of examples using the so-called **Pauli matrices**. These are the three Hermitian matrices

$$\sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad \sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}.$$

These matrices are important because:

Question 23.3.6. Show that these three matrices, plus the identity matrix, form a basis for the set of Hermitian 2×2 matrices.

So the Pauli matrices are a natural choice of basis.[1]

Their normalized eigenvectors are

$$|\uparrow\rangle = |0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad |\downarrow\rangle = |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$|\rightarrow\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad |\leftarrow\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$|\otimes\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix} \qquad |\odot\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}$$

which we call "z-up", "z-down", "x-up", "x-down", "y-up", "y-down" respectively. (The eigenvalues are $+1$ for "up" and -1 for "down".) So, given a state $|\psi\rangle \in \mathbb{C}^{\oplus 2}$ we can make a measurement with respect to any of these three bases by using the corresponding Pauli matrix.

In light of this, the previous examples were (a) measuring along σ_z, (b) measuring along id, and (c) measuring along $7\sigma_x$.

Notice that if we are given a state $|\psi\rangle$, and are told in advance that it is either $|\rightarrow\rangle$ or $|\leftarrow\rangle$ (or any other orthogonal states) then we are in what is more or less a classical situation. Specifically, if we make a measurement along σ_x, then we find out which state that $|\psi\rangle$ was in (with 100% certainty), and the state does not undergo any collapse. Thus, orthogonal states are reliably distinguishable.

§23.4 Entanglement

Prototypical example for this section: Singlet state: spooky action at a distance.

If you think that's weird, well, it gets worse.

Qubits don't just act independently: they can talk to each other by means of a *tensor product*. Explicitly, consider

$$H = \mathbb{C}^{\oplus 2} \otimes \mathbb{C}^{\oplus 2}$$

endowed with the norm described in Problem 13D*. One should think of this as a qubit A in a space H_A along with a second qubit B in a different space H_B, which have been

[1] Well, natural due to physics reasons.

allowed to interact in some way, and $H = H_A \otimes H_B$ is the set of possible states of *both* qubits. Thus

$$|0\rangle_A \otimes |0\rangle_B, \quad |0\rangle_A \otimes |1\rangle_B, \quad |1\rangle_A \otimes |0\rangle_B, \quad |1\rangle_A \otimes |1\rangle_B$$

is an orthonormal basis of H; here $|i\rangle_A$ is the basis of the first $\mathbb{C}^{\oplus 2}$ while $|i\rangle_B$ is the basis of the second $\mathbb{C}^{\oplus 2}$, so these vectors should be thought of as "unrelated" just as with any tensor product. The pure tensors mean exactly what you want: for example $|0\rangle_A \otimes |1\rangle_B$ means "0 for qubit A and 1 for qubit B".

As before, a measurement of a state in H requires a Hermitian map $H \to H$. In particular, if we only want to measure the qubit B along M_B, we can use the operator

$$\mathrm{id}_A \otimes M_B.$$

The eigenvalues of this operator coincide with the ones for M_B, and the eigenspace for λ will be the $H_A \otimes (H_B)_\lambda$, so when we take the projection the A qubit will be unaffected.

This does what you would hope for pure tensors in H:

Example 23.4.1 (Two non-entangled qubits)

Suppose we have qubit A in the state $\frac{i}{\sqrt{5}}|0\rangle_A + \frac{2}{\sqrt{5}}|1\rangle_A$ and qubit B in the state $\frac{1}{\sqrt{2}}|0\rangle_B + \frac{1}{\sqrt{2}}|1\rangle_B$. So, the two qubits in tandem are represented by the pure tensor

$$|\psi\rangle = \left(\frac{i}{\sqrt{5}}|0\rangle_A + \frac{2}{\sqrt{5}}|1\rangle_A\right) \otimes \left(\frac{1}{\sqrt{2}}|0\rangle_B + \frac{1}{\sqrt{2}}|1\rangle_B\right).$$

Suppose we measure $|\psi\rangle$ along

$$M = \mathrm{id}_A \otimes \sigma_z^B.$$

The eigenspace decomposition is

- $+1$ for the span of $|0\rangle_A \otimes |0\rangle_B$ and $|1\rangle_A \otimes |0\rangle_B$, and
- -1 for the span of $|0\rangle_A \otimes |1\rangle_B$ and $|1\rangle_A \otimes |1\rangle_B$.

(We could have used other bases, like $|{\to}\rangle_A \otimes |0\rangle_B$ and $|{\leftarrow}\rangle_A \otimes |0\rangle_B$ for the first eigenspace, but it doesn't matter.) Expanding $|\psi\rangle$ in the four-element basis, we find that we'll get the first eigenspace with probability

$$\left|\frac{i}{\sqrt{10}}\right|^2 + \left|\frac{2}{\sqrt{10}}\right|^2 = \frac{1}{2}.$$

and the second eigenspace with probability $\frac{1}{2}$ as well. (Note how the coefficients for A don't do anything!) After the measurement, we destroy the coefficients of the other eigenspace; thus (after re-normalization) we obtain the collapsed state

$$\left(\frac{i}{\sqrt{5}}|0\rangle_A + \frac{2}{\sqrt{5}}|1\rangle_A\right) \otimes |0\rangle_B \qquad \text{or} \qquad \left(\frac{i}{\sqrt{5}}|0\rangle_A + \frac{2}{\sqrt{5}}|1\rangle_A\right) \otimes |1\rangle_B$$

again with 50% probability each.

So this model lets us more or less work with the two qubits independently: when we make the measurement, we just make sure to not touch the other qubit (which corresponds to the identity operator).

Exercise 23.4.2. Show that if $\mathrm{id}_A \otimes \sigma_x^B$ is applied to the $|\psi\rangle$ in this example, there is no collapse at all. What's the result of this measurement?

Since the $\otimes$ is getting cumbersome to write, we say:

Abuse of Notation 23.4.3. From now on $|0\rangle_A \otimes |0\rangle_B$ will be abbreviated to just $|00\rangle$, and similarly for $|01\rangle$, $|10\rangle$, $|11\rangle$.

Example 23.4.4 (Simultaneously measuring a general 2-Qubit state)
Consider a normalized state $|\psi\rangle$ in $H = \mathbb{C}^{\oplus 2} \otimes \mathbb{C}^{\oplus 2}$, say

$$|\psi\rangle = \alpha\,|00\rangle + \beta\,|01\rangle + \gamma\,|10\rangle + \delta\,|11\rangle\,.$$

We can make a measurement along the diagonal matrix $T\colon H \to H$ with

$$T(|00\rangle) = 0\,|00\rangle\,, \quad T(|01\rangle) = 1\,|01\rangle\,, \quad T(|10\rangle) = 2\,|10\rangle\,, \quad T(|11\rangle) = 3\,|11\rangle\,.$$

Thus we get each of the eigenvalues 0, 1, 2, 3 with probability $|\alpha|^2$, $|\beta|^2$, $|\gamma|^2$, $|\delta|^2$. So if we like we can make "simultaneous" measurements on two qubits in the same way that we make measurements on one qubit.

However, some states behave very weirdly.

Example 23.4.5 (The singlet state)
Consider the state

$$|\Psi_-\rangle = \frac{1}{\sqrt{2}}\,|01\rangle - \frac{1}{\sqrt{2}}\,|10\rangle$$

which is called the **singlet state**. One can see that $|\Psi_-\rangle$ is not a simple tensor, which means that it doesn't just consist of two qubits side by side: the qubits in H_A and H_B have become *entangled*.
Now, what happens if we measure just the qubit A? This corresponds to making the measurement

$$T = \sigma_z^A \otimes \mathrm{id}_B.$$

The eigenspace decomposition of T can be described as:

- The span of $|00\rangle$ and $|01\rangle$, with eigenvalue $+1$.
- The span of $|10\rangle$ and $|11\rangle$, with eigenvalue -1.

So one of two things will happen:

- With probability $\frac{1}{2}$, we measure $+1$ and the collapsed state is $|01\rangle$.
- With probability $\frac{1}{2}$, we measure -1 and the collapsed state is $|10\rangle$.

But now we see that measurement along A has told us what the state of the bit B is completely!

By solely looking at measurements on A, we learn B; this paradox is called *spooky action at a distance*, or in Einstein's tongue, **spukhafte Fernwirkung**. Thus,

In tensor products of Hilbert spaces, states which are not pure tensors correspond to "entangled" states.

What this really means is that the qubits cannot be described independently; the state of the system must be given as a whole. That's what entangled states mean: the qubits somehow depend on each other.

§23.5 A few harder problems to think about

Problem 23A. We measure $|\Psi_-\rangle$ by $\sigma_x^A \otimes \mathrm{id}_B$, and hence obtain either $+1$ or -1. Determine the state of qubit B from this measurement.

Problem 23B (Greenberger-Horne-Zeilinger paradox)**.** Consider the state in $(\mathbb{C}^{\oplus 2})^{\otimes 3}$

$$|\Psi\rangle_{\text{GHZ}} = \frac{1}{\sqrt{2}} \left(|0\rangle_A |0\rangle_B |0\rangle_C - |1\rangle_A |1\rangle_B |1\rangle_C \right).$$

Find the value of the measurements along each of

$$\sigma_y^A \otimes \sigma_y^B \otimes \sigma_x^C, \quad \sigma_y^A \otimes \sigma_x^B \otimes \sigma_y^C, \quad \sigma_x^A \otimes \sigma_y^B \otimes \sigma_y^C, \quad \sigma_x^A \otimes \sigma_x^B \otimes \sigma_x^C.$$

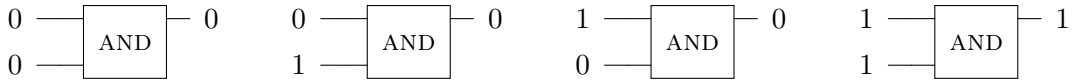
As for the paradox: what happens if you multiply all these measurements together?

24 Quantum circuits

Now that we've discussed qubits, we can talk about how to use them in circuits. The key change — and the reason that quantum circuits can do things that classical circuits cannot — is the fact that we are allowing linear combinations of 0 and 1.

§24.1 Classical logic gates

In classical logic, we build circuits which take in some bits for input, and output some more bits for input. These circuits are built out of individual logic gates. For example, the **AND gate** can be pictured as follows.



One can also represent the AND gate using the "truth table":

A	B	A and B
0	0	0
0	1	0
1	0	0
1	1	1

Similarly, we have the **OR gate** and the **NOT gate**:

A	B	A or B
0	0	0
0	1	1
1	0	1
1	1	1

A	not A
0	1
1	0

We also have a so-called **COPY gate**, which duplicates a bit.

Of course, the first theorem you learn about these gates is that:

Theorem 24.1.1 (AND, OR, NOT, COPY are universal)
The set of four gates AND, OR, NOT, COPY is universal in the sense that any boolean function $f\colon \{0,1\}^n \to \{0,1\}$ can be implemented as a circuit using only these gates.

Proof. Somewhat silly: we essentially write down a circuit that OR's across all input strings in $f^{\text{pre}}(1)$. For example, suppose we have $n = 3$ and want to simulate the function $f(abc)$ with $f(011) = f(110) = 1$ and 0 otherwise. Then the corresponding Boolean expression for f is simply

$$f(abc) = [(\text{not } a) \text{ and } b \text{ and } c] \text{ or } [a \text{ and } b \text{ and } (\text{not } c)].$$

Clearly, one can do the same for any other f, and implement this logic into a circuit. □

Remark 24.1.2 — Since x and $y = \text{not } ((\text{not } x) \text{ or } (\text{not } y))$, it follows that in fact, we can dispense with the AND gate.

§24.2 Reversible classical logic

Prototypical example for this section: CNOT gate, Toffoli gate.

For the purposes of quantum mechanics, this is not enough. To carry through the analogy we in fact need gates that are **reversible**, meaning the gates are bijections from the input space to the output space. In particular, such gates must take the same number of input and output gates.

Example 24.2.1 (Reversible gates)

(a) None of the gates AND, OR, COPY are reversible for dimension reasons.

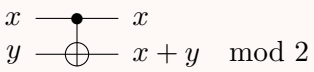
(b) The NOT gate, however, is reversible: it is a bijection $\{0,1\} \to \{0,1\}$.

Example 24.2.2 (The CNOT gate)

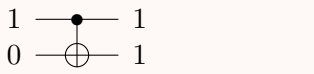
The controlled-NOT gate, or the **CNOT** gate, is a reversible 2-bit gate with the following truth table.

In		Out	
0	0	0	0
1	0	1	1
0	1	0	1
1	1	1	0

In other words, this gate XOR's the first bit to the second bit, while leaving the first bit unchanged. It is depicted as follows.



The first dot is called the "control", while the $\oplus$ is the "negation" operation: the first bit controls whether the second bit gets flipped or not. Thus, a typical application might be as follows.



So, NOT and CNOT are the only nontrivial reversible gates on two bits.

We now need a different definition of universal for our reversible gates.

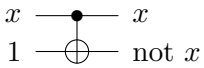
Definition 24.2.3. A set of reversible gates can **simulate** a Boolean function $f(x_1 \dots x_n)$, if one can implement a circuit which takes

- As input, $x_1 \dots x_n$ plus some fixed bits set to 0 or 1, called **ancilla bits**[1].
- As output, the input bits $x_1, \dots, x_n$, the output bit $f(x_1, \dots, x_n)$, and possibly some extra bits (called **garbage bits**).

[1]The English word "ancilla" means "maid".

The gate(s) are **universal** if they can simulate any Boolean function.

For example, the CNOT gate can simulate the NOT gate, using a single ancilla bit 1, according to the following circuit.



Unfortunately, it is not universal.

Proposition 24.2.4 (CNOT $\not\Rightarrow$ AND)

The CNOT gate cannot simulate the boolean function "x and y".

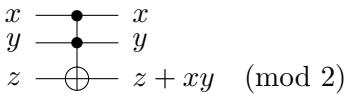
Sketch of Proof. One can see that any function simulated using only CNOT gates must be of the form

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n \pmod 2$$

because CNOT is the map $(x, y) \mapsto (x, x + y)$. Thus, even with ancilla bits, we can only create functions of the form $ax + by + c \pmod 2$ for fixed a, b, c. The AND gate is not of this form. □

So, we need at least a three-qubit gate. The most commonly used one is:

Definition 24.2.5. The three-bit **Toffoli gate**, also called the CCNOT gate, is given by



So the Toffoli has two controls, and toggles the last bit if and only if both of the control bits are 1.

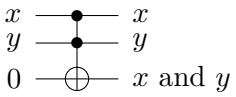
This replacement is sufficient.

Theorem 24.2.6 (Toffoli gate is universal)

The Toffoli gate is universal.

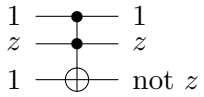
Proof. We will show it can *reversibly* simulate AND, NOT, hence OR, which we know is enough to show universality. (We don't need COPY because of reversibility.)

For the AND gate, we draw the circuit



with one ancilla bit, and no garbage bits.

For the NOT gate, we use two ancilla 1 bits and one garbage bit:



This completes the proof. □

Hence, in theory we can create any classical circuit we desire using the Toffoli gate alone. Of course, this could require exponentially many gates for even the simplest of functions. Fortunately, this is NO BIG DEAL because I'm a math major, and having 2^n gates is a problem best left for the CS majors.

§24.3 Quantum logic gates

In quantum mechanics, since we can have *linear combinations* of basis elements, our logic gates will instead consist of *linear maps*. Moreover, in quantum computation, gates are always reversible, which was why we took the time in the previous section to show that we can still simulate any function when restricted to reversible gates (e.g. using the Toffoli gate).

First, some linear algebra:

Definition 24.3.1. Let V be a finite dimensional inner product space. Then for a map $U \colon V \to V$, the following are equivalent:

- $\langle U(x), U(y) \rangle = \langle x, y \rangle$ for $x, y \in V$.
- $U^\dagger$ is the inverse of U.
- $\|x\| = \|U(x)\|$ for $x \in V$.

The map U is called **unitary** if it satisfies these equivalent conditions.

Then

Quantum logic gates are unitary matrices.

In particular, unlike the classical situation, quantum gates are always reversible (and hence they always take the same number of input and output bits).

For example, consider the CNOT gate. Its quantum analog should be a unitary map $U_{\text{CNOT}} \colon H \to H$, where $H = \mathbb{C}^{\oplus 2} \otimes \mathbb{C}^{\oplus 2}$, given on basis elements by

$$U_{\text{CNOT}}(|00\rangle) = |00\rangle, \quad U_{\text{CNOT}}(|01\rangle) = |01\rangle$$

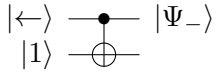
$$U_{\text{CNOT}}(|10\rangle) = |11\rangle, \quad U_{\text{CNOT}}(|11\rangle) = |10\rangle.$$

So pictorially, the quantum CNOT gate is given by

OK, so what? The whole point of quantum mechanics is that we allow linear qubits to be in linear combinations of $|0\rangle$ and $|1\rangle$, too, and this will produce interesting results. For example, let's take $|\leftarrow\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$ and plug it into the top, with $|1\rangle$ on the bottom, and see what happens:

$$U_{\text{CNOT}}\left(|\leftarrow\rangle \otimes |1\rangle\right) = U_{\text{CNOT}}\left(\frac{1}{\sqrt{2}}(|01\rangle - |11\rangle)\right) = \frac{1}{\sqrt{2}}\left(|01\rangle - |10\rangle\right) = |\Psi_-\rangle$$

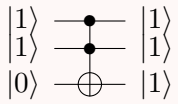
which is the fully entangled *singlet state*! Picture:



Thus, when we input mixed states into our quantum gates, the outputs are often entangled states, even when the original inputs are not entangled.

Example 24.3.2 (More examples of quantum gates)

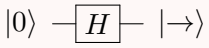
(a) Every reversible classical gate that we encountered before has a quantum analog obtained in the same way as CNOT: by specifying the values on basis elements. For example, there is a quantum Tofolli gate which for example sends



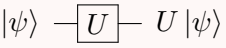
(b) The **Hadamard gate** on one qubit is a rotation given by

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Thus, it sends $|0\rangle$ to $|\rightarrow\rangle$ and $|1\rangle$ to $|\leftarrow\rangle$. Note that the Hadamard gate is its own inverse. It is depicted by an "H" box.



(c) More generally, if U is a 2×2 unitary matrix (i.e. a map $\mathbb{C}^{\oplus 2} \to \mathbb{C}^{\oplus 2}$) then there is **$U$-rotation gate** similar to the previous one, which applies U to the input.



For example, the classical NOT gate is represented by $U = \sigma_x$.

(d) A **controlled U-rotation gate** generalizes the CNOT gate. Let $U \colon \mathbb{C}^{\oplus 2} \to \mathbb{C}^{\oplus 2}$ be a rotation gate, and let $H = \mathbb{C}^{\oplus 2} \otimes \mathbb{C}^{\oplus 2}$ be a 2-qubit space. Then the controlled U gate has the following circuit diagrams.

Thus, U is applied when the controlling bit is 1, and CNOT is the special case $U = \sigma_x$. As before, we get interesting behavior if the control is mixed.

And now, some more counterintuitive quantum behavior. Suppose we try to use CNOT as a copy, with truth table.

In		Out	
0	0	0	0
1	0	1	1
0	1	0	1
1	1	1	0

The point of this gate is to be used with a garbage 0 at the bottom to try and simulate a "copy" operation. So indeed, one can check that

Thus we can copy $|0\rangle$ and $|1\rangle$. But as we've already seen if we input $|\leftarrow\rangle \otimes |0\rangle$ into U, we end up with the entangled state $|\Psi_-\rangle$ which is decisively *not* the $|\leftarrow\rangle \otimes |\leftarrow\rangle$ we wanted.

And in fact, the so-called **no-cloning theorem** implies that it's impossible to duplicate an arbitrary $|\psi\rangle$; the best we can do is copy specific orthogonal states as in the classical case. See also Problem 24B.

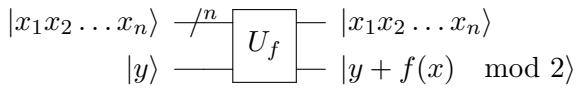
§24.4 Deutsch-Jozsa algorithm

The Deutsch-Jozsa algorithm is the first example of a nontrivial quantum algorithm which cannot be performed classically: it is a "proof of concept" that would later inspire Grover's search algorithm and Shor's factoring algorithm.

The problem is as follows: we're given a function $f\colon \{0,1\}^n \to \{0,1\}$, and promised that the function f is either

- A constant function, or
- A balanced function, meaning that exactly half the inputs map to 0 and half the inputs map to 1.

The function f is given in the form of a reversible black box U_f which is the control of a NOT gate, so it can be represented as the circuit diagram



i.e. if $f(x_1,\ldots,x_n)=0$ then the gate does nothing, otherwise the gate flips the y bit at the bottom. The slash with the n indicates that the top of the input really consists of n qubits, not just the one qubit drawn, and so the black box U_f is a map on $n+1$ qubits.

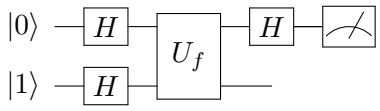
The problem is to determine, with as few calls to the black box U_f as possible, whether f is balanced or constant.

> **Question 24.4.1.** Classically, show that in the worst case we may need up to $2^{n-1}+1$ calls to the function f to answer the question.

So with only classical tools, it would take $O(2^n)$ queries to determine whether f is balanced or constant. However,

> **Theorem 24.4.2** (Deutsch-Jozsa)
> The Deutsch-Jozsa problem can be determined in a quantum circuit with only a single call to the black box.

Proof. For concreteness, we do the case $n=1$ explicitly; the general case is contained in Problem 24C. We claim that the necessary circuit is



Here the H's are Hadamard gates, and the meter at the end of the rightmost wire indicates that we make a measurement along the usual $|0\rangle$, $|1\rangle$ basis. This is not a typo! Even though classically the top wire is just a repeat of the input information, we are about to see that it's the top we want to measure.

Note that after the two Hadamard operations, the state we get is

$$\begin{aligned}|01\rangle &\xmapsto{H^{\otimes 2}} \left(\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\right) \otimes \left(\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)\right) \\ &= \frac{1}{2}\Big(|0\rangle \otimes (\,|0\rangle - |1\rangle\,) \;+\; |1\rangle \otimes (\,|0\rangle - |1\rangle\,)\Big).\end{aligned}$$

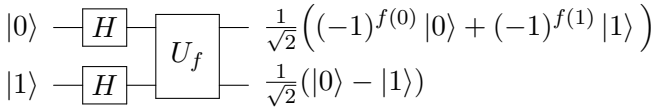
So after applying U_f, we obtain

$$\frac{1}{2}\Big(|0\rangle \otimes (\,|0 + f(0)\rangle - |1 + f(0)\rangle\,) \;+\; |1\rangle \otimes (\,|0 + f(1)\rangle - |1 + f(1)\rangle\,)\Big)$$

where the modulo 2 has been left implicit. Now, observe that the effect of going from $|0\rangle - |1\rangle$ to $|0 + f(x)\rangle - |1 + f(x)\rangle$ is merely to either keep the state the same (if $f(x) = 0$) or to negate it (if $f(x) = 1$). So we can simplify and factor to get

$$\frac{1}{2}\left((-1)^{f(0)}\,|0\rangle + (-1)^{f(1)}\,|1\rangle\right) \otimes (|0\rangle - |1\rangle)\,.$$

Thus, the picture so far is:



In particular, the resulting state is not entangled, and we can simply discard the last qubit (!). Now observe:

- If f is constant, then the upper-most state is $\pm|{\rightarrow}\rangle$.
- If f is balanced, then the upper-most state is $\pm|{\leftarrow}\rangle$.

So simply doing a measurement along σ_x will give us the answer. Equivalently, perform another H gate (so that $H|{\rightarrow}\rangle = |0\rangle$, $H|{\leftarrow}\rangle = |1\rangle$) and measuring along σ_z in the usual $|0\rangle$, $|1\rangle$ basis. Thus for $n = 1$ we only need a single call to the oracle. □

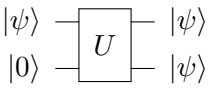
§24.5 A few harder problems to think about

Problem 24A (Fredkin gate)**.** The **Fredkin gate** (also called the controlled swap, or CSWAP gate) is the three-bit gate with the following truth table:

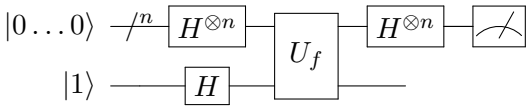
In			Out		
0	0	0	0	0	0
0	0	1	0	0	1
0	1	0	0	1	0
0	1	1	0	1	1
1	0	0	1	0	0
1	0	1	1	1	0
1	1	0	1	0	1
1	1	1	1	1	1

Thus the gate swaps the last two input bits whenever the first bit is 1. Show that this gate is also reversible and universal.

Problem 24B (Baby no-cloning theorem)**.** Show that there is no unitary map U on two qubits which sends $U(|\psi\rangle \otimes |0\rangle) = |\psi\rangle \otimes |\psi\rangle$ for any qubit $|\psi\rangle$, i.e. the following circuit diagram is impossible.



Problem 24C (Deutsch-Jozsa)**.** Given the black box U_f described in the Deutsch-Jozsa algorithm, consider the following circuit.



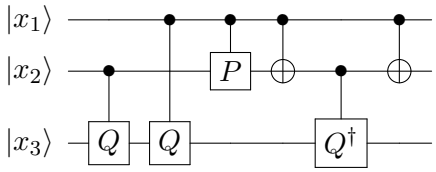
That is, take n copies of $|0\rangle$, apply the Hadamard rotation to all of them, apply U_f, reverse the Hadamard to all n input bits (again discarding the last bit), then measure all n bits in the $|0\rangle/|1\rangle$ basis (as in Example 23.4.4).

Show that the probability of measuring $|0\dots0\rangle$ is 1 if f is constant and 0 if f is balanced.

Problem 24D† (Barenco et al, 1995; arXiv:quant-ph/9503016v1)**.** Let

$$P = \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix} \qquad Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}$$

Verify that the quantum Toffoli gate can be implemented using just controlled rotations via the circuit



This was a big surprise to researchers when discovered, because classical reversible logic requires three-bit gates (e.g. Toffoli, Fredkind).

25 Shor's algorithm

OK, now for Shor's Algorithm: how to factor $M = pq$ in $O\left((\log M)^2\right)$ time.

This is arguably the reason agencies such as the US's National Security Agency have been diverting millions of dollars toward quantum computing.

§25.1 The classical (inverse) Fourier transform

The "crux move" in Shor's algorithm is the so-called quantum Fourier transform. The Fourier transform is used to extract *periodicity* in data, and it turns out the quantum analogue is a lot faster than the classical one.

Let me throw the definition at you first. Let N be a positive integer, and let $\omega_N = \exp\left(\frac{2\pi i}{N}\right)$.

Definition 25.1.1. Given a tuple of complex numbers

$$(x_0, x_1, \dots, x_{N-1})$$

its **discrete inverse Fourier transform** is the sequence $(y_0, y_1, \dots, y_{N-1})$ defined by

$$y_k = \frac{1}{N} \sum_{j=0}^{N-1} \omega_N^{jk} x_j.$$

Equivalently, one is applying the matrix

$$\frac{1}{N} \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega_N & \omega_N^2 & \dots & \omega_N^{N-1} \\ 1 & \omega_N^2 & \omega_N^4 & \dots & \omega_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_N^{N-1} & \omega_N^{2(N-1)} & \dots & \omega_N^{(N-1)^2} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{bmatrix}.$$

The reason this operation is important is because it lets us detect if the x_i are periodic. More generally, given a sequence of 1's appearing with period r, the amplitudes will peak at inputs which are divisible by $\frac{N}{\gcd(N,r)}$. Mathematically, we have that

$$x_k = \sum_{j=0}^{N-1} y_j \omega_N^{-jk}.$$

Example 25.1.2 (Example of discrete inverse Fourier transform)
Let $N = 6$, $\omega = \omega_6 = \exp(\frac{2\pi i}{6})$ and suppose $(x_0, x_1, x_2, x_3, x_4, x_5) = (0, 1, 0, 1, 0, 1)$

(hence x_i is periodic modulo 2). Thus,

$$\begin{aligned}
y_0 &= \tfrac{1}{6}\left(\omega^0 + \omega^0 + \omega^0\right) = 1/2 \\
y_1 &= \tfrac{1}{6}\left(\omega^1 + \omega^3 + \omega^5\right) = 0 \\
y_2 &= \tfrac{1}{6}\left(\omega^2 + \omega^6 + \omega^{10}\right) = 0 \\
y_3 &= \tfrac{1}{6}\left(\omega^3 + \omega^9 + \omega^{15}\right) = -1/2 \\
y_4 &= \tfrac{1}{6}\left(\omega^4 + \omega^{12} + \omega^{20}\right) = 0 \\
y_5 &= \tfrac{1}{6}\left(\omega^5 + \omega^{15} + \omega^{25}\right) = 0.
\end{aligned}$$

Thus, in the inverse transformation the "amplitudes" are all concentrated at multiples of 3; thus this reveals the periodicity of the original sequence by $\frac{N}{3} = 2$.

Remark 25.1.3 — The fact that this operation is called the "inverse" Fourier transform is mostly a historical accident (as my understanding goes). Confusingly, the corresponding quantum operation is the (not-inverted) Fourier transform.

If we apply the definition as written, computing the transform takes $O(N^2)$ time. It turns out that by a classical algorithm called the **fast Fourier transform** (whose details we won't discuss, but it effectively "reuses" calculations), one can reduce this to $O(N \log N)$ time. However, for Shor's algorithm this is also insufficient; we need something like $O\left((\log N)^2\right)$ instead. This is where the quantum Fourier transform comes in.

§25.2 The quantum Fourier transform

Note that to compute a Fourier transform, we need to multiply an $N \times N$ matrix with an N-vector, so this takes $O(N^2)$ multiplications. However, we are about to show that with a quantum computer, one can do this using $O((\log N)^2)$ quantum gates when $N = 2^n$, on a system with n qubits.

First, some more notation:

Abuse of Notation 25.2.1. In what follows, $|x\rangle$ will refer to $|x_n\rangle \otimes |x_{n-1}\rangle \otimes \cdots \otimes |x_1\rangle$ where $x = x_n x_{n-1} \ldots x_1$ in binary. For example, if $n = 3$ then $|6\rangle$ really means $|1\rangle \otimes |1\rangle \otimes |0\rangle$. Likewise, we refer to $0.x_1x_2\ldots x_n$ as binary.

Observe that the n-qubit space now has an orthonormal basis $|0\rangle$, $|1\rangle$, $\ldots$, $|N-1\rangle$

Definition 25.2.2. Consider an n-qubit state

$$|\psi\rangle = \sum_{k=0}^{N-1} x_k |k\rangle .$$

The **quantum Fourier transform** is defined by

$$U_{\text{QFT}}(|\psi\rangle) = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} \left(\sum_{k=0}^{N-1} \omega_N^{jk} x_k \right) |j\rangle .$$

In other words, using the basis $|0\rangle, \dots, |N-1\rangle$, U_{QFT} is given by the matrix

$$U_{\text{QFT}} = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega_N & \omega_N^2 & \dots & \omega_N^{N-1} \\ 1 & \omega_N^2 & \omega_N^4 & \dots & \omega_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_N^{N-1} & \omega_N^{2(N-1)} & \dots & \omega_N^{(N-1)^2} \end{bmatrix}$$

This is the exactly the same definition as before, except we have a $\sqrt{N}$ factor added so that U_{QFT} is unitary. But the trick is that in the quantum setup, the matrix can be rewritten:

Proposition 25.2.3 (Tensor representation)
Let $|x\rangle = |x_n x_{n-1} \dots x_1\rangle$. Then

$$\begin{aligned} U_{\text{QFT}}(|x_n x_{n-1} \dots x_1\rangle) = \frac{1}{\sqrt{N}} &(|0\rangle + \exp(2\pi i \cdot 0.x_1)\,|1\rangle) \\ &\otimes (|0\rangle + \exp(2\pi i \cdot 0.x_2 x_1)\,|1\rangle) \\ &\otimes \dots \\ &\otimes (|0\rangle + \exp(2\pi i \cdot 0.x_n \dots x_1)\,|1\rangle) \end{aligned}$$

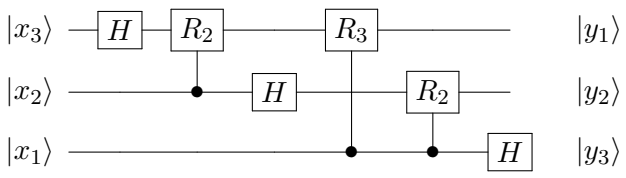
Proof. Direct (and quite annoying) computation. In short, expand everything. □

So by using mixed states, the quantum Fourier transform can use this "multiplication by tensor product" trick that isn't possible classically.

Now, without further ado, here's the circuit. Define the rotation matrices

$$R_k = \begin{bmatrix} 1 & 0 \\ 0 & \exp(2\pi i/2^k) \end{bmatrix}.$$

Then, for $n = 3$ the circuit is given by using controlled R_k's as follows:

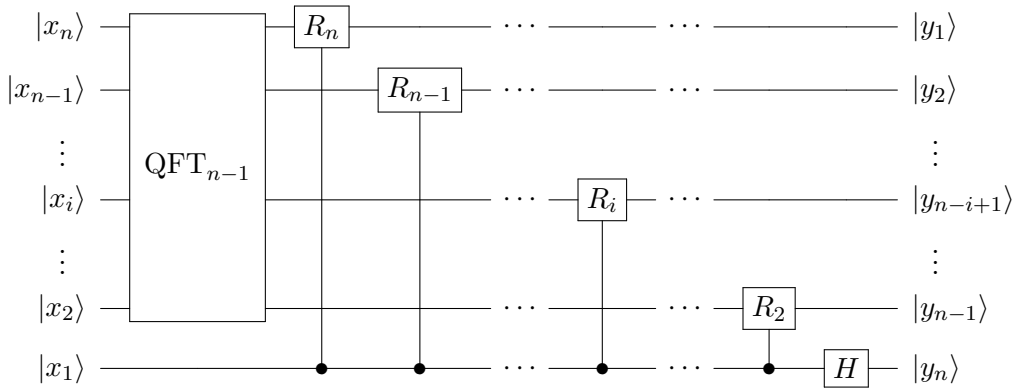


Exercise 25.2.4. Show that in this circuit, the image of $|x_3x_2x_1\rangle$ is

$$\Big(|0\rangle + \exp(2\pi i \cdot 0.x_1)\,|1\rangle\Big) \otimes \Big(|0\rangle + \exp(2\pi i \cdot 0.x_2x_1)\,|1\rangle\Big) \otimes \Big(|0\rangle + \exp(2\pi i \cdot 0.x_3x_2x_1)\,|1\rangle\Big)$$

as claimed.

For general n, we can write this as inductively as



Question 25.2.5. Convince yourself that when $n = 3$ the two circuits displayed are equivalent.

Thus, the quantum Fourier transform is achievable with $O(n^2)$ gates, which is enormously better than the $O(N \log N)$ operations achieved by the classical fast Fourier transform (where $N = 2^n$).

§25.3 Shor's algorithm

The quantum Fourier transform is the key piece of Shor's algorithm. Now that we have it, we can solve the factoring problem.

Let $p, q > 3$ be odd primes, and assume $p \neq q$. The main idea is to turn factoring an integer $M = pq$ into a problem about finding the order of $x \pmod M$; the latter is a "periodicity" problem that the quantum Fourier transform will let us solve. Specifically, say that an $x \pmod M$ is *good* if

(i) $\gcd(x, M) = 1$,

(ii) The order r of $x \pmod M$ is even, and

(iii) Factoring $0 \equiv (x^{r/2}-1)(x^{r/2}+1) \pmod M$, neither of the two factors is $0 \pmod M$. Thus one of them is divisible by p, and the other is divisible by q.

Exercise 25.3.1 (For contest number theory practice)**.** Show that for $M = pq$ at least half of the residues in $(\mathbb{Z}/M\mathbb{Z})^\times$ are good.

So if we can find the order of an arbitrary $x \in (\mathbb{Z}/M\mathbb{Z})^\times$, then we just keep picking x until we pick a good one (this happens more than half the time); once we do, we compute $\gcd(x^{r/2} - 1, M)$ using the Euclidean algorithm to extract one of the prime factors of M, and we're home free.

Now how do we do this? The idea is not so difficult: first we generate a sequence which is periodic modulo r.

Example 25.3.2 (Factoring 77: generating the periodic state)
Let's say we're trying to factor $M = 77$, and we randomly select $x = 2$, and want

to find its order r. Let $n = 13$ and $N = 2^{13}$, and start by initializing the state

$$|\psi\rangle = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} |k\rangle .$$

Now, build a circuit U_x (depending on x) which takes $|k\rangle |0\rangle$ to $|k\rangle |x^k \bmod M\rangle$. Applying this to $|\psi\rangle \otimes |0\rangle$ gives

$$U(|\psi\rangle |0\rangle) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} |k\rangle \otimes |x^k \bmod M\rangle .$$

Now suppose we measure the second qubit, and get a state of $|128\rangle$. That tells us that the collapsed state now, up to scaling, is

$$(|7\rangle + |7+r\rangle + |7+2r\rangle + \dots) \otimes |128\rangle .$$

The bottleneck is actually the circuit U_x; one can compute $x^k \pmod M$ by using repeated squaring, but it's still the clumsy part of the whole operation.

In general, the operation is:

- Pick a sufficiently large $N = 2^n$ (say, $N \geq M^2$).
- Generate $|\psi\rangle = \sum_{k=0}^{2^n-1} |k\rangle$.
- Build a circuit U_x which computes $|x^k \mod M\rangle$.
- Apply it to get a state $\frac{1}{\sqrt{N}} \sum_{k=0}^{2^n-1} |k\rangle \otimes |x^k \mod M\rangle$.
- Measure the second qubit to cause the first qubit to collapse to something which is periodic modulo r. Let $|\phi\rangle$ denote the left qubit.

Suppose we apply the quantum Fourier transform to the left qubit $|\phi\rangle$ now: since the left bit is periodic modulo r, we expect the transform will tell us what r is. Unfortunately, this doesn't quite work out, since N is a power of two, but we don't expect r to be.

Nevertheless, consider a state

$$|\phi\rangle = |k_0\rangle + |k_0 + r\rangle + \dots$$

so for example previously we had $k_0 = 7$ if we measured 128 on $x = 2$. Applying the quantum Fourier transform, we see that the coefficient of $|j\rangle$ in the transformed image is equal to

$$\omega_N^{k_0 j} \cdot \left(\omega_N^0 + \omega_N^{jr} + \omega_N^{2jr} + \omega_N^{3jr} + \dots \right)$$

As this is a sum of roots of unity, we realize we have destructive interference unless $\omega_N^{jr} = 1$ (since N is large). In other words, we approximately have

$$U_{\text{QFT}}(|\phi\rangle) \approx \sum_{\substack{0 \leq j < N \\ jr/N \in \mathbb{Z}}} |j\rangle$$

up to scaling as usual. The bottom line is that

If we measure $U_{\text{QFT}} |\phi\rangle$ we obtain a $|j\rangle$ such that $\frac{jr}{N}$ is close to an $s \in \mathbb{Z}$.

And thus given sufficient luck we can use continued fractions to extract the value of r.

Example 25.3.3 (Finishing the factoring of $M = 77$)

As before, we made an observation to the second qubit, and thus the first qubit collapses to the state $|\phi\rangle = |7\rangle + |7+r\rangle + \dots$. Now we make a measurement and obtain $j = 4642$, which means that for some integer s we have

$$\frac{4642r}{2^{13}} \approx s.$$

Now, we analyze the continued fraction of $\frac{4642}{2^{13}}$; we find the first few convergents are

$$0,\ 1,\ \frac{1}{2},\ \frac{4}{7},\ \frac{13}{23},\ \frac{17}{30},\ \frac{1152}{2033},\ \dots$$

So $\frac{17}{30}$ is a good approximation, hence we deduce $s = 17$ and $r = 30$ as candidates. And indeed, one can check that $r = 30$ is the desired order.

This won't work all the time[1] (for example, we could get unlucky and measure $j = 0$, i.e. $s = 0$, which would tell us no information at all).

But one can show that we succeed any time that

$$\gcd(s, r) = 1.$$

This happens at least $\frac{1}{\log r}$ of the time, and since $r < M$ this means that given sufficiently many trials, we will eventually extract the correct order r. This is Shor's algorithm.

[1] Not to mention the general issue of noise, but that's for engineers to worry about.

VIII

Calculus 101

Part VIII: Contents

26 Limits and series

Now that we have developed the theory of metric (and topological) spaces well, we give a three-chapter sequence which briskly covers the theory of single-variable calculus.

Much of the work has secretly already been done, For example, if x_n and y_n are real sequences with $\lim_n x_n = x$ and $\lim_n y_n = y$, then in fact $\lim_n (x_n + y_n) = x + y$ or $\lim_n (x_n y_n) = xy$, because we showed in Proposition 2.5.5 that arithmetic was continuous. We will also see that completeness plays a crucial role.

§26.1 Completeness and inf/sup

Prototypical example for this section: $\sup[0,1] = \sup(0,1) = 1$.

As $\mathbb{R}$ is a metric space, we may discuss continuity and convergence. There are two important facts about $\mathbb{R}$ which will make most of the following sections tick.

The first fact you have already seen before:

Theorem 26.1.1 ($\mathbb{R}$ is complete)
As a metric space, $\mathbb{R}$ is complete: sequences converge if and only if they are Cauchy.

The second one we have not seen before — it is the existence of inf and sup. Your intuition should be:

sup **is** max **adjusted slightly for infinite sets. (And** inf **is adjusted** min**.)**

Why the "adjustment"?

Example 26.1.2 (Why is max not good enough?)
Let's say we have the open interval $S = (0,1)$. The elements can get arbitrarily close to 1, so we would like to think "1 is the max of S"; except the issue is that $1 \notin S$. In general, infinite sets don't necessarily *have* a maximum, and we have to talk about bounds instead.
So we will define $\sup S$ in such a way that $\sup S = 1$. The definition is that "1 is the smallest number which is at least every element of S".

To write it out:

Definition 26.1.3. If S is a set of real numbers:

- An *upper bound* for S is a real number M such that $x \le M$ for all $x \in S$. If one exists, we say S is **bounded above**;
- A *lower bound* for S is a real number m such that $m \le x$ for all $x \in S$. If one exists, we say S is **bounded below**.
- If both upper and lower bounds exist, we say S is **bounded**.

Theorem 26.1.4 ($\mathbb{R}$ has inf's and sup's)

Let S be a nonempty set of real numbers.

- If S is bounded above then it has a *least* upper bound, which we denote by $\sup S$ and refer to as the **supremum** of S.
- If S is bounded below then it has a *greatest* lower bound, which we denote by $\inf S$ and refer to as the **infimum** of S.

Definition 26.1.5. For convenience, if S has not bounded above, we write $\sup S = +\infty$. Similarly, if S has not bounded below, we write $\inf S = -\infty$.

Example 26.1.6 (Supremums)

Since the examples for infimums are basically the same, we stick with supremums for now.

(a) If $S = \{1, 2, 3, \dots\}$ then S is not bounded above, so we have $\sup S = +\infty$.

(b) If $S = \{\dots, -2, -1\}$ denotes the set of negative integers, then $\sup S = -1$.

(c) Let $S = [0, 1]$ be a closed interval. Then $\sup S = 1$.

(d) Let $S = (0, 1)$ be an open interval. Then $\sup S = 1$ as well, even though 1 itself is not an element of S.

(e) Let $S = \mathbb{Q} \cap (0, 1)$ denote the set of rational numbers between 0 and 1. Then $\sup S = 1$ still.

(f) If S is a finite nonempty set, then $\sup S = \max S$.

Definition 26.1.7 (Porting definitions to sequences)**.** If $a_1, \dots$ is a sequence we will often write

$$\sup_n a_n \coloneqq \sup \{a_n \mid n \in \mathbb{N}\}$$
$$\inf_n a_n \coloneqq \inf \{a_n \mid n \in \mathbb{N}\}$$

for the supremum and infimum of the set of elements of the sequence. We also use the words "bounded above/below" for sequences in the same way.

Example 26.1.8 (Infimum of a sequence)

The sequence $a_n = \frac{1}{n}$ has infimum $\inf a_n = 0$.

§26.2 Proofs of the two key completeness properties of $\mathbb{R}$

Careful readers will note that we have not actually proven either Theorem 26.1.4 or Theorem 26.1.1. We will do so here.

First, we show that the ability to take infimums and supremums lets you prove completeness of $\mathbb{R}$.

Proof that Theorem 26.1.4 implies Theorem 26.1.1. Let $a_1, a_2, \dots$ be a Cauchy sequence. By discarding finitely many leading terms, we may as well assume that $|a_i - a_j| \le 100$ for all i and j. In particular, the sequence is now bounded; it lies between $[a_1 - 100, a_1 + 100]$ for example.

We want to show this sequence converges, so we have to first describe what the limit is. We know that to do this we are really going to have to use the fact that we live in $\mathbb{R}$. (For example we know in $\mathbb{Q}$ the limit of 1, 1.4, 1.41, 1.414, ... is nonexistent.)

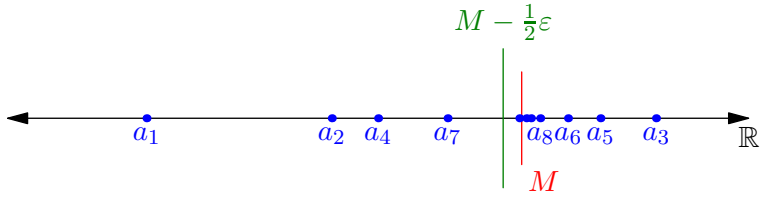
We propose the following: let

$$S = \{x \in \mathbb{R} \mid a_n \ge x \text{ for infinitely many } n\}.$$

We claim that the sequence converges to $M = \sup S$.

> **Exercise 26.2.1.** Show that this supremum makes sense by proving that $a_1 - 100 \in S$ (so S is nonempty) while all elements of S are at most $a_1 + 100$ (so S is bounded above). Thus we are allowed to actually take the supremum.

You can think of this set S with the following picture. We have a Cauchy sequence drawn in the real line which we think converges, which we can visualize as a bunch of dots on the real line, with some order on them. We wish to cut the line with a knife such that only finitely many dots are to the left of the knife. (For example, placing the knife all the way to the left always works.) The set S represents the places where we could put the knife, and M is "as far right" as we could go. Because of the way supremums work, M might not *itself* be a valid knife location, but certainly anything to its left is.



Let $\varepsilon > 0$ be given; we want to show eventually all terms are within ε of M. Because the sequence is Cauchy, there is an N such that eventually $|a_m - a_n| < \frac{1}{2}\varepsilon$ for $m \ge n \ge N$.

Now suppose we fix n and vary m. By the definition of M, it should be possible to pick the index m such that $a_m \ge M - \frac{1}{2}\varepsilon$ (there are infinitely many to choose from since $M - \frac{1}{2}\varepsilon$ is a valid knife location, and we only need $m \ge n$). In that case we have

$$|a_n - M| \le |a_n - a_m| + |a_m - M| < \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon$$

by the triangle inequality. This completes the proof. □

Therefore it is enough to prove the latter Theorem 26.1.4. To do this though, we would need to actually give a rigorous definition of the real numbers $\mathbb{R}$, since we have not done so yet!

One approach that makes this easy is to use the so-called **Dedekind cut** construction. Suppose we take the rational numbers $\mathbb{Q}$. Then one *defines* a real number to be a "cut" $A \mid B$ of the set of rational numbers: a pair of subsets of $\mathbb{Q}$ such that

- $\mathbb{Q} = A \sqcup B$ is a disjoint union;
- A and B are nonempty;
- we have $a < b$ for every $a \in A$ and $b \in B$, and

- A has no largest element (i.e. $\sup A \notin A$).

This can again be visualized by taking what you think of as the real line, and slicing at some real number. The subset $\mathbb{Q} \subset \mathbb{R}$ gets cut into two halves A and B. If the knife happens to land exactly at a rational number, by convention we consider that number to be in the right half (which explains the last fourth condition that $\sup A \notin A$).

With this definition Theorem 26.1.4 is easy: to take the supremum of a set of real numbers, we take the union of all the left halves. The hard part is then figuring out how to define $+$, $-$, $\times$, $\div$ and so on with this rather awkward construction. If you want to read more about this construction in detail, my favorite reference is [Pu02], in which all of this is done carefully in Chapter 1.

§26.3 Monotonic sequences

Here is a great exercise.

> **Exercise 26.3.1** (Mandatory)**.** Prove that if $a_1 \ge a_2 \ge \cdots \ge 0$ then the limit
>
> $$\lim_{n \to \infty} a_n$$
>
> exists. Hint: the idea in the proof of the previous section helps; you can also try to use completeness of $\mathbb{R}$. Second hint: if you are really stuck, wait until after Proposition 26.4.5, at which point you can use essentially copy its proof.

The proof here readily adapts by shifting.

Definition 26.3.2. A sequence a_n is **monotonic** if either $a_1 \ge a_2 \ge \dots$ or $a_1 \le a_2 \le \dots$.

Theorem 26.3.3 (Monotonic bounded sequences converge)
Let a_1, a_2, $\dots$ be a monotonic bounded sequence. Then $\lim_{n \to \infty} a_n$ exists.

Example 26.3.4 (Silly example of monotonicity)
Consider the sequence defined by

$$\begin{aligned} a_1 &= 1.2 \\ a_2 &= 1.24 \\ a_3 &= 1.248 \\ a_4 &= 1.24816 \\ a_5 &= 1.2481632 \\ &\vdots \end{aligned}$$

and so on, where in general we stuck on the decimal representation of the next power of 2. This will converge to *some* real number, although of course this number is quite unnatural and there is probably no good description for it.

In general, "infinite decimals" can now be defined as the limit of the truncated finite ones.

Example 26.3.5 ($0.9999\cdots = 1$)

In particular, I can finally make precise the notion you argued about in elementary school that

$$0.9999\cdots = 1.$$

We simply *define* a repeating decimal to be the limit of the sequence 0.9, 0.99, $0.999\ldots$. And it is obvious that the limit of this sequence is 1.

Some of you might be a little surprised since it seems like we really should have $0.9999 = 9\cdot 10^{-1} + 9\cdot 10^{-2} + \ldots$ — the limit of "partial sums". Don't worry, we're about to define those in just a moment.

Here is one other great use of monotonic sequences.

Definition 26.3.6. Let a_1, a_2, ... be a sequence (not necessarily monotonic) which is bounded below. We define

$$\limsup_{n\to\infty} a_n \coloneqq \lim_{N\to\infty} \sup_{n\geq N} a_n = \lim_{N\to\infty} \sup\left\{a_N, a_{N+1}, \ldots\right\}.$$

This is called the **limit supremum** of (a_n). We set $\limsup_{n\to\infty} a_n$ to be $+\infty$ if a_n is not bounded above.

If a_n is bounded above, the **limit infimum** $\liminf_{n\to\infty} a_n$ is defined similarly. In particular, $\liminf_{n\to\infty} a_n = -\infty$ if a_n is not bounded below.

Exercise 26.3.7. Show that these definitions make sense, by checking that the supremums are non-increasing, and bounded below.

We can think of $\limsup_n a_n$ as "supremum, but allowing finitely many terms to be discarded".

§26.4 Infinite series

Prototypical example for this section: $\sum_{k\geq 1}^{\infty} \frac{1}{k(k+1)} = \lim_{n\to\infty}\left(1 - \frac{1}{n+1}\right) = 1$.

We will actually begin by working with infinite series, since in the previous chapters we defined limits of sequences, and so this is actually the next closest thing to work with.[1]

This will give you a rigorous way to think about statements like

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

and help answer questions like "how can you add rational numbers and get an irrational one?".

[1] Conceptually: discrete things are easier to be rigorous about than continuous things, so series are actually "easier" than derivatives! I suspect the reason that most schools teach series last in calculus is that most calculus courses do not have proofs.

Definition 26.4.1. Consider a sequence $a_1, \dots$ of real numbers. The series $\sum_k a_k$ **converges** to a limit L if the sequence of "partial sums"

$$\begin{aligned} s_1 &= a_1 \\ s_2 &= a_1 + a_2 \\ s_3 &= a_1 + a_2 + a_3 \\ &\vdots \\ s_n &= a_1 + \cdots + a_n \end{aligned}$$

converges to the limit L. Otherwise it **diverges**.

Abuse of Notation 26.4.2 (Writing divergence as $+\infty$)**.** It is customary, if all the a_k are nonnegative, to write $\sum_k a_k = \infty$ to denote that the series diverges.

You will notice that by using the definition of sequences, we have masterfully sidestepped the issue of "adding infinitely many numbers" which would otherwise cause all sorts of problems.

An "infinite sum" is actually the *limit* of its partial sums. There is no infinite addition involved.

That's why it's for example okay to have $\sum_{n \ge 1} \frac{1}{n^2} = \frac{\pi^2}{6}$ be irrational; we have already seen many times that sequences of rational numbers can converge to irrational numbers. It also means we can gladly ignore all the irritating posts by middle schoolers about $1 + 2 + 3 + \cdots = -\frac{1}{12}$; the partial sums explode to $+\infty$, end of story, and if you want to assign a value to that sum it had better be a definition.

Example 26.4.3 (The classical telescoping series)
We can now prove the classic telescoping series

$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)}$$

in a way that doesn't just hand-wave the ending. Note that the kth partial sum is

$$\begin{aligned} \sum_{k=1}^{n} \frac{1}{k(k+1)} &= \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \cdots + \frac{1}{n(n+1)} \\ &= \left(\frac{1}{1} - \frac{1}{2} \right) + \cdots + \left(\frac{1}{n} - \frac{1}{n+1} \right) \\ &= 1 - \frac{1}{n+1}. \end{aligned}$$

The limit of this partial sum as $n \to \infty$ is 1.

Example 26.4.4 (Harmonic series diverges)
We can also make sense of the statement that $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$ (i.e. it diverges). We

may bound the 2^nth partial sums from below:

$$\begin{aligned}\sum_{k=1}^{2^n} \frac{1}{k} &= \frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{2^n} \\ &\ge \frac{1}{1} + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) \\ &\quad + \cdots + \underbrace{\left(\frac{1}{2^n} + \cdots + \frac{1}{2^n}\right)}_{2^{n-1} \text{ terms}} \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2} = 1 + \frac{n-1}{2}.\end{aligned}$$

A sequence satisfying $s_{2^n} \ge 1 + \frac{1}{2}(n-1)$ will never converge to a finite number!

I had better also mention that for nonnegative sums, convergence is just the same as having "finite sum" in the following sense.

Proposition 26.4.5 (Partial sums of nonnegatives bounded implies convergent)
Let $\sum_k a_k$ be a series of *nonnegative* real numbers. Then $\sum_k a_k$ converges to some limit if and only if there is a constant M such that

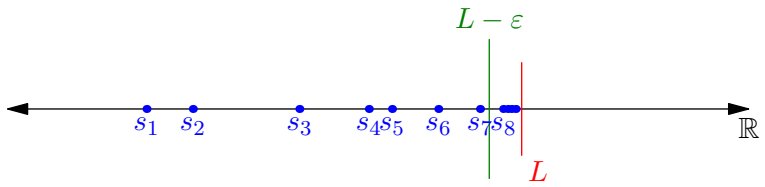
$$a_1 + \cdots + a_n < M$$

for every positive integer n.

Proof. This is actually just Theorem 26.3.3 in disguise, but since we left the proof as an exercise back then, we'll write it out this time.

Obviously if no such M exists then convergence will not happen, since this means the sequence s_n of partial sums is unbounded.

Conversely, if such M exists then we have $s_1 \le s_2 \le \cdots < M$. Then we contend the sequence s_n converges to $L := \sup_n s_n < \infty$. (If you read the proof that completeness implies Cauchy, the picture is nearly the same here, but simpler.)



Indeed, this means for any ε there are infinitely many terms of the sequence exceeding $L - \varepsilon$; but since the sequence is monotonic, once $s_n \ge L - \varepsilon$ then $s_{n'} \ge L - \varepsilon$ for all $n' \ge n$. This implies convergence. □

Abuse of Notation 26.4.6 (Writing $\sum < \infty$)**.** For this reason, if a_k are nonnegative real numbers, it is customary to write

$$\sum_k a_k < \infty$$

as a shorthand for "$\sum_k a_k$ converges to a finite limit", (or perhaps shorthand for "$\sum_k a_k$

is bounded" — as we have just proved these are equivalent). We will use this notation too.

§26.5 Series addition is not commutative: a horror story

One unfortunate property of the above definition is that it actually depends on the order of the elements. In fact, it turns out that there is an explicit way to describe when rearrangement is okay.

Definition 26.5.1. A series $\sum_k a_k$ of real numbers is said to **converge absolutely** if

$$\sum_k |a_k| < \infty$$

i.e. the series of absolute values converges to some limit. If the series converges, but not absolutely, we say it **converges conditionally**.

Proposition 26.5.2 (Absolute convergence $\implies$ convergence)
If a series $\sum_k a_k$ of real numbers converges absolutely, then it converges in the usual sense.

Exercise 26.5.3 (Great exercise)**.** Prove this by using the Cauchy criteria: show that if the partial sums of $\sum_k |a_k|$ are Cauchy, then so are the partial sums of $\sum_k a_k$.

Then, rearrangement works great.

Theorem 26.5.4 (Permutation of terms okay for absolute convergence)
Consider a series $\sum_k a_k$ which is absolutely convergent and has limit L. Then any permutation of the terms will also converge to L.

Proof. Suppose $\sum_k a_k$ converges to L, and b_n is a rearrangement. Let $\varepsilon > 0$. We will show that the partial sums of b_n are eventually within ε of L.

The hypothesis means that there is a large N in terms of ε such that

$$\left| \sum_{k=1}^{N} a_k - L \right| < \frac{1}{2}\varepsilon \quad \text{and} \quad \sum_{k=N+1}^{n} |a_k| < \frac{1}{2}\varepsilon$$

for every $n \geq N$ (the former from vanilla convergence of a_k and the latter from the fact that a_k converges absolutely, hence its partial sums are Cauchy).

Now suppose M is large enough that a_1, …, a_N are contained within the terms $\{b_1, \dots, b_M\}$. Then

$$\begin{aligned} b_1 + \cdots + b_M &= (a_1 + \cdots + a_N) \\ &\quad + \underbrace{a_{i_1} + a_{i_2} + \cdots + a_{i_{M-N}}}_{M-N \text{ terms with indices} > N} \end{aligned}$$

The terms in the first line sum up to within $\frac{1}{2}\varepsilon$ of L, and the terms in the second line have sum at most $\frac{1}{2}\varepsilon$ in absolute value, so the total $b_1 + \cdots + b_M$ is within $\frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon$ of L. □

In particular, when you have nonnegative terms, the world is great:

Nonnegative series can be rearranged at will.

And the good news is that actually, in practice, most of your sums will be nonnegative.

The converse is not true, and in fact, it is almost the worst possible converse you can imagine.

Theorem 26.5.5 (Riemann rearrangement theorem: Permutation of terms meaningless for conditional convergence)
Consider a series $\sum_k a_k$ which converges *conditionally* to some real number. Then, there exists a permutation of the series which converges conditionally to 1337. (Or any constant. You can also get it to diverge, too.)

So, permutation is as bad as possible for conditionally convergent series, and hence don't even bother to try.

§26.6 Limits of functions at points

Prototypical example for this section: $\lim_{x\to\infty} 1/x = 0$.

We had also better define the notion of a limit of a real function, which (surprisingly) we haven't actually defined yet. The definition will look like what we have seen before with continuity.

Definition 26.6.1. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function[2] and let $p \in \mathbb{R}$ be a point in the domain. Suppose there exists a real number L such that:

> For every $\varepsilon > 0$, there exists $\delta > 0$ such that if $|x - p| < \delta$ and $x \neq p$ then $|f(x) - L| < \varepsilon$.

Then we say L is the **limit** of f as $x \to p$, and write

$$\lim_{x\to p} f(x) = L.$$

There is an important point here: in this definition we *deliberately* require that $x \neq p$.

The value $\lim_{x\to p} f(x)$ does not depend on $f(p)$, and accordingly we often do not even bother to define $f(p)$.

Example 26.6.2 (Function with a hole)
Define the function $f\colon \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 3x & \text{if } x \neq 0 \\ 2019 & \text{otherwise.} \end{cases}$$

Then $\lim_{x\to 0} f(x) = 0$. The value $f(0) = 2019$ does not affect the limit. Obviously,

[2] Or $f\colon (a,b) \to \mathbb{R}$, or variants. We just need f to be defined on an open neighborhood of p.

because $f(0)$ was made up to be some artificial value that did not agree with the limit, this function is discontinuous at $x = 0$.

Question 26.6.3 (Mandatory)**.** Show that a function f is continuous at p if and only if $\lim_{x \to p} f(x)$ exists and equals $f(p)$.

Example 26.6.4 (Less trivial example: a rational piecewise function)
Define the function $f : \mathbb{R} \to \mathbb{R}$ as follows:

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ \frac{1}{q} & \text{if } x = \frac{p}{q} \text{ where } q > 0 \text{ and } \gcd(p,q) = 1 \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

For example, $f(\pi) = 0$, $f(2/3) = \frac{1}{3}$, $f(0.17) = \frac{1}{100}$. Then

$$\lim_{x \to 0} f(x) = 0.$$

For example, if $|x| < 1/100$ and $x \neq 0$ then $f(x)$ is either zero (for x irrational) or else is at most $\frac{1}{101}$ (if x is rational).
As $f(0) = 1$, this function is also discontinuous at $x = 0$. However, if we change the definition so that $f(0) = 0$ instead, then f becomes continuous at 0.

Example 26.6.5 (Famous example)
Let $f(x) = \frac{\sin x}{x}$, $f : \mathbb{R} \to \mathbb{R}$, where $f(0)$ is assigned any value. Then

$$\lim_{x \to 0} f(x) = 1.$$

We will not prove this here, since I don't want to get into trig yet. In general, I will basically only use trig functions for examples and not for any theory, so most properties of the trig functions will just be quoted.

Abuse of Notation 26.6.6 (The usual notation)**.** From now on, the above example will usually be abbreviated to just

$$\lim_{x \to 0} \frac{\sin x}{x} = 1.$$

The reason there is a slight abuse here is that I'm supposed to feed a function f into the limit, and instead I've written down an expression which is defined everywhere — except at $x = 0$. But that $f(0)$ value doesn't change anything. So the above means: "the limit of the function described by $f(x) = \frac{\sin x}{x}$, except $f(0)$ can be whatever it wants because it doesn't matter".

Remark 26.6.7 (For metric spaces) — You might be surprised that I didn't define the notion of $\lim_{x \to p} f(x)$ earlier for $f : M \to N$ a function on metric spaces. We can actually do so as above, but there is one nuance: what if our metric space M

is discrete, so p has no points nearby it? (Or even more simply, what if M is a one-point space?) We then cannot define $\lim_{x\to p} f(x)$ at all.
Thus if $f\colon M \to N$ and we want to define $\lim_{x\to p} f(x)$, we have the requirement that p should have a point within ε of it, for any $\varepsilon > 0$. In other words, p should not be an isolated point.

As usual, there are no surprises with arithmetic, we have $\lim_{x\to p}(f(x) \pm g(x)) = \lim_{x\to p} f(x) \pm \lim_{x\to p} g(x)$, and so on and so forth. We have effectively done this proof before so we won't repeat it again.

§26.7 Limits of functions at infinity

Annoyingly, we actually have to make this definition separately, even though it will not feel any different from earlier examples.

Definition 26.7.1. Let $f\colon \mathbb{R} \to \mathbb{R}$. Suppose there exists a real number L such that:

> For every $\varepsilon > 0$, there exists a constant M such that if $x > M$, then $|f(x) - L| < \varepsilon$.

Then we say L is the **limit** of f as x approaches ∞ and write

$$\lim_{x\to\infty} f(x) = L.$$

The limit $\lim_{x\to-\infty} f(x)$ is defined similarly, with $x > M$ replaced by $x < M$.

Fortunately, as ∞ is not an element of $\mathbb{R}$, we don't have to do the same antics about $f(\infty)$ like we had to do with "$f(p)$ set arbitrarily". So these examples can be more easily written down.

Example 26.7.2 (Limit at infinity)
The usual:

$$\lim_{x\to\infty} \frac{1}{x} = 0.$$

I'll even write out the proof: for any $\varepsilon > 0$, if $x > 1/\varepsilon$ then $\left|\frac{1}{x} - 0\right| < \varepsilon$.

There are no surprises with arithmetic: we have $\lim_{x\to\infty}(f(x)\pm g(x)) = \lim_{x\to\infty} f(x) \pm \lim_{x\to\infty} g(x)$, and so on and so forth. This is about the fourth time I've mentioned this, so I will not say more.

§26.8 A few harder problems to think about

Problem 26A. Define the sequence

$$a_n = (-1)^n + \frac{n^3}{2^n}$$

for every positive integer n. Compute the limit infimum and the limit supremum.

Problem 26B. For which bounded sequences a_n does $\liminf_n a_n = \limsup_n a_n$?

Problem 26C† (Comparison test)**.** Let $\sum a_n$ and $\sum b_n$ be two series. Assume $\sum b_n$ is absolutely convergent, and $|a_n| \le |b_n|$ for all integers n. Prove that $\sum_n a_n$ is absolutely convergent.

Problem 26D (Geometric series)**.** Let $-1 < r < 1$ be a real number. Show that the series

$$1 + r + r^2 + r^3 + \dots$$

converges absolutely and determine what it converges to.

Problem 26E (Alternating series test)**.** Let $a_0 \ge a_1 \ge a_2 \ge a_3 \ge \dots$ be a weakly decreasing sequence of nonnegative real numbers, and assume that $\lim_{n\to\infty} a_n = 0$. Show that the series $\sum_n (-1)^n a_n$ is convergent (it need not be absolutely convergent).

Problem 26F ([Pu02, Chapter 3, Exercise 55])**.** Let $(a_n)_{n\ge1}$ and $(b_n)_{n\ge1}$ be sequences of real numbers. Assume $a_1 \le a_2 \le \dots \le 1000$ and moreover that $\sum_n b_n$ converges. Prove that $\sum_n a_n b_n$ converges. (Note that in both the hypothesis and statement, we do not have absolute convergence.)

Problem 26G (Putnam 2016 B1)**.** Let $x_0, x_1, x_2, \dots$ be the sequence such that $x_0 = 1$ and for $n \ge 0$,

$$x_{n+1} = \log(e^{x_n} - x_n)$$

(as usual, log is the natural logarithm). Prove that the infinite series $x_0 + x_1 + \dots$ converges and determine its value.

Problem 26H. Consider again the function $f\colon \mathbb{R} \to \mathbb{R}$ in Example 26.6.4 defined by

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ \frac{1}{q} & \text{if } x = \frac{p}{q} \text{ where } q > 0 \text{ and } \gcd(p,q) = 1 \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

For every real number p, compute $\lim_{x\to p} f(x)$, if it exists. At which points is f continuous?

27 Bonus: A hint of p-adic numbers

This is a bonus chapter meant for those who have also read about **rings and fields**: it's a nice tidbit at the intersection of algebra and analysis.

In this chapter, we are going to redo most of the previous chapter with the absolute value $|-|$ replaced by the p-adic one. This will give us the p-adic integers $\mathbb{Z}_p$, and the p-adic numbers $\mathbb{Q}_p$. The one-sentence description is that these are "integers/rationals carrying full mod p^e information" (and only that information).

In everything that follows p is always assumed to denote a prime. The first four sections will cover the founding definitions culminating in a short solution to a USA TST problem. We will then state (mostly without proof) some more surprising results about continuous functions $f : \mathbb{Z}_p \to \mathbb{Q}_p$; finally we close with the famous proof of the Skolem-Mahler-Lech theorem using p-adic analysis.

§27.1 Motivation

Before really telling you what $\mathbb{Z}_p$ and $\mathbb{Q}_p$ are, let me tell you what you might expect them to do.

In elementary/olympiad number theory, we're already well-familiar with the following two ideas:

- Taking modulo a prime p or prime power p^e, and
- Looking at the exponent ν_p.

Let me expand on the first point. Suppose we have some Diophantine equation. In olympiad contexts, one can take an equation modulo p to gain something else to work with. Unfortunately, taking modulo p loses some information: the reduction $\mathbb{Z} \twoheadrightarrow \mathbb{Z}/p$ is far from injective.

If we want finer control, we could consider instead taking modulo p^2, rather than taking modulo p. This can also give some new information (cubes modulo 9, anyone?), but it has the disadvantage that $\mathbb{Z}/p^2$ isn't a field, so we lose a lot of the nice algebraic properties that we got if we take modulo p.

One of the goals of p-adic numbers is that we can get around these two issues I described. The p-adic numbers we introduce is going to have the following properties:

1. **You can "take modulo p^e for all e at once".** In olympiad contexts, we are used to picking a particular modulus and then seeing what happens if we take that modulus. But with p-adic numbers, we won't have to make that choice. An equation of p-adic numbers carries enough information to take modulo p^e.
2. **The numbers $\mathbb{Q}_p$ form a field**, the nicest possible algebraic structure: $1/p$ makes sense. Contrast this with $\mathbb{Z}/p^2$, which is not even an integral domain.
3. **It doesn't lose as much information** as taking modulo p does: rather than the surjective $\mathbb{Z} \twoheadrightarrow \mathbb{Z}/p$ we have an *injective* map $\mathbb{Z} \hookrightarrow \mathbb{Z}_p$.

4. **Despite this, you "ignore" some "irrelevant" data**. Just like taking modulo p, you want to zoom-in on a particular type of algebraic information, and this means necessarily losing sight of other things.[1]

So, you can think of p-adic numbers as the right tool to use if you only really care about modulo p^e information, but normal $\mathbb{Z}/p^e$ isn't quite powerful enough.

To be more concrete, I'll give a poster example now:

Example 27.1.1 (USA TST 2002/2)

For a prime p, show the value of

$$f_p(x) = \sum_{k=1}^{p-1} \frac{1}{(px+k)^2} \pmod{p^3}$$

does not depend on x.

Here is a problem where we *clearly* only care about p^e-type information. Yet it's a nontrivial challenge to do the necessary manipulations mod p^3 (try it!). The basic issue is that there is no good way to deal with the denominators modulo p^3 (in part $\mathbb{Z}/p^3$ is not even an integral domain).

However, with p-adic analysis we're going to be able to overcome these limitations and give a "straightforward" proof by using the identity

$$\left(1+\frac{px}{k}\right)^{-2} = \sum_{n\geq 0} \binom{-2}{n} \left(\frac{px}{k}\right)^n.$$

Such an identity makes no sense over $\mathbb{Q}$ or $\mathbb{R}$ for convergence reasons, but it will work fine over $\mathbb{Q}_p$, which is all we need.

§27.2 Algebraic perspective

Prototypical example for this section: $-1/2 = 1 + 3 + 3^2 + 3^3 + \cdots \in \mathbb{Z}_3$.

We now construct $\mathbb{Z}_p$ and $\mathbb{Q}_p$. I promised earlier that a p-adic integer will let you look at "all residues modulo p^e" at once. This definition will formalize this.

§27.2.i Definition of $\mathbb{Z}_p$

Definition 27.2.1 (Introducing $\mathbb{Z}_p$)**.** A **p-adic integer** is a sequence

$$x = (x_1 \bmod p,\ x_2 \bmod p^2,\ x_3 \bmod p^3,\ \ldots)$$

of residues x_e modulo p^e for each integer e, satisfying the compatibility relations $x_i \equiv x_j \pmod{p^i}$ for $i < j$.

The set $\mathbb{Z}_p$ of p-adic integers forms a ring under component-wise addition and multiplication.

[1] To draw an analogy: the equation $a^2+b^2+c^2+d^2 = -1$ has no integer solutions, because, well, squares are nonnegative. But you will find that this equation has solutions modulo any prime p, because once you take modulo p you stop being able to talk about numbers being nonnegative. The same thing will happen if we work in p-adics: the above equation has a solution in $\mathbb{Z}_p$ for every prime p.

Example 27.2.2 (Some 3-adic integers)

Let $p = 3$. Every usual integer n generates a (compatible) sequence of residues modulo p^e for each e, so we can view each ordinary integer as p-adic one:

$$50 = (2 \bmod 3,\ 5 \bmod 9,\ 23 \bmod 27,\ 50 \bmod 81,\ 50 \bmod 243,\ \dots).$$

On the other hand, there are sequences of residues which do not correspond to any usual integer despite satisfying compatibility relations, such as

$$(1 \bmod 3,\ 4 \bmod 9,\ 13 \bmod 27,\ 40 \bmod 81,\ \dots)$$

which can be thought of as $x = 1 + p + p^2 + \dots$.

In this way we get an injective map

$$\mathbb{Z} \hookrightarrow \mathbb{Z}_p \qquad n \mapsto \left(n \bmod p, n \bmod p^2, n \bmod p^3, \dots\right)$$

which is not surjective. So there are more p-adic integers than usual integers.

(Remark for experts: those of you familiar with category theory might recognize that this definition can be written concisely as

$$\mathbb{Z}_p \coloneqq \varprojlim \mathbb{Z}/p^e\mathbb{Z}$$

where the inverse limit is taken across $e \geq 1$.)

Exercise 27.2.3. Check that $\mathbb{Z}_p$ is an integral domain.

§27.2.ii Base p expansion

Here is another way to think about p-adic integers using "base p". As in the example earlier, every usual integer can be written in base p, for example

$$50 = \overline{1212}_3 = 2 \cdot 3^0 + 1 \cdot 3^1 + 2 \cdot 3^2 + 1 \cdot 3^3.$$

More generally, given any $x = (x_1, \dots) \in \mathbb{Z}_p$, we can write down a "base p" expansion in the sense that there are exactly p choices of x_k given x_{k-1}. Continuing the example earlier, we would write

$$\begin{aligned}(1 \bmod 3,\ 4 \bmod 9,\ 13 \bmod 27,\ 40 \bmod 81,\ \dots) &= 1 + 3 + 3^2 + \dots \\ &= \overline{\dots 1111}_3\end{aligned}$$

and in general we can write

$$x = \sum_{k \geq 0} a_k p^k = \overline{\dots a_2 a_1 a_0}_p$$

where $a_k \in \{0, \dots, p-1\}$, such that the equation holds modulo p^e for each e. Note the expansion is infinite to the *left*, which is different from what you're used to.

(Amusingly, negative integers also have infinite base p expansions: $-4 = \overline{\dots 222212}_3$, corresponding to $(2 \bmod 3,\ 5 \bmod 9,\ 23 \bmod 27,\ 77 \bmod 81 \dots)$.)

Thus you may often hear the advertisement that a p-adic integer is a "possibly infinite base p expansion". This is correct, but later on we'll be thinking of $\mathbb{Z}_p$ in a more and more "analytic" way, and so I prefer to think of this as

p-adic integers are Taylor series with base p.

Indeed, much of your intuition from generating functions $K[[X]]$ (where K is a field) will carry over to $\mathbb{Z}_p$.

§27.2.iii Constructing $\mathbb{Q}_p$

Here is one way in which your intuition from generating functions carries over:

Proposition 27.2.4 (Non-multiples of p are all invertible)
The number $x \in \mathbb{Z}_p$ is invertible if and only if $x_1 \neq 0$. In symbols,

$$x \in \mathbb{Z}_p^\times \iff x \not\equiv 0 \pmod{p}.$$

Contrast this with the corresponding statement for $K[[X]]$: a generating function $F \in K[[X]]$ is invertible iff $F(0) \neq 0$.

Proof. If $x \equiv 0 \pmod{p}$ then $x_1 = 0$, so clearly not invertible. Otherwise, $x_e \not\equiv 0 \pmod{p}$ for all e, so we can take an inverse y_e modulo p^e, with $x_e y_e \equiv 1 \pmod{p^e}$. As the y_e are themselves compatible, the element $(y_1, y_2, \dots)$ is an inverse. □

Example 27.2.5 (We have $-\frac{1}{2} = \overline{\dots 1111}_3 \in \mathbb{Z}_3$)
We claim the earlier example is actually

$$\begin{aligned} -\frac{1}{2} = (1 \bmod 3,\ 4 \bmod 9,\ 13 \bmod 27,\ 40 \bmod 81,\ \dots) &= 1 + 3 + 3^2 + \dots \\ &= \overline{\dots 1111}_3. \end{aligned}$$

Indeed, multiplying it by -2 gives

$$(-2 \bmod 3,\ -8 \bmod 9,\ -26 \bmod 27,\ -80 \bmod 81,\ \dots) = 1.$$

(Compare this with the "geometric series" $1 + 3 + 3^2 + \dots = \frac{1}{1-3}$. We'll actually be able to formalize this later, but not yet.)

Remark 27.2.6 ($\frac{1}{2}$ is an integer for $p > 2$) — The earlier proposition implies that $\frac{1}{2} \in \mathbb{Z}_3$ (among other things); your intuition about what is an "integer" is different here! In olympiad terms, we already knew $\frac{1}{2} \pmod 3$ made sense, which is why calling $\frac{1}{2}$ an "integer" in the 3-adics is correct, even though it doesn't correspond to any element of $\mathbb{Z}$.

Exercise 27.2.7 (Unimportant but tricky)**.** Rational numbers correspond exactly to eventually periodic base p expansions.

With this observation, here is now the definition of $\mathbb{Q}_p$.

Definition 27.2.8 (Introducing $\mathbb{Q}_p$)**.** Since $\mathbb{Z}_p$ is an integral domain, we let $\mathbb{Q}_p$ denote its field of fractions. These are the **p-adic numbers**.

Continuing our generating functions analogy:

$$\mathbb{Z}_p \text{ is to } \mathbb{Q}_p \quad \text{as} \quad K[[X]] \text{ is to } K((X)).$$

This means

$\mathbb{Q}_p$ can be thought of as Laurent series with base p.

and in particular according to the earlier proposition we deduce:

Proposition 27.2.9 ($\mathbb{Q}_p$ looks like formal Laurent series)
Every nonzero element of $\mathbb{Q}_p$ is uniquely of the form

$$p^k u \qquad \text{where } k \in \mathbb{Z},\ u \in \mathbb{Z}_p^\times.$$

Thus, continuing our base p analogy, elements of $\mathbb{Q}_p$ are in bijection with "Laurent series"

$$\sum_{k \geq -n} a_k p^k = \overline{\ldots a_2 a_1 a_0 . a_{-1} a_{-2} \ldots a_{-n}}_p$$

for $a_k \in \{0, \ldots, p-1\}$. So the base p representations of elements of $\mathbb{Q}_p$ can be thought of as the same as usual, but extending infinitely far to the left (rather than to the right).

Remark 27.2.10 (Warning) — The field $\mathbb{Q}_p$ has characteristic *zero*, not p.

Remark 27.2.11 (Warning on fraction field) — This result implies that you shouldn't think about elements of $\mathbb{Q}_p$ as x/y (for $x, y \in \mathbb{Z}_p$) in practice, even though this is the official definition (and what you'd expect from the name $\mathbb{Q}_p$). The only denominators you need are powers of p.
To keep pushing the formal Laurent series analogy, $K((X))$ is usually not thought of as quotient of generating functions but rather as "formal series with some negative exponents". You should apply the same intuition on $\mathbb{Q}_p$.

Remark 27.2.12 — At this point I want to make a remark about the fact $1/p \in \mathbb{Q}_p$, connecting it to the wish-list of properties I had before. In elementary number theory you can take equations modulo p, but if you do the quantity $n/p \bmod p$ doesn't make sense unless you know $n \bmod p^2$. You can't fix this by just taking modulo p^2 since then you need $n \bmod p^3$ to get $n/p \bmod p^2$, ad infinitum. You can work around issues like this, but the nice feature of $\mathbb{Z}_p$ and $\mathbb{Q}_p$ is that you have modulo p^e information for "all e at once": the information of $x \in \mathbb{Q}_p$ packages all the modulo p^e information simultaneously. So you can divide by p with no repercussions.

§27.3 Analytic perspective

§27.3.i Definition

Up until now we've been thinking about things mostly algebraically, but moving forward it will be helpful to start using the language of analysis. Usually, two real numbers are

considered "close" if they are close on the number of line, but for p-adic purposes we only care about modulo p^e information. So, we'll instead think of two elements of $\mathbb{Z}_p$ or $\mathbb{Q}_p$ as "close" if they differ by a large multiple of p^e.

For this we'll borrow the familiar ν_p from elementary number theory.

Definition 27.3.1 (p-adic valuation and absolute value)**.** We define the **p-adic valuation** $\nu_p \colon \mathbb{Q}_p^\times \to \mathbb{Z}$ in the following two equivalent ways:

- For $x = (x_1, x_2, \dots) \in \mathbb{Z}_p$ we let $\nu_p(x)$ be the largest e such that $x_e \equiv 0 \pmod{p^e}$ (or $e = 0$ if $x \in \mathbb{Z}_p^\times$). Then extend to all of $\mathbb{Q}_p^\times$ by $\nu_p(xy) = \nu_p(x) + \nu_p(y)$.
- Each $x \in \mathbb{Q}_p^\times$ can be written uniquely as $p^k u$ for $u \in \mathbb{Z}_p^\times$, $k \in \mathbb{Z}$. We let $\nu_p(x) = k$.

By convention we set $\nu_p(0) = +\infty$. Finally, define the **p-adic absolute value** $|\bullet|_p$ by

$$|x|_p = p^{-\nu_p(x)}.$$

In particular $|0|_p = 0$.

This fulfills the promise that x and y are close if they look the same modulo p^e for large e; in that case $\nu_p(x-y)$ is large and accordingly $|x-y|_p$ is small.

§27.3.ii Ultrametric space

In this way, $\mathbb{Q}_p$ and $\mathbb{Z}_p$ becomes a metric space with metric given by $|x-y|_p$.

> **Exercise 27.3.2.** Suppose $f \colon \mathbb{Z}_p \to \mathbb{Q}_p$ is continuous and $f(n) = (-1)^n$ for every $n \in \mathbb{Z}_{\geq 0}$. Prove that $p = 2$.

In fact, these spaces satisfy a stronger form of the triangle inequality than you are used to from $\mathbb{R}$.

Proposition 27.3.3 ($|\bullet|_p$ is an ultrametric)
For any $x, y \in \mathbb{Z}_p$, we have the **strong triangle inequality**

$$|x+y|_p \leq \max\left\{|x|_p, |y|_p\right\}.$$

Equality holds if (but not only if) $|x|_p \neq |y|_p$.

However, $\mathbb{Q}_p$ is more than just a metric space: it is a field, with its own addition and multiplication. This means we can do analysis just like in $\mathbb{R}$ or $\mathbb{C}$: basically, any notion such as "continuous function", "convergent series", et cetera has a p-adic analog. In particular, we can define what it means for an infinite sum to converge:

Definition 27.3.4 (Convergence notions)**.** Here are some examples of p-adic analogs of "real-world" notions.

- A sequence $s_1, \dots$ converges to a limit L if $\lim_{n\to\infty} |s_n - L|_p = 0$.
- The infinite series $\sum_k x_k$ converges if the sequence of partial sums $s_1 = x_1$, $s_2 = x_1 + x_2$, $\dots$, converges to some limit.
- $\dots$ et cetera $\dots$

With this definition in place, the "base p" discussion we had earlier is now true in the analytic sense: if $x = \overline{\dots a_2 a_1 a_0}_p \in \mathbb{Z}_p$ then

$$\sum_{k=0}^{\infty} a_k p^k \quad \text{converges to } x.$$

Indeed, the difference between x and the nth partial sum is divisible by p^n, hence the partial sums approach x as $n \to \infty$.

While the definitions are all the same, there are some changes in properties that should be true. For example, in $\mathbb{Q}_p$ convergence of partial sums is simpler:

Proposition 27.3.5 ($|x_k|_p \to 0$ iff convergence of series)
A series $\sum_{k=1}^{\infty} x_k$ in $\mathbb{Q}_p$ converges to some limit if and only if $\lim_{k\to\infty} |x_k|_p = 0$.

Contrast this with $\sum \frac{1}{n} = \infty$ in $\mathbb{R}$. You can think of this as a consequence of strong triangle inequality.

Proof. By multiplying by a large enough power of p, we may assume $x_k \in \mathbb{Z}_p$. (This isn't actually necessary, but makes the notation nicer.)

Observe that $x_k \pmod p$ must eventually stabilize, since for large enough n we have $|x_n|_p < 1 \iff \nu_p(x_n) \geq 1$. So let a_1 be the eventual residue modulo p of $\sum_{k=0}^{N} x_k \pmod p$ for large N. In the same way let a_2 be the eventual residue modulo p^2, and so on. Then one can check we approach the limit $a = (a_1, a_2, \dots)$. □

§27.3.iii More fun with geometric series

Let's finally state the p-adic analog of the geometric series formula.

Proposition 27.3.6 (Geometric series)
Let $x \in \mathbb{Z}_p$ with $|x|_p < 1$. Then

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots.$$

Proof. Note that the partial sums satisfy $1 + x + x^2 + \cdots + x^n = \frac{1-x^n}{1-x}$, and $x^n \to 0$ as $n \to \infty$ since $|x|_p < 1$. □

So, $1 + 3 + 3^2 + \cdots = -\frac{1}{2}$ is really a correct convergence in $\mathbb{Z}_3$. And so on.

If you buy the analogy that $\mathbb{Z}_p$ is generating functions with base p, then all the olympiad generating functions you might be used to have p-adic analogs. For example, you can prove more generally that:

Theorem 27.3.7 (Generalized binomial theorem)
If $x \in \mathbb{Z}_p$ and $|x|_p < 1$, then for any $r \in \mathbb{Q}$ we have the series convergence

$$\sum_{n \geq 0} \binom{r}{n} x^n = (1+x)^r.$$

(I haven't defined $(1+x)^r$, but it has the properties you expect.)

§27.3.iv Completeness

Note that the definition of $|\bullet|_p$ could have been given for $\mathbb{Q}$ as well; we didn't need $\mathbb{Q}_p$ to introduce it (after all, we have ν_p in olympiads already). The big important theorem I must state now is:

Theorem 27.3.8 ($\mathbb{Q}_p$ is complete)
The space $\mathbb{Q}_p$ is the completion of $\mathbb{Q}$ with respect to $|\bullet|_p$.

This is the definition of $\mathbb{Q}_p$ you'll see more frequently; one then defines $\mathbb{Z}_p$ in terms of $\mathbb{Q}_p$ (rather than vice-versa) according to

$$\mathbb{Z}_p = \left\{ x \in \mathbb{Q}_p : |x|_p \le 1 \right\}.$$

§27.3.v Philosophical notes

Let me justify why this definition is philosophically nice. Suppose you are an ancient Greek mathematician who is given:

> **Problem for Ancient Greeks.** Estimate the value of the sum
> $$S = \frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{10000^2}$$
> to within 0.001.

The sum S consists entirely of rational numbers, so the problem statement would be fair game for ancient Greece. But it turns out that in order to get a good estimate, it *really helps* if you know about the real numbers: because then you can construct the infinite series $\sum_{n\ge1} n^{-2} = \frac{1}{6}\pi^2$, and deduce that $S \approx \frac{\pi^2}{6}$, up to some small error term from the terms past $\frac{1}{10001^2}$, which can be bounded.

Of course, in order to have access to enough theory to prove that $S = \pi^2/6$, you need to have the real numbers; it's impossible to do calculus in $\mathbb{Q}$ (the sequence 1, 1.4, 1.41, 1.414, is considered "not convergent"!)

Now fast-forward to 2002, and suppose you are given

> **Problem from USA TST 2002.** Estimate the sum
> $$f_p(x) = \sum_{k=1}^{p-1} \frac{1}{(px+k)^2}$$
> to within mod p^3.

Even though $f_p(x)$ is a rational number, it still helps to be able to do analysis with infinite sums, and then bound the error term (i.e. take mod p^3). But the space $\mathbb{Q}$ is not complete with respect to $|\bullet|_p$ either, and thus it makes sense to work in the completion of $\mathbb{Q}$ with respect to $|\bullet|_p$. This is exactly $\mathbb{Q}_p$.

In any case, let's finally solve Example 27.1.1.

Example 27.3.9 (USA TST 2002)
We will now compute

$$f_p(x) = \sum_{k=1}^{p-1} \frac{1}{(px+k)^2} \pmod{p^3}.$$

Armed with the generalized binomial theorem, this becomes straightforward.

$$\begin{aligned} f_p(x) &= \sum_{k=1}^{p-1} \frac{1}{(px+k)^2} = \sum_{k=1}^{p-1} \frac{1}{k^2}\left(1+\frac{px}{k}\right)^{-2} \\ &= \sum_{k=1}^{p-1} \frac{1}{k^2} \sum_{n\geq 0} \binom{-2}{n} \left(\frac{px}{k}\right)^n \\ &= \sum_{n\geq 0} \binom{-2}{n} \sum_{k=1}^{p-1} \frac{1}{k^2} \left(\frac{x}{k}\right)^n p^n \\ &\equiv \sum_{k=1}^{p-1} \frac{1}{k^2} - 2x \left(\sum_{k=1}^{p-1} \frac{1}{k^3}\right) p + 3x^2 \left(\sum_{k=1}^{p-1} \frac{1}{k^4}\right) p^2 \pmod{p^3}. \end{aligned}$$

Using the elementary facts that $p^2 \mid \sum_k k^{-3}$ and $p \mid \sum_k k^{-4}$, this solves the problem.

§27.4 Mahler coefficients

One of the big surprises of p-adic analysis is that:

We can basically describe all continuous functions $\mathbb{Z}_p \to \mathbb{Q}_p$.

They are given by a basis of functions

$$\binom{x}{n} := \frac{x(x-1)\dots(x-(n-1))}{n!}$$

in the following way.

Theorem 27.4.1 (Mahler; see [Sc07, Theorem 51.1, Exercise 51.b])
Let $f\colon \mathbb{Z}_p \to \mathbb{Q}_p$ be continuous, and define

$$a_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} f(k). \tag{27.1}$$

Then $\lim_n a_n = 0$ and

$$f(x) = \sum_{n\geq 0} a_n \binom{x}{n}.$$

Conversely, if a_n is any sequence converging to zero, then $f(x) = \sum_{n\geq 0} a_n \binom{x}{n}$ defines a continuous function satisfying (27.1).

The a_i are called the *Mahler coefficients* of f.

Exercise 27.4.2. Last post we proved that if $f\colon \mathbb{Z}_p \to \mathbb{Q}_p$ is continuous and $f(n) = (-1)^n$ for every $n \in \mathbb{Z}_{\geq 0}$ then $p = 2$. Re-prove this using Mahler's theorem, and this time show conversely that a unique such f exists when $p = 2$.

You'll note that these are the same finite differences that one uses on polynomials in high school math contests, which is why they are also called "Mahler differences".

$$\begin{aligned} a_0 &= f(0) \\ a_1 &= f(1) - f(0) \\ a_2 &= f(2) - 2f(1) + f(0) \\ a_3 &= f(3) - 3f(2) + 3f(1) - f(0). \end{aligned}$$

Thus one can think of $a_n \to 0$ as saying that the values of $f(0)$, $f(1)$, ... behave like a polynomial modulo p^e for every $e \geq 0$.

The notion "analytic" also has a Mahler interpretation. First, the definition.

Definition 27.4.3. We say that a function $f\colon \mathbb{Z}_p \to \mathbb{Q}_p$ is **analytic** if it has a power series expansion

$$\sum_{n \geq 0} c_n x^n \quad c_n \in \mathbb{Q}_p \qquad \text{converging for } x \in \mathbb{Z}_p.$$

Theorem 27.4.4 ([Sc07, Theorem 54.4])
The function $f(x) = \sum_{n \geq 0} a_n \binom{x}{n}$ is analytic if and only if

$$\lim_{n \to \infty} \frac{a_n}{n!} = 0.$$

Analytic functions also satisfy the following niceness result:

Theorem 27.4.5 (Strassmann's theorem)
Let $f\colon \mathbb{Z}_p \to \mathbb{Q}_p$ be analytic. Then f has finitely many zeros.

To give an application of these results, we will prove the following result, which was interesting even before p-adics came along!

Theorem 27.4.6 (Skolem-Mahler-Lech)
Let $(x_i)_{i \geq 0}$ be an integral linear recurrence, meaning $(x_i)_{i \geq 0}$ is a sequence of integers

$$x_n = c_1 x_{n-1} + c_2 x_{n-2} + \cdots + c_k x_{n-k} \qquad n = 1, 2, \ldots$$

holds for some choice of integers $c_1, \ldots, c_k$. Then the set of indices $\{i \mid x_i = 0\}$ is eventually periodic.

Proof. According to the theory of linear recurrences, there exists a matrix A such that we can write x_i as a dot product

$$x_i = \left\langle A^i u, v \right\rangle.$$

Let p be a prime not dividing $\det A$. Let T be an integer such that $A^T \equiv \mathrm{id} \pmod{p}$ (with id denoting the identity matrix).

Fix any $0 \le r < N$. We will prove that either all the terms

$$f(n) = x_{nT+r} \qquad n = 0, 1, \ldots$$

are zero, or at most finitely many of them are. This will conclude the proof.

Let $A^T = \mathrm{id} + pB$ for some integer matrix B. We have

$$\begin{aligned} f(n) &= \left\langle A^{nT+r} u, v \right\rangle = \langle (\mathrm{id} + pB)^n A^r u, v \rangle \\ &= \sum_{k \ge 0} \binom{n}{k} \cdot p^n \langle B^n A^r u, v \rangle \\ &= \sum_{k \ge 0} a_n \binom{n}{k} \qquad \text{where } a_n = p^n \langle B^n A^r u, v \rangle \in p^n \mathbb{Z}. \end{aligned}$$

Thus we have written f in Mahler form. Initially, we define $f \colon \mathbb{Z}_{\ge 0} \to \mathbb{Z}$, but by Mahler's theorem (since $\lim_n a_n = 0$) it follows that f extends to a function $f \colon \mathbb{Z}_p \to \mathbb{Q}_p$. Also, we can check that $\lim_n \frac{a_n}{n!} = 0$ hence f is even analytic.

Thus by Strassman's theorem, f is either identically zero, or else it has finitely many zeros, as desired. □

§27.5 A few harder problems to think about

Problem 27A† ($\mathbb{Z}_p$ is compact)**.** Show that $\mathbb{Q}_p$ is not compact, but $\mathbb{Z}_p$ is. (For the latter, I recommend using sequential continuity.)

Problem 27B† (Totally disconnected)**.** Show that both $\mathbb{Z}_p$ and $\mathbb{Q}_p$ are *totally disconnected*: there are no connected sets other than the empty set and singleton sets.

Problem 27C (Mentioned in MathOverflow)**.** Let p be a prime. Find a sequence q_1, q_2, ... of rational numbers such that:

- the sequence q_n converges to 0 in the real sense;
- the sequence q_n converges to 2021 in the p-adic sense.

Problem 27D (USA TST 2011)**.** Let p be a prime. We say that a sequence of integers $\{z_n\}_{n=0}^{\infty}$ is a *p-pod* if for each $e \ge 0$, there is an $N \ge 0$ such that whenever $m \ge N$, p^e divides the sum

$$\sum_{k=0}^{m} (-1)^k \binom{m}{k} z_k.$$

Prove that if both sequences $\{x_n\}_{n=0}^{\infty}$ and $\{y_n\}_{n=0}^{\infty}$ are p-pods, then the sequence $\{x_n y_n\}_{n=0}^{\infty}$ is a p-pod.

28 Differentiation

§28.1 Definition

Prototypical example for this section: x^3 *has derivative* $3x^2$.

I suspect most of you have seen this before, but:

Definition 28.1.1. Let U be an open subset[1] of $\mathbb{R}$ and let $f : U \to \mathbb{R}$ be a function. Let $p \in U$. We say f is **differentiable** at p if the limit[2]

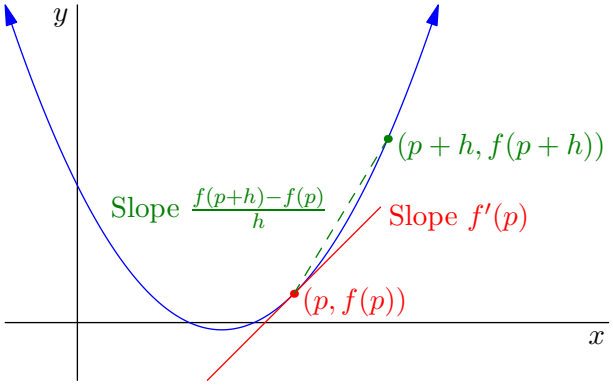
$$\lim_{h \to 0} \frac{f(p+h) - f(p)}{h}$$

exists. If so, we denote its value by $f'(p)$ and refer to this as the **derivative** of f at p.

The function f is differentiable if it is differentiable at every point. In that case, we regard the derivative $f' : (a,b) \to \mathbb{R}$ as a function it its own right.

> **Exercise 28.1.2.** Show that if f is differentiable at p then it is continuous at p too.

Here is the picture. Suppose $f : \mathbb{R} \to \mathbb{R}$ is differentiable (hence continuous). We draw a graph of f in the usual way and consider values of h. For any nonzero h, what we get is the slope of the *secant* line joining $(p, f(p))$ to $(p+h, f(p+h))$. However, as h gets close to zero, that secant line begins to approach a line which is tangent to the graph of the curve. A picture with f a parabola is shown below, with the tangent in red, and the secant in dashed green.



So the picture in your head should be that

> $f'(p)$ **looks like the slope of the tangent line at** $(p, f(p))$**.**

[1] We will almost always use $U = (a,b)$ or $U = \mathbb{R}$, and you will not lose much by restricting the definition to those.

[2] Remember we are following the convention in Abuse of Notation 26.6.6. So we mean "the limit of the function $h \mapsto \frac{f(p+h)-f(p)}{h}$ except the value at $h = 0$ can be anything". And this is important because that fraction does not have a definition at $h = 0$. As promised, we pay this no attention.

Remark 28.1.3 — Note that the derivatives are defined for functions on *open* intervals. This is important. If $f : [a, b] \to \mathbb{R}$ for example, we could still define the derivative at each interior point, but $f'(a)$ no longer makes sense since f is not given a value on any open neighborhood of a.

Let's do one computation and get on with this.

Example 28.1.4 (Derivative of x^3 is $3x^2$)
Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^3$. For any point p, and *nonzero* h we can compute

$$\begin{aligned} \frac{f(p+h) - f(p)}{h} &= \frac{(p+h)^3 - p^3}{h} \\ &= \frac{3p^2h + 3ph^2 + h^3}{h} \\ &= 3p^2 + 3ph + h^2. \end{aligned}$$

Thus,

$$\lim_{h \to 0} \frac{f(p+h) - f(p)}{h} = \lim_{h \to 0} (3p^2 + 3ph + h^2) = 3p^2.$$

Thus the slope at each point of f is given by the formula $3p^2$. It is customary to then write $f'(x) = 3x^2$ as the derivative of the entire function f.

Abuse of Notation 28.1.5. We will now be sloppy and write this as $(x^3)' = 3x^2$. This is shorthand for the significantly more verbose "the real-valued function x^3 on domain so-and-so has derivative $3p^2$ at every point p in its domain".

In general, a real-valued differentiable function $f : U \to \mathbb{R}$ naturally gives rise to derivative $f'(p)$ at every point $p \in U$, so it is customary to just give up on p altogether and treat f' as function itself $U \to \mathbb{R}$, even though this real number is of a "different interpretation": $f'(p)$ is meant to interpret a slope (e.g. your hourly pay rate) as opposed to a value (e.g. your total dollar worth at time t). If f is a function from real life, the units do not even match!

This convention is so deeply entrenched I cannot uproot it without more confusion than it is worth. But if you read the chapters on multivariable calculus you will see how it comes back to bite us, when I need to re-define the derivative to be a *linear map*, rather than a single real number.

§28.2 How to compute them

Same old, right? Sum rule, all that jazz.

Theorem 28.2.1 (Your friendly high school calculus rules)

In what follows f and g are differentiable functions, and U, V are open subsets of $\mathbb{R}$.

- (Sum rule) If $f, g \colon U \to \mathbb{R}$ then then $(f+g)'(x) = f'(x) + g'(x)$.
- (Product rule) If $f, g \colon U \to \mathbb{R}$ then then $(f \cdot g)'(x) = f'(x)g(x) + f(x)g'(x)$.
- (Chain rule) If $f \colon U \to V$ and $g \colon V \to \mathbb{R}$ then the derivative of the composed function $g \circ f \colon U \to \mathbb{R}$ is $g'(f(x)) \cdot f'(x)$.

Proof.

- Sum rule: trivial, do it yourself if you care.
- Product rule: for every nonzero h and point $p \in U$ we may write

$$\frac{f(p+h)g(p+h) - f(p)g(p)}{h} = \frac{f(p+h)-f(p)}{h} \cdot g(p+h) + \frac{g(p+h)-g(p)}{h} \cdot f(p)$$

  which as $h \to 0$ gives the desired expression.
- Chain rule: this is where Abuse of Notation 26.6.6 will actually bite us. Let $p \in U$, $q = f(p) \in V$, so that

$$(g \circ f)'(p) = \lim_{h \to 0} \frac{g(f(p+h)) - g(q)}{h}.$$

  We would like to write the expression in the limit as

$$\frac{g(f(p+h)) - g(q)}{h} = \frac{g(f(p+h)) - g(q)}{f(p+h) - q} \cdot \frac{f(p+h)-f(p)}{h}.$$

  The problem is that the denominator $f(p+h) - f(p)$ might be zero. So instead, we define the expression

$$Q(y) = \begin{cases} \frac{g(y)-g(q)}{y-q} & \text{if } y \neq q \\ g'(q) & \text{if } y = q \end{cases}$$

  which is continuous since g was differentiable at q. Then, we *do* have the equality

$$\frac{g(f(p+h)) - g(q)}{h} = Q\left(f(p+h)\right) \cdot \frac{f(p+h)-f(p)}{h}.$$

  because if $f(p+h) = q$ with $h \neq 0$, then both sides are equal to zero anyways.

  Then, in the limit as $h \to 0$, we have $\lim_{h\to 0} \frac{f(p+h)-f(p)}{h} = f'(p)$, while $\lim_{h \to 0} Q(f(p+h)) = Q(q) = g'(q)$ by continuity. This was the desired result. □

Exercise 28.2.2. Compute the derivative of the polynomial $f(x) = x^3 + 10x^2 + 2019$, viewed as a function $f \colon \mathbb{R} \to \mathbb{R}$.

Remark 28.2.3 — Quick linguistic point: the theorems above all hold at each individual point. For example the sum rule really should say that if $f, g \colon U \to \mathbb{R}$ are differentiable at the point p then so is $f+g$ and the derivative equals $f'(p) + g'(p)$. Thus if f and g are differentiable on all of U, then it of course follows that

$(f+g)' = f' + g'$. So each of the above rules has a "point-by-point" form which then implies the "whole U" form.
We only state the latter since that is what is used in practice. However, in the rare situations where you have a function differentiable only at certain points of U rather than the whole interval U, you can still use the below.

We next list some derivatives of well-known functions, but as we do not give rigorous definitions of these functions, we do not prove these here.

Proposition 28.2.4 (Derivatives of some well-known functions)

- The exponential function $\exp\colon \mathbb{R} \to \mathbb{R}$ defined by $\exp(x) = e^x$ is its own derivative.
- The trig functions sin and cos have $\sin' = \cos$, $\cos' = -\sin$.

Example 28.2.5 (A typical high-school calculus question)

This means that you can mechanically compute the derivatives of any artificial function obtained by using the above, which makes it a great source of busy work in American high schools and universities. For example, if

$$f(x) = e^x + x\sin(x^2) \qquad f\colon \mathbb{R} \to \mathbb{R}$$

then one can compute f' by:

$$\begin{aligned}
f'(x) &= (e^x)' + (x\sin(x^2))' && \text{sum rule} \\
&= e^x + (x\sin(x^2))' && \text{above table} \\
&= e^x + (x)'\sin(x^2) + x(\sin(x^2))' && \text{product rule} \\
&= e^x + \sin(x^2) + x(\sin(x^2))' && (x)' = 1 \\
&= e^x + \sin(x^2) + x \cdot 2x \cdot \cos(x^2) && \text{chain rule.}
\end{aligned}$$

Of course, this function f is totally artificial and has no meaning, which is why calculus is the topic of widespread scorn in the United States. That said, it is worth appreciating that calculations like this are possible: one could say we have a pseudo-theorem "derivatives can actually be computed in practice".

If we take for granted that $(e^x)' = e^x$, then we can derive two more useful functions to add to our library of functions we can differentiate.

Corollary 28.2.6 (Derivative of log is $1/x$)

The function $\log\colon \mathbb{R}_{>0} \to \mathbb{R}$ has derivative $(\log x)' = 1/x$.

Proof. We have that $x = e^{\log x}$. Differentiate both sides, and again use the chain rule[3]

$$1 = e^{\log x} \cdot (\log x)'.$$

[3]There is actually a small subtlety here: we are taking for granted that log is differentiable.

Thus $(\log x)' = \frac{1}{e^{\log x}} = 1/x$. □

Corollary 28.2.7 (Power rule)
Let r be a real number. The function $\mathbb{R}_{>0} \to \mathbb{R}$ by $x \mapsto x^r$ has derivative $(x^r)' = rx^{r-1}$.

Proof. We knew this for integers r already, but now we can prove it for any positive real number r. Write

$$f(x) = x^r = e^{r \log x}$$

considered as a function $f : \mathbb{R}_{>0} \to \mathbb{R}$. The chain rule (together with the fact that $(e^x)' = e^x$) now gives

$$\begin{aligned} f'(x) &= e^{r \log x} \cdot (r \log x)' \\ &= e^{r \log x} \cdot \frac{r}{x} = x^r \cdot \frac{r}{x} = rx^{r-1}. \end{aligned}$$

The reason we don't prove the formulas for e^x and $\log x$ is that we don't at the moment even have a rigorous definition for either, or even for 2^x if x is not rational. However it's nice to know that some things imply the other. □

§28.3 Local (and global) maximums

Prototypical example for this section: Horizontal tangent lines to the parabola are typically good pictures.

You may remember from high school that one classical use of calculus was to extract the minimum or maximum values of functions. We will give a rigorous description of how to do this here.

Definition 28.3.1. Let $f : U \to \mathbb{R}$ be a function. A **local maximum** is a point $p \in U$ such that there exists an open neighborhood V of p (contained inside U) such that $f(p) \geq f(x)$ for every $x \in V$.

A **local minimum** is defined similarly.[4]

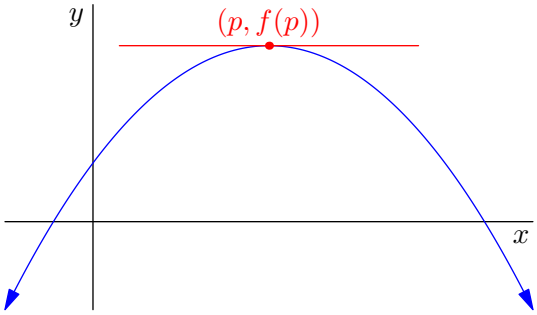
Definition 28.3.2. A point p is a **local extrema** if it satisfies either of these.

The nice thing about derivatives is that they pick up all extrema.

Theorem 28.3.3 (Fermat's theorem on stationary points)
Suppose $f : U \to \mathbb{R}$ is differentiable and $p \in U$ is a local extrema. Then $f'(p) = 0$.

If you draw a picture, this result is not surprising.

[4] Equivalently, it is a local maximum of $-f$.



(Note also: the converse is not true. Say, $f(x) = x^{2019}$ has $f'(0) = 0$ but $x = 0$ is not a local extrema for f.)

Proof. Assume for contradiction $f'(p) > 0$. Choose any $\varepsilon > 0$ with $\varepsilon < f'(p)$. Then for sufficiently small $|h|$ we should have

$$\frac{f(p+h) - f(p)}{h} > \varepsilon.$$

In particular $f(p+h) > f(p)$ for $h > 0$ while $f(p+h) < f(p)$ for $h < 0$. So p is not a local extremum.

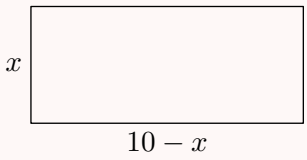
The proof for $f'(p) < 0$ is similar. □

However, this is not actually adequate if we want a complete method for optimization. The issue is that we seek *global* extrema, which may not even exist: for example $f(x) = x$ (which has $f'(x) = 1$) obviously has no local extrema at all. The key to resolving this is to use *compactness*: we change the domain to be a compact set Z, for which we know that f will achieve some global maximum. The set Z will naturally have some *interior* S, and calculus will give us all the extrema within S. Then we manually check all cases outside Z.

Let's see two extended examples. The one is simple, and you probably already know about it, but I want to show you how to use compactness to argue thoroughly, and how the "boundary" points naturally show up.

Example 28.3.4 (Rectangle area optimization)

Suppose we consider rectangles with perimeter 20 and want the rectangle with the smallest or largest area.



If we choose the legs of the rectangle to be x and $10 - x$, then we are trying to optimize the function

$$f(x) = x(10-x) = 10x - x^2 \qquad f \colon [0, 10] \to \mathbb{R}.$$

By compactness, there exists *some* global maximum and *some* global minimum. As f is differentiable on $(0, 10)$, we find that for any $p \in (0, 10)$, a global maximum will be a local maximum too, and hence should satisfy

$$0 = f'(p) = 10 - 2p \implies p = 5.$$

Also, the points $x = 0$ and $x = 10$ lie in the domain but not the interior $(0, 10)$. Therefore the global extrema (in addition to existing) must be among the three suspects $\{0, 5, 10\}$.
We finally check $f(0) = 0$, $f(5) = 25$, $f(10) = 0$. So the 5×5 square has the largest area and the degenerate rectangles have the smallest (zero) area.

Here is a non-elementary example.

Proposition 28.3.5 ($e^x \ge 1 + x$)
For all real numbers x we have $e^x \ge 1 + x$.

Proof. Define the differentiable function

$$f(x) = e^x - (x + 1) \qquad f : \mathbb{R} \to \mathbb{R}.$$

Consider the compact interval $Z = [-1, 100]$. If $x \le -1$ then obviously $f(x) > 0$. Similarly if $x \ge 100$ then obviously $f(x) > 0$ too. So we just want to prove that if $x \in Z$, we have $f(x) \ge 0$.

Indeed, there exists *some* global minimum p. It could be the endpoints -1 or 100. Otherwise, if it lies in $U = (-1, 100)$ then it would have to satisfy

$$0 = f'(p) = e^p - 1 \implies p = 0.$$

As $f(-1) > 0$, $f(100) > 0$, $f(0) = 0$, we conclude $p = 0$ is the global minimum of Z; and hence $f(x) \ge 0$ for all $x \in Z$, hence for all x. □

Remark 28.3.6 — If you are willing to use limits at $\pm\infty$, you can rewrite proofs like the above in such a way that you don't have to explicitly come up with endpoints like -1 or 100. We won't do so here, but it's nice food for thought.

§28.4 Rolle and friends

Prototypical example for this section: The racetrack principle, perhaps?

One corollary of the work in the previous section is Rolle's theorem.

Theorem 28.4.1 (Rolle's theorem)
Suppose $f : [a, b] \to \mathbb{R}$ is a continuous function, which is differentiable on the open interval (a, b), such that $f(a) = f(b)$. Then there is a point $c \in (a, b)$ such that $f'(c) = 0$.

Proof. Assume f is nonconstant (otherwise any c works). By compactness, there exists both a global maximum and minimum. As $f(a) = f(b)$, either the global maximum or the global minimum must lie inside the open interval (a, b), and then Fermat's theorem on stationary points finishes. □

I was going to draw a picture until I realized xkcd #2042 has one already.

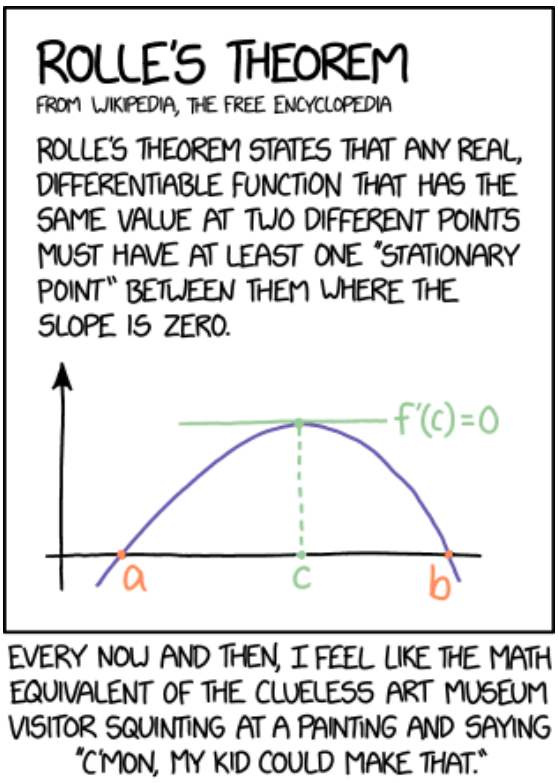


Image from [Mu]

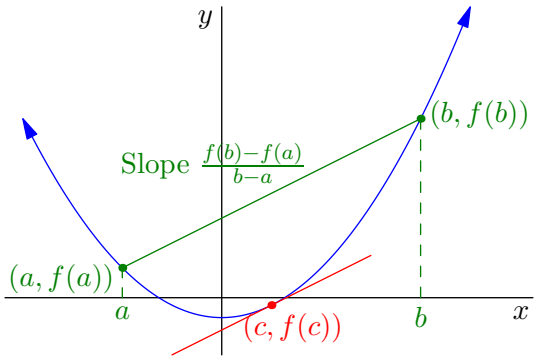
One can adapt the theorem as follows.

Theorem 28.4.2 (Mean value theorem)

Suppose $f\colon [a,b] \to \mathbb{R}$ is a continuous function, which is differentiable on the open interval (a,b). Then there is a point $c \in (a,b)$ such that

$$f'(c) = \frac{f(b)-f(a)}{b-a}.$$

Pictorially, there is a c such that the tangent at c has the same slope as the secant joining $(a, f(a))$, to $(b, f(b))$; and Rolle's theorem is the special case where that secant is horizontal.



Proof of mean value theorem. Let $s = \frac{f(b)-f(a)}{b-a}$ be the slope of the secant line, and define

$$g(x) = f(x) - sx$$

which intuitively shears f downwards so that the secant becomes horizontal. In fact $g(a) = g(b)$ now, so we apply Rolle's theorem to g. □

Remark 28.4.3 (For people with driver's licenses) — There is a nice real-life interpretation of this I should mention. A car is travelling along a one-dimensional road (with $f(t)$ denoting the position at time t). Suppose you cover 900 kilometers in your car over the course of 5 hours (say $f(0) = 0$, $f(5) = 900$). Then there is *some* point at time in which your speed at that moment was exactly 180 kilometers per hour, and so you cannot really complain when the cops pull you over for speeding.

The mean value theorem is important because it lets you relate **use derivative information to get information about the function** in a way that is really not possible without it. Here is one quick application to illustrate my point:

Proposition 28.4.4 (Racetrack principle)

Let $f, g \colon \mathbb{R} \to \mathbb{R}$ be two differentiable functions with $f(0) = g(0)$.

(a) If $f'(x) \geq g'(x)$ for every $x > 0$, then $f(x) \geq g(x)$ for every $x > 0$.

(b) If $f'(x) > g'(x)$ for every $x > 0$, then $f(x) > g(x)$ for every $x > 0$.

This proposition might seem obvious. You can think of it as a race track for a reason: if f and g denote the positions of two cars (or horses etc) and the first car is always faster than the second car, then the first car should end up ahead of the second car. At a special case $g = 0$, this says that if $f'(x) \geq 0$, i.e. "f is increasing", then, well, $f(x) \geq f(0)$ for $x > 0$, which had better be true. However, if you try to prove this by definition from derivatives, you will find that it is not easy! However, it's almost a prototype for the mean value theorem.

Proof of racetrack principle. We prove (a). Let $h = f - g$, so $h(0) = 0$. Assume for contradiction $h(p) < 0$ for some $p > 0$. Then the secant joining $(0, h(0))$ to $(p, h(p))$ has negative slope; in other words by mean value theorem there is a $0 < c < p$ such that

$$f'(c) - g'(c) = h'(c) = \frac{h(p) - h(0)}{p} = \frac{h(p)}{p} < 0$$

so $f'(c) < g'(c)$, contradiction. Part (b) is the same. □

Sometimes you will be faced with two functions which you cannot easily decouple; the following form may be more useful in that case.

Theorem 28.4.5 (Ratio mean value theorem)

Let $f, g \colon [a, b] \to \mathbb{R}$ be two continuous functions which are differentiable on (a, b), and such that $g(a) \neq g(b)$. Then there exists $c \in (a, b)$ such that

$$f'(c)(g(b) - g(a)) = g'(c)(f(b) - f(a))$$

Proof. Use Rolle's theorem on the function

$$h(x) = [f(x) - f(a)]\,[g(b) - g(a)] - [g(x) - g(a)]\,[f(b) - f(a)]\,.$$ □

This is also called Cauchy's mean value theorem or the extended mean value theorem.

§28.5 Smooth functions

Prototypical example for this section: All the functions you're used to.

Let $f\colon U \to \mathbb{R}$ be differentiable, thus giving us a function $f'\colon U \to \mathbb{R}$. If our initial function was nice enough, then we can take the derivative again, giving a function $f''\colon U \to \mathbb{R}$, and so on. In general, after taking the derivative n times, we denote the resulting function by $f^{(n)}$. By convention, $f^{(0)} = f$.

Definition 28.5.1. A function $f\colon U \to \mathbb{R}$ is **smooth** if it is infinitely differentiable; that is the function $f^{(n)}$ exists for all n.

> **Question 28.5.2.** Show that the absolute value function is not smooth.

Most of the functions we encounter, such as polynomials, e^x, log, sin, cos are smooth, and so are their compositions. Here is a weird example which we'll grow more next time.

Example 28.5.3 (A smooth function with all derivatives zero)
Consider the function

$$f(x) = \begin{cases} e^{-1/x} & x > 0 \\ 0 & x \le 0. \end{cases}$$

This function can be shown to be smooth, with $f^{(n)}(0) = 0$. So this function has every derivative at the origin equal to zero, despite being nonconstant!

§28.6 A few harder problems to think about

Problem 28A (Quotient rule)**.** Let $f\colon (a,b) \to \mathbb{R}$ and $g\colon (a,b) \to \mathbb{R}_{>0}$ be differentiable functions. Let $h = f/g$ be their quotient (also a function $(a,b) \to \mathbb{R}$). Show that the derivative of h is given by

$$h'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}.$$

Problem 28B. For real numbers $x > 0$, how small can x^x be?

Problem 28C (RMM 2018)**.** Determine whether or not there exist nonconstant polynomials $P(x)$ and $Q(x)$ with real coefficients satisfying

$$P(x)^{10} + P(x)^9 = Q(x)^{21} + Q(x)^{20}.$$

Problem 28D. Let $P(x)$ be a degree n polynomial with real coefficients. Prove that the equation $e^x = P(x)$ has at most $n+1$ real solutions in x.

Problem 28E (Jensen's inequality)**.** Let $f\colon (a,b) \to \mathbb{R}$ be a twice differentiable function such that $f''(x) \ge 0$ for all x (i.e. f is *convex*). Prove that

$$f\left(\frac{x+y}{2}\right) \le \frac{f(x)+f(y)}{2}$$

for all real numbers x and y in the interval (a,b).

Problem 28F (L'Hôpital rule, or at least one case)**.** Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be differentiable functions and let p be a real number. Suppose that

$$\lim_{x \to p} f(x) = \lim_{x \to p} g(x) = 0.$$

Prove that

$$\lim_{x \to p} \frac{f(x)}{g(x)} = \lim_{x \to p} \frac{f'(x)}{g'(x)}$$

provided the right-hand limit exists.

Problem 28G. Calculate the derivative of the function $f\colon (0, \infty) \to \mathbb{R}$ defined by $f(x) = x^x$.

29 Power series and Taylor series

Polynomials are very well-behaved functions, and are studied extensively for that reason. From an analytic perspective, for example, they are smooth, and their derivatives are easy to compute.

In this chapter we will study *power series*, which are literally "infinite polynomials" $\sum_n a_n x^n$. Armed with our understanding of series and differentiation, we will see three great things:

- Many of the functions we see in nature actually *are* given by power series. Among them are e^x, $\log x$, $\sin x$.
- Their convergence properties are actually quite well behaved: from the string of coefficients, we can figure out which x they converge for.
- The derivative of $\sum_n a_n x^n$ is actually just $\sum_n n a_n x^{n-1}$.

§29.1 Motivation

To get the ball rolling, let's start with one infinite polynomial you'll recognize: for any fixed number $-1 < x < 1$ we have the series convergence

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots$$

by the geometric series formula.

Let's pretend we didn't see this already in Problem 26D. So, we instead have a smooth function $f : (-1,1) \to \mathbb{R}$ by

$$f(x) = \frac{1}{1-x}.$$

Suppose we wanted to pretend that it was equal to an "infinite polynomial" near the origin, that is

$$(1-x)^{-1} = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + \cdots.$$

How could we find that polynomial, if we didn't already know?

Well, for starters we can first note that by plugging in $x = 0$ we obviously want $a_0 = 1$.

We have derivatives, so actually, we can then differentiate both sides to obtain that

$$(1-x)^{-2} = a_1 + 2a_2 x + 3a_3 x^2 + 4a_4 x^3 + \cdots.$$

If we now set $x = 0$, we get $a_1 = 1$. In fact, let's keep taking derivatives and see what we get.

$$\begin{aligned}
(1-x)^{-1} &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + \ldots \\
(1-x)^{-2} &= a_1 + 2a_2 x + 3a_3 x^2 + 4a_4 x^3 + 5a_5 x^4 + \ldots \\
2(1-x)^{-3} &= 2a_2 + 6a_3 x + 12a_4 x^2 + 20a_5 x^3 + \ldots \\
6(1-x)^{-4} &= 6a_3 + 24a_4 x + 60a_5 x^2 + \ldots \\
24(1-x)^{-5} &= 24a_4 + 120a_5 x + \ldots \\
&\vdots .
\end{aligned}$$

If we set $x = 0$ we find $1 = a_0 = a_1 = a_2 = \dots$ which is what we expect; the geometric series $\frac{1}{1-x} = 1 + x + x^2 + \cdots$. And so actually taking derivatives was enough to get the right claim!

§29.2 Power series

Prototypical example for this section: $\frac{1}{1-z} = 1 + z + z^2 + \cdots$, *which converges on* $(-1, 1)$.

Of course this is not rigorous, since we haven't described what the right-hand side is, much less show that it can be differentiated term by term. So we define the main character now.

Definition 29.2.1. A **power series** is a sum of the form

$$\sum_{n=0}^{\infty} a_n z^n = a_0 + a_1 z + a_2 z^2 + \cdots$$

where a_0, a_1, $\dots$ are real numbers, and z is a variable.

Abuse of Notation 29.2.2 ($0^0 = 1$)**.** If you are very careful, you might notice that when $z = 0$ and $n = 0$ we find 0^0 terms appearing. For this chapter the convention is that they are all equal to one.

Now, if I plug in a *particular* real number h, then I get a series of real numbers $\sum_{n=0}^{\infty} a_n h^n$. So I can ask, when does this series converge? It terms out there is a precise answer for this.

Definition 29.2.3. Given a power series $\sum_{n=0}^{\infty} a_n z^n$, the **radius of convergence** R is defined by the formula

$$\frac{1}{R} = \limsup_{n \to \infty} |a_n|^{1/n}.$$

with the convention that $R = 0$ if the right-hand side is ∞, and $R = \infty$ if the right-hand side is zero.

Theorem 29.2.4 (Cauchy-Hadamard theorem)

Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with radius of convergence R. Let h be a real number, and consider the infinite series

$$\sum_{n=0}^{\infty} a_n h^n$$

of real numbers. Then:

- The series converges absolutely if $|h| < R$.
- The series diverges if $|h| > R$.

Proof. This is not actually hard, but it won't be essential, so not included. □

Remark 29.2.5 — In the case $|h| = R$, it could go either way.

Example 29.2.6 ($\sum z^n$ has radius 1)
Consider the geometric series $\sum_n z^n = 1 + z + z^2 + \cdots$. Since $a_n = 1$ for every n, we get $R = 1$, which is what we expected.

Therefore, if $\sum_n a_n z^n$ is a power series with a nonzero radius $R > 0$ of convergence, then it can *also* be thought of as a function

$$(-R, R) \to \mathbb{R} \quad \text{by} \quad h \mapsto \sum_{n \geq 0} a_n h^n.$$

This is great. Note also that if $R = \infty$, this means we get a function $\mathbb{R} \to \mathbb{R}$.

Abuse of Notation 29.2.7 (Power series vs. functions)**.** There is some subtlety going on with "types" of objects again. Analogies with polynomials can help.

Consider $P(x) = x^3 + 7x + 9$, a polynomial. You *can*, for any real number h, plug in $P(h)$ to get a real number. However, in the polynomial *itself*, the symbol x is supposed to be a *variable* — which sometimes we will plug in a real number for, but that happens only after the polynomial is defined.

Despite this, "the polynomial $p(x) = x^3 + 7x + 9$" (which can be thought of as the coefficients) and "the real-valued function $x \mapsto x^3 + 7x + 9$" are often used interchangeably. The same is about to happen with power series: while they were initially thought of as a sequence of coefficients, the Cauchy-Hadamard theorem lets us think of them as functions too, and thus we blur the distinction between them.

§29.3 Differentiating them

Prototypical example for this section: We saw earlier $1 + x + x^2 + x^3 + \ldots$ *has derivative* $1 + 2x + 3x^2 + \ldots$.

As promised, differentiation works exactly as you want.

Theorem 29.3.1 (Differentiation works term by term)
Let $\sum_{n \geq 0} a_n z^n$ be a power series with radius of convergence $R > 0$, and consider the corresponding function

$$f \colon (-R, R) \to \mathbb{R} \quad \text{by} \quad f(x) = \sum_{n \geq 0} a_n x^n.$$

Then all the derivatives of f exist and are given by power series

$$\begin{aligned} f'(x) &= \sum_{n \geq 1} n a_n x^{n-1} \\ f''(x) &= \sum_{n \geq 2} n(n-1) a_n x^{n-2} \\ &\vdots \end{aligned}$$

which also converge for any $x \in (-R, R)$. In particular, f is smooth.

Proof. Also omitted. The right way to prove it is to define the notion "converges uniformly", and strengthen Cauchy-Hadamard to have this as a conclusion as well. □

Corollary 29.3.2 (A description of power series coefficients)

Let $\sum_{n\geq 0} a_n z^n$ be a power series with radius of convergence $R > 0$, and consider the corresponding function $f(x)$ as above. Then

$$a_n = \frac{f^{(n)}(0)}{n!}.$$

Proof. Take the nth derivative and plug in $x = 0$. □

§29.4 Analytic functions

Prototypical example for this section: The piecewise $e^{-1/x}$ or 0 function is not analytic, but is smooth.

With all these nice results about power series, we now have a way to do this process the other way: suppose that $f : U \to \mathbb{R}$ is a function. Can we express it as a power series?

Functions for which this *is* true are called analytic.

Definition 29.4.1. A function $f : U \to \mathbb{R}$ is **analytic** at the point $p \in U$ if there exists an open neighborhood V of p (inside U) and a power series $\sum_n a_n z^n$ such that

$$f(x) = \sum_{n\geq 0} a_n (x-p)^n$$

for any $x \in V$. As usual, the whole function is analytic if it is analytic at each point.

Question 29.4.2. Show that if f is analytic, then it's smooth.

Moreover, if f is analytic, then by the corollary above its coefficients are actually described exactly by

$$f(x) = \sum_{n\geq 0} \frac{f^{(n)}(p)}{n!}(x-p)^n.$$

Even if f is smooth but not analytic, we can at least write down the power series; we give this a name.

Definition 29.4.3. For smooth f, the power series $\sum_{n\geq 0} \frac{f^{(n)}(p)}{n!} z^n$ is called the **Taylor series** of f at p.

Example 29.4.4 (Examples of analytic functions)

(a) Polynomials, sin, cos, e^x, log all turn out to be analytic.

(b) The smooth function from before defined by

$$f(x) = \begin{cases} \exp(-1/x) & x > 0 \\ 0 & x \leq 0 \end{cases}$$

is *not* analytic. Indeed, suppose for contradiction it was. As all the derivatives are zero, its Taylor series would be $0 + 0x + 0x^2 + \cdots$. This Taylor series does *converge*, but not to the right value — as $f(\varepsilon) > 0$ for any $\varepsilon > 0$, contradiction.

Example (b) shows that if you have a function $f\colon \mathbb{R} \to \mathbb{R}$, then even knowing f is smooth and the full Taylor series at p, it's still impossible to recover any other values of f or deduce that f is analytic in any interval containing p.

However, it's at least true that:

Proposition 29.4.5 (Analytic at one point implies analytic on an interval)

Let $f\colon \mathbb{R} \to \mathbb{R}$ be smooth, and let $p \in \mathbb{R}$ be a point in the domain. Suppose that

- the Taylor series of f at p has radius of convergence $R > 0$; *and*
- that Taylor series actually does converge to the value $f(x)$ for every input $x \in (p-R, p+R)$ within the radius of convergence.

Then f is analytic on $(p-R, p+R)$.

This result is nontrivial because *a priori* we only know f is analytic at p; the result extends that to being analytic on the radius of convergence if $R > 0$. We'll use it for exp in just a moment, which is actually defined by a power series.

Like with differentiable functions:

Proposition 29.4.6 (All your usual closure properties for analytic functions)

The sums, products, compositions, nonzero quotients of analytic functions are analytic.

The upshot of this is that most of your usual functions that occur in nature, or even artificial ones like $f(x) = e^x + x\sin(x^2)$, will be analytic, hence describable locally by Taylor series.

§29.5 A definition of Euler's constant and exponentiation

We can actually give a definition of e^x using the tools we have now.

Definition 29.5.1. We define the map $\exp\colon \mathbb{R} \to \mathbb{R}$ by using the following power series, which has infinite radius of convergence:

$$\exp(x) = \sum_{n \geq 0} \frac{x^n}{n!}.$$

We then define Euler's constant as $e = \exp(1)$.

Question 29.5.2. Show that under this definition, $\exp' = \exp$. Also conclude from Proposition 29.4.5 that exp is analytic.

We are then settled with:

Proposition 29.5.3 (exp is multiplicative)

Under this definition,

$$\exp(x+y) = \exp(x)\exp(y).$$

Idea of proof. There is some subtlety here with switching the order of summation that we won't address. Modulo that:

$$\begin{aligned}\exp(x)\exp(y) &= \sum_{n\geq 0}\frac{x^n}{n!}\sum_{m\geq 0}\frac{y^m}{m!} = \sum_{n\geq 0}\sum_{m\geq 0}\frac{x^n}{n!}\frac{y^m}{m!} \\ &= \sum_{k\geq 0}\sum_{\substack{m+n=k\\ m,n\geq 0}}\frac{x^n y^m}{n!m!} = \sum_{k\geq 0}\sum_{\substack{m+n=k\\ m,n\geq 0}}\binom{k}{n}\frac{x^n y^m}{k!} \\ &= \sum_{k\geq 0}\frac{(x+y)^k}{k!} = \exp(x+y).\end{aligned}$$

□

Corollary 29.5.4 (exp is positive)

(a) We have $\exp(x) > 0$ for any real number x.

(b) The function exp is strictly increasing.

Proof. First

$$\exp(x) = \exp(x/2)^2 \geq 0$$

which shows exp is nonnegative. Also, $1 = \exp(0) = \exp(x)\exp(-x)$ implies $\exp(x) \neq 0$ for any x, proving (a).

(b) is just since $\exp'$ is strictly positive (racetrack principle). □

The log function then comes after.

Definition 29.5.5. We may define $\log\colon \mathbb{R}_{>0} \to \mathbb{R}$ to be the inverse function of exp.

Since its derivative is $1/x$ it is smooth; and then one may compute its coefficients to show it is analytic.

Note that this actually gives us a rigorous way to define a^r for any $a > 0$ and $r > 0$, namely

$$a^r := \exp\left(r \log a\right).$$

§29.6 This all works over complex numbers as well, except also complex analysis is heaven

We now mention that every theorem we referred to above holds equally well if we work over $\mathbb{C}$, with essentially no modifications.

- Power series are defined by $\sum_n a_n z^n$ with $a_n \in \mathbb{C}$, rather than $a_n \in \mathbb{R}$.
- The definition of radius of convergence R is unchanged! The series will converge if $|z| < R$.
- Differentiation still works great. (The definition of the derivative is unchanged.)
- Analytic still works great for functions $f\colon U \to \mathbb{C}$, with $U \subseteq \mathbb{C}$ open.

In particular, we can now even define complex exponentials, giving us a function

$$\exp\colon \mathbb{C} \to \mathbb{C}$$

since the power series still has $R = \infty$. More generally if $a > 0$ and $z \in \mathbb{C}$ we may still define

$$a^z \coloneqq \exp(z \log a).$$

(We still require the base a to be a positive real so that $\log a$ is defined, though. So this i^i issue is still there.)

However, if one tries to study calculus for complex functions as we did for the real case, in addition to most results carrying over, we run into a huge surprise:

If $f \colon \mathbb{C} \to \mathbb{C}$ is differentiable, it is analytic.

And this is just the beginning of the nearly unbelievable results that hold for complex analytic functions. But this is the part on real analysis, so you will have to read about this later!

§29.7 A few harder problems to think about

Problem 29A. Find the Taylor series of $\log(1 - x)$.

Problem 29B† (Euler formula)**.** Show that

$$\exp(i\theta) = \cos\theta + i \sin\theta$$

for any real number θ.

Problem 29C† (Taylor's theorem, Lagrange form)**.** Let $f \colon [a, b] \to \mathbb{R}$ be continuous and $n + 1$ times differentiable on (a, b). Define

$$P_n = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} \cdot (b - a)^k.$$

Prove that there exists $\xi \in (a, b)$ such that

$$f^{(n+1)}(\xi) = (n + 1)! \cdot \frac{f(b) - P_n}{(b - a)^{n+1}}.$$

This generalizes the mean value theorem (which is the special case $n = 0$, where $P_0 = f(a)$).

Problem 29D (Putnam 2018 A5)**.** Let $f \colon \mathbb{R} \to \mathbb{R}$ be smooth, and assume that $f(0) = 0$, $f(1) = 1$, and $f(x) \geq 0$ for every real number x. Prove that $f^{(n)}(x) < 0$ for some positive integer n and real number x.

Problem 29E. Let $f \colon \mathbb{R} \to \mathbb{R}$ be smooth. Suppose that for every point p, the Taylor series of f at p has positive radius of convergence. Prove that there exists at least one point at which f is analytic.

30 Riemann integrals

> "Trying to Riemann integrate discontinuous functions is kind of outdated."
> — Dennis Gaitsgory, [Ga15]

We will go ahead and define the Riemann integral, but we won't do very much with it. The reason is that the Lebesgue integral is basically better, so we will define it, check the fundamental theorem of calculus (or rather, leave it as a problem at the end of the chapter), and then always use Lebesgue integrals forever after.

§30.1 Uniform continuity

Prototypical example for this section: $f(x) = x^2$ is not uniformly continuous on $\mathbb{R}$, but functions on compact sets are always uniformly continuous.

Definition 30.1.1. Let $f \colon M \to N$ be a continuous map between two metric spaces. We say that f is **uniformly continuous** if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$d_M(p,q) < \delta \implies d_N(f(p), f(q)) < \varepsilon.$$

This difference is that given an $\varepsilon > 0$ we must specify a $\delta > 0$ which works for *every* choice p and q of inputs; whereas usually δ is allowed to depend on p and q. (Also, this definition can't be ported to a general topological space.)

Example 30.1.2 (Uniform continuity failure)

(a) The function $f \colon \mathbb{R} \to \mathbb{R}$ by $x \mapsto x^2$ is not uniformly continuous. Suppose we take $\varepsilon = 0.1$ for example. There is no δ such that if $|x-y| < \delta$ then $|x^2 - y^2| < 0.1$, since as x and y get large, the function f becomes increasingly sensitive to small changes.

(b) The function $(0,1) \to \mathbb{R}$ by $x \mapsto x^{-1}$ is not uniformly continuous.

(c) The function $\mathbb{R}_{>0} \to \mathbb{R}$ by $x \mapsto \sqrt{x}$ does turn out to be uniformly continuous (despite having unbounded derivatives!). Indeed, you can check that the assertion

$$|x-y| < \varepsilon^2 \implies |\sqrt{x} - \sqrt{y}| < \varepsilon$$

holds for any $x, y, \varepsilon > 0$.

The good news is that in the compact case all is well.

Theorem 30.1.3 (Uniform continuity free for compact spaces)

Let M be a compact metric space. Then any continuous map $f \colon M \to N$ is also uniformly continuous.

Proof. Assume for contradiction there is some bad $\varepsilon > 0$. Then taking $\delta = 1/n$, we find that for each integer n there exists points p_n and q_n which are within $1/n$ of each other,

but are mapped more than ε away from each other by f. In symbols, $d_M(p_n, q_n) \le 1/n$ but $d_N(f(p_n), f(q_n)) > \varepsilon$.

By compactness of M, we can find a convergent subsequence $p_{i_1}, p_{i_2}, \dots$ converging to some $x \in M$. Since the q_{i_n} is within $1/i_n$ of p_{i_n}, it ought to converge as well, to the same point $x \in M$. Then the sequences $f(p_{i_n})$ and $f(q_{i_n})$ should both converge to $f(x) \in N$, but this is impossible as they are always more than ε away from each other. □

This means for example that x^2 viewed as a continuous function $[0,1] \to \mathbb{R}$ is automatically uniformly continuous. Man, isn't compactness great?

§30.2 Dense sets and extension

Prototypical example for this section: Functions from $\mathbb{Q} \to N$ extend to $\mathbb{R} \to N$ if they're uniformly continuous and N is complete. See also counterexamples below.

Definition 30.2.1. Let S be a subset (or subspace) of a topological space X. Then we say that S is **dense** if every open subset of X contains a point of S.

Example 30.2.2 (Dense sets)

(a) $\mathbb{Q}$ is dense in $\mathbb{R}$.

(b) In general, any metric space M is dense in its completion $\overline{M}$.

Dense sets lend themselves to having functions completed. The idea is that if I have a continuous function $f \colon \mathbb{Q} \to N$, for some metric space N, then there should be *at most* one way to extend it to a function $\widetilde{f} \colon \mathbb{R} \to N$. For we can approximate each rational number by real numbers: if I know $f(1)$, $f(1.4)$, $f(1.41)$, … $\widetilde{f}(\sqrt{2})$ had better be the limit of this sequence. So it is certainly unique.

However, there are two ways this could go wrong:

Example 30.2.3 (Non-existence of extension)

(a) It could be that N is not complete, so the limit may not even exist in N. For example if $N = \mathbb{Q}$, then certainly there is no way to extend even the identity function $f \colon \mathbb{Q} \to N$ to a function $\widetilde{f} \colon \mathbb{R} \to N$.

(b) Even if N was complete, we might run into issues where f explodes. For example, let $N = \mathbb{R}$ and define

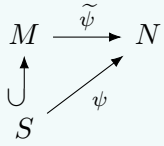
$$f(x) = \frac{1}{x - \sqrt{2}} \qquad f \colon \mathbb{Q} \to \mathbb{R}.$$

There is also no way to extend this due to the explosion of f near $\sqrt{2} \notin \mathbb{Q}$, which would cause $\widetilde{f}(\sqrt{2})$ to be undefined.

However, the way to fix this is to require f to be uniformly continuous, and in that case we do get a unique extension.

Theorem 30.2.4 (Extending uniformly continuous functions)

Let M be a metric space, N a *complete* metric space, and S a dense subspace of M. Suppose $\psi\colon S \to N$ is a *uniformly* continuous function. Then there exists a unique continuous function $\widetilde{\psi}\colon M \to N$ such that the diagram



commutes.

Outline of proof. As mentioned in the discussion, each $x \in M$ can be approximated by a sequence $x_1, x_2, \ldots$ in S with $x_i \to x$. The two main hypotheses, completeness and uniform continuity, are now used:

> **Exercise 30.2.5.** Prove that $\psi(x_1), \psi(x_2), \ldots$ converges in N by using uniform continuity to show that it is Cauchy, and then appealing to completeness of N.

Hence we define $\widetilde{\psi}(x)$ to be the limit of that sequence; this doesn't depend on the choice of sequence, and one can use sequential continuity to show $\widetilde{\psi}$ is continuous. □

§30.3 Defining the Riemann integral

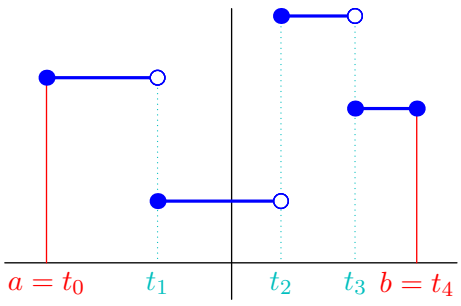
Extensions will allow us to define the Riemann integral. I need to introduce a bit of notation so bear with me.

Definition 30.3.1. Let $[a,b]$ be a closed interval.

- We let $C^0([a,b])$ denote the set of continuous functions on $[a,b] \to \mathbb{R}$.
- We let $R([a,b])$ denote the set of **rectangle functions** on $[a,b] \to \mathbb{R}$. These functions which are constant on the intervals $[t_0,t_1)$, $[t_1,t_2)$, $[t_2,t_3)$, $\ldots$, $[t_{n-2},t_{n-1})$, and also $[t_{n-1},t_n]$, for some $a = t_0 < t_1 < t_2 < \cdots < t_n = b$.
- We let $M([a,b]) = C^0([a,b]) \cup R([a,b])$.

Warning: only $C^0([a,b])$ is common notation, and the other two are made up.

See picture below for a typical a rectangle function. (It is irritating that we have to officially assign a single value to each t_i, even though there are naturally two values we want to use, and so we use the convention of letting the left endpoint be closed).



Definition 30.3.2. We can impose a metric on $M([a,b])$ by defining

$$d(f,g) = \sup_{x\in[a,b]} |f(x)-g(x)|.$$

Now, there is a natural notion of integral for rectangle functions: just sum up the obvious rectangles! Officially, this is the expression

$$f(a)(t_1-a) + f(t_1)(t_2-t_1) + +f(t_2)\,(t_3-t_2) + \cdots + f(t_n)\,(b-t_n).$$

We denote this function by

$$\Sigma\colon R([a,b]) \to \mathbb{R}.$$

Theorem 30.3.3 (The Riemann integral)

There exists a unique continuous map

$$\textstyle\int_a^b\colon M([a,b]) \to \mathbb{R}$$

such that the diagram

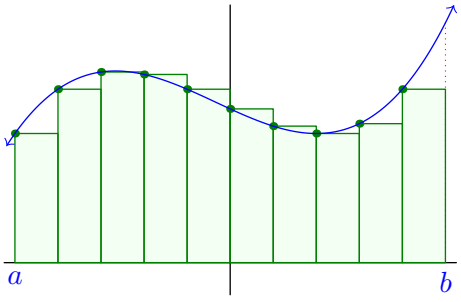
$$\begin{array}{ccc} M([a,b]) & \xrightarrow{\int_a^b} & \mathbb{R} \\ \uparrow & \nearrow_{\Sigma} & \\ R([a,b]) & & \end{array}$$

commutes.

Proof. We want to apply the extension theorem, so we just have to check a few things:

- We claim $R([a,b])$ is a dense subset of $M([a,b])$. In other words, for any continuous $f\colon [a,b]\to\mathbb{R}$ and $\varepsilon > 0$, we want there to exist a rectangle function that approximates f within ε.

  This follows by uniform continuity. We know there exists a $\delta > 0$ such that whenever $|x-y| < \delta$ we have $|f(x)-f(y)| < \varepsilon$. So as long as we select a rectangle function whose rectangles have width less than δ, and such that the upper-left corner of each rectangle lies on the graph of f, then we are all set.



- The "add-the-rectangles" map $\Sigma\colon R([a,b])\to\mathbb{R}$ is *uniformly* continuous. Actually this is pretty obvious: if two rectangle functions f and g have $d(f,g) < \varepsilon$, then $d(\Sigma f, \Sigma g) < \varepsilon(b-a)$.
- $\mathbb{R}$ is complete. □

§30.4 Meshes

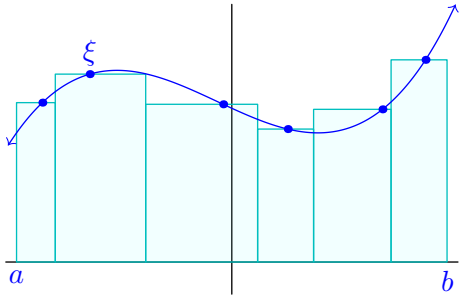
The above definition might seem fantastical, overcomplicated, hilarious, or terrible, depending on your taste. But if you unravel it, it's really the picture you are used to. What we have done is taking every continuous function $f\colon [a,b] \to \mathbb{R}$ and showed that it can be approximated by a rectangle function (which we phrased as a dense inclusion). Then we added the area of the rectangles. Nonetheless, we will give a definition that's more like what you're used to seeing in other places.

Definition 30.4.1. A *tagged partition* P of $[a,b]$ consists of a partition of $[a,b]$ into n intervals, with a point ξ_i in the nth interval, denoted

$$a = t_0 < t_1 < t_2 < \cdots < t_n = b \qquad \text{and} \qquad \xi_i \in [t_{i-1}, t_i] \quad \forall\, 1 \le i \le n.$$

The *mesh* of P is the width of the longest interval, i.e. $\max_i (t_i - t_{i-1})$.

Of course the point of this definition is that we add the rectangles, but the ξ_i are the sample points.



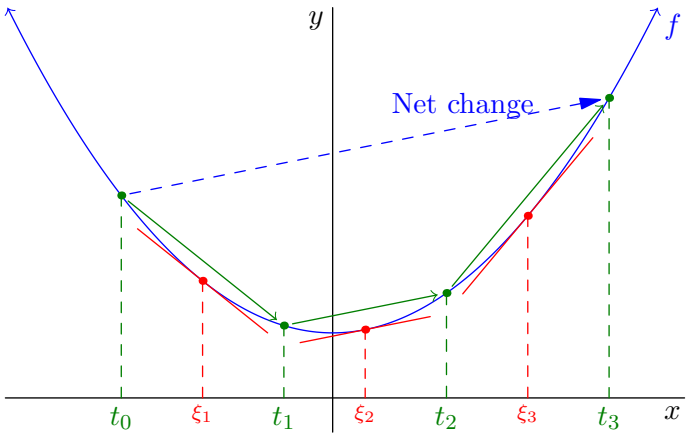
Theorem 30.4.2 (Riemann integral)
Let $f\colon [a,b] \to \mathbb{R}$ be continuous. Then

$$\int_a^b f(x)\,dx = \lim_{\substack{P \text{ tagged partition} \\ \text{mesh } P \to 0}} \left(\sum_{i=1}^{n} f(\xi_i)(t_i - t_{i-1}) \right).$$

Here the limit means that we can take any sequence of partitions whose mesh approaches zero.

Proof. The right-hand side corresponds to the areas of some rectangle functions g_1, g_2, ... with increasingly narrow rectangles. As in the proof Theorem 30.3.3, as the meshes of those rectangles approaches zero, by uniform continuity, we have $d(f, g_n) \to 0$ as well. Thus by continuity in the diagram of Theorem 30.3.3, we get $\lim_n \Sigma(g_n) = \int(f)$ as needed. □

Combined with the mean value theorem, this can be used to give a short proof of the fundamental theorem of calculus for functions f with a continuous derivative. The idea is that for any choice of partition $a \le t_0 < t_1 < t_2 < \cdots < t_n \le b$, using the Mean Value Theorem it should be possible to pick ξ_i in each interval to match with the slope of the secant: at which point the areas sum to the total change in f. We illustrate this situation with three points, and invite the reader to fill in the details as Problem 30B*.



One quick note is that although I've only defined the Riemann integral for continuous functions, there ought to be other functions for which it exists (including "piecewise continuous functions" for example, or functions "continuous almost everywhere"). The relevant definition is:

Definition 30.4.3. If $f\colon [a,b] \to \mathbb{R}$ is a function which is not necessarily continuous, but for which the limit

$$\lim_{\substack{P \text{ tagged partition} \\ \text{mesh } P \to 0}} \left(\sum_{i=1}^{n} f(\xi_i)(t_i - t_{i-1}) \right).$$

exists anyways, then we say f is **Riemann integrable** on $[a,b]$ and define its value to be that limit $\int_a^b f(x)\,dx$.

We won't really use this definition much, because we will see that every Riemann integrable function is Lebesgue integrable, and the Lebesgue integral is better.

Example 30.4.4 (Your AP calculus returns)

We had better mention that Problem 30B* implies that we can compute Riemann integrals in practice, although most of you may already know this from high-school calculus. For example, on the interval $(1,4)$, the derivative of the function $F(x) = \frac{1}{3}x^3$ is $F'(x) = x^2$. As $f(x) = x^2$ is a continuous function $f\colon [1,4] \to \mathbb{R}$, we get

$$\int_1^4 x^2\,dx = F(4) - F(1) = \frac{64}{3} - \frac{1}{3} = 21.$$

Note that we could also have picked $F(x) = \frac{1}{3}x^3 + 2019$; the function F is unique up to shifting, and this constant cancels out when we subtract. This is why it's common in high school to (really) abuse notation and write $\int x^2\,dx = \frac{1}{3}x^3 + C$.

§30.5 A few harder problems to think about

Problem 30A. Let $f\colon (a,b) \to \mathbb{R}$ be differentiable and assume f' is bounded. Show that f is uniformly continuous.

Problem 30B* (Fundamental theorem of calculus)**.** Let $f\colon [a,b] \to \mathbb{R}$ be continuous, differentiable on (a,b), and assume the derivative f' extends to a continuous function $f'\colon [a,b] \to \mathbb{R}$. Prove that

$$\int_a^b f'(x)\,dx = f(b) - f(a).$$

Problem 30C* (Improper integrals)**.** For each real number $r > 0$, evaluate the limit[1]

$$\lim_{\varepsilon \to 0^+} \int_\varepsilon^1 \frac{1}{x^r}\,dx$$

or show it does not exist.

This can intuitively be thought of as the "improper" integral $\int_0^1 x^{-r}\,dx$; it doesn't make sense in our original definition since we did not (and cannot) define the integral over the non-compact $(0, 1]$ but we can still consider the integral over $[\varepsilon, 1]$ for any $\varepsilon > 0$.

Problem 30D. Show that

$$\lim_{n\to\infty} \left(\frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n} \right) = \log 2.$$

[1]If you are not familiar with the notation $\varepsilon \to 0^+$, you can replace ε with $1/M$ for $M > 0$, and let $M \to \infty$ instead.

IX

Complex Analysis

Part IX: Contents

31 Holomorphic functions

Throughout this chapter, we denote by U an open subset of the complex plane, and by Ω an open subset which is also simply connected. The main references for this chapter were [Ya12; Ba10].

§31.1 The nicest functions on earth

In high school you were told how to differentiate and integrate real-valued functions. In this chapter on complex analysis, we'll extend it to differentiation and integration of complex-valued functions.

Big deal, you say. Calculus was boring enough. Why do I care about complex calculus?

Perhaps it's easiest to motivate things if I compare real analysis to complex analysis. In real analysis, your input lives inside the real line $\mathbb{R}$. This line is not terribly discerning – you can construct a lot of unfortunate functions. Here are some examples.

Example 31.1.1 (Optional: evil real functions)

You can skim over these very quickly: they're only here to make a point.

(a) The **Devil's Staircase** (or Cantor function) is a continuous function $H\colon [0,1] \to [0,1]$ which has derivative zero "almost everywhere", yet $H(0) = 0$ and $H(1) = 1$.

(b) The **Weierstraß function**

$$x \mapsto \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n \cos\left(2015^n \pi x\right)$$

is continuous *everywhere* but differentiable *nowhere*.

(c) The function

$$x \mapsto \begin{cases} x^{100} & x \geq 0 \\ -x^{100} & x < 0 \end{cases}$$

has the first 99 derivatives but not the 100th one.

(d) If a function has all derivatives (we call these **smooth** functions), then it has a Taylor series. But for real functions that Taylor series might still be wrong. The function

$$x \mapsto \begin{cases} e^{-1/x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

has derivatives at every point. But if you expand the Taylor series at $x = 0$, you get $0 + 0x + 0x^2 + \cdots$, which is wrong for *any* $x > 0$ (even $x = 0.0001$).

Let's even put aside the pathology. If I tell you the value of a real smooth function on the interval $[-1, 1]$, that still doesn't tell you anything about the function as a whole. It could be literally anything, because it's somehow possible to "fuse together" smooth functions.

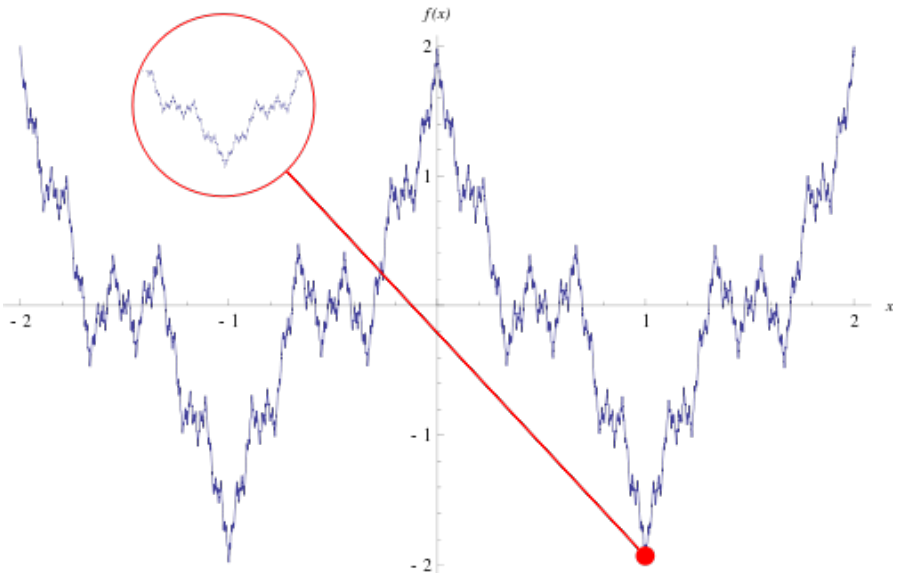


Figure 31.1: The Weierstraß Function (image from [Ee]).

So what about complex functions? If you consider them as functions $\mathbb{R}^2 \to \mathbb{R}^2$, you now have the interesting property that you can integrate along things that are not line segments: you can write integrals across curves in the plane. But $\mathbb{C}$ has something more: it is a *field*, so you can *multiply* and *divide* two complex numbers.

So we restrict our attention to differentiable functions called *holomorphic functions*. It turns out that the multiplication on $\mathbb{C}$ makes all the difference. The primary theme in what follows is that holomorphic functions are *really, really nice*, and that knowing tiny amounts of data about the function can determine all its values.

The two main highlights of this chapter, from which all other results are more or less corollaries:

- Contour integrals of loops are always zero.
- A holomorphic function is essentially given by its Taylor series; in particular, single-differentiable implies infinitely differentiable. Thus, holomorphic functions behave quite like polynomials.

Some of the resulting corollaries:

- It'll turn out that knowing the values of a holomorphic function on the boundary of the unit circle will tell you the values in its interior.
- Knowing the values of the function at 1, $\frac{1}{2}$, $\frac{1}{3}$, ... are enough to determine the whole function!
- Bounded holomorphic functions $\mathbb{C} \to \mathbb{C}$ must be constant.
- And more...

As [Pu02] writes: "Complex analysis is the good twin and real analysis is the evil one: beautiful formulas and elegant theorems seem to blossom spontaneously in the complex domain, while toil and pathology rule the reals".

§31.2 Complex differentiation

Prototypical example for this section: Polynomials are holomorphic; $\overline{z}$ is not.

Let $f\colon U \to \mathbb{C}$ be a complex function. Then for some $z_0 \in U$, we define the **derivative** at z_0 to be

$$\lim_{h \to 0} \frac{f(z_0+h) - f(z_0)}{h}.$$

Note that this limit may not exist; when it does we say f is **differentiable** at z_0.

What do I mean by a "complex" limit $h \to 0$? It's what you might expect: for every $\varepsilon > 0$ there should be a $\delta > 0$ such that

$$0 < |h| < \delta \implies \left| \frac{f(z_0+h) - f(z_0)}{h} - L \right| < \varepsilon.$$

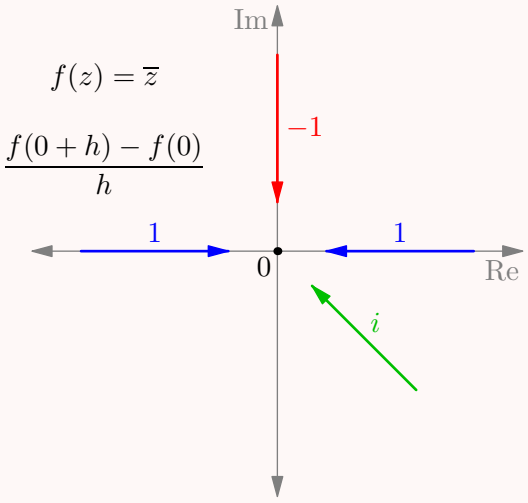
If you like topology, you are encouraged to think of this in terms of open neighborhoods in the complex plane. (This is why we require U to be open: it makes it possible to take δ-neighborhoods in it.)

But note that having a complex derivative is actually much stronger than a real function having a derivative. In the real line, h can only approach zero from below and above, and for the limit to exist we need the "left limit" to equal the "right limit". But the complex numbers form a *plane*: h can approach zero from many directions, and we need all the limits to be equal.

Example 31.2.1 (Important: conjugation is *not* holomorphic)
Let $f(z) = \overline{z}$ be complex conjugation, $f\colon \mathbb{C} \to \mathbb{C}$. This function, despite its simple nature, is not holomorphic! Indeed, at $z = 0$ we have,

$$\frac{f(h) - f(0)}{h} = \frac{\overline{h}}{h}.$$

This does not have a limit as $h \to 0$, because depending on "which direction" we approach zero from we have different values.



If a function $f\colon U \to \mathbb{C}$ is complex differentiable at all the points in its domain it is called **holomorphic**. In the special case of a holomorphic function with domain $U = \mathbb{C}$,

we call the function **entire**.[1]

Example 31.2.2 (Examples of holomorphic functions)
In all the examples below, the derivative of the function is the same as in their real analogues (e.g. the derivative of e^z is e^z).

(a) Any polynomial $z \mapsto z^n + c_{n-1}z^{n-1} + \cdots + c_0$ is holomorphic.

(b) The complex exponential $\exp\colon x+yi \mapsto e^x(\cos y + i \sin y)$ can be shown to be holomorphic.

(c) sin and cos are holomorphic when extended to the complex plane by $\cos z = \frac{e^{iz}+e^{-iz}}{2}$ and $\sin z = \frac{e^{iz}-e^{-iz}}{2i}$.

(d) As usual, the sum, product, chain rules and so on apply, and hence **sums, products, nonzero quotients, and compositions of holomorphic functions are also holomorphic**.

You are welcome to try and prove these results, but I won't bother to do so.

§31.3 Contour integrals

Prototypical example for this section: $\oint_\gamma z^m\,dz$ *around the unit circle.*

In the real line we knew how to integrate a function across a line segment $[a,b]$: essentially, we'd "follow along" the line segment adding up the values of f we see to get some area. Unlike in the real line, in the complex plane we have the power to integrate over arbitrary paths: for example, we might compute an integral around a unit circle. A contour integral lets us formalize this.

First of all, if $f\colon \mathbb{R} \to \mathbb{C}$ and $f(t) = u(t) + iv(t)$ for $u, v \in \mathbb{R}$, we can define an integral $\int_a^b$ by just adding the real and imaginary parts:

$$\int_a^b f(t)\,dt = \left(\int_a^b u(t)\,dt\right) + i\left(\int_a^b v(t)\,dt\right).$$

Now let $\alpha\colon [a,b] \to \mathbb{C}$ be a path, thought of as a complex differentiable[2] function. Such a path is called a **contour**, and we define its **contour integral** by

$$\oint_\alpha f(z)\,dz = \int_a^b f(\alpha(t)) \cdot \alpha'(t)\,dt.$$

You can almost think of this as a u-substitution (which is where the α' comes from). In particular, it turns out this integral does not depend on how α is "parametrized": a circle given by

$$[0, 2\pi] \to \mathbb{C}\colon t \mapsto e^{it}$$

and another circle given by

$$[0, 1] \to \mathbb{C}\colon t \mapsto e^{2\pi i t}$$

[1] Sorry, I know the word "holomorphic" sounds so much cooler. I'll try to do things more generally for that sole reason.

[2] This isn't entirely correct here: you want the path α to be continuous and mostly differentiable, but you allow a finite number of points to have "sharp bends"; in other words, you can consider paths which are combinations of n smooth pieces. But for this we also require that α has "bounded length".

and yet another circle given by

$$[0,1] \to \mathbb{C} \colon t \mapsto e^{2\pi i t^5}$$

will all give the same contour integral, because the paths they represent have the same geometric description: "run around the unit circle once".

In what follows I try to use α for general contours and γ in the special case of loops.

Let's see an example of a contour integral.

Theorem 31.3.1

Take $\gamma \colon [0, 2\pi] \to \mathbb{C}$ to be the unit circle specified by

$$t \mapsto e^{it}.$$

Then for any integer m, we have

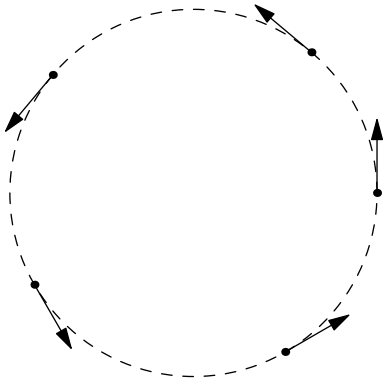
$$\oint_\gamma z^m \, dz = \begin{cases} 2\pi i & m = -1 \\ 0 & \text{otherwise} \end{cases}$$

Proof. The derivative of e^{it} is ie^{it}. So, by definition the answer is the value of

$$\begin{aligned} \int_0^{2\pi} (e^{it})^m \cdot (ie^{it}) \, dt &= \int_0^{2\pi} i(e^{it})^{1+m} \, dt \\ &= i \int_0^{2\pi} \cos[(1+m)t] + i \sin[(1+m)t] \, dt \\ &= -\int_0^{2\pi} \sin[(1+m)t] \, dt + i \int_0^{2\pi} \cos[(1+m)t] \, dt. \end{aligned}$$

This is now an elementary calculus question. One can see that this equals $2\pi i$ if $m = -1$ and otherwise the integrals vanish. □

Let me try to explain why this intuitively ought to be true for $m = 0$. In that case we have $\oint_\gamma 1 \, dz$. So as the integral walks around the unit circle, it "sums up" all the tangent vectors at every point (that's the direction it's walking in), multiplied by 1. And given the nice symmetry of the circle, it should come as no surprise that everything cancels out. The theorem says that even if we multiply by z^m for $m \neq -1$, we get the same cancellation.



Definition 31.3.2. Given $\alpha \colon [0,1] \to \mathbb{C}$, we denote by $\overline{\alpha}$ the "backwards" contour $\overline{\alpha}(t) = \alpha(1-t)$.

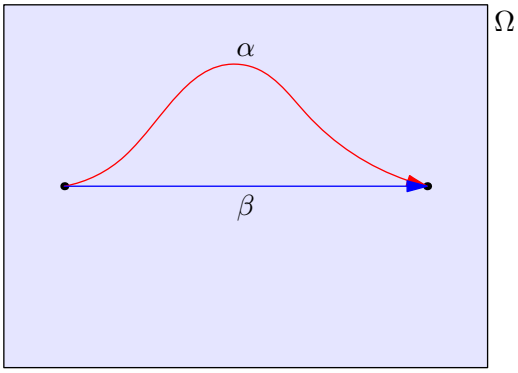
Question 31.3.3. What's the relation between $\oint_\alpha f\,dz$ and $\oint_{\overline{\alpha}} f\,dz$? Prove it.

This might seem a little boring. Things will get really cool really soon, I promise.

§31.4 Cauchy-Goursat theorem

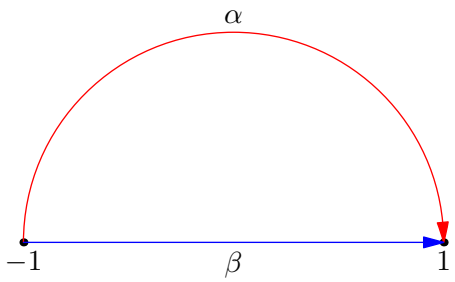
Prototypical example for this section: $\oint_\gamma z^m\,dz = 0$ *for* $m \ge 0$. *But if* $m < 0$, *Cauchy's theorem does not apply.*

Let $\Omega \subseteq \mathbb{C}$ be simply connected (for example, $\Omega = \mathbb{C}$), and consider two paths α, β with the same start and end points.



What's the relation between $\oint_\alpha f(z)\,dz$ and $\oint_\beta f(z)\,dz$? You might expect there to be some relation between them, considering that the space Ω is simply connected. But you probably wouldn't expect there to be *much* of a relation.

As a concrete example, let $\Psi\colon \mathbb{C} \to \mathbb{C}$ be the function $z \mapsto z - \operatorname{Re}[z]$ (for example, $\Psi(2015 + 3i) = 3i$). Let's consider two paths from -1 to 1. Thus β is walking along the real axis, and α which follows an upper semicircle.



Obviously $\oint_\beta \Psi(z)\,dz = 0$. But heaven knows what $\oint_\alpha \Psi(z)\,dz$ is supposed to equal. We can compute it now just out of non-laziness. If you like, you are welcome to compute it yourself (it's a little annoying but not hard). If I myself didn't mess up, it is

$$\oint_\alpha \Psi(z)\,dz = -\oint_{\overline{\alpha}} \Psi(z)\,dz = -\int_0^\pi (i\sin(t)) \cdot ie^{it}\,dt = \frac{1}{2}\pi i$$

which in particular is not zero.

But somehow Ψ is not a really natural function. It's not respecting any of the nice, multiplicative structure of $\mathbb{C}$ since it just rudely lops off the real part of its inputs. More precisely,

Question 31.4.1. Show that $\Psi(z) = z - \operatorname{Re}[z]$ is not holomorphic. (Hint: $\overline{z}$ is not holomorphic.)

Now here's a miracle: for holomorphic functions, the two integrals are *always equal*. Equivalently, (by considering α followed by $\overline{\beta}$) contour integrals of loops are always zero. This is the celebrated Cauchy-Goursat theorem (also called the Cauchy integral theorem, but later we'll have a "Cauchy Integral Formula" so blah).

Theorem 31.4.2 (Cauchy-Goursat theorem)
Let γ be a loop, and $f \colon \Omega \to \mathbb{C}$ a holomorphic function where Ω is open in $\mathbb{C}$ and simply connected. Then
$$\oint_\gamma f(z)\,dz = 0.$$

Remark 31.4.3 (Sanity check) — This might look surprising considering that we saw $\oint_\gamma z^{-1}\,dz = 2\pi i$ earlier. The subtlety is that z^{-1} is not even defined at $z = 0$. On the other hand, the function $\mathbb{C} \setminus \{0\} \to \mathbb{C}$ by $z \mapsto \frac{1}{z}$ *is* holomorphic! The defect now is that $\Omega = \mathbb{C} \setminus \{0\}$ is not simply connected. So the theorem passes our sanity checks, albeit barely.

The typical proof of Cauchy's Theorem assumes additionally that the partial derivatives of f are continuous and then applies the so-called Green's theorem. But it was Goursat who successfully proved the fully general theorem we've stated above, which assumed only that f was holomorphic.

Anyways, the theorem implies that $\oint_\gamma z^m\,dz = 0$ when $m \geq 0$. So much for all our hard work earlier. But so far we've only played with circles. This theorem holds for *any* contour which is a loop. So what else can we do?

§31.5 Cauchy's integral theorem

We now present a stunning application of Cauchy-Goursat, a "representation theorem": essentially, it says that values of f inside a disk are determined by just the values on the boundary! In fact, we even write down the exact formula. As [Ya12] says, "any time a certain type of function satisfies some sort of representation theorem, it is likely that many more deep theorems will follow." Let's pull back the curtain:

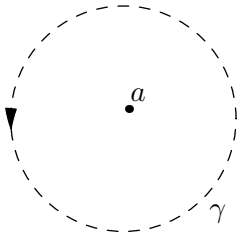
Theorem 31.5.1 (Cauchy's integral formula)
Let $\gamma \colon [0, 2\pi] \to \mathbb{C}$ be a circle in the plane given by $t \mapsto Re^{it}$, which bounds a disk D. Suppose $f \colon U \to \mathbb{C}$ is holomorphic such that U contains the circle and its interior. Then for any point a in the interior of D, we have
$$f(a) = \frac{1}{2\pi i} \oint_\gamma \frac{f(z)}{z - a}\,dz.$$

Note that we don't require U to be simply connected, but the reason is pretty silly: we're only going to ever integrate f over D, which is an open disk, and hence the disk is simply connected anyways.

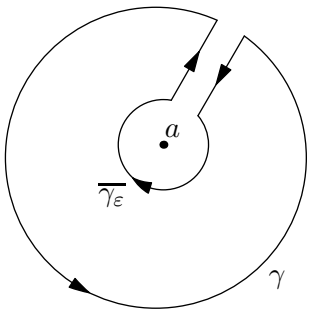
The presence of $2\pi i$, which you saw earlier in the form $\oint_{\text{circle}} z^{-1}\,dz$, is no accident. In fact, that's the central result we're going to use to prove the result.

> **Remark 31.5.2 —** With the introduction of meromorphic functions next chapter, we will see how to intuitively derive this formula in Remark 32.3.5.

Proof. There are several proofs out there, but I want to give the one that really draws out the power of Cauchy's theorem. Here's the picture we have: there's a point a sitting inside a circle γ, and we want to get our hands on the value $f(a)$.



We're going to do a trick: construct a **keyhole contour** $\Gamma_{\delta,\varepsilon}$ which has an outer circle γ, plus an inner circle $\overline{\gamma_\varepsilon}$, which is a circle centered at a with radius ε, running clockwise (so that γ_ε runs counterclockwise). The "width" of the corridor is δ. See picture:



Hence $\Gamma_{\delta,\varepsilon}$ consists of four smooth curves.

> **Question 31.5.3.** Draw a *simply connected* open set Ω which contains the entire $\Gamma_{\delta,\varepsilon}$ but does not contain the point a.

The function $\frac{f(z)}{z-a}$ manages to be holomorphic on all of Ω. Thus Cauchy's theorem applies and tells us that

$$0 = \oint_{\Gamma_{\delta,\varepsilon}} \frac{f(z)}{z-a}\,dz.$$

As we let $\delta \to 0$, the two walls of the keyhole will cancel each other (because f is continuous, and the walls run in opposite directions). So taking the limit as $\delta \to 0$, we are left with just γ and γ_ε, which (taking again orientation into account) gives

$$\oint_\gamma \frac{f(z)}{z-a}\,dz = -\oint_{\overline{\gamma_\varepsilon}} \frac{f(z)}{z-a}\,dz = \oint_{\gamma_\varepsilon} \frac{f(z)}{z-a}\,dz.$$

Thus **we've managed to replace γ with a much smaller circle γ_ε centered around** a, and the rest is algebra.

To compute the last quantity, write

$$\begin{aligned}\oint_{\gamma_\varepsilon} \frac{f(z)}{z-a}\,dz &= \oint_{\gamma_\varepsilon} \frac{f(z)-f(a)}{z-a}\,dz + f(a)\cdot\oint_{\gamma_\varepsilon}\frac{1}{z-a}\,dz \\ &= \oint_{\gamma_\varepsilon} \frac{f(z)-f(a)}{z-a}\,dz + 2\pi i f(a).\end{aligned}$$

where we've used Theorem 31.3.1. Thus, all we have to do is show that

$$\oint_{\gamma_\varepsilon} \frac{f(z)-f(a)}{z-a}\,dz = 0.$$

For this we can basically use the weakest bound possible, the so-called ML lemma which I'll cite without proof: it says "bound the function everywhere by its maximum".

Lemma 31.5.4 (ML estimation lemma)

Let f be a holomorphic function and α a path. Suppose $M = \max_{z \text{ on } \alpha} |f(z)|$, and let L be the length of α. Then

$$\left|\oint_\alpha f(z)\,dz\right| \le ML.$$

(This is straightforward to prove if you know the definition of length: $L = \int_a^b |\alpha'(t)|\,dt$, where $\alpha\colon [a,b] \to \mathbb{C}$.)

Anyways, as $\varepsilon \to 0$, the quantity $\frac{f(z)-f(a)}{z-a}$ approaches $f'(a)$, and so for small enough ε (i.e. z close to a) there's some upper bound M. Yet the length of γ_ε is the circumference $2\pi\varepsilon$. So the ML lemma says that

$$\left|\oint_{\gamma_\varepsilon} \frac{f(z)-f(a)}{z-a}\right| \le 2\pi\varepsilon \cdot M \to 0$$

as desired. □

§31.6 Holomorphic functions are analytic

Prototypical example for this section: Imagine a formal series $\sum_k c_k x^k$!

In the setup of the previous problem, we have a circle $\gamma\colon [0,2\pi] \to \mathbb{C}$ and a holomorphic function $f\colon U \to \mathbb{C}$ which contains the disk D. We can write

$$\begin{aligned}f(a) &= \frac{1}{2\pi i}\oint_\gamma \frac{f(z)}{z-a}\,dz \\ &= \frac{1}{2\pi i}\oint_\gamma \frac{f(z)/z}{1-\frac{a}{z}}\,dz \\ &= \frac{1}{2\pi i}\oint_\gamma f(z)/z \cdot \sum_{k\ge 0}\left(\frac{a}{z}\right)^k\,dz\end{aligned}$$

You can prove (using the so-called Weierstrass M-test) that the summation order can be switched:

$$\begin{aligned} f(a) &= \frac{1}{2\pi i} \sum_{k \ge 0} \oint_\gamma \frac{f(z)}{z} \cdot \left(\frac{a}{z}\right)^k \, dz \\ &= \frac{1}{2\pi i} \sum_{k \ge 0} \oint_\gamma a^k \cdot \frac{f(z)}{z^{k+1}} \, dz \\ &= \sum_{k \ge 0} \left(\frac{1}{2\pi i} \oint_\gamma \frac{f(z)}{z^{k+1}} \, dz \right) a^k. \end{aligned}$$

Letting $c_k = \frac{1}{2\pi i} \oint_\gamma \frac{f(z)}{z^{k+1}} \, dz$, and noting this is independent of a, this is

$$f(a) = \sum_{k \ge 0} c_k a^k$$

and that's the miracle: holomorphic functions are given by a **Taylor series**! This is one of the biggest results in complex analysis. Moreover, if one is willing to believe that we can take the derivative k times, we obtain

$$c_k = \frac{f^{(k)}(0)}{k!}$$

and this gives us $f^{(k)}(0) = k! \cdot c_k$.

Naturally, we can do this with any circle (not just one centered at zero). So let's state the full result below, with arbitrary center p.

Theorem 31.6.1 (Cauchy's differentiation formula)

Let $f : U \to \mathbb{C}$ be a holomorphic function and let D be a disk centered at point p bounded by a circle γ. Suppose D is contained inside U. Then f is given everywhere in D by a Taylor series

$$f(z) = c_0 + c_1(z-p) + c_2(z-p)^2 + \cdots$$

where

$$c_k = \frac{f^k(p)}{k!} = \frac{1}{2\pi i} \oint_\gamma \frac{f(w-p)}{(w-p)^{k+1}} \, dw$$

In particular,

$$f^{(k)}(p) = k! c_k = \frac{k!}{2\pi i} \oint_\gamma \frac{f(w-p)}{(w-p)^{k+1}} \, dw.$$

Most importantly,

Over any disk, a holomorphic function is given exactly by a Taylor series.

This establishes a result we stated at the beginning of the chapter: that a function being complex differentiable once means it is not only infinitely differentiable, but in fact equal to its Taylor series.

Remark 31.6.2 — If you're willing to assume this, you can see why Cauchy-Goursat theorem should be true: assuming

$$f(z) = c_0 + c_1 z + c_2 z^2 + \cdots$$

then, with γ the unit circle,

$$\begin{aligned}\oint_\gamma f(z)\,dz &= \oint_\gamma c_0 + c_1 z + c_2 z^2 + \ldots\,dz \\ &= \left(\oint_\gamma c_0\,dz\right) + \left(\oint_\gamma c_1 z\,dz\right) + \left(\oint_\gamma c_2 z^2\,dz\right) + \cdots\end{aligned}$$

We have already proven that each $\oint_\gamma z^m\,dz = 0$, so the sum ought to be 0 as well. Of course the argument is not completely rigorous, it exchanges the integration and the infinite sum without justification.

Remark 31.6.3 — You can see where the term $\frac{f(w-p)}{(w-p)^{k+1}}$ comes from in Remark 32.3.5. It is very intuitive that even if you forget it, you can derive it yourself as well!

I should maybe emphasize a small subtlety of the result: the Taylor series centered at p is only valid in a disk centered at p which lies entirely in the domain U. If $U = \mathbb{C}$ this is no issue, since you can make the disk big enough to accommodate any point you want. It's more subtle in the case that U is, for example, a square; you can't cover the entire square with a disk centered at some point without going outside the square. However, since U is open we can at any rate at least find some open neighborhood for which the Taylor series is correct – in stark contrast to the real case. Indeed, as you'll see in the problems, the existence of a Taylor series is incredibly powerful.

§31.7 Optional: Proof that holomorphic functions are analytic

It is recommended to read the next chapter first to understand the origin of the term $\frac{f(w-p)}{(w-p)^{k+1}}$ in Cauchy's differentiation formula above.

Each step of the proof is quite intuitive, if not a bit long. The outline is:

- We pretend that the function f is analytic. (Yes, this is not circular reasoning!)
- We use Cauchy's differentiation formula to write down a power series:[3]

$$c_0 + c_1 z + c_2 z^2 + \cdots$$

- We prove that the power series coincide with f using Cauchy-Goursat theorem.
- Note that the statement "f is analytic" literally means "for every $k \geq 0$, then $f^{(k)}$ is differentiable". So, we write down a power series for $f^{(k)}$, and show that it is differentiable. (We already did this for the real case in Proposition 29.4.5.)

[3] Assume $0 \in U$.

§31.7.i Proof of Cauchy-Goursat theorem

Suppose f is holomorphic i.e. differentiable. We wish to prove $\oint_\gamma f\,dz = 0$.

How may we attack this problem? Looking at the conclusion, we may want to stare at some function where $\oint_\gamma f\,dz \neq 0$.

We readily got an example from the previous chapter: $f(z) = \frac{1}{z}$.

> **Question 31.7.1.** What part of the hypothesis does not hold?

In any case, you see the problem is it's because f has a singularity at 0 (even though we haven't formally defined what a singularity is yet). So, we try to prove the contrapositive:

> **Theorem 31.7.2**
> Suppose $\oint_\gamma f\,dz \neq 0$. Then something weird happens to f somewhere inside γ.

(For arbitrary loops, it gets a bit more difficult, however. What does "inside γ" mean?)

Phrasing like this, it isn't that difficult. You may want to look at $f(z) = \frac{1}{z}$ a bit and try to figure out how the proof follows before continue reading.

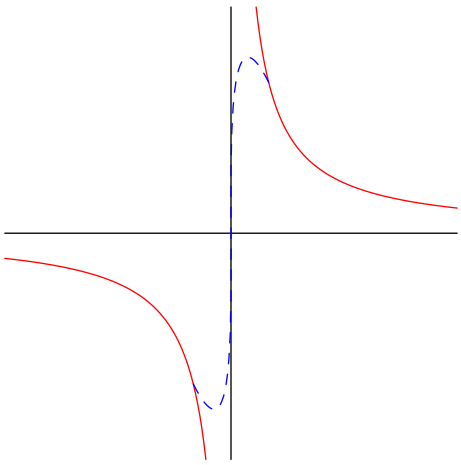
For simplicity, I will prove the statement for γ being a rectangle, leaving the case e.g. γ is a circle to the reader. The case of fully general γ will be handled later on.

As you may figured out, for $f(z) = \frac{1}{z-w}$, you can try to locate where the singularity w is by "binary search": compute $\oint_\gamma f\,dz$, if it is $2\pi i$, we know w is inside γ. We're going to do just that.

What should we search for? Let's see:

> **Exercise 31.7.3.** Suppose $\oint_\gamma f\,dz \neq 0$. Must there be a point where f blows up to infinity, like the point $z = 0$ in $\frac{1}{z}$?

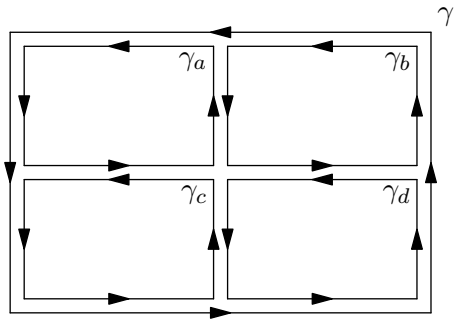
Answer: no, unfortunately. You can certainly take the function f above, and "smooth out" the singularity.



(Only real part depicted. You can imagine the imaginary part.)

The best we can hope for, then, is to find a point where f is not holomorphic (complex differentiable).

Construct 4 paths γ_a, γ_b, γ_c and γ_d as follows. The margin is only for illustration purpose, in reality the edges directly overlap on each other.



Notice that, because all the inner edges cancel out,

$$\oint_\gamma f\,dz = \oint_{\gamma_a} f\,dz + \oint_{\gamma_b} f\,dz + \oint_{\gamma_c} f\,dz + \oint_{\gamma_d} f\,dz.$$

Which means $\oint_{\gamma_i} f\,dz \neq 0$ for some $i \in \{a,b,c,d\}$. (Idea: we have more accurately located the singularity, now we know it is inside γ_i. Of course it's also possible that there are multiple singularities.)

We also have $|\oint_{\gamma_i} f\,dz| \geq \frac{1}{4} \cdot |\oint_\gamma f\,dz|$ for some i. The reason why we must carefully keep track of the magnitude (instead of just saying it's $\neq 0$) will become apparent later.

So, we keep doing that, and get a decreasing sequence of rectangles $\{\gamma_j\}$. Because the edge length gets halved each time, the rectangles converge to a single point p.

How would the rectangle perimeter decrease? Perhaps something like the following:

j	Perimeter of γ_j	$\lvert\oint_{\gamma_j} f\,dz\rvert$
0	1	1
1	$\frac{1}{2}$	$\geq \frac{1}{4}$
2	$\frac{1}{4}$	$\geq \frac{1}{16}$
3	$\frac{1}{8}$	$\geq \frac{1}{64}$

$|\oint_{\gamma_j} f\,dz|$ decreases quite quickly compared to the perimeter — as expected, we cannot hope for f to blow up at p, but this is sufficient to show f is not holomorphic.

For the sake of contradiction, assume otherwise. Then, by definition,

$$\lim_{h\to 0} \frac{f(p+h)-f(p)}{h} = f'(p)$$

where p is the point that the rectangles $\{\gamma_j\}$ converges to as defined above, and $f'(p) \in \mathbb{C}$ is the derivative. In other words, for $h \in \mathbb{C}$ close enough to 0,

$$f(p+h) = f(p) + f'(p)\cdot h + \varepsilon(h)\cdot h \text{ for } \varepsilon(h) \in o(1).$$

Why is this a problem? Notice that $f(p)$ and $f'(p)\cdot h$ are both polynomials, so

$$\oint_{\gamma_j} f(p) + f'(p)\cdot(z-p)\,dz = 0,$$

which means

$$\oint_{\gamma_j} f(z)\,dz = \oint_{\gamma_j} \varepsilon(h)\cdot(z-p)\,dz.$$

We know the left hand side decreases as 4^{-j}, but the integral on the right hand side is over a curve with length decreasing as 2^{-j}.

Exercise 31.7.4. Finish the proof. (Use the ML estimation lemma.)

Finally, what to do with arbitrary curve (which may not even have an interior[4])?

We construct the antiderivative $F\colon \Omega \to \mathbb{C}$ by integrating f across the side of a rectangle, prove $F' = f$, and get a "fundamental theorem of calculus", that is

$$\oint_\alpha f(z)\,dz = F(\alpha(b)) - F(\alpha(a))$$

where $\alpha\colon [a,b] \to \mathbb{C}$ is some path. Considering $\alpha = \gamma$, because the starting and ending point for a loop γ is the same, of course the integral would be 0.

§31.7.ii The rest

Next step, we should show the power series coincide with f, that is

$$f(z) = \oint_\gamma \frac{f(t)}{t}\,dt + \oint_\gamma \frac{f(t)}{t^2}\,dt \cdot z + \oint_\gamma \frac{f(t)}{t^3}\,dt \cdot z^2 + \cdots$$

Here we assume γ is the unit circle, the power series is centered at 0, and t is inside the unit disk.

> **Exercise 31.7.5.** Prove it. (You only need to know that you can interchange the infinite sum and the integral in this situation,[a] how to sum a geometric series, and Cauchy's integral formula)
>
> [a]Look at Example 38.1.4 for some horror stories where you cannot interchange a limit and an integral.

> **Remark 31.7.6 —** *Wait, where was Cauchy-Goursat theorem used?* If you forgot, it is used in the proof of Cauchy's integral formula.

After we have proven that f is a power series, then using Proposition 29.4.5 (suitably adapted for the case of complex holomorphic functions), the result follows.

§31.8 A few harder problems to think about

These aren't olympiad problems, but I think they're especially nice! In the next complex analysis chapter we'll see some more nice applications.

The first few results are the most important.

Problem 31A* (Liouville's theorem)**.** Let $f\colon \mathbb{C} \to \mathbb{C}$ be an entire function. Suppose that $|f(z)| < 1000$ for all complex numbers z. Prove that f is a constant function.

Problem 31B* (Zeros are isolated)**.** An **isolated set** in an open set U in the complex plane is a set of points S such that around each point in S, one can draw an open neighborhood not intersecting any other point of S.

Show that the zero set of any nonzero holomorphic function $f\colon U \to \mathbb{C}$ is an isolated set, unless there exists a nonempty open subset of U on which f is identically zero.

Problem 31C* (Identity theorem)**.** Let $f, g\colon U \to \mathbb{C}$ be holomorphic, and assume that U is connected. Prove that if f and g agree on some open neighborhood, then $f = g$.

[4]A space-filling curve is an example.

Problem 31D† (Maximums Occur On Boundaries)**.** Let $f\colon U \to \mathbb{C}$ be holomorphic, let $Y \subseteq U$ be compact, and let ∂Y be boundary[5] of Y. Show that

$$\max_{z \in Y} |f(z)| = \max_{z \in \partial Y} |f(z)| .$$

In other words, the maximum values of $|f|$ occur on the boundary. (Such maximums exist by compactness.)

Problem 31E (Harvard quals)**.** Let $f\colon \mathbb{C} \to \mathbb{C}$ be a nonconstant entire function. Prove that $f^{\text{img}}(\mathbb{C})$ is dense in $\mathbb{C}$. (In fact, a much stronger result is true: Little Picard's theorem says that the image of a nonconstant entire function omits at most one point.)

Problem 31F (Removable singularity theorem)**.** Let U be open, $p \in U$, and $f\colon U \setminus \{p\} \to \mathbb{C}$ be holomorphic. Suppose f is bounded. Show that $\lim_{z \to p} f(z)$ exists, and the extension $f\colon U \to \mathbb{C}$ is holomorphic at p.

[5]The boundary ∂Y is the set of points p such that no open neighborhood of p is contained in Y. It is also a compact set if Y is compact.

32 Meromorphic functions

§32.1 The second nicest functions on earth

If holomorphic functions are like polynomials, then *meromorphic* functions are like rational functions. Basically, a meromorphic function is a function of the form $\frac{A(z)}{B(z)}$ where $A, B \colon U \to \mathbb{C}$ are holomorphic and B is not zero. The most important example of a meromorphic function is $\frac{1}{z}$.

We are going to see that meromorphic functions behave like "almost-holomorphic" functions. Specifically, a meromorphic function A/B will be holomorphic at all points except the zeros of B (called *poles*). By the identity theorem, there cannot be too many zeros of B! So meromorphic functions can be thought of as "almost holomorphic" (like $\frac{1}{z}$, which is holomorphic everywhere but the origin). We saw that

$$\frac{1}{2\pi i} \oint_\gamma \frac{1}{z} \, dz = 1$$

for $\gamma(t) = e^{it}$ the unit circle. We will extend our results on contours to such situations.

It turns out that, instead of just getting $\oint_\gamma f(z) \, dz = 0$ like we did in the holomorphic case, the contour integrals will actually be used to *count the number of poles* inside the loop γ. It's ridiculous, I know.

§32.2 Meromorphic functions

Prototypical example for this section: $\frac{1}{z}$*, with a pole of order* 1 *and residue* 1 *at* $z = 0$.

Let U be an open subset of $\mathbb{C}$ again.

Definition 32.2.1. A function $f \colon U \to \mathbb{C}$ is **meromorphic** if there exists holomorphic functions $A, B \colon U \to \mathbb{C}$ with B not identically zero in any open neighborhood, and $f(z) = A(z)/B(z)$ whenever $B(z) \neq 0$.

Let's see how this function f behaves. If $z \in U$ has $B(z) \neq 0$, then in some small open neighborhood the function B isn't zero at all, and thus A/B is in fact *holomorphic*; thus f is holomorphic at z. (Concrete example: $\frac{1}{z}$ is holomorphic in any disk not containing 0.)

On the other hand, suppose $p \in U$ has $B(p) = 0$: without loss of generality, $p = 0$ to ease notation. By using the Taylor series at $p = 0$ we can put

$$B(z) = c_k z^k + c_{k+1} z^{k+1} + \cdots$$

with $c_k \neq 0$ (certainly some coefficient is nonzero since B is not identically zero!). Then we can write

$$\frac{1}{B(z)} = \frac{1}{z^k} \cdot \frac{1}{c_k + c_{k+1} z + \cdots}.$$

But the fraction on the right is a holomorphic function in this open neighborhood! So all that's happened is that we have an extra z^{-k} kicking around.

This gives us an equivalent way of viewing meromorphic functions:

Definition 32.2.2. Let $f : U \to \mathbb{C}$ as usual. A **meromorphic** function is a function which is holomorphic on U except at an isolated set S of points (meaning it is holomorphic as a function $U \setminus S \to \mathbb{C}$). For each $p \in S$, called a **pole** of f, the function f is further required to admit a **Laurent series**, meaning that

$$f(z) = \frac{c_{-m}}{(z-p)^m} + \frac{c_{-m+1}}{(z-p)^{m-1}} + \cdots + \frac{c_{-1}}{z-p} + c_0 + c_1(z-p) + \ldots$$

for all z in some open neighborhood of p, other than $z = p$. Here m is a positive integer, and $c_{-m} \neq 0$.

Note that the trailing end *must* terminate. By "isolated set", I mean that we can draw open neighborhoods around each pole in S, in such a way that no two open neighborhoods intersect.

Example 32.2.3 (Example of a meromorphic function)
Consider the function

$$\frac{z+1}{\sin z}.$$

It is meromorphic, because it is holomorphic everywhere except at the zeros of $\sin z$. At each of these points we can put a Laurent series: for example at $z = 0$ we have

$$\begin{aligned}
\frac{z+1}{\sin z} &= (z+1) \cdot \frac{1}{z - \frac{z^3}{3!} + \frac{z^5}{5!} - \cdots} \\
&= \frac{1}{z} \cdot \frac{z+1}{1 - \left(\frac{z^2}{3!} - \frac{z^4}{5!} + \frac{z^6}{7!} - \cdots \right)} \\
&= \frac{1}{z} \cdot (z+1) \sum_{k \geq 0} \left(\frac{z^2}{3!} - \frac{z^4}{5!} + \frac{z^6}{7!} - \cdots \right)^k.
\end{aligned}$$

If we expand out the horrible sum (which I won't do), then you get $\frac{1}{z}$ times a perfectly fine Taylor series, i.e. a Laurent series.

Abuse of Notation 32.2.4. We'll often say something like "consider the function $f : \mathbb{C} \to \mathbb{C}$ by $z \mapsto \frac{1}{z}$". Of course this isn't completely correct, because f doesn't have a value at $z = 0$. If I was going to be completely rigorous I would just set $f(0) = 2015$ or something and move on with life, but for all intents let's just think of it as "undefined at $z = 0$".

Why don't I just write $g : \mathbb{C} \setminus \{0\} \to \mathbb{C}$? The reason I have to do this is that it's still important for f to remember it's "trying" to be holomorphic on $\mathbb{C}$, even if isn't assigned a value at $z = 0$. As a function $\mathbb{C} \setminus \{0\} \to \mathbb{C}$ the function $\frac{1}{z}$ is actually holomorphic.

Remark 32.2.5 — I have shown that any function $A(z)/B(z)$ has this characterization with poles, but an important result is that the converse is true too: if $f : U \setminus S \to \mathbb{C}$ is holomorphic for some isolated set S, and moreover f admits a Laurent series at each point in S, then f can be written as a rational quotient of holomorphic functions. I won't prove this here, but it is good to be aware of.

Definition 32.2.6. Let p be a pole of a meromorphic function f, with Laurent series

$$f(z) = \frac{c_{-m}}{(z-p)^m} + \frac{c_{-m+1}}{(z-p)^{m-1}} + \cdots + \frac{c_{-1}}{z-p} + c_0 + c_1(z-p) + \ldots.$$

The integer m is called the **order** of the pole. A pole of order 1 is called a **simple pole**.

We also give the coefficient c_{-1} a name, the **residue** of f at p, which we write $\operatorname{Res}(f; p)$.

The order of a pole tells you how "bad" the pole is. The order of a pole is the "opposite" concept of the **multiplicity** of a **zero**. If f has a pole at zero, then its Laurent series near $z = 0$ might look something like

$$f(z) = \frac{1}{z^5} + \frac{8}{z^3} - \frac{2}{z^2} + \frac{4}{z} + 9 - 3z + 8z^2 + \cdots$$

and so f has a pole of order five. By analogy, if g has a zero at $z = 0$, it might look something like

$$g(z) = 3z^3 + 2z^4 + 9z^5 + \cdots$$

and so g has a zero of multiplicity three. These orders are additive: $f(z)g(z)$ still has a pole of order $5 - 3 = 2$, but $f(z)g(z)^2$ is completely patched now, and in fact has a **simple zero** now (that is, a zero of degree 1).

Exercise 32.2.7. Convince yourself that orders are additive as described above. (This is obvious once you understand that you are multiplying Taylor/Laurent series.)

Metaphorically, poles can be thought of as "negative zeros".

We can now give many more examples.

Example 32.2.8 (Examples of meromorphic functions)

(a) Any holomorphic function is a meromorphic function which happens to have no poles. Stupid, yes.

(b) The function $\mathbb{C} \to \mathbb{C}$ by $z \mapsto 100z^{-1}$ for $z \neq 0$ but undefined at zero is a meromorphic function. Its only pole is at zero, which has order 1 and residue 100.

(c) The function $\mathbb{C} \to \mathbb{C}$ by $z \mapsto z^{-3} + z^2 + z^9$ is also a meromorphic function. Its only pole is at zero, and it has order 3, and residue 0.

(d) The function $\mathbb{C} \to \mathbb{C}$ by $z \mapsto \frac{e^z}{z^2}$ is meromorphic, with the Laurent series at $z = 0$ given by

$$\frac{e^z}{z^2} = \frac{1}{z^2} + \frac{1}{z} + \frac{1}{2} + \frac{z}{6} + \frac{z^2}{24} + \frac{z^3}{120} + \cdots.$$

Hence the pole $z = 0$ has order 2 and residue 1.

Example 32.2.9 (A rational meromorphic function)

Consider the function $\mathbb{C} \to \mathbb{C}$ given by

$$\begin{aligned} z \mapsto \frac{z^4+1}{z^2-1} &= z^2 + 1 + \frac{2}{(z-1)(z+1)} \\ &= z^2 + 1 + \frac{1}{z-1} \cdot \frac{1}{1 + \frac{z-1}{2}} \\ &= \frac{1}{z-1} + \frac{3}{2} + \frac{9}{4}(z-1) + \frac{7}{8}(z-1)^2 - \ldots \end{aligned}$$

It has a pole of order 1 and residue 1 at $z = 1$. (It also has a pole of order 1 at $z = -1$; you are invited to compute the residue.)

Example 32.2.10 (Function with infinitely many poles)
The function $\mathbb{C} \to \mathbb{C}$ by

$$z \mapsto \frac{1}{\sin(z)}$$

has infinitely many poles: the numbers $z = \pi k$, where k is an integer. Let's compute the Laurent series at just $z = 0$:

$$\begin{aligned} \frac{1}{\sin(2\pi z)} &= \frac{1}{\frac{z}{1!} - \frac{z^3}{3!} + \frac{z^5}{5!} - \cdots} \\ &= \frac{1}{z} \cdot \frac{1}{1 - \left(\frac{z^2}{3!} - \frac{z^4}{5!} + \cdots\right)} \\ &= \frac{1}{z} \sum_{k \geq 0} \left(\frac{z^2}{3!} - \frac{z^4}{5!} + \cdots\right)^k. \end{aligned}$$

which is a Laurent series, though I have no clue what the coefficients are. You can at least see the residue; the constant term of that huge sum is 1, so the residue is 1. Also, the pole has order 1.

Example 32.2.11 (A function that is not meromorphic)
Consider the function

$$z \mapsto \frac{1}{\sin(1/z)}.$$

It is a holomorphic function on

$$U = \mathbb{C} \setminus \{0\} \setminus S$$

where we define $S = \{\frac{1}{\pi k} \mid k \in \mathbb{Z} \setminus \{0\}\}$. Similar to $z \mapsto \frac{1}{\sin(z)}$, each point in the set S has a pole of order 1.
However, at $z = 0$, the function admits no Laurent series — if it were, there would be a neighborhood around $z = 0$ where the function is defined, but there is no such set.
However, f is meromorphic on $\mathbb{C} \setminus \{0\}$ — the set S is isolated, but $S \cup \{0\}$ is not isolated.

The Laurent series, if it exists, is unique (as you might have guessed), and by our result on holomorphic functions it is actually valid for *any* disk centered at p (minus the point p). The part $\frac{c_{-1}}{z-p} + \cdots + \frac{c_{-m}}{(z-p)^m}$ is called the **principal part**, and the rest of the series $c_0 + c_1(z-p) + \cdots$ is called the **analytic part**.

§32.3 Winding numbers and the residue theorem

Recall that for a counterclockwise circle γ and a point p inside it, we had

$$\oint_\gamma (z-p)^m \, dz = \begin{cases} 0 & m \neq -1 \\ 2\pi i & m = -1 \end{cases}$$

where m is an integer. One can extend this result to in fact show that $\oint_\gamma (z-p)^m \, dz = 0$ for *any* loop γ, where $m \neq -1$. So we associate a special name for the nonzero value at $m = -1$.

Definition 32.3.1. For a point $p \in \mathbb{C}$ and a loop γ not passing through it, we define the **winding number**, denoted $\mathbf{I}(\gamma, p)$, by

$$\mathbf{I}(\gamma, p) = \frac{1}{2\pi i} \oint_\gamma \frac{1}{z-p} \, dz$$

For example, by our previous results we see that if γ is a circle, we have

$$\mathbf{I}(\text{circle}, p) = \begin{cases} 1 & p \text{ inside the circle} \\ 0 & p \text{ outside the circle.} \end{cases}$$

If you've read the chapter on fundamental groups, then this is just the fundamental group associated to $\mathbb{C} \setminus \{p\}$. In particular, the winding number is always an integer. (Essentially, it uses the complex logarithm to track how the argument of the function changes. The details are more complicated, so we omit them here). In the simplest case the winding numbers are either 0 or 1.

Definition 32.3.2. We say a loop γ is **regular** if $\mathbf{I}(\gamma, p) = 1$ for all points p in the interior of γ (for example, if γ is a counterclockwise circle).

With all these ingredients we get a stunning generalization of the Cauchy-Goursat theorem:

Theorem 32.3.3 (Cauchy's residue theorem)

Let $f \colon \Omega \to \mathbb{C}$ be meromorphic, where Ω is simply connected. Then for any loop γ not passing through any of its poles, we have

$$\frac{1}{2\pi i} \oint_\gamma f(z) \, dz = \sum_{\text{pole } p} \mathbf{I}(\gamma, p) \operatorname{Res}(f; p).$$

In particular, if γ is regular then the contour integral is the sum of all the residues, in the form

$$\frac{1}{2\pi i} \oint_\gamma f(z) \, dz = \sum_{\substack{\text{pole } p \\ \text{inside } \gamma}} \operatorname{Res}(f; p).$$

Question 32.3.4. Verify that this result coincides with what you expect when you integrate $\oint_\gamma cz^{-1} \, dz$ for γ a counter-clockwise circle.

The proof from here is not really too impressive – the "work" was already done in our statements about the winding number.

Proof. Let the poles with nonzero winding number be $p_1, \dots, p_k$ (the others do not affect the sum).[1] Then we can write f in the form

$$f(z) = g(z) + \sum_{i=1}^{k} P_i\left(\frac{1}{z-p_i}\right)$$

where $P_i\left(\frac{1}{z-p_i}\right)$ is the principal part of the pole p_i. (For example, if $f(z) = \frac{z^3-z+1}{z(z+1)}$ we would write $f(z) = (z-1) + \frac{1}{z} - \frac{1}{1+z}$.)

The point of doing so is that the function g is holomorphic (we've removed all the "bad" parts), so

$$\oint_\gamma g(z)\,dz = 0$$

by Cauchy-Goursat.

On the other hand, if $P_i(x) = c_1 x + c_2 x^2 + \cdots + c_d x^d$ then

$$\begin{aligned}\oint_\gamma P_i\left(\frac{1}{z-p_i}\right)\,dz &= \oint_\gamma c_1 \cdot \left(\frac{1}{z-p_i}\right)\,dz + \oint_\gamma c_2 \cdot \left(\frac{1}{z-p_i}\right)^2\,dz + \dots \\ &= c_1 \cdot \mathbf{I}(\gamma, p_i) + 0 + 0 + \dots \\ &= \mathbf{I}(\gamma, p_i)\operatorname{Res}(f; p_i).\end{aligned}$$

which gives the conclusion. □

Remark 32.3.5 (Intuition behind Cauchy's integral formula) — In the setting of Theorem 31.5.1, note that if f is meromorphic in the disk D, we can compute the Laurent series of f at the point a:

$$f(z) = \frac{c_{-m}}{(z-a)^m} + \frac{c_{-m+1}}{(z-a)^{m-1}} + \cdots + \frac{c_{-1}}{z-a} + c_0 + c_1(z-a) + \cdots$$

By the residue theorem, integrating $f(z)$ around the boundary of D results in the c_{-1} coefficient in the Laurent series:

$$\frac{1}{2\pi i}\oint_\gamma f(z)\,dz = \operatorname{Res}(f; a) = c_{-1}.$$

Of course, this is useless — f is holomorphic at a, so $c_{-1} = 0$. We want to compute $c_0 = f(a)$ instead.

Nevertheless, the trick is that *we can manipulate the function f* in order to move the coefficient we want to compute to the coefficient corresponding to $(z-a)^{-1}$. How are we going to do that? By dividing by $z-a$, of course!

So, $\frac{f(z)}{z-a}$ is meromorphic in the disk D, with Laurent series expansion around a being

$$\frac{f(z)}{z-a} = \frac{c_{-m}}{(z-a)^{m+1}} + \frac{c_{-m+1}}{(z-a)^m} + \cdots + \frac{c_{-1}}{(z-a)^2} + \frac{c_0}{z-a} + c_1 + c_2(z-a) + \cdots$$

Because $\frac{f(z)}{z-a}$ has no other poles in D except at a, the residue theorem immediately tells us the integral $\frac{1}{2\pi i}\oint_\gamma \frac{f(z)}{z-a}\,dz$ equals $\operatorname{Res}(\frac{f(z)}{z-a}; a)$, which equals c_0 looking at the Laurent series above.

[1] To show that there must be finitely many such poles: recall that all our contours $\gamma\colon [a,b] \to \mathbb{C}$ are in fact bounded, so there is some big closed disk D which contains all of γ. The poles outside D thus have winding number zero. Now we cannot have infinitely many poles inside the disk D, for D is compact and the set of poles is a closed and isolated set!

§32.4 Argument principle

One tricky application is as follows. Given a polynomial $P(x) = (x-a_1)^{e_1}(x-a_2)^{e_2}\dots(x-a_n)^{e_n}$, you might know that we have

$$\frac{P'(x)}{P(x)} = \frac{e_1}{x-a_1} + \frac{e_2}{x-a_2} + \dots + \frac{e_n}{x-a_n}.$$

The quantity P'/P is called the **logarithmic derivative**, as it is the derivative of $\log P$. This trick allows us to convert zeros of P into poles of P'/P with order 1; moreover the residues of these poles are the multiplicities of the roots.

In an analogous fashion, we can obtain a similar result for any meromorphic function f.

Proposition 32.4.1 (The logarithmic derivative)

Let $f\colon U \to \mathbb{C}$ be a meromorphic function. Then the logarithmic derivative f'/f is meromorphic as a function from U to $\mathbb{C}$; its only poles are:

(i) A pole at each zero of f whose residue is the multiplicity, and

(ii) A pole at each pole of f whose residue is the negative of the pole's order.

Again, you can almost think of a pole as a zero of negative multiplicity. This spirit is exemplified below.

Proof. Dead easy with Laurent series. Let a be a zero/pole of f, and WLOG set $a = 0$ for convenience. We take the Laurent series at zero to get

$$f(z) = c_k z^k + c_{k+1}z^{k+1} + \dots$$

where $k < 0$ if 0 is a pole and $k > 0$ if 0 is a zero. Taking the derivative gives

$$f'(z) = kc_k z^{k-1} + (k+1)c_{k+1}z^k + \dots.$$

Now look at f'/f; with some computation, it equals

$$\frac{f'(z)}{f(z)} = \frac{1}{z}\frac{kc_k + (k+1)c_{k+1}z + \dots}{c_k + c_{k+1}z + \dots}.$$

So we get a simple pole at $z = 0$, with residue k. □

Using this trick you can determine the number of zeros and poles inside a regular closed curve, using the so-called Argument Principle.[2]

[2]So-called because the *argument* of a complex number z is the angle formed by the real axis and the vector representing z, not because you need to use any argument. If $z \in \mathbb{C}$ is interpreted as a point in $\mathbb{R}^2$, the argument of z is the same as $\theta(z)$ defined in Example 44.7.4.

Theorem 32.4.2 (Argument principle)

Let γ be a regular curve. Suppose $f\colon U \to \mathbb{C}$ is meromorphic inside and on γ, and none of its zeros or poles lie on γ. Then

$$\frac{1}{2\pi i} \oint_\gamma \frac{f'}{f}\,dz = \frac{1}{2\pi i} \oint_{f\circ\gamma} \frac{1}{z}\,dz = Z - P$$

where Z is the number of zeros inside γ (counted with multiplicity) and P is the number of poles inside γ (again with multiplicity).

Proof. Immediate by applying Cauchy's residue theorem alongside the preceding proposition. In fact you can generalize to any curve γ via the winding number: the integral is

$$\frac{1}{2\pi i} \oint_\gamma \frac{f'}{f}\,dz = \sum_{\text{zero } z} \mathbf{I}(\gamma, z) - \sum_{\text{pole } p} \mathbf{I}(\gamma, p)$$

where the sums are with multiplicity. □

Thus the Argument Principle allows one to count zeros and poles inside any region of choice.

Computers can use this to get information on functions whose values can be computed but whose behavior as a whole is hard to understand. Suppose you have a holomorphic function f, and you want to understand where its zeros are. Then just start picking various circles γ. Even with machine rounding error, the integral will be close enough to the true integer value that we can decide how many zeros are in any given circle. Numerical evidence for the Riemann Hypothesis (concerning the zeros of the Riemann zeta function) can be obtained in this way.

§32.5 Digression: the Argument Principle viewed geometrically

There is another, more geometric, way to understand the Argument Principle.

Assume a function f is holomorphic on a connected open set U containing 0, and possibly has a zero or a pole at 0. Let $\gamma\colon [0, 2\pi] \to U$ be some curve contained in U, such that 0 is not in the image of the curve.

Let $a = \gamma(0)$ be the starting point of γ, and $b = \gamma(2\pi)$ be the ending point of γ.

We all know that $z \mapsto \log z$ is not an actual function — even ignoring the singularity at 0, it has a branch cut (we will formally handle this in Chapter 33).

Nevertheless, if we close our eyes and shuffling some symbols around, we get:

$$\begin{aligned}
\frac{1}{2\pi i} \oint_\gamma \frac{f'(z)}{f(z)}\,dz &= \frac{1}{2\pi i} \oint_\gamma \frac{d}{dz} \log f(z)\,dz \\
&= \frac{1}{2\pi i} \oint_\gamma d(\log f(z)) \\
&= \frac{1}{2\pi i} \cdot (\log f(b) - \log f(a)).
\end{aligned}$$

Miraculously, everything seems to cancel out so nicely! This is not a coincidence.

Now, if γ is a circle, then $a = b$, so the formula above seemingly states that the integral will be 0? Fortunately for us, no — log is in fact not a function.

So, does the formula above means anything? It does! While we won't prove this rigorously, the point is that:

If we let a point p smoothly moves from a to b, and let $\log f(p)$ follows the value, then $\log f(b) - \log f(a)$ represents the change in value of $\log f(p)$.

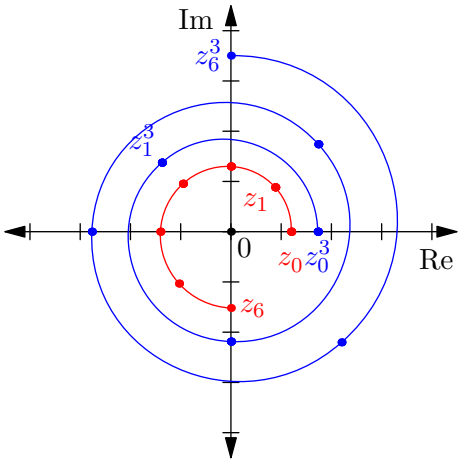
In the notation of Section 66.2, we have the mouse moves along γ from a to b, the first robot moves along $f \circ \gamma$ from $f(a)$ to $f(b)$, and the second robot moves from $\log f(a)$ to $\log f(b)$.

If we forget about the mouse for a moment, note that:

The quantity $\frac{1}{2\pi i} \oint_\gamma \frac{f'(z)}{f(z)}\, dz$ is equal to the number of times the *first* robot winds around the origin.

That is, $\mathbf{I}(f \circ \gamma, 0)$. (This is essentially obvious to see, because of all the work we have done to prove $\oint d\log z = \oint \frac{1}{z}\, dz$ equals the winding number.)

Finally, if we look at some simple examples like z^3:



We can immediately see the relation between the winding number and the multiplicity of a zero:

If z moves around the origin in a circle once, then z^n moves around the origin in a circle n times.

z^{-n} is not much different — it moves around the origin in a circle n times, just in the opposite direction.

Piecing all these pieces together, we get the Argument Principle — the logarithmic derivative can be used to count the multiplicity of the roots and the order of the poles.

§32.6 Philosophy: why are holomorphic functions so nice?

All the fun we've had with holomorphic and meromorphic functions comes down to the fact that complex differentiability is such a strong requirement. It's a small miracle that $\mathbb{C}$, which *a priori* looks only like $\mathbb{R}^2$, is in fact a field. Moreover, $\mathbb{R}^2$ has the nice property that one can draw nontrivial loops (it's also true for real functions that $\int_a^a f\, dx = 0$, but this is not so interesting!), and this makes the theory much more interesting.

As another piece of intuition from Siu[3]: If you try to get (left) differentiable functions over *quaternions*, you find yourself with just linear functions.

§32.7 A few harder problems to think about

Problem 32A (Fundamental theorem of algebra)**.** Prove that if f is a nonzero polynomial of degree n then it has n roots.

Problem 32B† (Rouché's theorem)**.** Let $f, g \colon U \to \mathbb{C}$ be holomorphic functions, where U contains the unit disk. Suppose that $|f(z)| > |g(z)|$ for all z on the unit circle. Prove that f and $f+g$ have the same number of zeros which lie strictly inside the unit circle (counting multiplicities).

Problem 32C (Wedge contour)**.** For each odd integer $n \geq 3$, evaluate the improper integral

$$\int_0^\infty \frac{1}{1+x^n}\,dx.$$

Problem 32D (Another contour)**.** Prove that the integral

$$\int_{-\infty}^\infty \frac{\cos x}{x^2+1}\,dx$$

converges and determine its value.

Problem 32E***.** Let $f \colon U \to \mathbb{C}$ be a nonconstant holomorphic function.

(a) (Open mapping theorem) Prove that $f^{\text{img}}(U)$ is open in $\mathbb{C}$.[4]

(b) (Maximum modulus principle) Show that $|f|$ cannot have a maximum over U. That is, show that for any $z \in U$, there is some $z' \in U$ such that $|f(z)| < |f(z')|$.

[3]Harvard professor.

[4]Thus the image of *any* open set $V \subseteq U$ is open in $\mathbb{C}$ (by repeating the proof for the restriction of f to V).

33 Holomorphic square roots and logarithms

In this chapter we'll make sense of a holomorphic square root and logarithm. The main results are Theorem 33.3.2, Theorem 33.4.2, Corollary 33.5.1, and Theorem 33.5.2. If you like, you can read just these four results, and skip the discussion of how they came to be.

Let $f\colon U \to \mathbb{C}$ be a holomorphic function. A **holomorphic nth root** of f is a function $g\colon U \to \mathbb{C}$ such that $f(z) = g(z)^n$ for all $z \in U$. A **logarithm** of f is a function $g\colon U \to \mathbb{C}$ such that $f(z) = e^{g(z)}$ for all $z \in U$. The main question we'll try to figure out is: when do these exist? In particular, what if $f = \mathrm{id}$?

§33.1 Motivation: square root of a complex number

To start us off, can we define $\sqrt{z}$ for any complex number z?

The first obvious problem that comes up is that for any z, there are *two* numbers w such that $w^2 = z$. How can we pick one to use? For our ordinary square root function, we had a notion of "positive", and so we simply took the positive root.

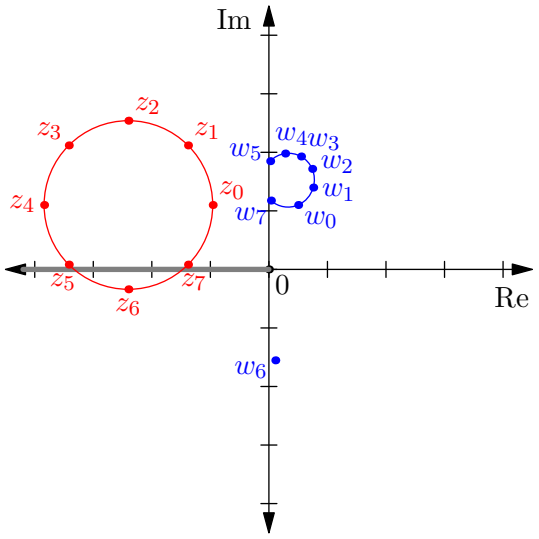
Let's expand on this: given $z = r(\cos\theta + i\sin\theta)$ (here $r \geq 0$) we should take the root to be

$$w = \sqrt{r}\left(\cos\alpha + i\sin\alpha\right).$$

such that $2\alpha \equiv \theta \pmod{2\pi}$; there are two choices for $\alpha \pmod{2\pi}$, differing by π.

For complex numbers, we don't have an obvious way to pick α. Nonetheless, perhaps we can also get away with an arbitrary distinction: let's see what happens if we just choose the α with $-\frac{1}{2}\pi < \alpha \leq \frac{1}{2}\pi$.

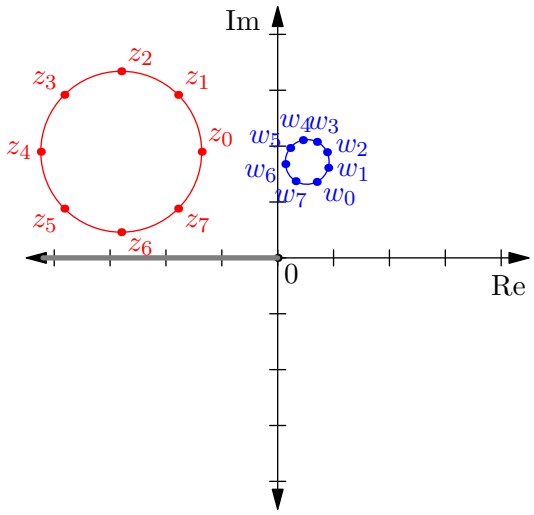
Pictured below are some points (in red) and their images (in blue) under this "upper-half" square root. The condition on α means we are forcing the blue points to lie on the right-half plane.



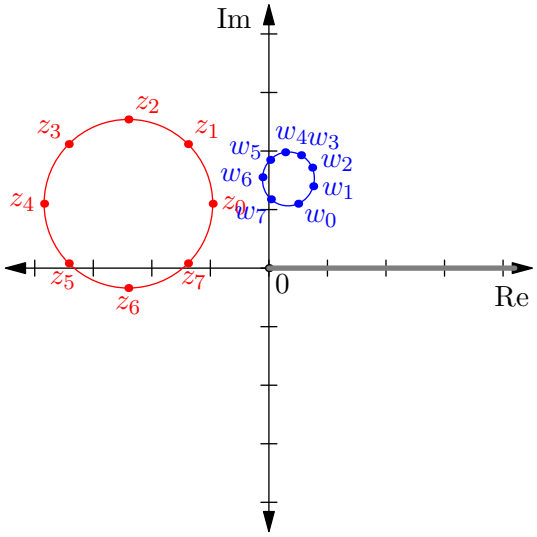
Here, $w_i^2 = z_i$ for each i, and we are constraining the w_i to lie in the right half of the complex plane. We see there is an obvious issue: there is a big discontinuity near

the points w_5 and w_7! The nearby point w_6 has been mapped very far away. This discontinuity occurs since the points on the negative real axis are at the "boundary". For example, given -4, we send it to $-2i$, but we have hit the boundary: in our interval $-\frac{1}{2}\pi \le \alpha < \frac{1}{2}\pi$, we are at the very left edge.

The negative real axis that we must not touch is what we will later call a *branch cut*, but for now I call it a **ray of death**. It is a warning to the red points: if you cross this line, you will die! However, if we move the red circle just a little upwards (so that it misses the negative real axis) this issue is avoided entirely, and we get what seems to be a "nice" square root.



In fact, the ray of death is fairly arbitrary: it is the set of "boundary issues" that arose when we picked $-\frac{1}{2}\pi < \alpha \le \frac{1}{2}\pi$. Suppose we instead insisted on the interval $0 \le \alpha < \pi$; then the ray of death would be the *positive* real axis instead. The earlier circle we had now works just fine.



What we see is that picking a particular α-interval leads to a different set of edge cases, and hence a different ray of death. The only thing these rays have in common is their starting point of zero. In other words, given a red circle and a restriction of α, I can make a nice "square rooted" blue circle as long as the ray of death misses it.

So, what exactly is going on?

§33.2 Square roots of holomorphic functions

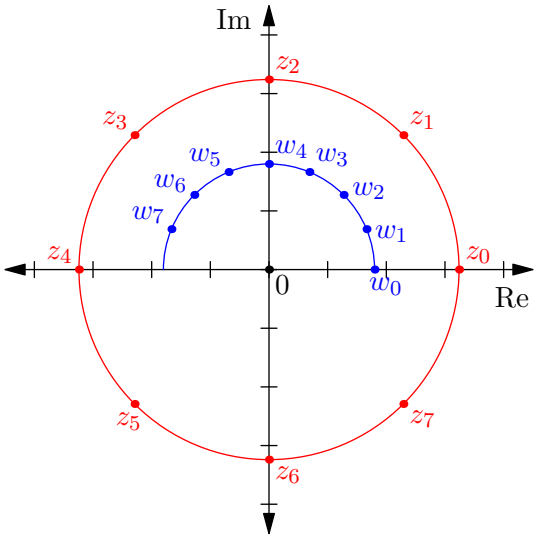
To get a picture of what's happening, we would like to consider a more general problem: let $f\colon U \to \mathbb{C}$ be holomorphic. Then we want to decide whether there is a holomorphic $g\colon U \to \mathbb{C}$ such that

$$f(z) = g(z)^2.$$

Our previous discussion with $f = \mathrm{id}$ tells us we cannot hope to achieve this for $U = \mathbb{C}$; there is a "half-ray" which causes problems. However, there are certainly functions $f\colon \mathbb{C} \to \mathbb{C}$ such that a g exists. As a simplest example, $f(z) = z^2$ should definitely have a square root!

Now let's see if we can fudge together a square root. Earlier, what we did was try to specify a rule to force one of the two choices at each point. This is unnecessarily strict. Perhaps we can do something like: start at a point in $z_0 \in U$, pick a square root w_0 of $f(z_0)$, and then try to "fudge" from there the square roots of the other points. What do I mean by fudge? Well, suppose z_1 is a point very close to z_0, and we want to pick a square root w_1 of $f(z_1)$. While there are two choices, we also would expect w_0 to be close to w_1. Unless we are highly unlucky, this should tell us which choice of w_1 to pick. (Stupid concrete example: if I have taken the square root $-4.12i$ of -17 and then ask you to continue this square root to -16, which sign should you pick for $\pm 4i$?)

There are two possible ways we could get unlucky in the scheme above: first, if $w_0 = 0$, then we're sunk. But even if we avoid that, we have to worry that if we run a full loop in the complex plane, we might end up in a different place from where we started. For concreteness, consider the following situation, again with $f = \mathrm{id}$:



We started at the point z_0, with one of its square roots as w_0. We then wound a full red circle around the origin, only to find that at the end of it, the blue arc is at a different place where it started!

The interval construction from earlier doesn't work either: no matter how we pick the interval for α, any ray of death must hit our red circle. The problem somehow lies with the fact that we have enclosed the very special point 0.

Nevertheless, we know that if we take $f(z) = z^2$, then we don't run into any problems with our "make it up as you go" procedure. So, what exactly is going on?

§33.3 Covering projections

By now, if you have read the part on algebraic topology, this should all seem quite familiar. The "fudging" procedure exactly describes the idea of a lifting.

More precisely, recall that there is a covering projection

$$(-)^2 \colon \mathbb{C} \setminus \{0\} \to \mathbb{C} \setminus \{0\}.$$

Let $V = \{z \in U \mid f(z) \neq 0\}$. For $z \in U \setminus V$, we already have the square root $g(z) = \sqrt{f(z)} = \sqrt{0} = 0$. So the burden is completing $g \colon V \to \mathbb{C}$.

Then essentially, what we are trying to do is construct a lifting g in the diagram

$$\begin{array}{ccc} & & E = \mathbb{C} \setminus \{0\} \\ & \overset{g}{\nearrow} & \big\downarrow {\scriptstyle p = \bullet^2} \\ V & \xrightarrow[f]{} & B = \mathbb{C} \setminus \{0\}. \end{array}$$

Our map p can be described as "winding around twice". Our Theorem 66.2.5 now tells us that this lifting exists if and only if

$$f_*^{\text{img}}(\pi_1(V)) \subseteq p_*^{\text{img}}(\pi_1(E))$$

is a subset of the image of $\pi_1(E)$ by p. Since B and E are both punctured planes, we can identify them with S^1.

> **Question 33.3.1.** Show that the image under p is exactly $2\mathbb{Z}$ once we identify $\pi_1(B) = \mathbb{Z}$.

That means that for any loop γ in V, we need $f \circ \gamma$ to have an *even* winding number around $0 \in B$. This amounts to

$$\frac{1}{2\pi} \oint_\gamma \frac{f'}{f} \, dz \in 2\mathbb{Z}$$

since f has no poles.

Replacing 2 with n and carrying over the discussion gives the first main result.

Theorem 33.3.2 (Existence of holomorphic nth roots)
Let $f \colon U \to \mathbb{C}$ be holomorphic. Then f has a holomorphic nth root if and only if

$$\frac{1}{2\pi i} \oint_\gamma \frac{f'}{f} \, dz \in n\mathbb{Z}$$

for every contour γ in U.

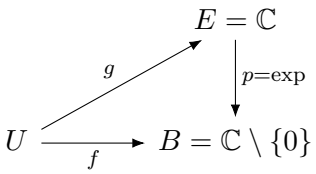
§33.4 Complex logarithms

The multivalued nature of the complex logarithm comes from the fact that

$$\exp(z + 2\pi i) = \exp(z).$$

So if $e^w = z$, then any complex number $w + 2\pi i k$ is also a solution.

We can handle this in the same way as before: it amounts to a lifting of the following diagram.



There is no longer a need to work with a separate V since:

Question 33.4.1. Show that if f has any zeros then g can't possibly exist.

In fact, the map $\exp\colon \mathbb{C} \to \mathbb{C} \setminus \{0\}$ is a universal cover, since $\mathbb{C}$ is simply connected. Thus, $p^{\text{img}}(\pi_1(\mathbb{C}))$ is *trivial*. So in addition to being zero-free, f cannot have any winding number around $0 \in B$ at all. In other words:

Theorem 33.4.2 (Existence of logarithms)
Let $f\colon U \to \mathbb{C}$ be holomorphic. Then f has a logarithm if and only if

$$\frac{1}{2\pi i} \oint_\gamma \frac{f'}{f}\, dz = 0$$

for every contour γ in U.

§33.5 Some special cases

The most common special case is

Corollary 33.5.1 (Nonvanishing functions from simply connected domains)
Let $f\colon \Omega \to \mathbb{C}$ be continuous, where Ω is simply connected. If $f(z) \neq 0$ for every $z \in \Omega$, then f has both a logarithm and holomorphic nth root.

Finally, let's return to the question of $f = \text{id}$ from the very beginning. What's the best domain U such that

$$\sqrt{-}\colon U \to \mathbb{C}$$

is well-defined? Clearly $U = \mathbb{C}$ cannot be made to work, but we can do almost as well. For note that the only zero of $f = \text{id}$ is at the origin. Thus if we want to make a logarithm exist, all we have to do is make an incision in the complex plane that renders it impossible to make a loop around the origin. The usual choice is to delete negative half of the real axis, our very first ray of death; we call this a **branch cut**, with **branch point** at $0 \in \mathbb{C}$ (the point which we cannot circle around). This gives

Theorem 33.5.2 (Branch cut functions)
There exist holomorphic functions

$$\log : \mathbb{C} \setminus (-\infty, 0] \to \mathbb{C}$$
$$\sqrt[n]{-} : \mathbb{C} \setminus (-\infty, 0] \to \mathbb{C}$$

satisfying the obvious properties.

There are many possible choices of such functions (n choices for the nth root and infinitely many for log); a choice of such a function is called a **branch**. So this is what is meant by a "branch" of a logarithm.

The **principal branch** is the "canonical" branch, analogous to the way we arbitrarily pick the positive branch to define $\sqrt{-}\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. For log, we take the w such that $e^w = z$ and the imaginary part of w lies in $(-\pi, \pi]$ (since we can shift by integer multiples of $2\pi i$). Often, authors will write $\operatorname{Log} z$ to emphasize this choice.

§33.6 A few harder problems to think about

Problem 33A. Show that a holomorphic function $f\colon U \to \mathbb{C}$ has a holomorphic logarithm if and only if it has a holomorphic nth root for every integer n.

Problem 33B. Show that the function $f\colon U \to \mathbb{C}$ by $z \mapsto z(z-1)$ has a holomorphic square root, where U is the entire complex plane minus the closed interval $[0,1]$.

34 Bonus: Topological Abel-Ruffini Theorem

We've already shown the Fundamental Theorem of Algebra. Now, with our earlier intuition on holomorphic nth roots, we can now show that there is no general formula for the roots of a quintic polynomial.

§34.1 The Game Plan

Firstly, what do we even mean by "formula" here?

Definition 34.1.1. A **quintic formula** would be a formula taking in the coefficients $(a_0, \dots, a_5)$ of a degree 5 polynomial P, using the operations $+$, $-$, $\times$, $\div$, $\sqrt[n]{}$ finitely many times, that maps to the five roots $(z_1, \cdots, z_5)$ of P.

Now, any proposed quintic formula F receives the same coefficients when the roots are the same, and thus gives the same output. This is fine at first glance, but swapping two roots continuously might pose more issues. F must create and preserve some order of the roots under these permutations.

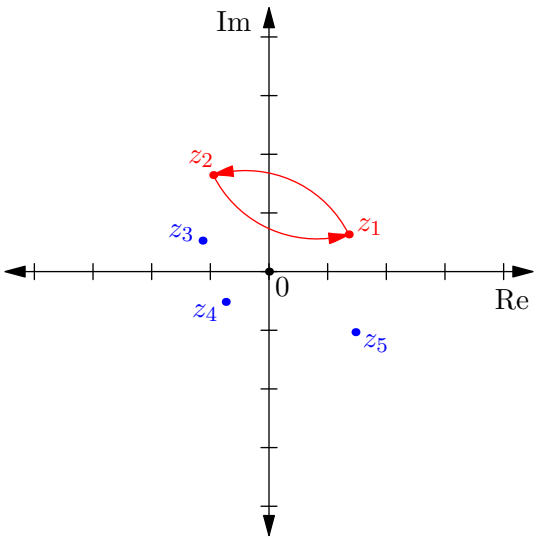
> **Question 34.1.2.** Convince yourself any F indeed must track which root is which when moving roots along smooth paths.

> **Remark 34.1.3 —** This isn't true if we bring even more complicated functions such as **Bring Radicals** to the table. But this wasn't really considered "fair game."

§34.2 Step 1: The Simplest Case

Let's first ignore the $\sqrt[n]{}$ operator for motivation. Suppose I told you that some rational function R always finds a root of a quintic polynomial $P(z) = (z-z_1)(z-z_2)(z-z_3)(z-z_4)(z-z_5)$. For simplicity, let all the roots be distinct.

Suppose that initially R outputs z_1. Consider what happens we smoothly swap the roots z_1 and z_2 along two non-intersecting paths that doesn't go through other roots.



Since R is continuous, it must be tracking the same root. However, once we finish swapping z_1 and z_2, the coefficients of P are the same as they were initially. But this means that R has been tricked into changing the root it outputs, contradiction!

The bigger picture here is that we were able to find an operation that fixes R while changing the order of the roots in S_5.

§34.3 Step 2: Nested Roots

Once we add $\sqrt[n]{}$ back to the picture, this idea no longer works right out of the box.

Example 34.3.1 (Quadratic Formula)

If you've done any competition math, you know that for a polynomial $P(z) = az^2 + bz + c = (z - z_1)(z - z_2)$, it follows that the two branches of

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

give z_1 and z_2.

So why can't swapping z_1 and z_2 yield a contradiction here? It's because while all the coefficients end up in the starting position, the liftings of how $\sqrt{b^2 - 4ac}$ travels may not.

Exercise 34.3.2. Consider the polynomial $z^2 - 1$. Then smoothly swap the roots to get the intermediary polynomials of $(z - e^{it})(z + e^{it})$. See that the two roots given by the quadratic formula also swap position.

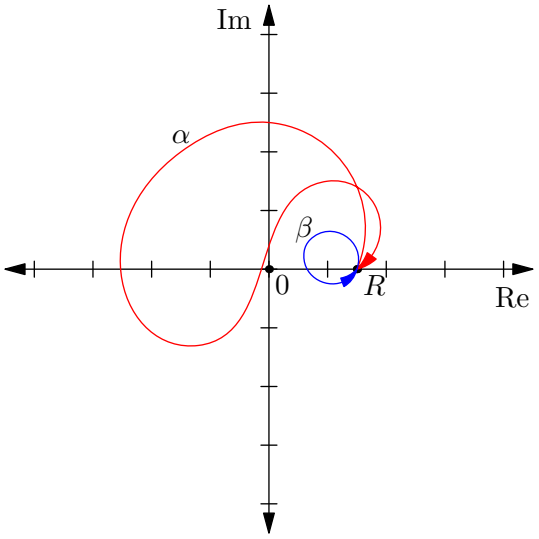
Let's now consider the next simplest case of the nth root of a rational function $\sqrt[n]{R}$, and try to fix it with a nontrivial permutation of the roots.

Swapping the roots z_1 and z_2, we keep R the same, but R's path α around the origin may have accumulated some change in phase $2\pi a$. If we were to unswap z_1 and z_2 in the same manner, we'd undo the change in phase, but we'd also be back to doing nothing.

However, while changes in phase are *abelian*, permutations are not. Let's consider another operation of swapping the roots z_2 and z_3. Taking a commutator of the

two operations, we keep all the phases the same, but end up with a permutation $(12)(23)(12)^{-1}(23)^{-1}$.

If we mark the second operation's path with β, this corresponds to $\alpha\beta\alpha^{-1}\beta^{-1}$.



Exercise 34.3.3. Show that this permutation operation is nontrivial.

We now have better tools: We have permutations in S_5 that fix the nth roots of rational functions, and their compositions under $+$, $-$, $\times$, $\div$.

How do we handle the nested radicals now?

Example 34.3.4 (Cubic Formula)

The cubic formula contains a nasty term

$$\sqrt[3]{\frac{2b^3 - 9abc + 27a^2d + \sqrt{(2b^3 - 9abc + 27a^d)^2 - 4(b^3 - 3ac)^3}}{2}}.$$

Here, we've taken multiple roots.

Definition 34.3.5. Define the degree of a nested radical as the maximum number of times radicals can be found in other radicals.

Let's now consider nested radicals of degree 2, such as say $\sqrt[3]{\sqrt{ab+c} - \sqrt{d}}$. We know that we have nontrivial commutators σ and ρ that fix the interior of the cube root, but once again the phase may not be preserved under each operation individually. Once again, we can again consider the *commutators* of these commutators, say $\sigma\rho\sigma^{-1}\rho^{-1}$ which by the same logic fixes the issues with phase.

There's no reason, we can't consider the commutators of commutators of commutators to fix radicals of degree 3 and so on. It thus just remains that we always can keep getting nontrivial commutators.

§34.4 Step 3: Normal Groups

We've reduced this to a group theory problem. Given a chain of commutators

$$S_5 = G \supseteq G^{(1)} \supseteq G^{(2)} \supseteq \dots$$

where each group is the commutator subgroup of the next, we want to show that $G^{(n)}$ never becomes trivial. This chain is called the **derived series**.

> **Exercise 34.4.1.** Show that for the commutator subgroup $[G, G]$ of a group G, we have that $[G, G] \trianglelefteq G$, and that $G/[G, G]$ is Abelian.

Definition 34.4.2. A group G is **solvable** if its derived series is nontrivial.

So all that remains is showing that S_5 is not solvable. This is a calculation that isn't relevant to the topology ideas in this chapter, so we defer it to Problem 34A.

§34.5 Summary

While this is indeed a valid proof, it has some pros and cons. As a con, we haven't shown that any polynomial such as $z^5 - z - 1$ has a root that can't be expressed using nested nth roots. We've only that we don't have a formula for all degree 5 polynomials.

As a pro, this argument makes it easy to add even more functions such as exp, sin, and cos to the mix and show even then that no such formula exists. It also allows you to broadly understand what people mean when they compare this theorem to a fact that A_5 is not solvable.

§34.6 A few harder problems to think about

Problem 34A. Show that A_5 is not solvable.

X

Measure Theory

Part X: Contents

35 Measure spaces

Here is an outline of where we are going next. Our *goal* over the next few chapters is to develop the machinery to state (and in some cases prove) the law of large numbers and the central limit theorem. For these purposes, the scant amount of work we did in Calculus 101 is going to be awfully insufficient: integration over $\mathbb{R}$ (or even $\mathbb{R}^n$) is just not going to cut it.

This chapter will develop the theory of "measure spaces", which you can think of as "spaces equipped with a notion of size". We will then be able to integrate over these with the so-called Lebesgue integral (which in some senses is almost strictly better than the Riemann one).

§35.1 Letter connotations

There are a lot of "types" of objects moving forward, so here are the letter connotations we'll use throughout the next several chapters. This makes it easier to tell what the "type" of each object is just by which letter is used.

- Measure spaces denoted by Ω, their elements denoted by ω.
- Algebras and σ-algebras denoted by script $\mathscr{A}$, $\mathscr{B}$, ... Sets in them denoted by early capital Roman A, B, C, D, E, ...
- Measures (i.e. functions assigning sets to reals) denoted usually by μ or ρ.
- Random variables (functions sending worlds to reals) denoted usually by late capital Roman X, Y, Z, ...
- Functions from $\mathbb{R} \to \mathbb{R}$ by Roman letters like f and g for pdf's and F and G for cdf's.
- Real numbers denoted by lower Roman letters like x, y, z.

§35.2 Motivating measure spaces via random variables

To motivate *why* we want to construct measure spaces, I want to talk about a (real) **random variable**, which you might think of as

- the result of a coin flip,
- the high temperature in Boston on Saturday,
- the possibility of rain on your 18.725 date next weekend.

Why does this need a long theory to develop well? For a simple coin flip one intuitively just thinks "50% heads, 50% tails" and is done with it. The situation is a little trickier with temperature since it is continuous rather than discrete, but if all you care about is that one temperature, calculus seems like it might be enough to deal with this.

But it gets more slippery once the variables start to "talk to" each other: the high temperature tells you a little bit about whether it will rain, because e.g. if the temperature

is very high it's quite likely to be sunny. Suddenly we find ourselves wishing we could talk about conditional probability, but this is a whole can of worms — the relations between these sorts of things can get very complicated very quickly.

The big idea to getting a formalism for this is that:

Our measure spaces Ω will be thought of as a space of entire worlds, with each $\omega \in \Omega$ representing a world. Random variables are functions from worlds to $\mathbb{R}$.

This way, the space of "worlds" takes care of all the messy interdependence.

Then, we can assign "measures" to sets of worlds: for example, to be a fair coin means that if you are only interested in that one coin flip, the "fraction" of worlds in which that coin showed heads should be $\frac{1}{2}$. This is in some ways backwards from what you were told in high-school: officially, we start with the space of worlds, rather than starting with the probabilities.

It will soon be clear that there is no way we can assign a well-defined measure to every single one of the 2^Ω subsets. Fortunately, in practice, we won't need to, and the notion of a σ-algebra will capture the idea of "enough measur-*able* sets for us to get by".

Remark 35.2.1 (Random seeds) — Another analogy if you do some programming: each $\omega \in \Omega$ is a *random seed*, and everything is determined from there.

§35.3 Motivating measure spaces geometrically

So, we have a set Ω of possible points (which in the context of the previous discussion can be thought of as the set of worlds), and we want to assign a *measure* (think volume) to subsets of points in Ω. We will now describe some of the obstacles that we will face, in order to motivate *how* measure spaces are defined (as the previous section only motivated *why* we want such things).

If you try to do this naïvely, you basically immediately run into set-theoretic issues. A good example to think about why this might happen is if $\Omega = \mathbb{R}^2$ with the measure corresponding to area. You can define the area of a triangle as in high school, and you can then try and define the area of a circle, maybe by approximating it with polygons. But what area would you assign to the subset $\mathbb{Q}^2$, for example? (It turns out "zero" is actually a working answer.) Or, a unit disk is composed of infinitely many points; each of the points better have measure zero, but why does their union have measure π then? Blah blah blah.

We'll say more about this later, but you might have already heard of the **Banach-Tarski paradox** which essentially shows there is no good way that you can assign a measure to every single subset of $\mathbb{R}^3$ and still satisfy basic sanity checks. There are just too many possible subsets of Euclidean space.

However, the good news is that most of these sets are not ones that we will ever care about, and it's enough to define measures for certain "sufficiently nice sets". The adjective we will use is *measurable*, and it will turn out that this will be way, way more than good enough for any practical purposes.

We will generally use A, B, ... for measurable sets and denote the entire family of measurable sets by curly $\mathscr{A}$.

§35.4 σ-algebras and measurable spaces

Here's the machine code.

Definition 35.4.1. A **measurable space** consists of a space Ω of points, and a **σ-algebra** $\mathscr{A}$ of subsets of Ω (the "measurable sets" of Ω). The set $\mathscr{A}$ is required to satisfy the following axioms:

- $\mathscr{A}$ contains $\varnothing$ and Ω.
- $\mathscr{A}$ should be closed under complements and *countable* unions/intersections. (Hint on nomenclature: σ usually indicates some sort of "countably finite" condition.)

(Complaint: this terminology is phonetically confusing, because it can be confused with "measure space" later. The way to think about is that "measur*able* spaces have a σ-algebra, so we *could* try to put a measure on it, but we *haven't*, yet.")

Though this definition is how we actually think about it in a few select cases, for the most part, and we will usually instantiate $\mathscr{A}$ in practice in a different way:

Definition 35.4.2. Let Ω be a set, and consider some family of subsets $\mathscr{F}$ of Ω. Then the **σ-algebra generated by $\mathscr{F}$** is the smallest σ-algebra $\mathscr{A}$ which contains $\mathscr{F}$.

As is commonplace in math, when we see "generated", this means we sort of let the definition "take care of itself". So, if $\Omega = \mathbb{R}$, maybe I want $\mathscr{A}$ to contain all open sets. Well, then the definition means it should contain all complements too, so it contains all the closed sets. Then it has to contain all the half-open intervals too, and then.... Rather than try to reason out what exactly the final shape $\mathscr{A}$ looks like (which basically turns out to be impossible), we just give up and say "$\mathscr{A}$ is all the sets you can get if you start with the open sets and apply repeatedly union/complement operations". Or even more bluntly: "start with open sets, shake vigorously".[1]

I've gone on too long with no examples.

Example 35.4.3 (Examples of measurable spaces)

The first two examples actually say what $\mathscr{A}$ is; the third example (most important) will use generation.

(a) If Ω is any set, then the power set $\mathscr{A} = 2^{\Omega}$ is obviously a σ-algebra. This will be used if Ω is countably finite, but it won't be very helpful if Ω is huge.

(b) If Ω is an uncountable set, then we can declare $\mathscr{A}$ to be all subsets of Ω which are either countable, or which have countable complement. (You should check this satisfies the definitions.) This is a very "coarse" algebra.

(c) If Ω is a topological space, the **Borel σ-algebra** is defined as the σ-algebra generated by all the open sets of Ω. We denote it by $\mathscr{B}(\Omega)$, and call the space a **Borel space**. As warned earlier, it is basically impossible to describe what it looks like, and instead you should think of it as saying "we can measure the open sets".

[1] As will be mentioned later in Section 36.4, an explicit construction using transfinite induction is possible. That construction is also useful for, for example, proving $|\mathscr{B}(\mathbb{R})| = |\mathbb{R}|$.

Question 35.4.4. Show that the closed sets are in $\mathscr{B}(\Omega)$ for any topological space Ω. Show that $[0,1)$ is also in $\mathscr{B}(\mathbb{R})$.

§35.5 Measure spaces

Definition 35.5.1. Measurable spaces $(\Omega, \mathscr{A})$ are then equipped with a function $\mu\colon \mathscr{A} \to [0,+\infty]$ called the **measure**, which is required to satisfy the following axioms:

- $\mu(\varnothing) = 0$
- **Countable additivity**: If A_1, A_2, ... are disjoint sets in $\mathscr{A}$, then
$$\mu\left(\bigsqcup_n A_n\right) = \sum_n \mu(A_n).$$

The triple $(\Omega, \mathscr{A}, \mu)$ is called a **measure space**. It's called a **probability space** if $\mu(\Omega) = 1$.

Exercise 35.5.2 (Weaker equivalent definitions)**.** I chose to give axioms for $\mathscr{A}$ and μ that capture how people think of them in practice, which means there is some redundancy: for example, being closed under complements and unions is enough to get intersections, by de Morgan's law. Here are more minimal definitions, which are useful if you are trying to prove something satisfies them to reduce the amount of work you have to do:

(a) The axioms on $\mathscr{A}$ can be weakened to (i) $\varnothing \in \mathscr{A}$ and (ii) $\mathscr{A}$ is closed under complements and countable unions.

(b) The axioms on μ can be weakened to (i) $\mu(\varnothing) = 0$, (ii) $\mu(A \sqcup B) = \mu(A) + \mu(B)$, and (iii) for $A_1 \supseteq A_2 \supseteq \cdots$ with $\mu(A_1) < \infty$, we have $\mu\left(\bigcap_n A_n\right) = \lim_n \mu(A_n)$.

Remark 35.5.3 — Here are some immediate remarks on these definitions.

- If $A \subseteq B$ are measurable, then $\mu(A) \le \mu(B)$ since $\mu(B) = \mu(A) + \mu(B - A)$.
- In particular, in a probability space all measures are in $[0,1]$. On the other hand, for general measure spaces we'll allow $+\infty$ as a possible measure (hence the choice of $[0,+\infty]$ as codomain for μ).
- We want to allow at least countable unions / additivity because with finite unions it's too hard to make progress: it's too hard to estimate the area of a circle without being able to talk about limits of countably infinitely many triangles.

We *don't* want to allow uncountable unions and additivity, because uncountable sums basically never work out. In particular, there is a nice elementary exercise as follows:

Exercise 35.5.4 (Tricky)**.** Let S be an uncountable set of positive real numbers. Show that some finite subset $T \subseteq S$ has sum greater than 10^{2019}. Colloquially, "uncountably many positive reals cannot have finite sum".

So countable sums are as far as we'll let the infinite sums go. This is the reason why we considered σ-algebras in the first place.

Example 35.5.5 (Measures)
We now discuss measures on each of the spaces in our previous examples.

(a) If $\mathscr{A} = 2^\Omega$ (or for that matter any $\mathscr{A}$) we may declare $\mu(A) = |A|$ for each $A \in \mathscr{A}$ (even if $|A| = \infty$). This is called the **counting measure**, simply counting the number of elements.

This is useful if Ω is countably infinite, and optimal if Ω is finite (and nonempty). In the latter case, we will often normalize by $\mu(A) = \frac{|A|}{|\Omega|}$ so that Ω becomes a probability space.

(b) Suppose Ω was uncountable and we took $\mathscr{A}$ to be the countable sets and their complements. Then

$$\mu(A) = \begin{cases} 0 & A \text{ is countable} \\ 1 & \Omega \setminus A \text{ is countable} \end{cases}$$

is a measure. (Check this.)

(c) Elephant in the room: defining a measure on $\mathscr{B}(\Omega)$ is hard even for $\Omega = \mathbb{R}$, and is done in the next chapter. So you will have to hold your breath. Right now, all you know is that by declaring my *intent* to define a measure $\mathscr{B}(\Omega)$, I am hoping that at least every open set will have a volume.

§35.6 A hint of Banach-Tarski

I will now try to convince you that $\mathscr{B}(\Omega)$ is a necessary concession, and for general topological spaces like $\Omega = \mathbb{R}^n$, there is no hope of assigning a measure to 2^Ω. (In the literature, this example is called a Vitali set.)

Example 35.6.1 (A geometric example why $\mathscr{A} = 2^\Omega$ is unsuitable)
Let Ω denote the unit circle in $\mathbb{R}^2$ and $\mathscr{A} = 2^\Omega$. We will show that any measure μ on Ω with $\mu(\Omega) = 1$ will have undesirable properties.
Let $\sim$ denote an equivalence relation on Ω defined as follows: two points are equivalent if they differ by a rotation around the origin by a rational multiple of π. We may pick a representative from each equivalence class, letting X denote the set of representatives. Then

$$\Omega = \bigsqcup_{\substack{q \in \mathbb{Q} \\ 0 \le q < 2}} (X \text{ rotated by } q\pi \text{ radians}).$$

Since we've only rotated X, each of the rotations should have the same measure m. But $\mu(\Omega) = 1$, and there is no value we can assign that measure: if $m = 0$ we get $\mu(\Omega) = 0$ and $m > 0$ we get $\mu(\Omega) = \infty$.

Remark 35.6.2 (Choice) — Experts may recognize that picking a representative (i.e. creating set X) technically requires the Axiom of Choice. That is why, when people talk about Banach-Tarski issues, the Axiom of Choice almost always gets

honorable mention as well.

Stay tuned to actually see a construction for $\mathscr{B}(\mathbb{R}^n)$ in the next chapter.

§35.7 Measurable functions

Prototypical example for this section: For $S \subseteq \Omega$, $\mathbf{1}_S \colon \Omega \to \mathbb{R}$ is a measurable function if and only if S is a measurable set.

In the past, when we had topological spaces, we considered continuous functions. The analog here is:

Definition 35.7.1. Let $(X, \mathscr{A})$ and $(Y, \mathscr{B})$ be measurable spaces (or measure spaces). A function $f \colon X \to Y$ is **measurable** if for any measurable set $S \subseteq Y$ (i.e. $S \in \mathscr{B}$) we have $f^{\text{pre}}(S)$ is measurable (i.e. $f^{\text{pre}}(S) \in \mathscr{A}$).

In most cases Y is actually a topological space with the Borel σ-algebra (e.g. $Y = \mathbb{R}$) and in that case we can replace "measurable set S" with "open set S". You can take this as a standing assumption for the rest of this text.

Apart from the obvious symmetry with the definition of continuous function, as we will see in Section 37.2, this definition is such that for a nonnegative function $f \colon \Omega \to \mathbb{R}_{\geq 0}$, $\int_\Omega f \, d\mu$ exists if and only if f is measurable.

Remark 35.7.2 — By symmetry, you might have guessed that a function $f : X \to Y$ is measurable if for any measurable $S \subseteq Y$, we have $f^{\text{pre}}(S) \subseteq X$ is measurable. Nevertheless, this definition doesn't work the way we expect — even continuous function can fail this definition.

Example 35.7.3 (Continuous function with non-measurable preimage of measurable set)

Let $f \colon [0,1] \to [0,1]$ be the Devil's Staircase (or Cantor function). This function is continuous, and has the property that, let $C \subseteq [0,1]$ be the Cantor set, then $|C| = 0$, yet $f^{\text{img}}(C) = [0,1]$ with measure 1.
Let $g \colon [0,1] \to [0,2]$ be defined by $g(x) = f(x) + x$. Then,

- For each open interval (a,b) that is removed from the Cantor set C, then $|g^{\text{img}}((a,b))| = |(a,b)|$.
- $g^{\text{img}}(C)$ has measure 1.

Note that g is bijective, let $h \colon [0,2] \to [0,1]$, $h = g^{-1}$. Then h is continuous, however:

- $h^{\text{pre}}(C) = g^{\text{img}}(C)$ has measure 1, so it has some non-measurable subset,
- C has measure 0, so every subset of C is (Lebesgue) measurable,
- thus, $h^{\text{pre}}(D)$ is non-measurable for some measurable subset $D \subseteq C$.

In practice, most functions you encounter will be continuous anyways, and in that case we are fine.

Proposition 35.7.4 (Continuous implies Borel measurable)

Suppose X and Y are topological spaces and we pick the Borel measures on both. A function $f\colon X \to Y$ which is continuous as a map of topological spaces is also measurable.

Proof. Follows from the fact that pre-images of open sets are open, thus Borel measurable. □

§35.8 On the word "almost"

In later chapters we will begin seeing the phrase "almost everywhere" and "almost surely" start to come up, and it seems prudent to take the time to talk about it now.

Definition 35.8.1. We say that property P occurs **almost everywhere** or **almost surely** if the set

$$\{\omega \in \Omega \mid P \text{ does not hold for } \omega\}$$

has measure zero.

For example, if we say "$f = g$ almost everywhere" for some functions f and g defined on a measure space Ω, then we mean that $f(\omega) = g(\omega)$ for all $\omega \in \Omega$ other than a measure-zero set.

There, that's the definition. The main thing to now update your instincts on is that

In measure theory, we basically only care about things up to almost-everywhere.

Here are some examples:

- If $f = g$ almost everywhere, then measure theory will basically not tell these functions apart. For example, $\int_\Omega f \, d\omega = \int_\Omega g \, d\omega$ will hold for two functions agreeing almost everywhere.
- As another example, if we prove "there exists a unique function f such that so-and-so", the uniqueness is usually going to be up to measure-zero sets.

You can think of this sort of like group isomorphism, where two groups are considered "basically the same" when they are isomorphic, except this one might take a little while to get used to.[2]

§35.9 A few harder problems to think about

Problem 35A†. Let $(\Omega, \mathscr{A}, \mu)$ be a probability space. Show that the intersection of countably many sets of measure 1 also has measure 1.

Problem 35B (On countable σ-algebras)**.** Let $\mathscr{A}$ be a σ-algebra on a set Ω. Suppose that $\mathscr{A}$ has countable cardinality. Prove that $|\mathscr{A}|$ is finite and equals a power of 2.

[2] In fact, some people will even define functions on measure spaces as *equivalence classes* of maps, modded out by agreement outside a measure zero set.

36 Constructing the Borel and Lebesgue measure

It's very difficult to define in one breath a measure on the Borel space $\mathscr{B}(\mathbb{R}^n)$. It is easier if we define a weaker notion first. There are two such weaker notions that we will define:

- A **pre-measure**: satisfies the axioms of a measure, but defined on *fewer* sets than a measure: they'll be defined on an "algebra" rather than the full-fledged "σ-algebra".
- An **outer measure**: defined on 2^Ω but satisfies weaker axioms.

It will turn out that pre-measures yield outer measures, and outer measures yield measures.

§36.1 Pre-measures

Prototypical example for this section: Let $\Omega = \mathbb{R}^2$. Then we take $\mathscr{A}_0$ generated by rectangles, with μ_0 the usual area.

The way to define a pre-measure is to weaken the σ-algebra to an algebra.

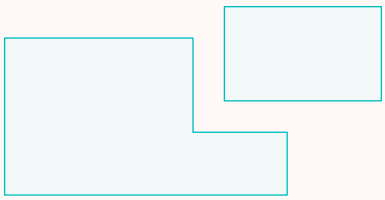
Definition 36.1.1. Let Ω be a set. We define notions of an **algebra**, which is the same as σ-algebra except with "countable" replaced by finite everywhere.

That is: an algebra $\mathscr{A}_0$ on Ω is a nonempty subset of 2^Ω, which is closed under complement and *finite* union. The smallest algebra containing a subset $\mathscr{F} \subseteq 2^\Omega$ is the **algebra generated by $\mathscr{F}$**.

In practice, we will basically always use generation for algebras.

Example 36.1.2

When $\Omega = \mathbb{R}^n$, we can let $\mathcal{L}_0$ be the algebra generated by $[a_1, b_1] \times \cdots \times [a_n, b_n]$. A typical element might look like:



Unsurprisingly, since we have *finitely* many rectangles and their complements involved, in this case we actually *can* unambiguously assign an area, and will do so soon.

Definition 36.1.3. A **pre-measure** μ_0 on a algebra $\mathscr{A}_0$ is a function $\mu_0 \colon \mathscr{A}_0 \to [0, +\infty]$ which satisfies the axioms

- $\mu_0(\varnothing) = 0$, and

- **Countable additivity**: if A_1, A_2, ... are disjoint sets in $\mathscr{A}_0$ and *moreover* the disjoint union $\bigsqcup A_i$ is contained in $\mathscr{A}_0$ (not guaranteed by algebra axioms!), then

$$\mu_0\left(\bigsqcup_n A_n\right) = \sum_n \mu_0(A_n).$$

Example 36.1.4 (The pre-measure on $\mathbb{R}^n$)
Let $\Omega = \mathbb{R}^2$. Then, let $\mathcal{L}_0$ be the algebra generated by rectangles $[a_1, a_2] \times [b_1, b_2]$. We then let

$$\mu_0([a_1, a_2] \times [b_1, b_2]) = (a_2 - a_1)(b_2 - b_1)$$

the area of the rectangle. As elements of $\mathcal{L}_0$ are simply *finite* unions of rectangles and their complements (picture drawn earlier), it's not difficult to extend this to a pre-measure λ_0 which behaves as you expect — although we won't do this.

Since we are sweeping something under the rug that turns out to be conceptually important, I'll go ahead and blue-box it.

Proposition 36.1.5 (Geometry sanity check that we won't prove)
For $\Omega = \mathbb{R}^n$ and $\mathcal{L}_0$ the algebra generated by rectangular prisms, one can define a pre-measure λ_0 on $\mathcal{L}_0$.

From this point forwards, we will basically do almost no geometry[1] whatsoever in defining the measure $\mathscr{B}(\mathbb{R}^n)$, and only use set theory to extend our measure. So, Proposition 36.1.5 is the only sentry which checks to make sure that our "initial definition" is sane.

To put the point another way, suppose an **insane scientist**[2] tried to define a notion of area in which every rectangle had area 1. Intuitively, this shouldn't be possible: every rectangle can be dissected into two halves and we ought to have $1 + 1 \neq 1$. However, the only thing that would stop them is that they couldn't extend their pre-measure on the algebra $\mathcal{L}_0$. If they somehow got past that barrier and got a pre-measure, nothing in the rest of the section would prevent them from getting an entire *bona fide* measure with this property. Thus, in our construction of the Lebesgue measure, most of the geometric work is captured in the (omitted) proof of Proposition 36.1.5.

§36.2 Outer measures

Prototypical example for this section: Keep taking $\Omega = \mathbb{R}^2$; see the picture to follow.

The other way to weaken a measure is to relax the countable additivity, and this yields the following:

Definition 36.2.1. An **outer measure** μ^* on a set Ω is a function $\mu^* \colon 2^\Omega \to [0, +\infty]$ satisfying the following axioms:

- $\mu^*(\varnothing) = 0$;

[1] White lie. Technically, we will use one more fact: that open sets of $\mathbb{R}^n$ can be covered by countably infinitely many rectangles, as in Exercise 36.5.1. This step doesn't involve any area assignments, though.

[2] Because "mad scientists" are overrated.

- if $E \subseteq F$ and $E, F \in 2^\Omega$ then $\mu^*(E) \le \mu^*(F)$;
- for any subsets E_1, E_2, ... of Ω we have

$$\mu^*\left(\bigcup_n E_n\right) \le \sum_n \mu^*(E_n).$$

(I don't really like the word "outer measure", since I think it is a bit of a misnomer: I would rather call it "fake measure", since it's not a measure either.)

The reason for the name "outer measure" is that you almost always obtain outer measures by approximating them from "outside" sets. Officially, the result is often stated as follows (as Problem 36A†).

> For a set Ω, let $\mathcal{E}$ be *any* subset of 2^Ω and let $\rho\colon \mathcal{E} \to [0, +\infty]$ be *any* function. Then
> $$\mu^*(E) = \inf\left\{\sum_{n=1}^{\infty} \rho(E_n) \mid E_n \in \mathcal{E},\ E \subseteq \bigcup_{n=1}^{\infty} E_n\right\}$$
> is an outer measure.

However, I think the above theorem is basically always wrong to use in practice, because it is *way too general*. As I warned with the insane scientist, we really do want some sort of sanity conditions on ρ: otherwise, if we apply the above result as stated, there is no guarantee that μ^* will be compatible with ρ in any way.

So, I think it is really better to apply the theorem to pre-measures μ_0 for which one *does* have some sort of guarantee that the resulting μ^* is compatible with μ_0. In practice, this is always how we will want to construct our outer measures.

Theorem 36.2.2 (Constructing outer measures from pre-measures)

Let μ_0 be a pre-measure on an algebra $\mathscr{A}_0$ on a set Ω.

(a) The map $\mu^*\colon 2^\Omega \to [0, +\infty]$ defined by

$$\mu^*(E) = \inf\left\{\sum_{n=1}^{\infty} \mu_0(A_n) \mid A_n \in \mathscr{A}_0,\ E \subseteq \bigcup_{n=1}^{\infty} A_n\right\}$$

is an outer measure.

(b) Moreover, this measure agrees with μ_0 on sets in $\mathscr{A}_0$.

Intuitively, what is going on is that $\mu^*(A)$ is the infimum of coverings of A by countable unions of elements in $\mathscr{A}_0$. Part (b) is the first half of the compatibility condition I promised; the other half appears later as Proposition 36.3.2.

Proof of Theorem 36.2.2. As alluded to already, part (a) is a special case of Problem 36A† (and proving it in this generality is actually easier, because you won't be distracted by unnecessary properties).

We now check (b), that $\mu^*(A) = \mu_0(A)$ for $A \in \mathscr{A}_0$. One bound is quick:

Question 36.2.3. Show that $\mu^*(A) \le \mu_0(A)$.

For the reverse, suppose that $A \subseteq \bigcup_n A_n$. Then, define the sets

$$\begin{aligned} B_1 &= A \cap A_1 \\ B_2 &= (A \cap A_2) \setminus B_1 \\ B_3 &= (A \cap A_3) \setminus (B_1 \cup B_2) \\ &\vdots \end{aligned}$$

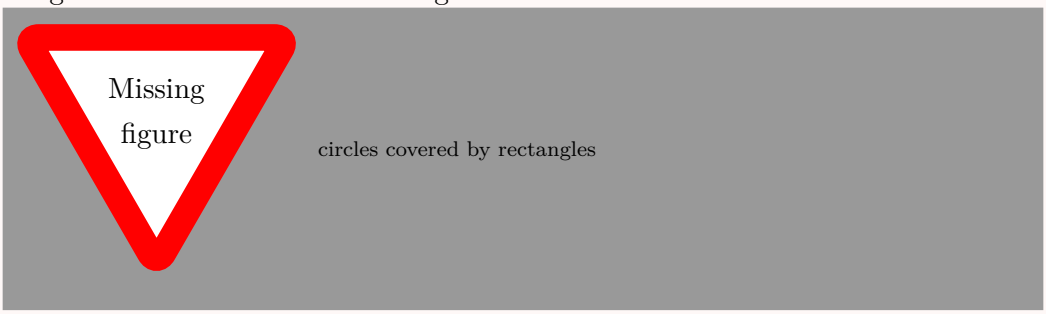
and so on. Then the B_n are disjoint elements of $\mathscr{A}_0$ with $B_n \subset A_n$, and we have rigged the definition so that $\bigsqcup_n B_n = A$. Thus by definition of pre-measure,

$$\mu_0(A) = \sum_n \mu_0(B_n) \le \sum_n \mu_0(A_n)$$

as desired. □

Example 36.2.4

Let $\Omega = \mathbb{R}^2$ and λ_0 the pre-measure from before. Then $\lambda^*(A)$ is, intuitively, the infimum of coverings of the set A by rectangles. Here is a picture you might use to imagine the situation with A being the unit disk.



§36.3 Carathéodory extension for outer measures

We will now take any outer measure and turn it into a proper measure. To do this, we first need to specify the σ-algebra on which we will define the measure.

Definition 36.3.1. Let μ^* be an outer measure. We say a set A is **Carathéodory measurable with respect to μ^***, or just **μ^*-measurable**, if the following condition holds: for any set $E \in 2^\Omega$,

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \setminus A).$$

This definition is hard to motivate, but turns out to be the right one. One way to motivate is this: it turns out that in $\mathbb{R}^n$, it will be equivalent to a reasonable geometric condition (which I will state in Proposition 36.4.3), but since that geometric definition requires information about $\mathbb{R}^n$ itself, this is the "right" generalization for general measure spaces.

Since our goal was to extend our $\mathscr{A}_0$, we had better make sure this definition lets us measure the initial sets that we started with!

Proposition 36.3.2 (Carathéodory measurability is compatible with the initial $\mathscr{A}_0$)
Suppose μ^* was obtained from a pre-measure μ_0 on an algebra $\mathscr{A}_0$, as in Theorem 36.2.2. Then every set in $\mathscr{A}_0$ is μ^*-measurable.

This is the second half of the compatibility condition that we get if we make sure our initial μ_0 at least satisfies the pre-measure axioms. (The first half was (b) of Theorem 36.2.2.)

Proof. Let $A \in \mathscr{A}_0$ and $E \in 2^\Omega$; we wish to prove $\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \setminus A)$. The definition of outer measure already requires $\mu^*(E) \le \mu^*(E \cap A) + \mu^*(E \setminus A)$ and so it's enough to prove the reverse inequality.

By definition of infimum, for any $\varepsilon > 0$, there is a covering $E \subset \bigcup_n A_n$ with $\mu^*(E) + \varepsilon \ge \sum_n \mu_0(A_n)$. But

$$\sum_n \mu_0(A_n) = \sum_n (\mu_0(A_n \cap A) + \mu_0(A_n \setminus A)) \ge \mu^*(E \cap A) + \mu^*(E \setminus A)$$

with the first equality being the definition of pre-measure on $\mathscr{A}_0$, the second just being by definition of μ^* (since $A_n \cap A$ certainly covers $E \cap A$, for example). Thus $\mu^*(E) + \varepsilon \ge \mu^*(E \cap A) + \mu^*(E \setminus A)$. Since the inequality holds for any $\varepsilon > 0$, we're done. □

To add extra icing onto the cake, here is one more niceness condition which our constructed measure will happen to satisfy.

Definition 36.3.3. A **null set** of a measure space $(\Omega, \mathscr{A}, \mu)$ is a set $A \in \mathscr{A}$ with $\mu(A) = 0$. A measure space $(\Omega, \mathscr{A}, \mu)$ is **complete** if whenever A is a null set, then all subsets of A are in $\mathscr{A}$ as well (and hence null sets).

This is a nice property to have, for obvious reasons. Visually, if I have a bunch of dust which I *already* assigned weight zero, and I blow away some of the dust, then the remainder should still have an assigned weight — zero. The extension theorem will give us σ-algebras with this property.

Theorem 36.3.4 (Carathéodory extension theorem for outer measures)
If μ^* is an outer measure, and $\mathscr{A}^{\text{cm}}$ is the set of μ^*-measurable sets with respect to μ^*, then $\mathscr{A}^{\text{cm}}$ is a σ-algebra on Ω, and the restriction μ^{cm} of μ^* to $\mathscr{A}^{\text{cm}}$ gives a *complete* measure space.

(Phonetic remark: you can think of the superscript $^{\text{cm}}$ as standing for either "Carathéodory measurable" or "complete". Both are helpful for remembering what this represents. This notation is not standard but the pun was too good to resist.)

Thus, if we compose Theorem 36.2.2 with Theorem 36.3.4, we find that every pre-measure μ_0 on an algebra $\mathscr{A}_0$ naturally gives a σ-algebra $\mathscr{A}^{\text{cm}}$ with a complete measure μ^{cm}, and our two compatibility results (namely (b) of Theorem 36.2.2, together with Proposition 36.3.2) means that $\mathscr{A}^{\text{cm}} \supset \mathscr{A}_0$ and μ^{cm} agrees with μ.

Here is a table showing the process, where going down each row of the table corresponds to restriction process.

		Construct order	Notes
2^Ω	μ^*	Step 2	μ^* is outer measure obtained from μ_0
$\mathscr{A}^{\text{cm}}$	μ^{cm}	Step 3	$\mathscr{A}^{\text{cm}}$ defined as μ^*-measurable sets, $(\mathscr{A}^{\text{cm}}, \mu^{\text{cm}})$ is complete.
$\mathscr{A}_0$	μ_0	Step 1	μ_0 is a pre-measure

§36.4 Defining the Lebesgue measure

This lets us finally define the Lebesgue measure on $\mathbb{R}^n$. We wrap everything together at once now.

Definition 36.4.1. We create a measure on $\mathbb{R}^n$ by the following procedure.

- Start with the algebra $\mathcal{L}_0$ generated by rectangular prisms, and define a *pre-measure* λ_0 on this $\mathcal{L}_0$ (this was glossed over in the example).
- By Theorem 36.2.2, this gives the **Lebesgue outer measure** λ^* on $2^{\mathbb{R}^n}$, which is compatible on all the rectangular prisms.
- By Carathéodory (Theorem 36.3.4), this restricts to a complete measure λ on the σ-algebra $\mathcal{L}(\mathbb{R}^n)$ of λ^*-measurable sets (which as promised contains all rectangular prisms).[3]

The resulting complete measure, denoted λ, is called the **Lebesgue measure**.

The algebra $\mathcal{L}(\mathbb{R}^n)$ we obtained will be called the **Lebesgue σ-algebra**; sets in it are said to be **Lebesgue measurable**.

Here is the same table from before, with the values filled in for the special case $\Omega = \mathbb{R}^n$, which gives us the Lebesgue algebra.

		Construct order	Notes
$2^{\mathbb{R}^n}$	λ^*	Step 2	λ^* is Lebesgue outer measure
$\mathcal{L}(\mathbb{R}^n)$	λ	Step 3	Lebesgue σ-algebra (complete)
$\mathcal{L}_0$	λ_0	Step 1	Define pre-measure on rectangles

Of course, now that we've gotten all the way here, if we actually want to *compute* any measures, we can mostly gleefully forget about how we actually constructed the measure and just use the properties. The hard part was to showing that there *is* a way to assign measures consistently; actually figuring out what that measure's value is *given that it exists* is often much easier. Here is an example.

Example 36.4.2 (The Cantor set has measure zero)

The standard **middle-thirds Cantor set** is the subset $[0,1]$ obtained as follows: we first delete the open interval $(1/3, 2/3)$. This leaves two intervals $[0,1/3]$ and $[2/3,1]$ from which we delete the middle thirds again from both, i.e. deleting $(1/9, 2/9)$ and $(7/9, 8/9)$. We repeat this procedure indefinitely and let C denote the result. An illustration is shown below.

Image from [1207]

It is a classic fact that C is uncountable (it consists of ternary expansions omitting the digit 1). But it is measurable (it is an intersection of closed sets!) and we

[3] If I wanted to be consistent with the previous theorems, I might prefer to write $\mathcal{L}^{\text{cm}}$ and λ^{cm} for emphasis. It seems no one does this, though, so I won't.

contend it has measure zero. Indeed, at the nth step, the result has measure $(2/3)^n$ leftover. So $\mu(C) \le (2/3)^n$ for every n, forcing $\mu(C) = 0$.

This is fantastic, but there is one elephant in the room: how are the Lebesgue σ-algebra and the Borel σ-algebra related? To answer this question briefly, I will state two results (but another answer is given in the next section). The first is a geometric interpretation of the strange Carathéodory measurable hypothesis.

Proposition 36.4.3 (A geometric interpretation of Lebesgue measurability)

A set $A \subseteq \mathbb{R}^n$ is Lebesgue measurable if and only if for every $\varepsilon > 0$, there is an open set $U \supset A$ such that

$$\lambda^*(U \setminus A) < \varepsilon$$

where λ^* is the Lebesgue outer measure.

I want to say that this was Lebesgue's original formulation of "measurable", but I'm not sure about that. In any case, we won't need to use this, but it's good to see that our definition of Lebesgue measurable has a down-to-earth geometric interpretation.

Question 36.4.4. Deduce that every open set is Lebesgue measurable. Conclude that the Lebesgue σ-algebra contains the Borel σ-algebra. (A different proof is given later on.)

However, the containment is proper: there are more Lebesgue measurable sets than Borel ones. Indeed, it can actually be proven using transfinite induction (though we won't) that $|\mathscr{B}(\mathbb{R})| = |\mathbb{R}|$.[4] Using this, one obtains:

Exercise 36.4.5. Show the Borel σ-algebra is not complete. (Hint: consider the Cantor set. You won't be able to write down an example of a non-measurable set, but you can use cardinality arguments.) Thus the Lebesgue σ-algebra strictly contains the Borel one.

Nonetheless, there is a great way to describe the Lebesgue σ-algebra, using the idea of completeness.

Definition 36.4.6. Let $(\Omega, \mathscr{A}, \mu)$ be a measure space. The **completion** $(\Omega, \overline{\mathscr{A}}, \overline{\mu})$ is defined as follows: we let

$$\overline{\mathscr{A}} = \{A \cup N \mid A \in \mathscr{A}, N \text{ subset of null set}\}.$$

and $\overline{\mu}(A \cup N) = \mu(A)$. One can check this is well-defined, and in fact $\overline{\mu}$ is the unique extension of μ from $\mathscr{A}$ to $\overline{\mathscr{A}}$.

This looks more complicated than it is. Intuitively, all we are doing is "completing" the measure by telling $\overline{\mu}$ to regard any subset of a null set as having measure zero, too.

Then, the saving grace:

Theorem 36.4.7 (Lebesgue is completion of Borel)

For $\mathbb{R}^n$, the Lebesgue measure is the completion of the Borel measure.

[4] See https://math.stackexchange.com/a/70891 for a sketch.

Proof. This actually follows from results in the next section, namely Exercise 36.5.1 and part (c) of Carathéodory for pre-measures (Theorem 36.5.5). □

§36.5 A fourth row: Carathéodory for pre-measures

Prototypical example for this section: The fourth row for the Lebesgue measure is $\mathscr{B}(\mathbb{R}^n)$.

In many cases, $\mathscr{A}^{\text{cm}}$ is actually bigger than our original goal, and instead we only need to extend μ_0 on $\mathscr{A}_0$ to μ on $\mathscr{A}$, where $\mathscr{A}$ is the σ-algebra generated by $\mathscr{A}_0$. Indeed, our original goal was to get $\mathscr{B}(\mathbb{R}^n)$, and in fact:

Exercise 36.5.1. Show that $\mathscr{B}(\mathbb{R}^n)$ is the σ-algebra generated by the $\mathcal{L}_0$ we defined earlier.

Fortunately, this restriction is trivial to do.

Question 36.5.2. Show that $\mathscr{A}^{\text{cm}} \supset \mathscr{A}$, so we can just restrict μ^{cm} to $\mathscr{A}$.

We will in a moment add this as the fourth row in our table.

However, if this is the end goal, than a somewhat different Carathéodory theorem can be stated because often one more niceness condition holds:

Definition 36.5.3. A pre-measure or measure μ on Ω is **σ-finite** if Ω can be written as a countable union $\Omega = \bigcup_n A_n$ with $\mu(A_n) < \infty$ for each n.

Question 36.5.4. Show that the pre-measure λ_0 we had, as well as the Borel measure $\mathscr{B}(\mathbb{R}^n)$, are both σ-finite.

Actually, for us, σ-finite is basically always going to be true, so you can more or less just take it for granted.

Theorem 36.5.5 (Carathéodory extension theorem for pre-measures)

Let μ_0 be a pre-measure on an algebra $\mathscr{A}_0$ of Ω, and let $\mathscr{A}$ denote the σ-algebra generated by $\mathscr{A}_0$. Let $\mathscr{A}^{\text{cm}}$, μ^{cm} be as in Theorem 36.3.4. Then:

(a) The restriction of μ^{cm} to $\mathscr{A}$ gives a measure μ extending μ_0.

(b) If μ_0 was σ-finite, then μ is the unique extension of μ_0 to $\mathscr{A}$.

(c) If μ_0 was σ-finite, then μ^{cm} is the completion of μ, hence the unique extension of μ_0 to $\mathscr{A}^{\text{cm}}$.

Here is the updated table, with comments if μ_0 was indeed σ-finite.

		Construct order	Notes
2^Ω	μ^*	Step 2	μ^* is outer measure obtained from μ_0
$\mathscr{A}^{\text{cm}}$	μ^{cm}	Step 3	$(\mathscr{A}^{\text{cm}}, \mu^{\text{cm}})$ is completion $(\mathscr{A}, \mu)$, $\mathscr{A}^{\text{cm}}$ defined as μ^*-measurable sets
$\mathscr{A}$	μ	Step 4	$\mathscr{A}$ defined as σ-alg. generated by $\mathscr{A}_0$
$\mathscr{A}_0$	μ_0	Step 1	μ_0 is a pre-measure

And here is the table for $\Omega = \mathbb{R}^n$, with Borel and Lebesgue in it.

		Construct order	Notes
$2^{\mathbb{R}^n}$	λ^*	Step 2	λ^* is Lebesgue outer measure
$\mathcal{L}(\mathbb{R}^n)$	λ	Step 3	Lebesgue σ-algebra, completion of Borel one
$\mathscr{B}(\mathbb{R}^n)$	μ	Step 4	Borel σ-algebra, generated by $\mathcal{L}_0$
$\mathcal{L}_0$	λ_0	Step 1	Define pre-measure on rectangles

Going down one row of the table corresponds to restriction, while each of $\mu_0 \to \mu \to \mu^{\text{cm}}$ is a unique extension when μ_0 is σ-finite.

Proof of Theorem 36.5.5. For (a): this is just Theorem 36.2.2 and Theorem 36.3.4 put together, combined with the observation that $\mathscr{A}^* \supset \mathscr{A}_0$ and hence $\mathscr{A}^* \supset \mathscr{A}$. Parts (b) and (c) are more technical, and omitted. □

§36.6 From now on, we assume the Borel measure

explain why

§36.7 A few harder problems to think about

Problem 36A† (Constructing outer measures from arbitrary ρ)**.** For a set Ω, let $\mathcal{E}$ be *any* subset of 2^Ω and let $\rho\colon \mathcal{E} \to [0, +\infty]$ be *any* function. Prove that

$$\mu^*(E) = \inf\left\{ \sum_{n=1}^{\infty} \rho(E_n) \mid E_n \in \mathcal{E},\ E \subseteq \bigcup_{n=1}^{\infty} E_n \right\}$$

is an outer measure.

Problem 36B (The insane scientist)**.** Let $\Omega = \mathbb{R}^2$, and let $\mathcal{E}$ be the set of (non-degenerate) rectangles. Let $\rho(E) = 1$ for every rectangle $E \in \mathcal{E}$. Ignoring my advice, the insane scientist uses ρ to construct an outer measure μ^*, as in Problem 36A†.

(a) Find $\mu^*(S)$ for each subset S of $\mathbb{R}^2$.

(b) Which sets are μ^*-measurable?

You should find that no rectangle is μ^*-measurable, unsurprisingly foiling the scientist.

Problem 36C. A function $f\colon \mathbb{R} \to \mathbb{R}$ is continuous. Must f be measurable with respect to the Lebesgue measure on $\mathbb{R}$?

37 Lebesgue integration

On any measure space $(\Omega, \mathscr{A}, \mu)$ we can then, for a function $f\colon \Omega \to [0,\infty]$ define an integral

$$\int_\Omega f \, d\mu.$$

This integral may be $+\infty$ (even if f is finite). As the details of the construction won't matter for us later on, we will state the relevant definitions, skip all the proofs, and also state all the properties that we actually care about. Consequently, this chapter will be quite short.

§37.1 The definition

The construction is done in four steps.

Definition 37.1.1. If A is a measurable set of Ω, then the **indicator function** $\mathbf{1}_A\colon \Omega \to \mathbb{R}$ is defined by

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

Step 1 (Indicator functions) — For an indicator function, we require

$$\int_\Omega \mathbf{1}_A \, d\mu \coloneqq \mu(A)$$

(which may be infinite).

We extend this linearly now for nonnegative functions which are sums of indicators: these functions are called **simple functions**.

Step 2 (Simple functions) — Let $A_1, \dots, A_n$ be a finite collection of measurable sets. Let $c_1, \dots, c_n$ be either nonnegative real numbers or $+\infty$. Then we define

$$\int_\Omega \left(\sum_{i=1}^n c_i \mathbf{1}_{A_i} \right) d\mu \coloneqq \sum_{i=1}^n c_i \mu(A_i).$$

If $c_i = \infty$ and $\mu(A_i) = 0$, we treat $c_i\mu(A_i) = 0$.

One can check the resulting sum does not depend on the representation of the simple function as $\sum c_i \mathbf{1}_{A_i}$. In particular, it is compatible with the previous step.

Conveniently, this is already enough to define the integral for $f\colon \Omega \to [0,+\infty]$. Note that $[0,+\infty]$ can be thought of as a topological space where we add new open sets $(a,+\infty]$ for each real number a to our usual basis of open intervals. Thus we can equip it with the Borel sigma-algebra.[1]

[1] We *could* also try to define a measure on it, but we will not: it is a good enough for us that it is a measurable space.

Step 3 (Nonnegative functions) — For each measurable function $f : \Omega \to [0, +\infty]$, let

$$\int_\Omega f \, d\mu \coloneqq \sup_{0 \le s \le f} \left(\int_\Omega s \, d\mu \right)$$

where the supremum is taken over all *simple* s such that $0 \le s \le f$. As before, this integral may be $+\infty$.

That is,

We define the integral $\int_\Omega f \, d\mu$ by approximating it from below with simple functions, for which we know how to integrate.

One can check this is compatible with the previous definitions. At this point, we introduce an important term.

Definition 37.1.2. A measurable (nonnegative) function $f : \Omega \to [0, +\infty]$ is **absolutely integrable** or just **integrable** if $\int_\Omega f \, d\mu < \infty$.

Warning: I find "integrable" to be *really* confusing terminology. Indeed, *every* measurable function from Ω to $[0, +\infty]$ can be assigned a Lebesgue integral, it's just that this integral may be $+\infty$. So the definition is far more stringent than the name suggests. Even constant functions can fail to be integrable:

Example 37.1.3 (We really should call it "finitely integrable")
The constant function 1 is *not* integrable on $\mathbb{R}$, since $\int_{\mathbb{R}} 1 \, d\mu = \mu(\mathbb{R}) = +\infty$.

For this reason, I will usually prefer the term "absolutely integrable". (If it were up to me, I would call it "finitely integrable", and usually do so privately.)

Remark 37.1.4 (Why don't we approximate the integral from above?) — For bounded functions on measure spaces with $|\Omega| < \infty$, we can equivalently define

$$\int_\Omega f \, d\mu \coloneqq \inf_{0 \le f \le s} \left(\int_\Omega s \, d\mu \right)$$

where the infimum is taken over all simple s such that $f \le s$. However, if the functions are unbounded or $|\Omega| = \infty$, the situation is not that simple:

- The function $f(x) = x^{-2}$ defined over $\Omega = (1, \infty)$ is absolutely integrable, yet for all simple s such that $f \le s$ we have $\int_\Omega s \, d\mu = \infty$.
- The function $f(x) = x^{-0.5}$ defined over $\Omega = (0, 1)$ is absolutely integrable, yet there's no simple s such that $f \le s$ and s is finite almost everywhere.

Finally, this lets us integrate general functions.

Definition 37.1.5. In general, a measurable function $f : \Omega \to [-\infty, \infty]$ is **absolutely integrable** or just **integrable** if $|f|$ is.

Since we'll be using the first word, this is easy to remember: "absolutely integrable" requires taking absolute values.

Step 4 (Absolutely integrable functions) — If $f\colon \Omega \to [-\infty, \infty]$ is absolutely integrable, then we define

$$f^+(x) = \max\{f(x), 0\}$$
$$f^-(x) = \min\{f(x), 0\}$$

and set

$$\int_\Omega f \, d\mu = \int_\Omega |f^+| \, d\mu - \int_\Omega |f^-| \, d\mu$$

which in particular is finite.

That said, calling it "finitely integrable" here would also make it as easy to remember:

Exercise 37.1.6. Show that $\int_\Omega |f| d\mu < \infty$ if and only if $\int_\Omega |f^+| d\mu < \infty$ and $\int_\Omega |f^-| d\mu < \infty$.

You may already start to see that we really like nonnegative functions: with the theory of measures, it is possible to integrate them, and it's even okay to throw in $+\infty$'s everywhere. But once we start dealing with functions that can be either positive or negative, we have to start adding finiteness restrictions — actually essentially what we're doing is splitting the function into its positive and negative part, requiring both are finite, and then integrating.

To finish this section, we state for completeness some results that you probably could have guessed were true. Fix $\Omega = (\Omega, \mathscr{A}, \mu)$, and let f and g be measurable real-valued functions such that $f(x) = g(x)$ almost everywhere.

- (Almost-everywhere preservation) The function f is absolutely integrable if and only if g is, and if so, their Lebesgue integrals match.
- (Additivity) If f and g are absolutely integrable then

  $$\int_\Omega f + g \, d\mu = \int_\Omega f \, d\mu + \int_\Omega g \, d\mu.$$

  The "absolutely integrable" hypothesis can be dropped if f and g are nonnegative.
- (Scaling) If f is absolutely integrable and $c \in \mathbb{R}$ then cf is absolutely integrable and

  $$\int_\Omega cf \, d\mu = c \int_\Omega f \, d\mu.$$

  The "absolutely integrable" hypothesis can be dropped if f is nonnegative and $c > 0$.
- (Monotoncity) If f and g are absolutely integrable and $f \le g$, then

  $$\int_\Omega f \, d\mu \le \int_\Omega g \, d\mu.$$

  The "absolutely integrable" hypothesis can be dropped if f and g are nonnegative.

There are more famous results like monotone/dominated convergence that are also true, but we won't state them here as we won't really have a use for them in the context of probability. (They appear later on in a bonus chapter.)

§37.2 An equivalent definition

The Lebesgue integral can also be defined as follows — which should be more intuitive on the various choices of the definitions we made in the steps.

In this definition,

The integral $\int_\Omega f\ d\mu$ is just the volume of the region under the graph of f.

Let us define it:

Step 1 (The region under the graph) — For a nonnegative function $f\colon \Omega \to \mathbb{R}$, define the **region under the function** f, $R(f)$, to be $\{(x,y) \in \Omega \times \mathbb{R}, 0 \le y \le f(x)\}$.

Remark 37.2.1 — It should be clear why we only define this for nonnegative function initially — for general function f, the only way we could sensibly define the region would be something like the following:

$$R^+(f) = \{(x,y) \in \Omega \times \mathbb{R}, f(x) \ge 0, 0 \le y \le f(x)\},$$
$$R^-(f) = \{(x,y) \in \Omega \times \mathbb{R}, f(x) \le 0, 0 \ge y \ge f(x)\}.$$

Nevertheless, notice that $R^+(f)$ is simply the region under the function $f^+(x) = \max\{f(x), 0\}$, and $R^-(f)$ has the same measure as the region under the function $f^-(x) = \min\{f(x), 0\}$, so defining $\int_\Omega f\ d\mu$ for nonnegative functions first would actually simplify the definition.

Step 2 (Making $\Omega \times \mathbb{R}$ into a measure space) — We define a pre-measure on $\Omega \times \mathbb{R}$ the obvious way: if $X \subseteq \Omega$ and $Y \subseteq \mathbb{R}$ are measurable subsets respectively, then assign $|X \times Y| = |X| \times |Y|$.
The pre-measure can be extended to a measure, as we have done in the previous chapter.

Step 3 (Nonnegative functions) — For each function $f\colon \Omega \to [0, +\infty]$, let

$$\int_\Omega f\ d\mu \coloneqq |R(f)|.$$

The integral is well-defined whenever $R(f)$ is measurable.

As promised in Section 35.7, the definition of measurable function satisfies:

A nonnegative function f is measurable if and only if we can "measure" the region below the graph of f.

The last step is exactly the same as in the previous section.

Step 4 (Absolutely integrable functions) — If $f\colon \Omega \to [-\infty,\infty]$ is absolutely integrable, then we define

$$\int_\Omega f\,d\mu = \int_\Omega |f^+|\,d\mu - \int_\Omega |f^-|\,d\mu.$$

§37.3 Relation to Riemann integrals (or: actually computing Lebesgue integrals)

For closed intervals, this actually just works out of the box.

Theorem 37.3.1 (Lebesgue integral generalizes Riemann integral)
Let $f\colon [a,b] \to \mathbb{R}$ be a Riemann integrable function (where $[a,b]$ is equipped with the Borel measure). Then f is also Lebesgue integrable and the integrals agree:

$$\int_a^b f(x)\,dx = \int_{[a,b]} f\,d\mu.$$

Note that a Riemann integrable function *must be bounded*, which means if you try to construct a function $f\colon [0,1] \to \mathbb{R}$ in the same vein as Problem 37B† by

$$f(x) = \begin{cases} \frac{\sin(1/x)}{x} & x > 0 \\ 0 & x = 0 \end{cases}$$

the function f will in fact *not* be Riemann integrable! Although of course, the improper Riemann integral $\lim_{\varepsilon\to 0^+} \int_\varepsilon^1 f(x)\,dx$ exists.

Thus in practice, we do all theory with Lebesgue integrals (they're nicer), but when we actually need to compute $\int_{[1,4]} x^2\,d\mu$ we just revert back to our usual antics with the Fundamental Theorem of Calculus.

Example 37.3.2 (Integrating x^2 over $[1,4]$)
Reprising our old example:

$$\int_{[1,4]} x^2\,d\mu = \int_1^4 x^2\,dx = \frac{1}{3}\cdot 4^3 - \frac{1}{3}\cdot 1^3 = 21.$$

This even works for *improper* integrals, if the functions are nonnegative. The statement is a bit cumbersome to write down, but here it is.

Theorem 37.3.3 (Improper integrals are nice Lebesgue ones)
Let $f \ge 0$ be a *nonnegative* continuous function defined on $(a,b) \subseteq \mathbb{R}$, possibly allowing $a = -\infty$ or $b = \infty$. Then

$$\int_{(a,b)} f\,d\mu = \lim_{\substack{a'\to a^+ \\ b'\to b^-}} \int_{a'}^{b'} f(x)\,dx$$

where we allow both sides to be $+\infty$ if f is not absolutely integrable.

The right-hand side makes sense since $[a', b'] \subsetneq (a, b)$ is a compact interval on which f is continuous. This means that improper Riemann integrals of nonnegative functions can just be regarded as Lebesgue ones over the corresponding open intervals.

It's probably better to just look at an example though.

Example 37.3.4 (Integrating $1/\sqrt{x}$ on $(0,1)$)

For example, you might be familiar with improper integrals like

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx \coloneqq \lim_{\varepsilon \to 0^+} \int_\varepsilon^1 \frac{1}{\sqrt{x}}\,dx = \lim_{\varepsilon \to 0^+} \left(2\sqrt{1} - 2\sqrt{\varepsilon}\right) = 2.$$

(Note this appeared before as Problem 30C*.) In the Riemann integration situation, we needed the limit as $\varepsilon \to 0^+$ since otherwise $\frac{1}{\sqrt{x}}$ is not defined as a function $[0,1] \to \mathbb{R}$. However, it is a *measurable nonnegative* function $(0,1) \to [0, +\infty]$, and hence

$$\int_{(0,1)} \frac{1}{\sqrt{x}}\,d\mu = 2.$$

If f is not nonnegative, then all bets are off. Indeed Problem 37B† is the famous counterexample.

§37.4 A few harder problems to think about

Problem 37A* (The indicator of the rationals)**.** Take the indicator function $\mathbf{1}_{\mathbb{Q}} \colon \mathbb{R} \to \{0,1\} \subseteq \mathbb{R}$ for the rational numbers.

(a) Prove that $\mathbf{1}_{\mathbb{Q}}$ is not Riemann integrable.

(b) Show that $\int_{\mathbb{R}} \mathbf{1}_{\mathbb{Q}}$ exists and determine its value — the one you expect!

Problem 37B† (An improper Riemann integral with sign changes)**.** Define $f \colon (1, \infty) \to \mathbb{R}$ by $f(x) = \frac{\sin(x)}{x}$. Show that f is not absolutely integrable, but that the improper Riemann integral

$$\int_1^\infty f(x)\,dx \coloneqq \lim_{b \to \infty} \int_1^b f(x)\,dx$$

nonetheless exists.

38 Swapping order with Lebesgue integrals

§38.1 Motivating limit interchange

Prototypical example for this section: $\mathbf{1}_{\mathbb{Q}}$ *is good!*

One of the issues with the Riemann integral is that it behaves badly with respect to convergence of functions, and the Lebesgue integral deals with this. This is therefore often given as a poster child on why the Lebesgue integral has better behaviors than the Riemann one.

We technically have already seen this: consider the indicator function $\mathbf{1}_{\mathbb{Q}}$, which is not Riemann integrable by Problem 37A*. But we can readily compute its Lebesgue integral over $[0,1]$, as

$$\int_{[0,1]} \mathbf{1}_{\mathbb{Q}} \, d\mu = \mu\left([0,1] \cap \mathbb{Q}\right) = 0$$

since it is countable.

This *could* be thought of as a failure of existence for the Riemann integral.

Example 38.1.1 ($\mathbf{1}_{\mathbb{Q}}$ is a limit of finitely supported functions)
We can define the sequence of functions g_1, g_2, ... by

$$g_n(x) = \begin{cases} 1 & (n!)x \text{ is an integer} \\ 0 & \text{else.} \end{cases}$$

Then each g_n is piecewise continuous and hence Riemann integrable on $[0,1]$ (with integral zero), but $\lim_{n\to\infty} g_n = \mathbf{1}_{\mathbb{Q}}$ is not.

The limit here is defined in the following sense:

Definition 38.1.2. Let f and $f_1, f_2, \ldots : \Omega \to \mathbb{R}$ be a sequence of functions. Suppose that for each $\omega \in \Omega$, the sequence

$$f_1(\omega),\ f_2(\omega),\ f_3(\omega),\ , \ldots$$

converges to $f(\omega)$. Then we say f_1, f_2, ... **converges pointwise** to the limit f, written $\lim_{n\to\infty} f_n = f$.

We can define $\liminf_{n\to\infty} f_n$ and $\limsup_{n\to\infty} f_n$ similarly.

By "the Lebesgue integral has better behavior", we means the following:

Proposition 38.1.3
If $f_1, f_2, \ldots : \Omega \to \mathbb{R}$ are measurable functions, then $\liminf_{n\to\infty} f_n$ and $\limsup_{n\to\infty} f_n$ are measurable.

When f_n are all nonnegative, this means $\int_\Omega \liminf_{n\to\infty} f_n d\mu$ and $\int_\Omega \limsup_{n\to\infty} f_n d\mu$ exists. (If they can be negative, the behavior is not that nice. Problem 37B† gives an example.)

Unfortunately, even if the integral exists, we can't always exchange pointwise limit with Lebesgue integral.

Why would we want to? For instance, if we face this problem:

> Compute $\lim_{k\to\infty} \int_1^\infty \frac{1}{k} e^{-x^2}\ dx$.

While the integral $\int e^{-x^2} dx$ is not computable by elementary means, we would like to say the limit is simply 0 (why wouldn't it be?)

Unfortunately, pointwise convergence is actually a fairly weak notion of convergence.

Example 38.1.4

In all of these examples, we cannot interchange the limit and the integral without changing the result.

- The sequence $f_k(x) = \frac{\sin(x)}{x} \cdot \mathbf{1}_{(1,k)}$ converges pointwise to $f(x) = \frac{\sin(x)}{x} \cdot \mathbf{1}_{(1,\infty)}$ as $k \to \infty$, and the limit $\lim_{k\to\infty} \int f_k(x)\ dx$ exists, but f is not integrable.
- Similarly, $f_k(x) = \frac{\sin(1/x)}{x} \cdot \mathbf{1}_{(1/k,\infty)}$ converges pointwise to $f(x) = \frac{\sin(1/x)}{x} \cdot \mathbf{1}_{(0,\infty)}$ as $k \to \infty$, the limit $\lim_{k\to\infty} \int f_k(x)\ dx$ exists and is finite, but f is not integrable.
- The sequence $f_k(x) = \frac{\mathbf{1}_{(0,k)}}{k}$ converges pointwise to $f(x) = 0$ as $k \to \infty$, for every k then $\int f_k(x)\ dx = 1$, but $\int f(x)\ dx = 0$.

  Note that, in this case, the convergence is actually uniform!
- We don't even need k in the denominator — the sequence $f_k(x) = \mathbf{1}_{(0,k)}$ also converges pointwise to $f(x) = 0$, but this time, for every k then $\int f_k(x)\ dx = \infty$!
- The sequence $f_k(x) = k \cdot \mathbf{1}_{(0,1/k)}$ converges pointwise to $f(x) = 0$ as $k \to \infty$. But similar to above, $\int f_k(x)\ dx = 1$ for every k, but $\int f(x)\ dx = 0$.

The last example is similar in behavior to an example known as the Witch's hat.[a]

[a] https://www.geogebra.org/m/dv7ctmed has an animation.

As such, the convergence theorems stated below is an attempt to classify all the possible anomalies, and to show that in "usual" cases, interchanging limit and integral just works.

As mentioned earlier, we choose to use the Lebesgue integral instead of the Riemann integral, because in such cases, the Lebesgue integral will usually just exist.

§38.2 Overview

The three big-name results for exchanging pointwise limits with Lebesgue integrals is:

- Fatou's lemma: the most general statement possible, for any nonnegative measurable functions.
- Monotone convergence: "increasing limits" just work.

- Dominated convergence (actually Fatou-Lebesgue): limits that are not too big (bounded by some absolutely integrable function) just work.

§38.3 Fatou's lemma

In all the above examples, we see that:

- The failure of the interchange of limit and integral is caused by the functions in the sequence have too much room to "wiggle around", and
- as such, the integrals $\int f_k(x)dx$ are all greater than the integral of the limit $\int f(x)dx$.

Of course, by negating all the functions $f_k(x)$ we can get $\lim_{k\to\infty} \int f_k(x)dx < \int f(x)dx$. But, as it turns out, for nonnegative functions, *this sort of behavior is the only behavior possible*. In other words,

For nonnegative functions, if limit of integral is not equal to integral of limit, the former one is always larger.

Lemma 38.3.1 (Fatou's lemma, preliminary version)

Let $f_1, f_2, \ldots : \Omega \to [0, +\infty]$ be a sequence of *nonnegative* measurable functions, converging pointwise to f. Then f is nonnegative, measurable, and

$$\int_\Omega f \, d\mu \le \lim_{n\to\infty} \left(\int_\Omega f_n \, d\mu \right).$$

Here we allow either side to be $+\infty$.

As it turns out, this lemma can significantly be generalized as follows. If you compare the two statements, you can see the two lim operators are changed to lim inf — when the sequence actually converges, lim inf and lim equals.

Lemma 38.3.2 (Fatou's lemma)

Let $f_1, f_2, \ldots : \Omega \to [0, +\infty]$ be a sequence of *nonnegative* measurable functions. Then $\liminf_{n\to\infty} f_n : \Omega \to [0, +\infty]$ is measurable and

$$\int_\Omega \left(\liminf_{n\to\infty} f_n \right) \, d\mu \le \liminf_{n\to\infty} \left(\int_\Omega f_n \, d\mu \right).$$

Here we allow either side to be $+\infty$.

Notice that there are *no extra hypothesis* on f_n other than nonnegative: which makes this quite surprisingly versatile if you ever are trying to prove some general result.

§38.4 Everything else

The big surprise is how quickly all the "big-name" theorem follows from Fatou's lemma. Here is the so-called "monotone convergence theorem".

Corollary 38.4.1 (Monotone convergence theorem)
Let f and $f_1, f_2, \ldots : \Omega \to [0, +\infty]$ be a sequence of *nonnegative* measurable functions such that $\lim_n f_n = f$ and $f_n(\omega) \le f(\omega)$ for each n. Then f is measurable and

$$\lim_{n\to\infty} \left(\int_\Omega f_n \, d\mu \right) = \int_\Omega f \, d\mu.$$

Here we allow either side to be $+\infty$.

Proof. We have

$$\begin{aligned} \int_\Omega f \, d\mu &= \int_\Omega \left(\liminf_{n\to\infty} f_n \right) \, d\mu \\ &\le \liminf_{n\to\infty} \int_\Omega f_n \, d\mu \\ &\le \limsup_{n\to\infty} \int_\Omega f_n \, d\mu \\ &\le \int_\Omega f \, d\mu \end{aligned}$$

where the first $\le$ is by Fatou lemma, and the third by the fact that $\int_\Omega f_n \le \int_\Omega f$ for every n. This implies all the inequalities are equalities and we are done. □

You can see how short the proof is — proving $\limsup_{n\to\infty} \int_\Omega f_n \, d\mu \le \int_\Omega f \, d\mu$ is the easy half, and the difficult half is automatically taken care of by Fatou's lemma.

Remark 38.4.2 (The monotone convergence theorem does not require monotonicity!) — In the literature it is much more common to see the hypothesis $f_1(\omega) \le f_2(\omega) \le \cdots \le f(\omega)$ rather than just $f_n(\omega) \le f(\omega)$ for all n, which is where the theorem gets its name. However as we have shown this hypothesis is superfluous! This is pointed out in https://mathoverflow.net/a/296540/70654, as a response to a question entitled "Do you know of any very important theorems that remain unknown?".

Example 38.4.3 (Monotone convergence gives $\mathbf{1}_\mathbb{Q}$)
This already implies Example 38.1.1. Letting g_n be the indicator function for $\frac{1}{n!}\mathbb{Z}$ as described in that example, we have $g_n \le \mathbf{1}_\mathbb{Q}$ and $\lim_{n\to\infty} g_n(x) = \mathbf{1}_\mathbb{Q}(x)$, for each individual x. So since $\int_{[0,1]} g_n \, d\mu = 0$ for each n, this gives $\int_{[0,1]} \mathbf{1}_\mathbb{Q} = 0$ as we already knew.

The most famous result, though is the following.

Corollary 38.4.4 (Fatou-Lebesgue theorem)
Let f and $f_1, f_2, \ldots : \Omega \to \mathbb{R}$ be a sequence of measurable functions. Assume that $g \colon \Omega \to \mathbb{R}$ is an *absolutely integrable* function for which $|f_n(\omega)| \le |g(\omega)|$ for all $\omega \in \Omega$. Then the inequality

$$\begin{aligned}\int_\Omega \left(\liminf_{n\to\infty} f_n\right)\, d\mu &\le \liminf_{n\to\infty} \left(\int_\Omega f_n\, d\mu\right) \\ &\le \limsup_{n\to\infty} \left(\int_\Omega f_n\, d\mu\right) \le \int_\Omega \left(\limsup_{n\to\infty} f_n\right)\, d\mu.\end{aligned}$$

Proof. There are three inequalities:

- The first inequality follows by Fatou on $g + f_n$ which is nonnegative.
- The second inequality is just $\liminf \le \limsup$. (This makes the theorem statement easy to remember!)
- The third inequality follows by Fatou on $g - f_n$ which is nonnegative. □

Exercise 38.4.5. Where is the fact that g is absolutely integrable used in this proof?

Corollary 38.4.6 (Dominated convergence theorem)
Let $f_1, f_2, \ldots : \Omega \to \mathbb{R}$ be a sequence of measurable functions such that $f = \lim_{n\to\infty} f_n$ exists. Assume that $g \colon \Omega \to \mathbb{R}$ is an *absolutely integrable* function for which $|f_n(\omega)| \le |g(\omega)|$ for all $\omega \in \Omega$. Then

$$\int_\Omega f\, d\mu = \lim_{n\to\infty} \left(\int_\Omega f_n\, d\mu\right).$$

In other words,

If there's only finite "space" for the functions f_k to "wiggle around", then no anomaly can happen.

Proof. If $f(\omega) = \lim_{n\to\infty} f_n(\omega)$, then $f(\omega) = \liminf_{n\to\infty} f_n(\omega) = \limsup_{n\to\infty} f_n(\omega)$. So all the inequalities in the Fatou-Lebesgue theorem become equalities, since the leftmost and rightmost sides are equal. □

Note this gives yet another way to verify Example 38.1.1. In general, the dominated convergence theorem is a favorite cliché for undergraduate exams, because it is easy to create questions for it. Here is one example showing how they all look.

Example 38.4.7 (The usual Lebesgue dominated convergence examples)
Suppose one wishes to compute

$$\lim_{n\to\infty} \left(\int_{(0,1)} \frac{n \sin(n^{-1}x)}{\sqrt{x}}\right)\, dx$$

then one starts by observing that the inner term is bounded by the absolutely

integrable function $x^{-1/2}$. Therefore it equals

$$\begin{aligned}\int_{(0,1)} \lim_{n\to\infty} \left(\frac{n\sin(n^{-1}x)}{\sqrt{x}} \right) dx &= \int_{(0,1)} \frac{x}{\sqrt{x}}\, dx \\ &= \int_{(0,1)} \sqrt{x}\, dx = \frac{2}{3}.\end{aligned}$$

We can also say something else about the behavior of the anomalies — that is, when $|\Omega| < \infty$, the anomaly only happens in a set of *small measure*.

Theorem 38.4.8 (Egorov's theorem)

Let $f_1, f_2, \ldots : \Omega \to \mathbb{R}$ be a sequence of measurable functions, on a measure space Ω with $|\Omega| < \infty$, such that $f = \lim_{n\to\infty} f_n$ exists and is finite almost everywhere. Then, for any $\varepsilon > 0$, we can find a subset $U \subseteq \Omega$, such that the remainder has small measure:

$$|\Omega \smallsetminus U| < \varepsilon,$$

and the convergence is uniform on U: the sequence

$$f_1|_U, f_2|_U, \ldots$$

converges to f_U uniformly.

This is because of the following theorem.

Theorem 38.4.9 (Uniform convergence theorem)

Let $f_1, f_2, \ldots : \Omega \to \mathbb{R}$ be a sequence of integrable functions, on a measure space Ω with $|\Omega| < \infty$, such that $\lim_{n\to\infty} f_n = f$, and the convergence is uniform. Then f is integrable and,

$$\lim_{n\to\infty} \left(\int_\Omega f_n \, d\mu \right) = \int_\Omega f \, d\mu.$$

In other words,

The fact that $\int f\, d\mu \neq \lim \int f_k\, d\mu$ is only caused by $\int_{\Omega\setminus U} f\, d\mu \neq \lim \int_{\Omega\setminus U} f\, d\mu$.

Example 38.4.10 (Removing a set of small measure will allow interchanging the integral and the limit)

We take a few examples from Example 38.1.4, and see what happens if we remove a set of small measure here.

- Consider the sequence $f_k(x) = k \cdot \mathbf{1}_{(0,1/k)}$. If, for any $\varepsilon > 0$, we delete a segment $(0, \varepsilon)$ from the domain of f_k, then we will have that f_k converges uniformly to f as $k \to \infty$, and that $\lim_{k\to\infty} \int f_k(x)\, dx = \int f(x)\, dx = 0$.
- Similarly, the sequence $f_k(x) = \frac{\sin(1/x)}{x} \cdot \mathbf{1}_{(1/k,1)}$ converges pointwise to $f(x) =$

$\frac{\sin(1/x)}{x} \cdot \mathbf{1}_{(0,1)}$, and if we delete a segment $(0, \varepsilon)$, then everything checks out.

Remark 38.4.11 — Just because we only need to delete a set of small measure, doesn't mean the set is concentrated in a small interval. The reader is invited to construct a sequence $f_k \colon [0,1] \to \mathbb{R}^+$ that converges pointwise to f, but in order to make the convergence uniform, a dense subset of $[0,1]$ need to be removed. (Hint: take any discontinuous everywhere nonnegative function f, and set $f_k = \min(k, f)$.)

§38.5 Fubini and Tonelli

TO BE WRITTEN

§38.6 A few harder problems to think about

problems

39 Bonus: A hint of Pontryagin duality

In this short chapter we will give statements about how to generalize our Fourier analysis (a bonus chapter Chapter 14) to a much wider class of groups G.

§39.1 LCA groups

Prototypical example for this section: $\mathbb{T}$, $\mathbb{R}$.

Earlier we played with $\mathbb{R}$, which is nice because in addition to being a topological space, it is also an abelian group under addition. These sorts of objects which are both groups and spaces have a name.

Definition 39.1.1. A group G is a **topological group** is a Hausdorff[1] topological space equipped also with a group operation $(G, \cdot)$, such that both maps

$$\begin{aligned} G \times G \to G \quad &\text{by} \quad (x,y) \mapsto xy \\ G \to G \quad &\text{by} \quad x \mapsto x^{-1} \end{aligned}$$

are continuous.

For our Fourier analysis, we need some additional conditions.

Definition 39.1.2. A **locally compact abelian (LCA) group** G is one for which the group operation is abelian, and moreover the topology is *locally compact*: for every point p of G, there exists a compact subset K of G such that $K \ni p$, and K contains some open neighborhood of p.

Our previous examples all fall into this category:

Example 39.1.3 (Examples of locally compact abelian groups)

- Any finite group Z with the discrete topology is LCA.
- The circle group $\mathbb{T}$ is LCA and also in fact compact.
- The real numbers $\mathbb{R}$ are an example of an LCA group which is *not* compact.

These conditions turn out to be enough for us to define a measure on the space G. The relevant theorem, which we will just quote:

[1]Some authors omit the Hausdorff condition.

Theorem 39.1.4 (Haar measure)

Let G be a locally compact abelian group. We regard it as a measurable space using its Borel σ-algebra $\mathscr{B}(G)$. There exists a measure $\mu\colon \mathscr{B}(G) \to [0,\infty]$, called the **Haar measure**, satisfying the following properties:

- $\mu(gS) = \mu(S)$ for every $g \in G$ and measurable S. That means that μ is "translation-invariant" under translation by G.
- $\mu(K)$ is finite for any compact set K.
- if S is measurable, then $\mu(S) = \inf\{\mu(U) \mid U \supseteq S \text{ open}\}$.
- if U is open, then $\mu(U) = \sup\{\mu(K) \mid K \subseteq U \text{ compact}\}$.

Moreover, it is unique up to scaling by a positive constant.

Remark 39.1.5 — Note that if G is compact, then $\mu(G)$ is finite (and positive). For this reason the Haar measure on a LCA group G is usually normalized so $\mu(G) = 1$.

For this chapter, we will only use the first two properties at all, and the other two are just mentioned for completeness. Note that this actually generalizes the chapter where we constructed a measure on $\mathscr{B}(\mathbb{R}^n)$, since $\mathbb{R}^n$ is an LCA group!

So, in short: if we have an LCA group, we have a measure μ on it.

§39.2 The Pontryagin dual

Now the key definition is:

Definition 39.2.1. Let G be an LCA group. Then its **Pontryagin dual** is the abelian group

$$\widehat{G} \coloneqq \{\text{continuous group homomorphisms } \xi\colon G \to \mathbb{T}\}.$$

The maps ξ are called **characters**. It can be itself made into an LCA group.[2]

Example 39.2.2 (Examples of Pontryagin duals)

- $\widehat{\mathbb{Z}} \cong \mathbb{T}$, since group homomorphisms $\mathbb{Z} \to \mathbb{T}$ are determined by the image of 1.
- $\widehat{\mathbb{T}} \cong \mathbb{Z}$. The characters are given by $\theta \mapsto n\theta$ for $n \in \mathbb{Z}$.
- $\widehat{\mathbb{R}} \cong \mathbb{R}$. This is because a nonzero continuous homomorphism $\mathbb{R} \to S^1$ is determined by the fiber above $1 \in S^1$. (Algebraic topologists might see covering projections here.)
- $\widehat{\mathbb{Z}/n\mathbb{Z}} \cong \mathbb{Z}/n\mathbb{Z}$, characters ξ being determined by the image $\xi(1) \in \mathbb{T}$.
- $\widehat{G \times H} \cong \widehat{G} \times \widehat{H}$.

[2] If you must know the topology, it is the **compact-open topology**: for any compact set $K \subseteq G$ and open set $U \subseteq \mathbb{T}$, we declare the set of all ξ with $\xi^{\text{img}}(K) \subseteq U$ to be open, and then take the smallest topology containing all such sets. We won't use this at all.

Exercise 39.2.3 ($\widehat{Z} \cong Z$, for those who read Section 18.1)**.** If Z is a finite abelian group, show that $\widehat{\widehat{Z}} \cong Z$, using the results of the previous example. You may now recognize that the bilinear form $\cdot\colon Z \times Z \to \mathbb{T}$ is exactly a choice of isomorphism $Z \to \widehat{Z}$. It is not "canonical".

True to its name as the dual, and in analogy with $(V^\vee)^\vee \cong V$ for vector spaces V, we have:

Theorem 39.2.4 (Pontryagin duality theorem)
For any LCA group G, there is an isomorphism

$$G \cong \widehat{\widehat{G}} \qquad \text{by} \qquad x \mapsto (\xi \mapsto \xi(x))\,.$$

The compact case is especially nice.

Proposition 39.2.5 (G compact $\iff$ $\widehat{G}$ discrete)
Let G be an LCA group. Then G is compact if and only if $\widehat{G}$ is discrete.

Proof. Problem 39B. □

§39.3 The orthonormal basis in the compact case

Let G be a compact LCA group, and work with its Haar measure. We may now let $L^2(G)$ be the space of square-integrable functions to $\mathbb{C}$, i.e.

$$L^2(G) = \left\{ f\colon G \to \mathbb{C} \quad \text{such that} \quad \int_G |f|^2 < \infty \right\}.$$

Thus we can equip it with the inner form

$$\langle f, g \rangle = \int_G f \cdot \overline{g}.$$

In that case, we get all the results we wanted before:

Theorem 39.3.1 (Characters of $\widehat{G}$ form an orthonormal basis)
Assume G is LCA and compact (so $\widehat{G}$ is discrete). Then the characters

$$(e_\xi)_{\xi \in \widehat{G}} \qquad \text{by} \qquad e_\xi(x) = e(\xi(x)) = \exp(2\pi i \xi(x))$$

form an orthonormal basis of $L^2(G)$. Thus for each $f \in L^2(G)$ we have

$$f = \sum_{\xi \in \widehat{G}} \widehat{f}(\xi) e_\xi$$

where

$$\widehat{f}(\xi) = \langle f, e_\xi \rangle = \int_G f(x) \exp(-2\pi i \xi(x))\, d\mu.$$

The sum $\sum_{\xi \in \widehat{G}}$ makes sense since $\widehat{G}$ is discrete. In particular,

- Letting $G = Z$ for a finite group G gives "Fourier transform on finite groups".
- The special case $G = \mathbb{Z}/n\mathbb{Z}$ has its own Wikipedia page: the "discrete-time Fourier transform".
- Letting $G = \mathbb{T}$ gives the "Fourier series" earlier.

§39.4 The Fourier transform of the non-compact case

If G is LCA but not compact, then Theorem 39.3.1 becomes false. On the other hand, it's still possible to define $\widehat{G}$. We can then try to write the Fourier coefficients anyways: let

$$\widehat{f}(\xi) = \int_G f \cdot \overline{e_\xi} \, d\mu$$

for $\xi \in \widehat{G}$ and $f \colon G \to \mathbb{C}$. The results are less fun in this case, but we still have, for example:

Theorem 39.4.1 (Fourier inversion formula in the non-compact case)

Let μ be a Haar measure on G. Then there exists a unique Haar measure ν on $\widehat{G}$ (called the **dual measure**) such that: whenever $f \in L^1(G)$ and $\widehat{f} \in L^1(\widehat{G})$, we have

$$f(x) = \int_{\widehat{G}} \widehat{f}(\xi)\xi(x) \, d\nu$$

for almost all $x \in G$ (with respect to μ). If f is continuous, this holds for all x.

So while we don't have the niceness of a full inner product from before, we can still in some situations at least write f as integral in sort of the same way as before.

In particular, they have special names for a few special G:

- If $G = \mathbb{R}$, then $\widehat{G} = \mathbb{R}$, yielding the "(continuous) Fourier transform".
- If $G = \mathbb{Z}$, then $\widehat{G} = \mathbb{T}$, yielding the "discrete time Fourier transform".

§39.5 Summary

We summarize our various flavors of Fourier analysis from the previous sections in the following table. In the first part G is compact, in the second half G is not.

Name	Domain G	Dual $\widehat{G}$	Characters
Binary Fourier analysis	$\{\pm 1\}^n$	$S \subseteq \{1, \dots, n\}$	$\prod_{s \in S} x_s$
Fourier transform on finite groups	Z	$\xi \in \widehat{Z} \cong Z$	$e(i\xi \cdot x)$
Discrete Fourier transform	$\mathbb{Z}/n\mathbb{Z}$	$\xi \in \mathbb{Z}/n\mathbb{Z}$	$e(\xi x/n)$
Fourier series	$\mathbb{T} \cong [-\pi, \pi]$	$n \in \mathbb{Z}$	$\exp(inx)$
Continuous Fourier transform	$\mathbb{R}$	$\xi \in \mathbb{R}$	$e(\xi x)$
Discrete time Fourier transform	$\mathbb{Z}$	$\xi \in \mathbb{T} \cong [-\pi, \pi]$	$\exp(i\xi n)$

You might notice that the **various names are awful**. This is part of the reason I got confused as a high school student: every type of Fourier series above has its own Wikipedia article. If it were up to me, we would just use the term "G-Fourier transform", and that would make everyone's lives a lot easier.

§39.6 A few harder problems to think about

Problem 39A. If G is compact, so $\widehat{G}$ is discrete, describe the dual measure ν.

Problem 39B. Show that an LCA group G is compact if and only if $\widehat{G}$ is discrete. (You will need the compact-open topology for this.)

XI

Probability (TO DO)

Part XI: Contents

40 Random variables (TO DO)

write chapter

Having properly developed the Lebesgue measure and the integral on it, we can now proceed to develop random variables.

§40.1 Random variables

With all this set-up, random variables are going to be really quick to define.

Definition 40.1.1. A (real) **random variable** X on a probability space $\Omega = (\Omega, \mathscr{A}, \mu)$ is a measurable function $X \colon \Omega \to \mathbb{R}$, where $\mathbb{R}$ is equipped with the Borel σ-algebra.

In particular, addition of random variables, etc. all makes sense, as we can just add. Also, we can integrate X over Ω, by previous chapter.

Definition 40.1.2 (First properties of random variables)**.** Given a random variable X, the **expected value** of X is defined by the Lebesgue integral

$$\mathbb{E}[X] = \int_\Omega X(\omega)\, d\mu.$$

Confusingly, the letter μ is often used for expected values.

The **kth moment** of X is defined as $\mathbb{E}[X^k]$, for each positive integer $k \geq 1$. The **variance** of X is then defined as

$$\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Question 40.1.3. Show that $\mathbf{1}_A$ is a random variable (just check that it is Borel measurable), and its expected value is $\mu(A)$.

An important property of expected value you probably already know:

Theorem 40.1.4 (Linearity of expectation)
If X and Y are random variables on Ω then

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Proof. $\mathbb{E}[X+Y] = \int_\Omega X(\omega) + Y(\omega)\, d\mu = \int_\Omega X(\omega)\, d\mu + \int_\Omega Y(\omega)\, d\mu = \mathbb{E}[X] + \mathbb{E}[Y]$. □

Note that X and Y do not have to be "independent" here: a notion we will define shortly.

§40.2 Distribution functions

§40.3 Examples of random variables

§40.4 Characteristic functions

§40.5 Independent random variables

§40.6 A few harder problems to think about

Problem 40A (Equidistribution)**.** Let X_1, X_2, ... be i.i.d. uniform random variables on $[0,1]$. Show that almost surely the X_i are equidistributed, meaning that

$$\lim_{N\to\infty} \frac{\#\{1 \le i \le N \mid a \le X_i(\omega) \le b\}}{N} = b-a \qquad \forall 0 \le a < b \le 1$$

holds for almost all choices of ω.

Problem 40B (Side length of triangle independent from median)**.** Let X_1, Y_1, X_2, Y_2, X_3, Y_3 be six independent standard Gaussians. Define triangle ABC in the Cartesian plane by $A = (X_1, Y_1)$, $B = (X_2, Y_2)$, $C = (X_3, Y_3)$. Prove that the length of side BC is independent from the length of the A-median.

41 Large number laws (TO DO)

write chapter

§41.1 Notions of convergence

§41.1.i Almost sure convergence

Definition 41.1.1. Let X, X_n be random variables on a probability space Ω. We say X_n **converges almost surely** to X if

$$\mu\left(\omega \in \Omega : \lim_n X_n(\omega) = X(\omega)\right) = 1.$$

This is a very strong notion of convergence: it says in almost every *world*, the values of X_n converge to X. In fact, it is almost better for me to give a *non-example*.

Example 41.1.2 (Non-example of almost sure convergence)
Imagine an immortal skeleton archer is practicing shots, and on the nth shot, he scores a bulls-eye with probability $1 - \frac{1}{n}$ (which tends to 1 because the archer improves over time). Let $X_n \in \{0, 1, \ldots, 10\}$ be the score of the nth shot.
Although the skeleton is gradually approaching perfection, there are *almost no worlds* in which the archer misses only finitely many shots: that is

$$\mu\left(\omega \in \Omega : \lim_n X_n(\omega) = 10\right) = 0.$$

§41.1.ii Convergence in probability

Therefore, for many purposes we need a weaker notion of convergence.

Definition 41.1.3. Let X, X_n be random variables on a probability space Ω. We say X_n **converges in probability** to X if if for every $\varepsilon > 0$ and $\delta > 0$, we have

$$\mu\left(\omega \in \Omega : |X_n(\omega) - X(\omega)| < \varepsilon\right) \geq 1 - \delta$$

for n large enough (in terms of ε and δ).

In this sense, our skeleton archer does succeed: for any $\delta > 0$, if $n > \delta^{-1}$ then the skeleton archer does hit a bulls-eye in a $1 - \delta$ fraction of the worlds. In general, you can think of this as saying that for any $\delta > 0$, the chance of an ε-anomaly event at the nth stage eventually drops below δ.

Remark 41.1.4 — To mask δ from the definition, this is sometimes written instead as: for all ε

$$\lim_{n\to\infty} \mu\left(\omega \in \Omega : |X_n(\omega) - X(\omega)| < \varepsilon\right) = 1.$$

I suppose it doesn't make much difference, though I personally don't like the asymmetry.

§41.1.iii Convergence in law

§41.2 Weak law of large numbers

As the name implies, this is a direct corollary of the strong law of large numbers. Nevertheless, the proof of this law is simpler, and some applications only require the weak law.

write

§41.2.i Application: Weierstrass approximation

§41.3 Strong law of large numbers

§41.3.i Motivation: Biased random walk

Consider a random walk defined as follows:

- Let $X_0 = 1$.
- For each $i \geq 1$, define X_i to be $X_{i-1} - 1$ with probability $p = 0.6$ or $X_{i-1} + 1$ with probability $1 - p = 0.4$.

Then we can ask: What's the expected number of steps until some X_i equals 0? A naive attempt might be the following.

> Let $f(i)$ be the expected number of steps starting to reach 0 starting from $X_0 = i$.
>
> Then:
>
> - $f(0) = 0$,
> - $f(1) = 1 + 0.6f(0) + 0.4f(2)$,
> - $f(2) = 1 + 0.6f(1) + 0.4f(3)$,
> - $\vdots$

This isn't getting anywhere because there are infinitely many terms. A better attempt is the following:

> Let the answer be x. If we start from $X_0 = 2$, let i be the first time such that $X_i = 1$ and j be the first time after i such that $X_j = 0$. Then
>
> $$\mathbb{E}[i] = \mathbb{E}[j - i] = x.$$
>
> Therefore,
>
> $$x = 1 + 0.6 \cdot 0 + 0.4 \cdot (2x)$$
>
> Solving the equation, we get $x = 5$.

It gives the correct result — however, the same method gives $x = -5$ when the probability of going down is $p = 0.4$, which is clearly nonsense.

What went wrong? The problem is when $p = 0.4$, there is a nonzero probability[1] that the sequence never reaches 0, so the expected value is undefined and we're subtracting ∞ from ∞ in the proof.

In this case, the strong law of large numbers can help us patch this hole, by showing that in almost every world, the sequence X_i eventually reaches 0.

[1] Preview: Using martingale theory next chapter, you will be able to prove the probability the sequence never reaches 0 is exactly $1 - \frac{0.4}{0.6}$.

§41.3.ii Statement

Theorem 41.3.1 (Strong law of large numbers)

Let X_1, X_2, ... be i.i.d. random variables with mean 0. Define the partial mean

$$M_n = \frac{X_1 + \cdots + X_n}{n}.$$

Then, in almost every world, $M_n \to 0$.

In other words, M_n converges almost surely to 0.

The requirement that the mean is 0 is only to simplify the proof, as long as the mean exists, we can subtract the mean from the random variables and apply the theorem.

Example 41.3.2 (The hypothesis $\mathbb{E}[X_i] = 0$ is important)

Consider an example where $M_n \to 0$ does not hold — this is a minor variation of the St. Petersburg paradox.

Let the distribution of each X_i be as follows:

$$X_i = \begin{cases} 1 & \text{with probability } \frac{1}{4} \\ -1 & \text{with probability } \frac{1}{4} \\ 2 & \text{with probability } \frac{1}{8} \\ -2 & \text{with probability } \frac{1}{8} \\ 4 & \text{with probability } \frac{1}{16} \\ -4 & \text{with probability } \frac{1}{16} \\ & \vdots \end{cases}$$

Formally, X_i takes each of the value in $\{2^k, -2^k\}$ with probability 2^{-k-2}.

In this case, the mean $\mathbb{E}[X_i] = \int_\Omega X_i(\omega)$ is actually undefined. Furthermore, as symmetric as the distribution may look, in almost no world will M_n converge to 0. Intuitively, you can see why:

- Within the first 16 values, on average there's one X_i with $|X_i| = 4$, this will skew M_{16} by $\frac{1}{4}$.
- Within the first 32 values, on average there's one X_i with $|X_i| = 8$, this will skew M_{32} by $\frac{1}{4}$.
- Et cetera.

In other words, just like our skeleton archer, there are almost no worlds in which the M_n got skewed by more than $\frac{1}{4}$ only finitely many times.

§41.3.iii Proof for finite-variance case

In practice, most distribution we ever come across has finite variance, it may be better to give a counterexample.

Example 41.3.3 (A distribution with finite mean but infinite variance)
Taking $Y_i = \operatorname{sgn}(X_i)\sqrt{|X_i|}$ where X_i is the St. Petersburg paradox example above suffices. If you do the math, you will see $\mathbb{E}[Y_i] = 0$, but $\mathbb{E}[Y_i^2] = \infty$.

We will give a proof when $\mathbb{E}[X_i^2]$ is finite first.

First, we define a seemingly unrelated series as follows:

$$T_n = X_1 + \frac{X_2}{2} + \frac{X_3}{3} + \cdots + \frac{X_n}{n}.$$

This step is a bit difficult to motivate. On the positive side, it is easy to show the following:

Claim 41.3.4. In almost every world, the sequence T_n converges.

That is the same as saying the series

$$X_1 + \frac{X_2}{2} + \frac{X_3}{3} + \cdots$$

converges.

The key idea is to show that the total variance of the summands are finite. Indeed:

$$\begin{aligned}\operatorname{Var}[X_1] + \operatorname{Var}\left[\frac{X_2}{2}\right] + \operatorname{Var}\left[\frac{X_3}{3}\right] + \cdots &= \operatorname{Var}[X_1] + \frac{1}{4}\operatorname{Var}[X_2] + \frac{1}{9}\operatorname{Var}[X_3] + \cdots \\ &= \operatorname{Var}[X_1] \cdot \left(1 + \frac{1}{4} + \frac{1}{9} + \cdots\right)\end{aligned}$$

which is finite.

Why should finite total variance imply almost surely convergence? Intuitively, we recall:

Theorem 41.3.5 (Chebyshev's inequality)
Let X be a random variable with mean 0 and variance σ^2. Then

$$\Pr[|X| \geq k\sigma] \leq \frac{1}{k^2}.$$

Or equivalently we can write it in the following form, which avoid the $\sqrt{-}$ implicit in the definition of σ:

$$\Pr[|X| \geq a] \leq \frac{1}{a^2}\operatorname{Var}[X].$$

So if we look at, say, T_{1000} and T_{2000}:

$$\operatorname{Var}[T_{2000} - T_{1000}] = \sum_{i=1001}^{2000} \frac{\operatorname{Var}[X_i]}{i^2}$$

Because $\sum_{i=1}^{\infty} \frac{\operatorname{Var}[X_i]}{i^2}$ is finite, we expect $\sum_{i=1001}^{2000} \frac{\operatorname{Var}[X_i]}{i^2}$ to be very small, which means T_{2000} should deviate very little from T_{1000}.

To show convergence, we need something stronger, however.

Theorem 41.3.6 (Kolmogorov's inequality)

Let $X_1, \ldots, X_n$ be independent random variables with mean 0. Define $S_i = X_1 + \cdots + X_i$ for each $1 \le i \le n$. Then

$$\Pr[|S_i| \ge a \text{ for any } 1 \le i \le n] \le \frac{1}{a^2} \operatorname{Var}[S_n].$$

You can see why this theorem is stronger — with Chebyshev's inequality, we can only show

$$\Pr[|S_n| \ge a] \le \frac{1}{a^2} \operatorname{Var}[S_n].$$

So, with the same right hand side, we can also bound the probability of $|S_1| \ge a \vee |S_2| \ge a \vee \cdots$ for free!

Proof. Define A_i be the event that i is the smallest value such that $|S_i| \ge a$. Then the left hand side above equals

$$\Pr[|S_i| \ge a \text{ for any } 1 \le i \le n] = \Pr[A_1] + \Pr[A_2] + \cdots + \Pr[A_n].$$

Intuitively, if the events $|S_i| \ge a$ were independent, the best we can do is to use Chebyshev's inequality to bound individual probability values:

$$\Pr[|S_i| \ge a] \le \frac{1}{a^2} \operatorname{Var}[S_i]$$

However, they're very much not independent — in fact, they are positively correlated! For example, we have:

$$\mathbb{E}[S_n \mid S_1 = a] = a$$

because $\mathbb{E}[X_2 + \cdots + X_n] = 0$. So if each X_i is symmetrically distributed, $\Pr[S_n \ge a \mid S_1 = a] \ge \frac{1}{2}$, which is much larger than $\frac{1}{a^2} \operatorname{Var}[S_n]$ for large a.

Here is the formal proof. For each $1 \le i \le n$, we have

$$\mathbb{E}[S_i^2 \mid A_i] \ge a^2$$

which is equivalent to

$$\Pr[A_i] \le \frac{1}{a^2} \mathbb{E}[S_i^2 \cdot \mathbf{1}_{A_i}]$$

and

$$\begin{aligned} \mathbb{E}[S_n^2 \cdot \mathbf{1}_{A_i}] &= \mathbb{E}[(S_i + (S_n - S_i))^2 \cdot \mathbf{1}_{A_i}] \\ &= \mathbb{E}[S_i^2 \cdot \mathbf{1}_{A_i}] + \mathbb{E}[S_i \cdot \mathbf{1}_{A_i}(S_n - S_i)] + \mathbb{E}[(S_n - S_i)^2 \cdot \mathbf{1}_{A_i}] \end{aligned}$$

The middle term $\mathbb{E}[S_i \cdot \mathbf{1}_{A_i}(S_n - S_i)]$ is 0 because $S_i \cdot \mathbf{1}_{A_i}$ and $S_n - S_i = X_{i+1} + \cdots + X_n$ are independent and $\mathbb{E}[X_{i+1} + \cdots + X_n] = 0$, and the last term is ≥ 0.

Combining the inequalities, we get

$$a^2 \Pr[A_i] \le \mathbb{E}[S_n^2 \cdot \mathbf{1}_{A_i}].$$

Summing over all i gives the final result. □

Generalizing:

Corollary 41.3.7

Let $X_1, \dots$ be independent random variables with mean 0. Define S_i as above. Then

$$\Pr[|S_i| \geq a \text{ for any } 1 \leq i] \leq \frac{1}{a^2} \sum_{1 \leq i} \operatorname{Var}[X_i].$$

Proof. The event

$$|S_i| \geq a \text{ for any } 1 \leq i \leq n$$

is a subset of the event

$$|S_i| \geq a \text{ for any } 1 \leq i \leq n+1$$

therefore we have

$$\Pr[|S_i| \geq a \text{ for any } 1 \leq i] = \lim_{n \to \infty} \Pr[|S_i| \geq a \text{ for any } 1 \leq i \leq n].$$

Applying Kolmogorov's inequality on each $\Pr[|S_i| \geq a \text{ for any } 1 \leq i \leq n]$, we get the result. □

Now, the idea is to apply this on the *tails* of the sequence

$$X_1, \frac{X_2}{2}, \frac{X_3}{3}, \dots$$

By the corollary, we know that for every $\varepsilon > 0$, in almost every world, there exists n_ε such that for arbitrary $i \geq n_\varepsilon$, $|T_i - T_{n_\varepsilon}| < \frac{\varepsilon}{2}$. By triangle inequality, for arbitrary $i, j \geq n_\varepsilon$, then $|T_i - T_j| < \varepsilon$.

By Cauchy's criterion for convergence, this implies the sequence T_n is convergent in almost every world.

Finally, here is the relation with the original goal:

Claim 41.3.8 (Relation with the original series)**.** In every world where T_n converges, then M_n converges to 0.

Proof. Just a bit of algebraic manipulation. We try to write M_n in terms of T_n.

We have

$$X_n = n \cdot (T_n - T_{n-1})$$

so

$$\begin{aligned} M_n &= \frac{(T_1 - T_0) + 2(T_2 - T_1) + \dots + n(T_n - T_{n-1})}{n} \\ &= \frac{nT_n - (T_0 + T_1 + \dots + T_{n-1})}{n} \\ &= T_n - \frac{T_0 + T_1 + \dots + T_{n-1}}{n}. \end{aligned}$$

Now this is easy: given T_n converges, $\frac{T_0+T_1+\dots+T_{n-1}}{n}$ must also converge to the same value (Cesàro mean), so $M_n \to 0$ as required. □

Exercise 41.3.9. The converse is not true: if $M_n \to 0$, T_n does not necessarily converge. Find a counterexample. (Write T_n in terms of M_n, and see what happens.)

§41.3.iv The general proof

write

The basic idea is to truncate the value of each X_i so that each of them has finite variance.

§41.4 A few harder problems to think about

Problem 41A (Quantifier hell)**.** In the definition of convergence in probability suppose we allowed $\delta = 0$ (rather than $\delta > 0$). Show that the modified definition is equivalent to almost sure convergence.

Problem 41B (Almost sure convorgence is not topologizable)**.** Consider the space of all random variables on $\Omega = [0, 1]$. Prove that it's impossible to impose a metric on this space which makes the following statement true:

> A sequence X_1, X_2, ..., of random variables converges almost surely to X if and only if X_i converge to X in the metric.

42 Stopped martingales (TO DO)

§42.1 How to make money almost surely

We now take our newfound knowledge of measure theory to a casino.

Here's the most classical example that shows up: a casino lets us play a game where we can bet any amount of on a fair coin flip, but with bad odds: we win $\$n$ if the coin is heads, but lose $\$2n$ if the coin is tails, for a value of n of our choice. This seems like a game that no one in their right mind would want to play.

Well, if we have unbounded time and money, we actually can almost surely make a profit.

Example 42.1.1 (Being even greedier than 18th century France)
In the game above, we start by betting \$1.

- If we win, we leave having made \$1.
- If we lose, we then bet \$10 instead, and
  - If we win, then we leave having made $\$10 - \$2 = \$8$, and
  - If we lose then we bet \$100 instead, and
    - If we win, we leave having made $\$1000 - \$20 - \$2 = \$978$, and
    - If we lose then we bet \$1000 instead, and so on...

Since the coin will almost surely show heads eventually, we make money whenever that happens. In fact, the expected amount of time until a coin shows heads is only 2 flips! What could go wrong?

This chapter will show that under sane conditions such as "finite time" or "finite money", one cannot actually make money in this way — the *optional stopping theorem*. This will give us an excuse to define conditional probabilities, and then talk about martingales (which generalize the fair casino).

Once we realize that trying to extract money from Las Vegas is a lost cause, we will stop gambling and then return to solving math problems, by showing some tricky surprises, where problems that look like they have nothing to do with gambling can be solved by considering a suitable martingale.

In everything that follows, $\Omega = (\Omega, \mathscr{A}, \mu)$ is a probability space.

§42.2 Sub-σ-algebras and filtrations

Prototypical example for this section: σ-algebra generated by a random variable, and coin flip filtration.

We considered our Ω as a space of worlds, equipped with a σ-algebra $\mathscr{A}$ that lets us integrate over Ω. However, it is a sad fact of life that at any given time, you only know partial information about the world. For example, at the time of writing, we know that the

world did not end in 2012 (see https://en.wikipedia.org/wiki/2012_phenomenon), but the fate of humanity in future years remains at slightly uncertain.

Let's write this measure-theoretically: we could consider

$$\begin{aligned}\Omega &= A \sqcup B \\ A &= \{\omega \text{ for which world ends in 2012}\} \\ B &= \{\omega \text{ for which world does not end in 2012}\}.\end{aligned}$$

We will assume that A and B are measurable sets, that is, $A, B \in \mathscr{A}$. That means we could have good fun arguing about what the values of $\mu(A)$ and $\mu(B)$ should be ("a priori probability that the world ends in 2012"), but let's move on to a different silly example.

We will now introduce a new notion that we will need when we define conditional probabilities later.

Definition 42.2.1. Let $\Omega = (\Omega, \mathscr{A}, \mu)$ be a probability space. A **sub-σ-algebra** $\mathscr{F}$ on Ω is exactly what it sounds like: a σ-algebra $\mathscr{F}$ on the set Ω such that each $A \in \mathscr{F}$ is measurable (i.e., $\mathscr{F} \subseteq \mathscr{A}$).

The motivation is that $\mathscr{F}$ is the σ-algebra of sets which let us ask questions about some piece of information. For example, in the 2012 example we gave above, we might take $\mathscr{F} = \{\varnothing, A, B, \Omega\}$, which are the sets we care about if we are thinking only about 2012.

Here are some more serious examples.

Example 42.2.2 (Examples of sub-σ-algebras)

(a) Let $X \colon \Omega \to \{0, 1, 2\}$ be a random variable taking on one of three values. If we're interested in X then we could define

$$\begin{aligned}A &= \{\omega \mid X(\omega) = 1\} \\ B &= \{\omega \mid X(\omega) = 2\} \\ C &= \{\omega \mid X(\omega) = 3\}\end{aligned}$$

then we could write

$$\mathscr{F} = \{\varnothing,\ A,\ B,\ C,\ A \cup B,\ B \cup C,\ C \cup A,\ \Omega\}.$$

This is a sub-σ-algebra on $\mathscr{F}$ that lets us ask questions about X like "what is the probability $X \neq 3$", say.

(b) Now suppose $Y \colon \Omega \to [0, 1]$ is another random variable. If we are interested in Y, the $\mathscr{F}$ that captures our curiosity is

$$\mathscr{F} = \{Y^{\text{pre}}(B) \mid B \subseteq [0, 1] \text{ is measurable } \}.$$

You might notice a trend here which we formalize now:

Definition 42.2.3. Let $X \colon \Omega \to \mathbb{R}$ be a random variable. The **sub-σ-algebra generated by X** is defined by

$$\sigma(X) \coloneqq \{X^{\text{pre}}(B) \mid B \subseteq \mathbb{R} \text{ is measurable } \}.$$

If $X_1, \ldots$ is a sequence (finite or infinite) of random variables, the sub-σ-algebra generated by them is the smallest σ-algebra which contains $\sigma(X_i)$ for each i.

Finally, we can put a lot of these together — since we're talking about time, we learn more as we grow older, and this can be formalized.

Definition 42.2.4. A **filtration** on $\Omega = (\Omega, \mathscr{A}, \mu)$ is a nested sequence[1]

$$\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \dots$$

of sub-σ-algebras on Ω.

Example 42.2.5 (Filtration)

Suppose you're bored in an infinitely long class and start flipping a fair coin to pass the time. (Accordingly, we could let $\Omega = \{H, T\}^\infty$ consist of infinite sequences of heads H and tails T.) We could let $\mathscr{F}_n$ denote the sub-σ-algebra generated by the values of the first n coin flips. So:

- $\mathscr{F}_0 = \{\varnothing, \Omega\}$,
- $\mathscr{F}_1 = \{\varnothing, \text{first flip } H, \text{first flip } T, \Omega\}$,
- $\mathscr{F}_2 = \{\varnothing, \text{first flips } HH, \text{second flip } T, \Omega, \text{first flip and second flip differ}, \dots\}$.
- and so on, with $\mathscr{F}_n$ being the measurable sets "determined" only by the first n coin flips.

Exercise 42.2.6. In the previous example, compute the cardinality $|\mathscr{F}_n|$ for each integer n.

More importantly,

X is $\mathscr{F}$-measurable if X is determined only by the information given in $\mathscr{F}$.

Example 42.2.7

In the example above, let X_3 be the value of the third coin flip. Then:

- X_3 is not $\mathscr{F}_2$-measurable. (That is, we don't know X_3 from the knowledge of the first 2 coin flips.)
- But it is $\mathscr{F}_3$-measurable.

Exercise 42.2.8. Check this! (Recall that a function is measurable if it lifts open sets to measurable sets. So you need to show e.g. $X_3^{\text{pre}}(\{H\}) \notin \mathscr{F}_2$.)

So, not only can we formalize partial information about the world, we can also formalize what it means for something to only depend on that partial information.

[1] For convenience, we will restrict ourselves to $\mathbb{Z}_{\geq 0}$-indexed filtrations, though really any index set is okay.

§42.3 Conditional expectation

Prototypical example for this section: $\mathbb{E}(X \mid X+Y)$ *for* X *and* Y *distributed over* $[0,1]$.

We'll need the definition of conditional probability to define a martingale, but this turns out to be surprisingly tricky. Let's consider the following simple example to see why.

Example 42.3.1 (Why high-school methods aren't enough here)
Suppose we have two independent random variables X, Y distributed uniformly over $[0,1]$ (so we may as well take $\Omega = [0,1]^2$). We might try to ask the question:

> *"what is the expected value of* X *given that* $X+Y = 0.6$*"?*

Intuitively, we know the answer has to be 0.3. However, if we try to write down a definition, we quickly run into trouble. Ideally we want to say something like

$$\mathbb{E}[X \text{ given } X+Y=0.6] = \frac{\int_S X}{\int_S 1} \text{ where } S = \{\omega \in \Omega \mid X(\omega)+Y(\omega) = 0.6\}.$$

The problem is that S is a set of measure zero, so we quickly run into $\frac{0}{0}$, meaning a definition of this shape will not work out.

The way that this is typically handled in measure theory is to use the notion of sub-σ-algebra that we defined.

But first, we should explain what $\mathbb{E}(X \mid X+Y)$ means first — why are we conditioning on another random variable instead of an event?

To motivate conditioning on a random variable, consider the following situation. Suppose that the weather tomorrow depends on the weather today and the random fluctuations. So we may have statements such as:

$$\Pr(\text{it rains tomorrow} \mid \text{it rains today}) = 0.6,$$
$$\Pr(\text{it rains tomorrow} \mid \text{it doesn't rain today}) = 0.3.$$

This is the standard conditional probability: $\Pr(A \mid B) = \frac{\Pr(A \wedge B)}{\Pr(B)}$.

Note that "the weather today" is itself a random variable.

Let Z be the weather forecast tonight's prediction of the probability, suppose it works as above. Then Z is a random real variable, defined by:

$$Z \colon \Omega \to \mathbb{R}$$
$$Z(\omega) = \Pr(\text{it rains tomorrow} \mid \text{weather today} = \text{weather today}(\omega))$$

It would only be reasonable to write

$$Z = \Pr(\text{it rains tomorrow} \mid \text{weather today}).$$

We're conditioning on a random variable, and $\Pr(\cdots \mid \cdots)$ is itself a random variable instead of a single value in $\mathbb{R}$, but that's perfectly okay.

Similarly, if Ω is finite and every subset of it is measurable, for random real variables X and Y it would be sensible for us to define random real variable $Z = \mathbb{E}(X \mid Y)$ by

$$Z \colon \Omega \to \mathbb{R}$$
$$Z(\omega) = \mathbb{E}[X \mid Y = Y(\omega)].$$

Example 42.3.2

Let X and Y be the result of rolling two dices. Then:

- $\mathbb{E}[X \mid X+Y=3] = 1.5$, as you can easily calculate.
- $\mathbb{E}(X \mid X+Y)$ is a random variable, which would takes the value 1.5 in any world whether $X+Y=3$.
- More generally, we have in fact

$$\mathbb{E}(X \mid X+Y) = \frac{X+Y}{2}.$$

Notice how the random variable $\mathbb{E}(X \mid X+Y)$ depends only on the value of $X+Y$ — by definition.

Of course, as we explained earlier, this naive attempts will give us division-by-zero everywhere for the continuous case — so, enters the sub-σ-algebra.

Proposition 42.3.3 (Conditional expectation definition)

Let $X\colon \Omega \to \mathbb{R}$ be an *absolutely integrable* random variable (meaning $\mathbb{E}[|X|] < \infty$) over a probability space Ω, and let $\mathscr{F}$ be a sub-σ-algebra on it.
Then there exists a function $\eta\colon \Omega \to \mathbb{R}$ satisfying the following two properties:

- η is $\mathscr{F}$-measurable (that is, measurable as a function $(\Omega, \mathscr{F}, \mu) \to \mathbb{R}$); and
- for any set $A \in \mathscr{F}$ we have $\mathbb{E}[\eta \cdot \mathbf{1}_A] = \mathbb{E}[X \cdot \mathbf{1}_A]$.

Moreover, this random variable is unique up to almost sureness.

Proof. Omitted, but relevant buzzword used is "Radon-Nikodym derivative". □

Definition 42.3.4. Let η be as in the previous proposition.

- We denote η by $\mathbb{E}(X \mid \mathscr{F})$ and call it the **conditional expectation** of X with respect to $\mathscr{F}$.
- If Y is a random variable then $\mathbb{E}(X \mid Y)$ denotes $\mathbb{E}(X \mid \sigma(Y))$, i.e. the conditional expectation of X with respect to the σ-algebra generated by Y.

Example 42.3.5

As we can expect, $\eta = \frac{X+Y}{2}$ satisfies the condition of $\mathbb{E}(X \mid X+Y)$ above.
The way to motivate doing all this is the following. We want to be able to say something like:

$$\mathbb{E}[X \mid X+Y=0.6] = \lim_{\varepsilon \to 0} \mathbb{E}[X \mid 0.6-\varepsilon < X+Y < 0.6+\varepsilon]$$

Unfortunately, this setup does not work in general where $\mathscr{F}$ might not be generated by just one random real variable. Let's see how the definition above helps us.

- Let $A = \{\omega \in \Omega \mid 0.6-\varepsilon < X(\omega)+Y(\omega) < 0.6+\varepsilon\}$, this set certainly belongs to the sub-σ-algebra generated by $X+Y$ (because it is $(X+Y)^{\text{pre}}((0.6-$

$\varepsilon, 0.6+\varepsilon))$).

- Recall that η is $\mathscr{F}$-measurable means η only depends on the information in $\mathscr{F} = \sigma(X+Y)$, that is, on $X+Y$. This makes sense.
- Look at the right hand side:

$$\mathbb{E}[X \mid 0.6-\varepsilon < X+Y < 0.6+\varepsilon] = \mathbb{E}[X \cdot \mathbf{1}_A].$$

The law of total expectation says that $\mathbb{E}[\mathbb{E}(X \mid Y)] = \mathbb{E}[X]$. So, intuitively, the second property above simply requires this law to hold over all set $A \in \mathscr{F}$.

In our case, we have the following:

$$\mathbb{E}[\eta \mid 0.6-\varepsilon < X+Y < 0.6+\varepsilon] = \mathbb{E}[X \mid 0.6-\varepsilon < X+Y < 0.6+\varepsilon].$$

More fine print:

Remark 42.3.6 (This notation is terrible) — The notation $\mathbb{E}(X \mid \mathscr{F})$ is admittedly confusing, since it is actually an entire function $\Omega \to \mathbb{R}$, rather than just a real number like $\mathbb{E}[X]$ — though, as you can see, it has its merits. For this reason I try to be careful to remember to use parentheses rather than square brackets for conditional expectations; not everyone does this.

Abuse of Notation 42.3.7. In addition, when we write $Y = \mathbb{E}(X \mid \mathscr{F})$, there is some abuse of notation happening here since $\mathbb{E}(X \mid \mathscr{F})$ is defined only up to some reasonable uniqueness (i.e. up to measure zero changes). So this really means that "Y satisfies the hypothesis of Proposition 42.3.3", but this is so pedantic that no one bothers.

For example, in the example above, if we change η to be

$$\eta(\omega) = \begin{cases} 0 & \text{if } X(\omega)+Y(\omega) = 0.6 \\ \frac{X(\omega)+Y(\omega)}{2} & \text{otherwise} \end{cases}$$

then $\mathbb{E}[\eta \cdot \mathbf{1}_A] = \mathbb{E}[X \cdot \mathbf{1}_A]$ still holds for every set A, but now it seems to be saying $\mathbb{E}[X \mid X+Y = 0.6] = 0$?

Nevertheless, we must agree that we must sacrifice a measure zero set, since otherwise if we have

$$T(\omega) = \begin{cases} 1 & \text{if } Y(\omega) > 0.5 \text{ or } (Y(\omega) = 0.5 \text{ and } X(\omega) \in \mathbb{Q}) \\ 0 & \text{otherwise} \end{cases}$$

then it is certainly measurable (i.e. a random variable), $\mathbb{E}[T \mid Y = 0.4] = 0$ and $\mathbb{E}[T \mid Y = 0.6] = 1$, but what is $\mathbb{E}[T \mid Y = 0.5]$? (You may argue it should be 0, but what if $\mathbb{Q}$ is changed to something more non-measurable? Besides, why should the conditional expectation change when we only modify T on a probability-zero set anyway?)

properties

§42.4 Supermartingales

Prototypical example for this section: Visiting a casino is a supermartingale, assuming house odds.

Definition 42.4.1. Let X_0, X_1, ... be a sequence of random variables on a probability space Ω, and let $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots$ be a filtration.

Then $(X_n)_{n\geq 0}$ is a **supermartingale** with respect to $(\mathscr{F}_n)_{n\geq 0}$ if the following conditions hold:

- X_n is absolutely integrable for every n;
- X_n is measurable with respect to $\mathscr{F}_n$; and
- for each $n = 1, 2, \ldots$ the inequality
$$\mathbb{E}(X_n \mid \mathscr{F}_{n-1}) \leq X_{n-1}$$
holds for all $\omega \in \Omega$.

In a **submartingale** the inequality $\leq$ is replaced with $\geq$, and in a **martingale** it is replaced by $=$.

what's the etymology of the term?

Abuse of Notation 42.4.2 (No one uses that filtration thing anyways)**.** We will always take $\mathscr{F}_n$ to be the σ-algebra generated by the previous variables X_0, X_1, ..., X_{n-1}, and do so without further comment. Nonetheless, all the results that follow hold in the more general setting of a supermartingale with respect to some filtration.

We will prove all our theorems for supermartingales; the analogous versions for submartingales can be obtained by replacing $\leq$ with $\geq$ everywhere (since X_n is a martingale iff $-X_n$ is a supermartingale) and for martingales by replacing $\leq$ with $=$ everywhere (since X_n is a martingale iff it is both a supermartingale and a submartingale).

Let's give examples.

Example 42.4.3 (Supermartingales)

(a) **Random walks**: an ant starts at the position 0 on the number line. Every minute, it flips a fair coin and either walks one step left or one step right. If X_t is the position at the tth time, then X_t is a martingale, because
$$\mathbb{E}(X_t \mid X_0, \ldots, X_{t-1}) = \frac{(X_{t-1}+1) + (X_{t-1}-1)}{2} = X_{t-1}.$$

(b) **Casino game**: Consider a gambler using the strategy described at the beginning of the chapter. This is a martingale, since every bet the gambler makes has expected value 0.

(c) **Multiplying independent variables**: Let X_1, X_2, ..., be independent (not necessarily identically distributed) integrable random variables with mean 1. Then the sequence Y_1, Y_2, ... defined by
$$Y_n \coloneqq X_1 X_2 \cdots X_n$$
is a martingale; as $\mathbb{E}(Y_n \mid Y_1, \ldots, Y_{n-1}) = \mathbb{E}[Y_n] \cdot Y_{n-1} = Y_{n-1}$.

(d) **Iterated blackjack**: Suppose one shows up to a casino and plays infinitely many games of blackjack. If X_t is their wealth at time t, then X_t is a supermartingale. This is because each game has negative expected value (house edge).

Example 42.4.4 (Frivolous/inflamamtory example — real life is a supermartingale)
Let X_t be your happiness on day t of your life. Life has its ups and downs, so it is not the case that $X_t \le X_{t-1}$ for every t. For example, you might win the lottery one day.
However, on any given day, many things can go wrong (e.g. zombie apocalypse), and by Murphy's Law this is more likely than things going well. Also, as you get older, you have an increasing number of responsibilities and your health gradually begins to deteriorate.
Thus it seems that

$$\mathbb{E}(X_t \mid X_0, \dots, X_{t-1}) \le X_{t-1}$$

is a reasonable description of the future — *in expectation*, each successive day is slightly worse than the previous one. (In particular, if we set $X_t = -\infty$ on death, then as long as you have a positive probability of dying, the displayed inequality is obviously true.)

Before going on, we will state without proof one useful result: if a martingale is bounded, then it will almost certainly converge.

Theorem 42.4.5 (Doob's martingale convergence theorem)
Let $X_0, \dots$ be a supermartingale on a probability space Ω such that

$$\sup_{n \in \mathbb{Z}_{\ge 0}} \mathbb{E}\left[|X_n|\right] < \infty.$$

Then, there exists a random variable $X_\infty \colon \Omega \to \mathbb{R}$ such that

$$X_n \xrightarrow{\text{a.s.}} X_\infty.$$

§42.5 Optional stopping

Prototypical example for this section: Las Vegas.

In the first section we described how to make money almost surely. The key advantage the gambler had was the ability to quit whenever he wanted (equivalently, an ability to control the size of the bets; betting \$0 forever is the same as quitting.) Let's formalize a notion of stopping time.

The idea is we want to define a function $\tau \colon \Omega \to \{0, 1, 2, \dots\} \cup \{\infty\}$ such that

- $\tau(\omega)$ specifies the index after which we *stop* the martingale. Note that the decisions to stop after time n must be made with only the information available at that time — i.e., with respect to $\mathscr{F}_n$ of the filtration.
- $X_{\tau \wedge n}$ is the random value representing the value at time n of the stopped martingale, where if n is *after* the stopping time, we just take it to be the our currently value after we leave.

  So for example in a world ω where we stopped at time 3, then $X_{\tau\wedge 0}(\omega) = X_0(\omega)$, $X_{\tau\wedge 1}(\omega) = X_1(\omega)$, $X_{\tau\wedge 2}(\omega) = X_2(\omega)$, $X_{\tau\wedge 3}(\omega) = X_3(\omega)$, but then

  $$X_3(\omega) = X_{\tau\wedge 4}(\omega) = X_{\tau\wedge 5}(\omega) = X_{\tau\wedge 6}(\omega) = \dots$$

  since we have stopped — the value stops changing.

- X_τ denotes the eventual value after we stop (or the limit X_∞ if we never stop).

Here's the compiled machine code.

Definition 42.5.1. Let $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots$ be a filtration on a probability space Ω.

- A **stopping time** is a function

$$\tau \colon \Omega \to \{0, 1, 2, \dots\} \cup \{\infty\}$$

with the property that for each integer n, the set

$$\{\omega \in \Omega \mid \tau(\omega) = n\}$$

is $\mathscr{F}_n$-measurable (i.e., is in $\mathscr{F}_n$).

- For each $n \geq 0$ we define $X_{\tau \wedge n} \colon \Omega \to \mathbb{R}$ by

$$X_{\tau \wedge n}(\omega) = X_{\min\{\tau(\omega), n\}}(\omega)$$

- Finally, we let the eventual outcome be denoted by

$$X_\tau(\omega) = \begin{cases} X_{\tau(\omega)}(\omega) & \tau(\omega) \neq \infty \\ \lim_{n \to \infty} X_n(\omega) & \tau(\omega) = \infty \text{ and } \lim_{n \to \infty} X_n(\omega) \text{ exists} \\ \text{undefined} & \text{otherwise.} \end{cases}$$

We require that the "undefined" case occurs only for a set of measure zero (for example, if Theorem 42.4.5 applies). Otherwise we don't allow X_τ to be defined.

Proposition 42.5.2 (Stopped supermartingales are still supermartingales)
Let $X_0, X_1, \dots$ be a supermartingale. Then the sequence

$$X_{\tau \wedge 0}, \ X_{\tau \wedge 1}, \ \dots$$

is itself a supermartingale.

Proof. We have almost everywhere the inequalities

$$\begin{aligned} \mathbb{E}\left(X_{\tau \wedge n} \mid \mathscr{F}_{n-1}\right) &= \mathbb{E}\left(X_{n-1} + \mathbf{1}_{\tau(\omega) = n-1}(X_n - X_{n-1}) \mid \mathscr{F}_{n-1}\right) \\ &= \mathbb{E}\left(X_{n-1} \mid \mathscr{F}_{n-1}\right) + \mathbb{E}\left(\mathbf{1}_{\tau(\omega) = n-1} \cdot (X_n - X_{n-1}) \mid \mathscr{F}_{n-1}\right) \\ &= X_{n-1} + \mathbf{1}_{\tau(\omega) = n-1} \cdot \mathbb{E}\left(X_n - X_{n-1} \mid \mathscr{F}_{n-1}\right) \leq X_{n-1} \end{aligned}$$

as functions from $\Omega \to \mathbb{R}$. □

Theorem 42.5.3 (Doob's optional stopping theorem)

Let X_0, X_1, ... be a supermartingale on a probability space Ω, with respect to a filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots$. Let τ be a stopping time with respect to this filtration. Suppose that *any* of the following hypotheses are true, for some constant C:

(a) **Finite time**: $\tau(\omega) \le C$ for almost all ω.

(b) **Finite money**: for each $n \ge 1$, $|X_{\tau \wedge n}(\omega)| \le C$ for almost all ω.

(c) **Finite bets**: we have $\mathbb{E}[\tau] < \infty$, and for each $n \ge 1$, the conditional expectation

$$\mathbb{E}\left(|X_n - X_{n-1}| \mid \mathscr{F}_n\right)$$

takes on values at most C for almost all $\omega \in \Omega$ satisfying $\tau(\omega) \ge n$.

Then X_τ is well-defined almost everywhere, and more importantly,

$$\mathbb{E}[X_\tau] \le \mathbb{E}[X_0].$$

The last equation can be cheekily expressed as "the only winning move is not to play".

do later tonight

Proof. □

Exercise 42.5.4. Conclude that going to Las Vegas with the strategy described in the first section is a really bad idea. What goes wrong?

While this is useful to make us stop gambling, it doesn't allow us to compute anything — we don't know anything about $\mathbb{E}[X_\tau]$ other than it's $\le \mathbb{E}[X_0]$. However:

Corollary 42.5.5

With the same hypothesis as above:

- If X_0, X_1, ... is a submartingale, then $\mathbb{E}[X_\tau] \ge \mathbb{E}[X_0]$.
- If it is a martingale, then $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$.

Proof. If X_0, X_1, ... is a submartingale, then Y_0, Y_1, ... defined by $Y_i = -X_i$ is a supermartingale, and the hypothesis is still satisfied. Apply the theorem to Y_τ we get the result.

If X_0, X_1, ... is a martingale, then it is both a supermartingale and a submartingale, the result follows immediately. □

This finally let us calculate something — if we can compute $\mathbb{E}[X_0]$ and write the result as $\mathbb{E}[X_\tau]$ for some martingale, then we can solve the problem!

§42.6 Fun applications of optional stopping (TO DO)

We now give three problems which showcase some of the power of the results we have developed so far.

§42.6.i The ballot problem

Suppose Alice and Bob are racing in an election; Alice received a votes total while Bob received b votes total, and $a > b$. If the votes are chosen in random order, one could ask: what is the probability that Alice remains strictly ahead of Bob in the election?

Proposition 42.6.1 (Ballot problem)
This occurs with probability $\frac{a-b}{a+b}$.

We should try to model this as a martingale. A natural way to do it is the random walk, as in Example 42.4.3:

$$X_0 = 0$$
$$X_i = X_{i-1} + \begin{cases} 1 \text{ with probability } \frac{1}{2} \\ -1 \text{ otherwise.} \end{cases}$$

Here, each 1 represents Alice getting a vote, and each -1 represents Bob getting a vote.

Then, we need to compute

$$\Pr[X_i > 0 \text{ for all } 1 \le i \le a+b \mid X_{a+b} = a-b].$$

While this is natural, the fact that the probability is conditioned on $X_{a+b} = a - b$ makes us unable to apply the optional stopping theorem.

Instead, the following will work: We start with all the votes, and *remove* them in random order.

$$(A_0, B_0) = (a, b),$$
$$(A_i, B_i) = \begin{cases} (A_{i-1} - 1, B_{i-1}) \text{ with probability } \frac{A_{i-1}}{A_{i-1}+B_{i-1}} \\ (A_{i-1}, B_{i-1} - 1) \text{ otherwise} \end{cases} \quad \text{for } 1 \le i \le a+b.$$

Of course, here A_i represents the number of votes Alice has left, and B_i represents the number of votes Bob has left.

Now we just need to compute

$$\Pr[A_i > B_i \text{ for all } 0 \le i \le a+b-1].$$

There's no longer any conditional expectation!

Next, we need to construct a martingale. We could try to define $X_i = A_i - B_i$ similar to above, but that will not be a martingale.

Here is the trick: we *modify* X_i to make it a martingale. (More examples where a sequence of random variables is modified to create a martingale can be found in Problem 42A.)

Example 42.6.2

Suppose $a = 2$ and $b = 1$. Then:

$$(A_0, B_0) = (2, 1),$$
$$(A_1, B_1) = \begin{cases} (1,1) \text{ with probability } \frac{2}{3} \\ (2,0) \text{ otherwise.} \end{cases}$$

So $\mathbb{E}[A_0 - B_0] = 1$ while $\mathbb{E}[A_1 - B_1] = \frac{2}{3} \cdot 0 + \frac{1}{3} \cdot 2 = \frac{2}{3}$, if we define X_i as above of course it wouldn't be a martingale.
However, it's not difficult to find ways to modify it to form a martingale. For example:

- If we define $X_1 = A_1 - B_1 + \frac{1}{3}$, then $\mathbb{E}[X_1] = 1$, so we're fine.
- Similarly, we can also define $X_1 = \frac{3}{2} \cdot (A_1 - B_1)$.
- Or $X_1 = \frac{9}{4} \cdot (A_1 - B_1)^2$.
- ... et cetera...

We need to make $\mathbb{E}[X_i \mid X_0 = x_0, X_1 = x_1, \dots, X_{i-1} = x_{i-1}] = x_{i-1}$ for *every* i and every $(x_0, x_1, \dots, x_{i-1})$. (You can try to find out yourself what modification will work yourself before continue reading.)

Turns out, the following will work:

$$X_i = \frac{A_i - B_i}{a + b - i} \text{ for } 0 \le i \le a + b - 1.$$

We cannot extend it to $i \ge a+b$, but it is fine. (If you're worried, just define $X_i = X_{a+b-1}$ for $i \ge a + b$.)

Exercise 42.6.3. Check that it works. (The math is very similar to the problem about Pólya's urn in Problem 42A.)

For the stopping time, there is only one natural way to define it: τ is $a + b - 1$ if Alice remains ahead of Bob i.e. $X_i > 0$ for every $0 \le i \le a + b - 1$, otherwise τ is the smallest i such that $X_i = 0$.

Then, the optional stopping theorem states:

$$\mathbb{E}[X_\tau] = \mathbb{E}[X_0].$$

$\mathbb{E}[X_0]$ is easy to calculate, it is $\frac{A_0 - B_0}{a+b} = \frac{a-b}{a+b}$. What is $\mathbb{E}[X_\tau]$?

- If Alice remains ahead of Bob, $X_\tau = X_{a+b-1} = 1$.
- Otherwise, $X_\tau = 0$.

Therefore $\mathbb{E}[X_\tau]$ is exactly the probability we need to calculate, so we're done.

Remark 42.6.4 (This is cheating a little) — Note that you can equivalently write

$$X_i = \frac{A_i - B_i}{A_i + B_i}.$$

> Which is exactly the form of the answer.
> That is to say, if you already know the form of the answer, martingale theory can help you to check it. But if you don't...

§42.6.ii ABRACADABRA

To be written.
https://www.jeremykun.com/2014/03/03/martingales-and-the-optional-stopping-theorem/

§42.6.iii USA TST 2018

To be written.
https://aops.com/community/p9513099

§42.7 A few harder problems to think about

Problem 42A (Examples of martingales)**.** We give some more examples of martingales.

(a) **(Simple random walk)** Let X_1, X_2, ... be i.i.d. random variables which equal $+1$ with probability $1/2$, and -1 with probability $1/2$. Prove that

$$Y_n = (X_1 + \cdots + X_n)^2 - n$$

is a martingale.

(b) **(de Moivre's martingale)** Fix real numbers p and q such that $p, q > 0$ and $p+q = 1$. Let X_1, X_2, ... be i.i.d. random variables which equal $+1$ with probability p, and -1 with probability q. Show that

$$Y_n = \left(qp^{-1}\right)^{X_1+X_2+\cdots+X_n}$$

is a martingale.

(c) **(Pólya's urn)** An urn contains one red and one blue marble initially. Every minute, a marble is randomly removed from the urn, and two more marbles of the same color are added to the urn. Thus after n minutes, the urn will have $n+2$ marbles.

Let r_n denote the fraction of marbles which are red. Show that r_n is a martingale.

Problem 42B. A deck has 52 cards; of them 26 are red and 26 are black. The cards are drawn and revealed one at a time. At any point, if there is at least one card remaining in the deck, you may stop the dealer; you win if (and only if) the next card in the deck is red. If all cards are dealt, then you lose. Across all possible strategies, determine the maximal probability of winning.

Problem 42C (Wald's identity)**.** Let μ be a real number. Let X_1, X_2, ... be independent random variables on a probability space Ω with mean μ. Finally let $\tau\colon \Omega \to \{1, 2, \dots\}$ be a stopping time such that $\mathbb{E}[\tau] < \infty$, and the event $\tau = n$ depends only on X_1, ..., X_n.

Prove that

$$\mathbb{E}[X_1 + X_2 + \cdots + X_\tau] = \mu\mathbb{E}[\tau].$$

Problem 42D (Unbiased drunkard's walk)**.** An ant starts at 0 on a number line, and walks left or right one unit with probability $1/2$. It stops once it reaches either -17 or $+8$.

(a) Find the probability it reaches $+8$ before -17.

(b) Find the expected value of the amount of time it takes to reach either endpoint.

Problem 42E (Biased drunkard's walk)**.** Let $0 < p < 1$ be a real number. An ant starts at 0 on a number line, and walks left or right one unit with probability p. It stops once it reaches either -17 or $+8$. Find the probability it reaches $+8$ first.

Problem 42F. The number 1 is written on a blackboard. Every minute, if the number a is written on the board, it's erased and replaced by a real number in the interval $[0, 2.01a]$ selected uniformly at random. What is the probability that the resulting sequence of numbers approaches 0?

XII

Differential Geometry

Part XII: Contents

43 Multivariable calculus done correctly

As I have ranted about before, linear algebra is done wrong by the extensive use of matrices to obscure the structure of a linear map. Similar problems occur with multivariable calculus, so here I would like to set the record straight.

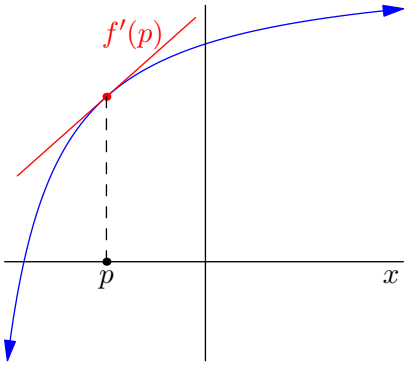
Since we are doing this chapter using morally correct linear algebra, it's imperative you're comfortable with linear maps, and in particular the dual space $V^\vee$ which we will repeatedly use.

In this chapter, all vector spaces have norms and are finite-dimensional over $\mathbb{R}$. So in particular every vector space is also a metric space (with metric given by the norm), and we can talk about open sets as usual.

§43.1 The total derivative

Prototypical example for this section: *If* $f(x,y) = x^2+y^2$, *then* $(Df)_{(x,y)} = 2x\mathbf{e}_1^\vee + 2y\mathbf{e}_2^\vee$.

First, let $f\colon [a,b] \to \mathbb{R}$. You might recall from high school calculus that for every point $p \in \mathbb{R}$, we defined $f'(p)$ as the derivative at the point p (if it existed), which we interpreted as the *slope* of the "tangent line".



That's fine, but I claim that the "better" way to interpret the derivative at that point is as a *linear map*, that is, as a *function*. If $f'(p) = 1.5$, then the derivative tells me that if I move ε away from p then I should expect f to change by about 1.5ε. In other words,

The derivative of f at p approximates f near p by a *linear function*.

What about more generally? Suppose I have a function like $f\colon \mathbb{R}^2 \to \mathbb{R}$, say

$$f(x,y) = x^2+y^2$$

for concreteness or something. For a point $p \in \mathbb{R}^2$, the "derivative" of f at p ought to represent a linear map that approximates f at that point p. That means I want a linear map $T\colon \mathbb{R}^2 \to \mathbb{R}$ such that

$$f(p+v) \approx f(p) + T(v)$$

for small displacements $v \in \mathbb{R}^2$.

Even more generally, if $f\colon U \to W$ with $U \subseteq V$ open (in the $\|\bullet\|_V$ metric as usual), then the derivative at $p \in U$ ought to be so that

$$f(p+v) \approx f(p) + T(v) \in W.$$

(We need U open so that for small enough v, $p+v \in U$ as well.) In fact this is exactly what we were doing earlier with $f'(p)$ in high school.

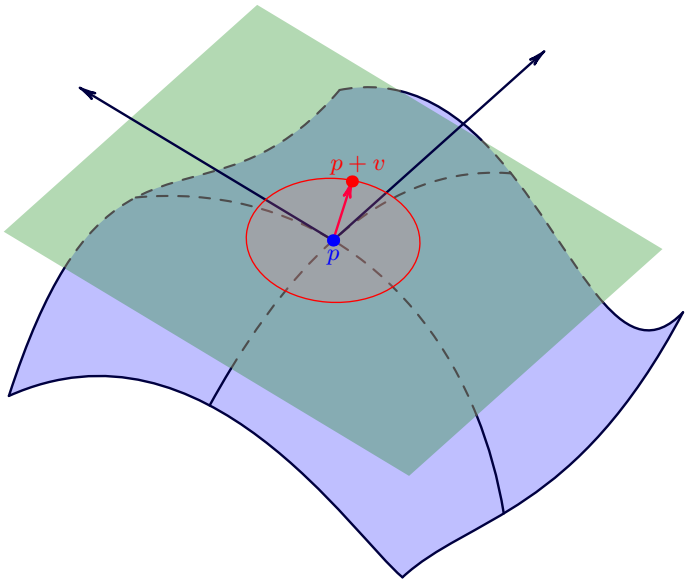


Image derived from [gk]

The only difference is that, by an unfortunate coincidence, a linear map $\mathbb{R} \to \mathbb{R}$ can be represented by just its slope. And in the unending quest to make everything a number so that it can be AP tested, we immediately forgot all about what we were trying to do in the first place and just defined the derivative of f to be a *number* instead of a *function*.

The fundamental idea of Calculus is the local approximation of functions by linear functions. The derivative does exactly this.

Jean Dieudonné as quoted in [Pu02] continues:

> In the classical teaching of Calculus, this idea is immediately obscured by the accidental fact that, on a one-dimensional vector space, there is a one-to-one correspondence between linear forms and numbers, and therefore the derivative at a point is defined as a number instead of a linear form. This **slavish subservience to the shibboleth of numerical interpretation at any cost** becomes much worse . . .

So let's do this right. The only thing that we have to do is say what "$\approx$" means, and for this we use the norm of the vector space.

Definition 43.1.1. Let $U \subseteq V$ be open. Let $f\colon U \to W$ be a continuous function, and $p \in U$. Suppose there exists a linear map $T\colon V \to W$ such that

$$\lim_{\|v\|_V \to 0} \frac{\|f(p+v) - f(p) - T(v)\|_W}{\|v\|_V} = 0.$$

Then T is the **total derivative** of f at p. We denote this by $(Df)_p$, and say f is **differentiable at** p.

If $(Df)_p$ exists at every point, we say f is **differentiable**.

Question 43.1.2. Check if that $V = W = \mathbb{R}$, this is equivalent to the single-variable definition. (What are the linear maps from V to W?)

Example 43.1.3 (Total derivative of $f(x,y) = x^2 + y^2$)
Let $V = \mathbb{R}^2$ with standard basis $\mathbf{e}_1$, $\mathbf{e}_2$ and let $W = \mathbb{R}$, and let $f(x\mathbf{e}_1 + y\mathbf{e}_2) = x^2 + y^2$. Let $p = a\mathbf{e}_1 + b\mathbf{e}_2$. Then, we claim that

$$(Df)_p \colon \mathbb{R}^2 \to \mathbb{R} \quad \text{by} \quad v \mapsto 2a \cdot \mathbf{e}_1^\vee(v) + 2b \cdot \mathbf{e}_2^\vee(v).$$

Here, the notation $\mathbf{e}_1^\vee$ and $\mathbf{e}_2^\vee$ makes sense, because by definition $(Df)_p \in V^\vee$: these are functions from V to $\mathbb{R}$!

Let's check this manually with the limit definition. Set $v = xe_1 + ye_2$, and note that the norm on V is $\|(x,y)\|_V = \sqrt{x^2+y^2}$ while the norm on W is just the absolute value $\|c\|_W = |c|$. Then we compute

$$\begin{aligned}
\frac{\|f(p+v) - f(p) - T(v)\|_W}{\|v\|_V} &= \frac{\left|(a+x)^2 + (b+y)^2 - (a^2+b^2) - (2ax+2by)\right|}{\sqrt{x^2+y^2}} \\
&= \frac{x^2+y^2}{\sqrt{x^2+y^2}} = \sqrt{x^2+y^2} \\
&\to 0
\end{aligned}$$

as $\|v\| \to 0$. Thus, for $p = ae_1 + be_2$ we indeed have $(Df)_p = 2a \cdot \mathbf{e}_1^\vee + 2b \cdot \mathbf{e}_2^\vee$.

Remark 43.1.4 — As usual, differentiability implies continuity.

Remark 43.1.5 — Although $U \subseteq V$, it might be helpful to think of vectors from U and V as different types of objects (in particular, note that it's possible for $0_V \notin U$). The vectors in U are "inputs" on our space while the vectors coming from V are "small displacements". For this reason, I deliberately try to use $p \in U$ and $v \in V$ when possible.

§43.2 The projection principle

You may have learned single-variable calculus as the topic of doing differentiation and integration on single-variable functions $\mathbb{R} \to \mathbb{R}$. So "multivariable calculus" ought to be calculus with functions $\mathbb{R}^n \to \mathbb{R}^m$. You might notice there are *two* upgrades happening here:

- The domain got upgraded from $\mathbb{R}$ to $\mathbb{R}^n$.
- The codomain got upgraded from $\mathbb{R}$ to $\mathbb{R}^m$.

The point of this section is that the second upgrade is *super easy* in comparison to the first upgrade, and basically doesn't require doing anything new. All the interesting actions happens because we upgraded the domain, not the codomain. Here's why:

Theorem 43.2.1 (Projection principle)

Let U be an open subset of the vector space V. Let W be an m-dimensional real vector space with basis $w_1, \dots, w_m$. Then there is a bijection between continuous functions $f\colon U \to W$ and m-tuples of continuous $f_1, f_2, \dots, f_m\colon U \to \mathbb{R}$ by projection onto the ith basis element, i.e.

$$f(v) = f_1(v)w_1 + \cdots + f_m(v)w_m.$$

Proof. Obvious. □

The theorem remains true if one replaces "continuous" by "differentiable", "smooth", "arbitrary", or most other reasonable words. Translation:

To think about a function $f\colon U \to \mathbb{R}^m$, it suffices to think about each coordinate separately.

For this reason, we'll most often be interested in functions $f\colon U \to \mathbb{R}$. That's why the dual space $V^\vee$ is so important.

§43.3 Total and partial derivatives

Prototypical example for this section: *If $f(x,y) = x^2 + y^2$, then $(Df)\colon (x,y) \mapsto 2x \cdot \mathbf{e}_1^\vee + 2y \cdot \mathbf{e}_2^\vee$, and $\frac{\partial f}{\partial x} = 2x$, $\frac{\partial f}{\partial y} = 2y$.*

Let $U \subseteq V$ be open and let V have a basis $\mathbf{e}_1, \dots, \mathbf{e}_n$. Suppose $f\colon U \to \mathbb{R}$ is a function which is differentiable everywhere, meaning $(Df)_p \in V^\vee$ exists for every p. In that case, one can consider Df as *itself* a function:

$$\begin{aligned} Df\colon U &\to V^\vee \\ p &\mapsto (Df)_p. \end{aligned}$$

This is a little crazy: to every *point* in U we associate a *function* in $V^\vee$. We say Df is the **total derivative** of f, to reflect how much information we're dealing with. We say $(Df)_p$ is the total derivative at p.

Let's apply the projection principle now to Df. Since we picked a basis $\mathbf{e}_1, \dots, \mathbf{e}_n$ of V, there is a corresponding dual basis $\mathbf{e}_1^\vee, \mathbf{e}_2^\vee, \dots, \mathbf{e}_n^\vee$. The Projection Principle tells us that Df can thus be thought of as just n functions, so we can write

$$Df = \psi_1 \mathbf{e}_1^\vee + \cdots + \psi_n \mathbf{e}_n^\vee.$$

In fact, we can even describe what the ψ_i are.

Definition 43.3.1. The **i^{th} partial derivative** of $f\colon U \to \mathbb{R}$, denoted

$$\frac{\partial f}{\partial \mathbf{e}_i}\colon U \to \mathbb{R}$$

is defined by

$$\frac{\partial f}{\partial \mathbf{e}_i}(p) \coloneqq \lim_{t \to 0} \frac{f(p + te_i) - f(p)}{t}.$$

You can think of it as "f' along $\mathbf{e}_i$".

Question 43.3.2. Check that if Df exists, then

$$(Df)_p(\mathbf{e}_i) = \frac{\partial f}{\partial \mathbf{e}_i}(p).$$

Remark 43.3.3 — Of course you can write down a definition of $\frac{\partial f}{\partial v}$ for any v (rather than just the $\mathbf{e}_i$).

From the above remarks, we can derive that

$$\boxed{Df = \frac{\partial f}{\partial \mathbf{e}_1} \cdot \mathbf{e}_1^\vee + \cdots + \frac{\partial f}{\partial \mathbf{e}_n} \cdot \mathbf{e}_n^\vee.}$$

and so given a basis of V, we can think of Df as just the n partials.

Remark 43.3.4 — Keep in mind that each $\frac{\partial f}{\partial \mathbf{e}_i}$ is a function from U to the *reals*. That is to say,

$$(Df)_p = \underbrace{\frac{\partial f}{\partial \mathbf{e}_1}(p)}_{\in\mathbb{R}} \cdot \mathbf{e}_1^\vee + \cdots + \underbrace{\frac{\partial f}{\partial \mathbf{e}_n}(p)}_{\in\mathbb{R}} \cdot \mathbf{e}_n^\vee \in V^\vee.$$

Example 43.3.5 (Partial derivatives of $f(x,y) = x^2 + y^2$)
Let $f \colon \mathbb{R}^2 \to \mathbb{R}$ by $(x,y) \mapsto x^2 + y^2$. Then in our new language,

$$Df \colon (x,y) \mapsto 2x \cdot \mathbf{e}_1^\vee + 2y \cdot \mathbf{e}_2^\vee.$$

Thus the partials are

$$\frac{\partial f}{\partial x} \colon (x,y) \mapsto 2x \in \mathbb{R} \quad \text{and} \quad \frac{\partial f}{\partial y} \colon (x,y) \mapsto 2y \in \mathbb{R}$$

With all that said, I haven't really said much about how to find the total derivative itself. For example, if I told you

$$f(x,y) = x \sin y + x^2 y^4$$

you might want to be able to compute Df without going through that horrible limit definition I told you about earlier.

Fortunately, it turns out you already know how to compute partial derivatives, because you had to take AP Calculus at some point in your life. It turns out for most reasonable functions, this is all you'll ever need.

Theorem 43.3.6 (Continuous partials implies differentiable)
Let $U \subseteq V$ be open and pick any basis $\mathbf{e}_1, \dots, \mathbf{e}_n$. Let $f \colon U \to \mathbb{R}$ and suppose that $\frac{\partial f}{\partial \mathbf{e}_i}$ is defined for each i and moreover is *continuous*. Then f is differentiable and Df is given by

$$Df = \sum_{i=1}^{n} \frac{\partial f}{\partial \mathbf{e}_i} \cdot \mathbf{e}_i^\vee.$$

Proof. Not going to write out the details, but... given $v = t_1e_1 + \cdots + t_ne_n$, the idea is to just walk from p to $p+t_1e_1$, $p+t_1e_1+t_2e_2$, ..., up to $p+t_1e_1+t_2e_2+\cdots+t_ne_n = p+v$, picking up the partial derivatives on the way. Do some calculation. □

Remark 43.3.7 — The continuous condition cannot be dropped. The function

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0). \end{cases}$$

is the classic counterexample – the total derivative Df does not exist at zero, even though both partials do.

Example 43.3.8 (Actually computing a total derivative)
Let $f(x,y) = x\sin y + x^2y^4$. Then

$$\begin{aligned} \frac{\partial f}{\partial x}(x,y) &= \sin y + y^4 \cdot 2x \\ \frac{\partial f}{\partial y}(x,y) &= x\cos y + x^2 \cdot 4y^3. \end{aligned}$$

So Theorem 43.3.6 applies, and $Df = \frac{\partial f}{\partial x}\mathbf{e}_1^\vee + \frac{\partial f}{\partial y}\mathbf{e}_2^\vee$, which I won't bother to write out.

The example $f(x,y) = x^2+y^2$ is the same thing. That being said, who cares about $x\sin y + x^2y^4$ anyways?

§43.4 (Optional) A word on higher derivatives

Let $U \subseteq V$ be open, and take $f\colon U \to W$, so that $Df\colon U \to \operatorname{Hom}(V,W)$.

Well, $\operatorname{Hom}(V,W)$ can also be thought of as a normed vector space in its own right: it turns out that one can define an operator norm on it by setting

$$\|T\| \coloneqq \sup\left\{ \frac{\|T(v)\|_W}{\|v\|_V} \mid v \neq 0_V \right\}.$$

So $\operatorname{Hom}(V,W)$ can be thought of as a normed vector space as well. Thus it makes sense to write

$$D(Df)\colon U \to \operatorname{Hom}(V, \operatorname{Hom}(V,W))$$

which we abbreviate as D^2f. Dropping all doubt and plunging on,

$$D^3f\colon U \to \operatorname{Hom}(V, \operatorname{Hom}(V, \operatorname{Hom}(V,W))).$$

I'm sorry. As consolation, we at least know that $\operatorname{Hom}(V,W) \cong V^\vee \otimes W$ in a natural way, so we can at least condense this to

$$D^kf\colon V \to (V^\vee)^{\otimes k} \otimes W$$

rather than writing a bunch of Hom's.

Remark 43.4.1 — If $k = 2$, $W = \mathbb{R}$, then $D^2 f(v) \in (V^\vee)^{\otimes 2}$, so it can be represented as an $n \times n$ matrix, which for some reason is called a **Hessian**.

The most important property of the second derivative is that

Theorem 43.4.2 (Symmetry of $D^2 f$)
Let $f : U \to W$ with $U \subseteq V$ open. If $(D^2 f)_p$ exists at some $p \in U$, then it is symmetric, meaning
$$(D^2 f)_p(v_1, v_2) = (D^2 f)_p(v_2, v_1).$$

I'll just quote this without proof (see e.g. [Pu02, §5, theorem 16]), because double derivatives make my head spin. An important corollary of this theorem:

Corollary 43.4.3 (Clairaut's theorem: mixed partials are symmetric)
Let $f : U \to \mathbb{R}$ with $U \subseteq V$ open be twice differentiable. Then for any point p such that the quantities are defined,
$$\frac{\partial}{\partial \mathbf{e}_i}\frac{\partial}{\partial \mathbf{e}_j} f(p) = \frac{\partial}{\partial \mathbf{e}_j}\frac{\partial}{\partial \mathbf{e}_i} f(p).$$

§43.5 Towards differential forms

This concludes the exposition of what the derivative really is: the key idea I want to communicate in this chapter is that Df should be thought of as a map from $U \to V^\vee$.

The next natural thing to do is talk about *integration*. The correct way to do this is through a so-called *differential form*: you'll finally know what all those stupid dx's and dy's really mean. (They weren't just there for decoration!)

§43.6 A few harder problems to think about

Problem 43A* (Chain rule)**.** Let $U_1 \xrightarrow{f} U_2 \xrightarrow{g} U_3$ be differentiable maps between open sets of normed vector spaces V_i, and let $h = g \circ f$. Prove the Chain Rule: for any point $p \in U_1$, we have
$$(Dh)_p = (Dg)_{f(p)} \circ (Df)_p.$$

Problem 43B. Let $U \subseteq V$ be open, and $f : U \to \mathbb{R}$ be differentiable k times. Show that $(D^k f)_p$ is symmetric in its k arguments, meaning for any $v_1, \dots, v_k \in V$ and any permutation σ on $\{1, \dots, k\}$ we have
$$(D^k f)_p(v_1, \dots, v_k) = (D^k f)_p(v_{\sigma(1)}, \dots, v_{\sigma(k)}).$$

44 Differential forms

In this chapter, all vector spaces are finite-dimensional real inner product spaces. We first start by (non-rigorously) drawing pictures of all the things that we will define in this chapter. Then we re-do everything again in its proper algebraic context.

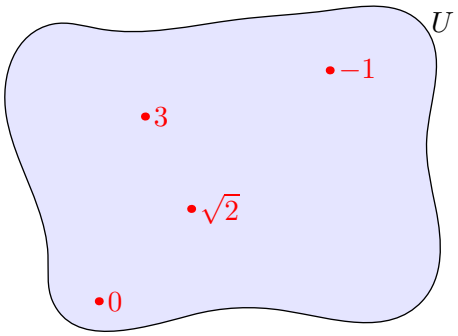
§44.1 Pictures of differential forms

Before defining a differential form, we first draw some pictures. The key thing to keep in mind is

> **"The definition of a differential form is: something you can integrate."**
> **— Joe Harris**

We'll assume that all functions are **smooth**, i.e. infinitely differentiable.

Let $U \subseteq V$ be an open set of a vector space V. Suppose that we have a function $f : U \to \mathbb{R}$, i.e. we assign a value to every point of U.



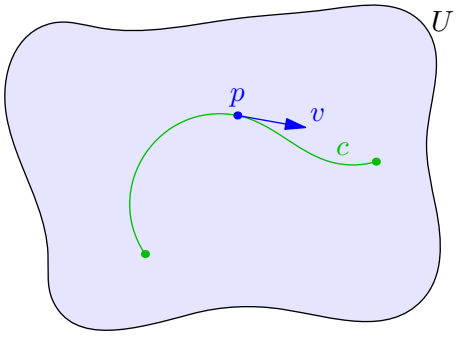
Definition 44.1.1. A **0-form** f on U is just a smooth function $f : U \to \mathbb{R}$.

Thus, if we specify a finite set S of points in U we can "integrate" over S by just adding up the values of the points:

$$0 + \sqrt{2} + 3 + (-1) = 2 + \sqrt{2}.$$

So, **a 0-form f lets us integrate over 0-dimensional "cells"**.

But this is quite boring, because as we know we like to integrate over things like curves, not single points. So, by analogy, we want a 1-form to let us integrate over 1-dimensional cells: i.e. over curves. What information would we need to do that? To answer this, let's draw a picture of a curve c, which can be thought of as a function $c : [0, 1] \to U$.



We might think that we could get away with just specifying a number on every point of U (i.e. a 0-form f), and then somehow "add up" all the values of f along the curve. We'll use this idea in a moment, but we can in fact do something more general. Notice how when we walk along a smooth curve, at every point p we also have some extra information: a *tangent vector* v. So, we can define a 1-form α as follows. A 0-form just took a point and gave a real number, but **a 1-form will take both a point *and* a tangent vector at that point, and spit out a real number.** So a 1-form α is a smooth function on pairs (p, v), where v is a tangent vector at p, to $\mathbb{R}$. Hence

$$\alpha \colon U \times V \to \mathbb{R}.$$

Actually, for any point p, we will require that $\alpha(p, -)$ is a linear function in terms of the vectors: i.e. we want for example that $\alpha(p, 2v) = 2\alpha(p, v)$. So it is more customary to think of α as:

Definition 44.1.2. A 1-form α is a smooth function

$$\alpha \colon U \to V^\vee.$$

Like with Df, we'll use α_p instead of $\alpha(p)$. So, at every point p, α_p is some linear functional that eats tangent vectors at p, and spits out a real number. Thus, we think of α_p as an element of $V^\vee$;

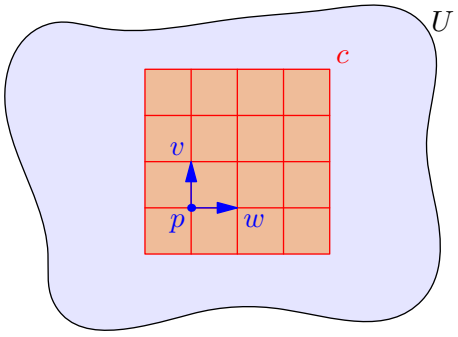
$$\alpha_p \in V^\vee.$$

Remark 44.1.3 (Warning: arc length isn't a 1-form) — You might recall that, in high school calculus, the "arc-length element" ds can be integrated along a curve: $\int_c ds$ is the length of the curve c.
This is *not* a 1-form! More on this later. (To be brief: basically, the issue is that it's not a linear function. In some places you'll see $ds = \sqrt{dx^2 + dy^2}$ written colloquially, which should give you a sense that ds does not behave like a linear thing in dx and dy.)

Next, we draw pictures of 2-forms. This should, for example, let us integrate over a blob (a so-called 2-cell) of the form

$$c \colon [0,1] \times [0,1] \to U$$

i.e. for example, a square in U. In the previous example with 1-forms, we looked at tangent vectors to the curve c. This time, at points we will look at *pairs* of tangent vectors in U: in the same sense that lots of tangent vectors approximate the entire curve, lots of tiny squares will approximate the big square in U.



So what should a 2-form β be? As before, it should start by taking a point $p \in U$, so β_p is now a linear functional: but this time, it should be a linear map on two vectors v and w. Here v and w are not tangent so much as their span cuts out a small parallelogram. So, the right thing to do is in fact consider

$$\beta_p \in V^\vee \wedge V^\vee.$$

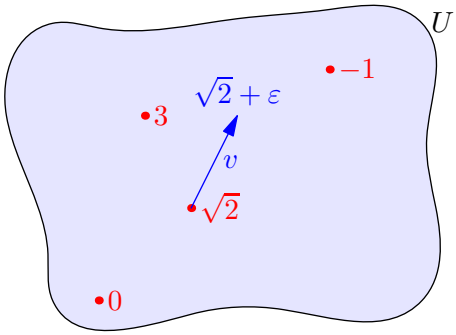
That is, to use the wedge product to get a handle on the idea that v and w span a parallelogram. Another valid choice would have been $(V \wedge V)^\vee$; in fact, the two are isomorphic,[1] but it will be more convenient to write it in the former.[2]

§44.2 Pictures of exterior derivatives

Next question:

How can we build a 1-form from a 0-form?

Let f be a 0-form on U; thus, we have a function $f : U \to \mathbb{R}$. Then in fact there is a very natural 1-form on U arising from f, appropriately called df. Namely, given a point p and a tangent vector v, the differential form $(df)_p$ returns the *change in f along v*. In other words, it's just the total derivative $(Df)_p(v)$.



Thus, df measures "the change in f".

Now, even if I haven't defined integration yet, given a curve c from a point a to b, what do you think

$$\int_c df$$

should be equal to? Remember that df is the 1-form that measures "infinitesimal change in f". So if we add up all the change in f along a path from a to b, then the answer we get should just be

$$\int_c df = f(b) - f(a).$$

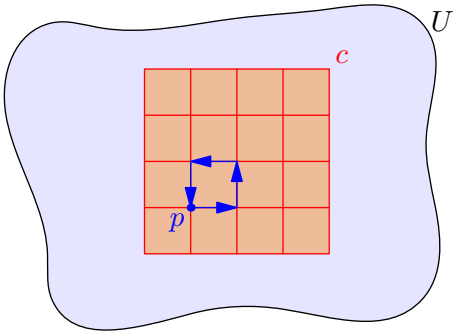
This is the first case of something we call Stokes' theorem.

Generalizing, how should we get from a 1-form to a 2-form? At each point p, the 2-form β gives a β_p which takes in a "parallelogram" and returns a real number. Now suppose we have a 1-form α. Then along each of the edges of a parallelogram, with an appropriate sign convention the 1-form α gives us a real number. So, given a 1-form α, we define $d\alpha$ to be the 2-form that takes in a parallelogram spanned by v and w, and returns **the measure of α along the boundary**.

[1] We only consider finite-dimensional V.

[2] See Section 44.5.

Now, what happens if you integrate $d\alpha$ along the entire square c? The right picture is that, if we think of each little square as making up the big square, then the adjacent boundaries cancel out, and all we are left is the main boundary. This is again just a case of the so-called Stokes' theorem.



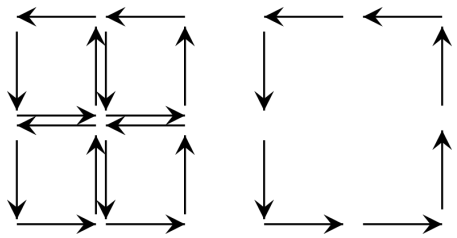


Image from [Na]

§44.3 Differential forms

Prototypical example for this section: Algebraically, something that looks like $f\mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee + \dots$, and geometrically, see the previous section.

Let's now get a handle on what dx means. Fix a real vector space V of dimension n, and let $\mathbf{e}_1, \dots, \mathbf{e}_n$ be a standard basis. Let U be an open set.

Definition 44.3.1. We define a **differential k-form** α on U to be a smooth (infinitely differentiable) map $\alpha \colon U \to \bigwedge^k(V^\vee)$. (Here $\bigwedge^k(V^\vee)$ is the wedge product.)

Like with Df, we'll use α_p instead of $\alpha(p)$.

Example 44.3.2 (k-forms for $k = 0, 1$)

(a) A 0-form is just a function $U \to \mathbb{R}$.

(b) A 1-form is a function $U \to V^\vee$. For example, the total derivative Df of a function $V \to \mathbb{R}$ is a 1-form.

(c) Let $V = \mathbb{R}^3$ with standard basis $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$. Then a typical 2-form is given by

$$\alpha_p = f(p) \cdot \mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee + g(p) \cdot \mathbf{e}_1^\vee \wedge \mathbf{e}_3^\vee + h(p) \cdot \mathbf{e}_2^\vee \wedge \mathbf{e}_3^\vee \in \bigwedge^2(V^\vee)$$

where $f, g, h \colon V \to \mathbb{R}$ are smooth functions.

Now, by the projection principle (Theorem 43.2.1) we only have to specify a function on each of $\binom{n}{k}$ basis elements of $\bigwedge^k(V^\vee)$. So, take any basis $\{\mathbf{e}_i\}$ of V, and take the usual basis for $\bigwedge^k(V^\vee)$ of elements

$$\mathbf{e}_{i_1}^\vee \wedge \mathbf{e}_{i_2}^\vee \wedge \dots \wedge \mathbf{e}_{i_k}^\vee.$$

Thus, a general k-form takes the shape

$$\alpha_p = \sum_{1 \le i_1 < \dots < i_k \le n} f_{i_1,\dots,i_k}(p) \cdot \mathbf{e}_{i_1}^\vee \wedge \mathbf{e}_{i_2}^\vee \wedge \dots \wedge \mathbf{e}_{i_k}^\vee.$$

Since this is a huge nuisance to write, we will abbreviate this to just

$$\alpha = \sum_I f_I \cdot d\mathbf{e}_I$$

where we understand the sum runs over $I = (i_1, \dots, i_k)$, and $d\mathbf{e}_I$ represents $\mathbf{e}_{i_1}^\vee \wedge \cdots \wedge \mathbf{e}_{i_k}^\vee$.

Now that we have an element $\bigwedge^k(V^\vee)$, what can it do? Well, first let me get the definition on the table, then tell you what it's doing.

Definition 44.3.3 (How to evaluate a differential form at a point)**.** For linear functions $\xi_1, \dots, \xi_k \in V^\vee$ and vectors $v_1, \dots, v_k \in V$, set

$$(\xi_1 \wedge \cdots \wedge \xi_k)(v_1, \dots, v_k) \coloneqq \det \begin{bmatrix} \xi_1(v_1) & \dots & \xi_1(v_k) \\ \vdots & \ddots & \vdots \\ \xi_k(v_1) & \dots & \xi_k(v_k) \end{bmatrix}.$$

You can check that this is well-defined under e.g. $v \wedge w = -w \wedge v$ and so on.

Example 44.3.4 (Evaluation of a differential form)
Set $V = \mathbb{R}^3$. Suppose that at some point p, the 2-form α returns

$$\alpha_p = 2\mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee + \mathbf{e}_1^\vee \wedge \mathbf{e}_3^\vee.$$

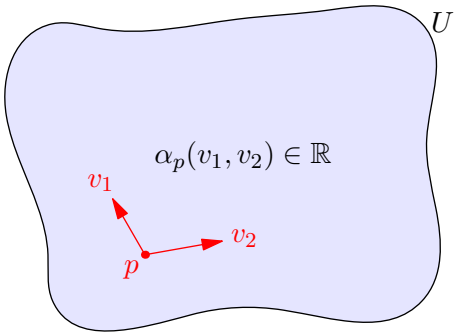
Let $v_1 = 3\mathbf{e}_1 + \mathbf{e}_2 + 4\mathbf{e}_3$ and $v_2 = 8\mathbf{e}_1 + 9\mathbf{e}_2 + 5\mathbf{e}_3$. Then

$$\alpha_p(v_1, v_2) = 2 \det \begin{bmatrix} 3 & 8 \\ 1 & 9 \end{bmatrix} + \det \begin{bmatrix} 3 & 8 \\ 4 & 5 \end{bmatrix} = 21.$$

What does this definition mean? One way to say it is that

If I walk to a point $p \in U$, a k-form α will take in k vectors $v_1, \dots, v_k$ and spit out a number, which is to be interpreted as a (signed) volume.

Picture:



In other words, at every point p, we get a function α_p. Then I can feed in k vectors to α_p and get a number, which I interpret as a signed volume of the parallelepiped spanned by the $\{v_i\}$'s in some way (e.g. the flux of a force field). That's why α_p as a "function" is contrived to lie in the wedge product: this ensures that the notion of "volume" makes sense, so that for example, the equality $\alpha_p(v_1, v_2) = -\alpha_p(v_2, v_1)$ holds.

This is what makes differential forms so fit for integration.

§44.4 Exterior derivatives

Prototypical example for this section: Possibly $(dx_1)_p = \mathbf{e}_1^\vee$.

We now define the exterior derivative[3] df that we gave pictures of at the beginning of the chapter. It turns out that the exterior derivative is easy to compute given explicit coordinates to work with.

Firstly, we define the exterior derivative of a function $f : U \to \mathbb{R}$, as

$$df \coloneqq Df = \sum_i \frac{\partial f}{\partial \mathbf{e}_i} \mathbf{e}_i^\vee$$

In particular, suppose $V = \mathbb{R}^n$ and $f(x_1, \dots, x_n) = x_1$ (i.e. $f = \mathbf{e}_1^\vee$). Then:

> **Question 44.4.1.** Show that for any $p \in U$,
>
> $$(d(\mathbf{e}_1^\vee))_p = \mathbf{e}_1^\vee.$$

Abuse of Notation 44.4.2. Unfortunately, someone somewhere decided it would be a good idea to use "x_1" to denote $\mathbf{e}_1^\vee$ (because *obviously*[4] x_1 means "the function that takes $(x_1, \dots, x_n) \in \mathbb{R}^n$ to x_1") and then decided that

$$dx_1 \coloneqq d(\mathbf{e}_1^\vee).$$

This notation is so entrenched that I have no choice but to grudgingly accept it. Note that it's not even right, since technically it's $(dx_1)_p = \mathbf{e}_1^\vee$; dx_1 is a 1-form.

> **Remark 44.4.3 —** This is the reason why we use the notation $\frac{df}{dx}$ in calculus now: given, say, $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$, it is indeed true that
>
> $$df = 2x \cdot \mathbf{e}_1^\vee = 2x \cdot dx$$
>
> and so by (more) abuse of notation we write $df/dx = 2x$.

More generally, we can define the exterior derivative in terms of our basis $\mathbf{e}_1, \dots, \mathbf{e}_n$ as follows:

Definition 44.4.4. If $\alpha = \sum_I f_I d\mathbf{e}_I$ then we define the **exterior derivative** as

$$d\alpha \coloneqq \sum_I df_I \wedge d\mathbf{e}_I = \sum_I \sum_j \frac{\partial f_I}{\partial \mathbf{e}_j} d\mathbf{e}_j \wedge d\mathbf{e}_I.$$

It turns out this doesn't depend on the choice of basis; we'll mention a basis-free definition at the end of this section.

> **Example 44.4.5** (Computing some exterior derivatives)
> Let $V = \mathbb{R}^3$ with standard basis $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$. Let $f(x, y, z) = x^4 + y^3 + 2xz$. Then we compute
>
> $$df = Df = (4x^3 + 2z)\, dx + 3y^2\, dy + 2x\, dz.$$

[3] The name "exterior derivative" comes from the wedge product $\wedge$ being alternatively called the exterior product.

[4] Sarcasm.

Next, we can evaluate $d(df)$ as prescribed: it is

$$\begin{aligned} d^2 f &= (12x^2\,dx + 2dz) \wedge dx + (6y\,dy) \wedge dy + 2(dx \wedge dz) \\ &= 12x^2(dx \wedge dx) + 2(dz \wedge dx) + 6y(dy \wedge dy) + 2(dx \wedge dz) \\ &= 2(dz \wedge dx) + 2(dx \wedge dz) \\ &= 0. \end{aligned}$$

So surprisingly, $d^2 f$ is the zero map. Here, we have exploited Abuse of Notation 44.4.2 for the first time, in writing dx, dy, dz.

And in fact, this is always true in general:

Theorem 44.4.6 (Exterior derivative vanishes)
Let α be any k-form. Then $d^2(\alpha) = 0$. Even more succinctly,

$$d^2 = 0.$$

The proof is left as Problem 44B.

Exercise 44.4.7. Compare the statement $d^2 = 0$ to the geometric picture of a 2-form given at the beginning of this chapter. Why does this intuitively make sense?

Here are some other properties of d:

- As we just saw, $d^2 = 0$.
- For a k-form α and ℓ-form β, one can show that
$$d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^k(\alpha \wedge d\beta).$$
- If $f \colon U \to \mathbb{R}$ is smooth, then $df = Df$.

In fact, one can show that df as defined above is the *unique* map sending k-forms to $(k+1)$-forms with these properties. So, one way to *define* df is to take as axioms the bulleted properties above and then declare d to be the unique solution to this functional equation. In any case, this tells us that our definition of d does not depend on the basis chosen.

Recall that df measures the change in boundary. In that sense, $d^2 = 0$ is saying something like "the boundary of the boundary is empty". We'll make this precise when we see Stokes' theorem in the next chapter.

§44.5 Digression: $\bigwedge^k(V^\vee)$ versus $\left(\bigwedge^k(V)\right)^\vee$

Earlier on, we remarked that $\bigwedge^k(V^\vee) \cong \left(\bigwedge^k(V)\right)^\vee$ canonically, but we use the former for convenience.

The former notation is indeed more convenient (wedge product of two differential form is natural), but it's not clear why Definition 44.3.3 is defined in such a way.

If we used $\left(\bigwedge^k(V)\right)^\vee$ instead, it's trivial to evaluate a differential form: For $f \in \left(\bigwedge^k(V)\right)^\vee$ and vectors $v_1, \dots, v_k \in V$, then

$$f(v_1, \dots, v_k) \coloneqq f(v_1 \wedge \cdots \wedge v_k).$$

This is because f naturally takes in an element of $\bigwedge^k(V)$ and returns a real number.

But now, it is not clear how we can take $f \in (\bigwedge^1(V))^\vee$ and $g \in (\bigwedge^1(V))$, and return something like $f \wedge g \in (\bigwedge^2(V))^\vee$: The natural choice $(v \wedge w \mapsto f(v)g(w))$ isn't even well-defined![5]

To figure out what to do, we have to take a step back and consider the tensor product. For a vector space V, define $T^k(V) = \underbrace{V \otimes V \otimes \cdots \otimes V}_{k \text{ times}}$.

We have the following diagram:

$$\begin{array}{ccc} T^k(V^\vee) & \xhookrightarrow{\phi} & (T^k(V))^\vee \\ \iota \uparrow \downarrow q & & q^\vee \uparrow \\ \bigwedge^k(V^\vee) & & (\bigwedge^k(V))^\vee \end{array}$$

What is going on?

First, there is a natural map $T^k(V^\vee) \to (T^k(V))^\vee$ given by

$$\phi(\xi_1 \otimes \cdots \otimes \xi_k) = v_1 \otimes \cdots \otimes v_k \mapsto \xi_1(v_1)\xi_2(v_2)\cdots\xi_k(v_k)$$

and extends to all of $T^k(V^\vee)$ by linearity the obvious way.

Unlike the situation with the wedge product above, this map is indeed well-defined.

With some manual work, we can check ϕ is injective. Because both $T^k(V^\vee)$ and $(T^k(V))^\vee$ has dimension $(\dim V)^k$, ϕ is bijective.

Next, note that $\bigwedge^k(V)$ is just "$T^k(V)$ but with more relations imposed", there is a natural quotient map $q\colon T^k(V) \twoheadrightarrow \bigwedge^k(V)$. So, the tensors are divided into equivalence classes.

Example 44.5.1

If $V = \mathbb{R}^2$, then $T^2(V)$ would have elements such as $\mathbf{e}_1 \otimes \mathbf{e}_1$, $\mathbf{e}_1 \otimes \mathbf{e}_2$ or $-\mathbf{e}_2 \otimes \mathbf{e}_1$. Mapping these elements to $\bigwedge^2(V)$, we get $\mathbf{e}_1 \wedge \mathbf{e}_1 = 0$, and $\mathbf{e}_1 \wedge \mathbf{e}_2 = -\mathbf{e}_2 \wedge \mathbf{e}_1$, i.e. $\mathbf{e}_1 \otimes \mathbf{e}_2$ and $-\mathbf{e}_2 \otimes \mathbf{e}_1$ belongs to the same equivalence class.

The map q induces a dual map $q^\vee\colon (\bigwedge^k(V))^\vee \to (T^k(V))^\vee$.

Question 44.5.2. Convince yourself that a function $f \in (T^k(V))^\vee$ belongs to $\operatorname{im} q^\vee$ if and only if f assigns the same value for every element in each equivalence class, as defined above.

Thus, we get an isomorphism $q \circ \phi^{-1} \circ q^\vee\colon (\bigwedge^k(V))^\vee \to \bigwedge^k(V^\vee)$.[6]

To check this is indeed an isomorphism, we will construct its inverse map. As defined above, each equivalence class in $T^k(V^\vee)$ (fiber of $g \in \bigwedge^k(V^\vee)$) has multiple elements, but we can find a canonical element by the following:

Definition 44.5.3. For vector space V, and element $f = f_1 \otimes f_2 \otimes \cdots \otimes f_k \in T^k(V)$, we define the alternation of f as follows:

$$\operatorname{Alt} f = \frac{1}{k!} \sum_{\sigma \in S_k} \operatorname{sgn}(\sigma) \cdot (f_{\sigma(1)} \otimes f_{\sigma(2)} \otimes \cdots \otimes f_{\sigma(k)})$$

and extend it to all of $T^k(V)$.

[5] You can try it with $f = \mathbf{e}_1^\vee$ and $g = \mathbf{e}_2^\vee$, evaluate it at $\mathbf{e}_1 \wedge \mathbf{e}_2$ and $-\mathbf{e}_2 \wedge \mathbf{e}_1$, which we know is equal.

[6] We're using q for both the map $T^k(V) \to \bigwedge^k(V)$ and $T^k(V^\vee) \to \bigwedge^k(V^\vee)$, by abuse of notation.

Here, S_k is the permutation group. Notice the similarity with the definition of the determinant.

Example 44.5.4

As above, $V = \mathbb{R}^2$. Then we get:

$$\operatorname{Alt}(\mathbf{e}_1 \otimes \mathbf{e}_2) = \frac{\mathbf{e}_1 \otimes \mathbf{e}_2 - \mathbf{e}_2 \otimes \mathbf{e}_1}{2}.$$

Notice that if we swap the first and second component of $\mathbf{e}_1 \otimes \mathbf{e}_2$, we get $\mathbf{e}_2 \otimes \mathbf{e}_1$ which has little to do with the original tensor. However, if we swap the first and second component of $\frac{\mathbf{e}_1 \otimes \mathbf{e}_2 - \mathbf{e}_2 \otimes \mathbf{e}_1}{2}$, we get $\frac{\mathbf{e}_2 \otimes \mathbf{e}_1 - \mathbf{e}_1 \otimes \mathbf{e}_2}{2}$, which is exactly the negation of the original tensor!

We see that $\operatorname{Alt} f$ is a desirable representative of the equivalence class of f because:

- $\operatorname{Alt}(\operatorname{Alt} f) = \operatorname{Alt} f$;
- $q(f) = q(\operatorname{Alt} f)$ where q is the quotient map $T^k(V) \twoheadrightarrow \bigwedge^k(V)$;
- $\operatorname{Alt} f$ is an *alternating tensor* — that is, if we swap two adjacent components of $\operatorname{Alt} f$ for each pure tensor, then the whole tensor gets negated.

Thus it makes sense for us to define $\iota\colon \bigwedge^k(V^\vee) \hookrightarrow T^k(V^\vee)$ that takes each element to the alternating tensor in $T^k(V^\vee)$.

Example 44.5.5

With the same example as above, $V = \mathbb{R}^2$, then we get

$$\iota(\mathbf{e}_1 \wedge \mathbf{e}_2) = \operatorname{Alt}(\mathbf{e}_1 \otimes \mathbf{e}_2) = \frac{\mathbf{e}_2 \otimes \mathbf{e}_1 - \mathbf{e}_1 \otimes \mathbf{e}_2}{2}.$$

Finally,

Exercise 44.5.6. Show that $\operatorname{im}(\phi \circ \iota) = \operatorname{im} q^\vee$, and that $\iota^\vee \circ \phi \circ \iota$ and $q \circ \phi^{-1} \circ q^\vee$ are inverses of each other.

It is common notation that we want to define the wedge product such that $\mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee$ takes in $\mathbf{e}_1 \wedge \mathbf{e}_2$ (that is, the square formed by $\mathbf{e}_1$ and $\mathbf{e}_2$), and returns 1. However, if we define the wedge product naturally by the method above, we get

$$\iota(\mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee) = \frac{\mathbf{e}_1^\vee \otimes \mathbf{e}_2^\vee - \mathbf{e}_2^\vee \otimes \mathbf{e}_1^\vee}{2}$$

which means

$$\phi(\iota(\mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee))(\mathbf{e}_1, \mathbf{e}_2) = \frac{1 \cdot 1 - 0 \cdot 0}{2} = \frac{1}{2}.$$

So, a corrective factor $k!$ is needed.

To see how "difficult" the wedge product will be if we use the second notation, let $V = \mathbb{R}^3$, $\alpha = dx \wedge dy \in \bigwedge^2(V^\vee)$, and $\beta = dz \in \bigwedge^1(V^\vee)$.

Then, we know:

- $\alpha(\mathbf{e}_1 \wedge \mathbf{e}_2) = 1$.

- $\beta(\mathbf{e}_3) = 1$.
- We should have $(\alpha \wedge \beta)(\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3) = 1$.

The last point is obvious if we let the wedge product be the map $\wedge\colon \bigwedge^2(V^\vee) \times \bigwedge^1(V^\vee) \to \bigwedge^3(V^\vee)$.

However, if we're given α and β as elements of $(\bigwedge^2(V))^\vee$ and $(\bigwedge^1(V))^\vee$ respectively (that is, we can only evaluate α at $v \wedge w$ for $v, w \in V$; and we can only evaluate β at v for $v \in V$), then it would be much more difficult to write down what $\alpha \wedge \beta$ should be. In fact,

$$(\alpha \wedge \beta)(v_1 \wedge v_2 \wedge v_3) = \alpha(v_1 \wedge v_2)\beta(v_3) - \alpha(v_1 \wedge v_3)\beta(v_2) + \alpha(v_2 \wedge v_3)\beta(v_1).$$

You can see that this is a variant of the alternation operator (or the evaluation operation), where we compute a weighted average in order to force $\alpha \wedge \beta$ to be alternating.

§44.6 Tangential remark: Arc length ds is not a 1-form

As mentioned in a remark earlier, the arc length ds is not a 1-form.[7]

We said earlier that differential form is something you can integrate. You can certainly integrate ds, but it's not considered a 1-form!

While we can easily check against the definition that ds is not linear (Problem 45E), it still raises the question that why we would want to define differential form to exclude ds. What's going on here?

In fact, the true story is that the objects that are integrable over a smooth curve are **1-densities**. We will define this later.

For simplicity, we work over $\mathbb{R}^2$ in this section. Given a (smooth) 1-density ω that can be integrated over a smooth curve c, we would like the integral $\int_c \omega$ to have the following properties:

- It is additive: if c is the concatenation of two curves c_1 and c_2, then $\int_c \omega = \int_{c_1} \omega + \int_{c_2} \omega$.
- Because everything is smooth, we would expect that if c is a tiny line segment, then in fact $\int_{c_1} \omega \approx \int_{c_2} \omega$ if we divide c into two segments c_1 and c_2 of equal length.

  Thus, it's natural to require $\int_c \omega$ to be "approximately linear" when the length of c is small enough.

  In symbols: for $\varepsilon > 0$, let c_ε be the initial segment of the curve c with length ε, then
  $$\lim_{\varepsilon \to 0^+} \frac{\int_{c_\varepsilon} \omega}{\varepsilon} = h$$
  for some finite constant h.

We certainly can formalize a 1-density ω to be simply a function that takes smooth curves c and returns the value $\int_c \omega$ satisfying the two conditions above, but this definition is clunky.

A better way to do it is to observe that, if we know $\int_c \omega$ for tiny curves c, then we can integrate ω over any smooth curves c by chopping it up into tiny curves. But this isn't completely formal — of course, as the length of a curve tends to 0, the integral $\int_c \omega$ also tends to 0 — so instead, we consider the limit above:

$$\lim_{\varepsilon \to 0^+} \frac{\int_{c_\varepsilon} \omega}{\varepsilon}.$$

[7] https://mathoverflow.net/q/90455 has a discussion on this.

Question 44.6.1. Convince yourself that, given two curves $c\colon [0,1] \to \mathbb{R}^2$ and $c_2\colon [0,1] \to \mathbb{R}^2$ that starts at the same point $c(0) = c_2(0) = p$, and moves in the same direction $c'(0) = c_2'(0) = v$, then basic smoothness condition of $\int_c \omega$ would guarantee that the limit above is the same.

Thus,

We can define a 1-density ω to be a function that takes in a point p and the initial direction $v \in \mathbb{R}^2$, which is understood as a tangent vector of $\mathbb{R}^2$ at p, and returns the limit.

Formally:

Definition 44.6.2 (1-density)**.** A 1-density ω is a function $\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$.

We write $\omega_p(v) \in \mathbb{R}$.

Since only the direction matters, it makes sense to make ω satisfy $\omega_p(\lambda v) = \lambda \omega_p(v)$ for $\lambda \geq 0$. In particular, $\omega_p(0) = 0$.

Then, ds is the differential form $ds_p(v) = \|v\|$. While we have not rigorously defined how to integrate over a curve (we will do this next chapter), you can intuitively see how it works.

With this definition, a 1-form is just a 1-density that is in addition linear in the second argument — $\omega_p(v + w) = \omega_p(v) + \omega_p(w)$.

So, what is the special properties that differential forms enjoys? For one, if ω is a differential form, we have:

> Let $c\colon [0,1] \to \mathbb{R}^2$ be a smooth curve, then for any sequence of smooth curves c_k that converges uniformly to c, then $\int_{c_k} \omega$ converges to $\int_c \omega$.

You can easily imagine how this can fail for ds — a sequence of piecewise smooth curves that consist of only horizontal and vertical lines can approximate a circle, but the arc length of these jagged curves can never converge to the circumference of the circle.[8]

§44.7 Closed and exact forms

Let α be a k-form.

Definition 44.7.1. We say α is **closed** if $d\alpha = 0$.

Definition 44.7.2. We say α is **exact** if for some $(k-1)$-form β, $d\beta = \alpha$. If $k = 0$, α is exact only when $\alpha = 0$.

Question 44.7.3. Show that exact forms are closed.

A natural question arises: are there closed forms which are not exact? Surprisingly, the answer to this question is tied to topology. Here is one important example.

[8] In fact, using the same argument, you can also prove that, conversely, any smooth density that satisfies the latter property must in fact be linear!

Example 44.7.4 (The angle form)
Let $U = \mathbb{R}^2 \setminus \{0\}$, and let $\theta(p)$ be the angle formed by the x-axis and the line from the origin to p.
The 1-form $\alpha \colon U \to (\mathbb{R}^2)^\vee$ defined by

$$\alpha = \frac{-y\,dx + x\,dy}{x^2+y^2}$$

is called the **angle form**: given $p \in U$ it measures the change in angle $\theta(p)$ along a tangent vector. So intuitively, "$\alpha = d\theta$". Indeed, one can check directly that the angle form is closed.
However, α is not exact: there is no global smooth function $\theta \colon U \to \mathbb{R}$ having α as a derivative. This reflects the fact that one can actually perform a full 2π rotation around the origin, i.e. θ only makes sense mod 2π. Thus existence of the angle form α reflects the possibility of "winding" around the origin.

So the key idea is that the failure of a closed form to be exact corresponds quite well with "holes" in the space: the same information that homotopy and homology groups are trying to capture. To draw another analogy, in complex analysis Cauchy-Goursat only works when U is simply connected. The "hole" in U is being detected by the existence of a form α. The so-called de Rham cohomology will make this relation explicit.

§44.8 A few harder problems to think about

Problem 44A. Show directly that the angle form

$$\alpha = \frac{-y\,dx + x\,dy}{x^2+y^2}$$

is closed.

Problem 44B. Establish Theorem 44.4.6, which states that $d^2 = 0$.

45 Integrating differential forms

We now show how to integrate differential forms over cells, and state Stokes' theorem in this context. In this chapter, all vector spaces are finite-dimensional and real.

§45.1 Motivation: line integrals

Given a function $g\colon [a,b] \to \mathbb{R}$, we know by the fundamental theorem of calculus that

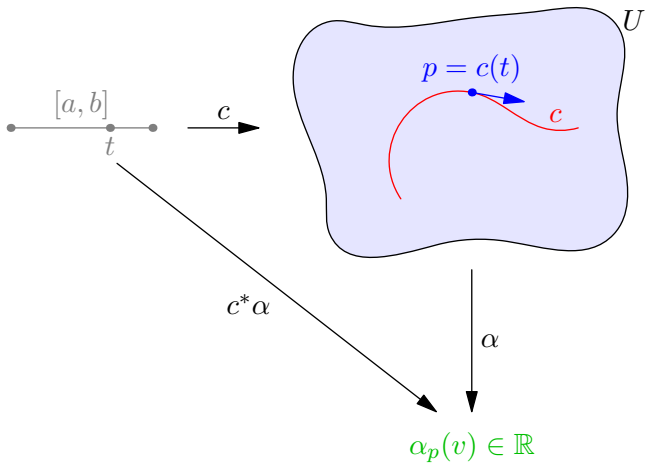
$$\int_{[a,b]} g(t)\,dt = f(b) - f(a)$$

where f is a function such that $g = df/dt$. Equivalently, for $f\colon [a,b] \to \mathbb{R}$,

$$\int_{[a,b]} g\,dt = \int_{[a,b]} df = f(b) - f(a)$$

where df is the exterior derivative we defined earlier.

Cool, so we can integrate over $[a,b]$. Now suppose more generally, we have U an open subset of our real vector space V and a 1-form $\alpha\colon U \to V^\vee$. We consider a **parametrized curve**, which is a smooth function $c\colon [a,b] \to U$. Picture:



We want to define an $\int_c \alpha$ such that:

The integral $\int_c \alpha$ should add up all the α along the curve c.

Our differential form α first takes in a point p to get $\alpha_p \in V^\vee$. Then, it eats a tangent vector $v \in V$ to the curve c to finally give a real number $\alpha_p(v) \in \mathbb{R}$. We would like to "add all these numbers up", using only the notion of an integral over $[a,b]$.

Exercise 45.1.1. Try to guess what the definition of the integral should be. (By type-checking, there's only one reasonable answer.)

So, the definition we give is

$$\int_c \alpha \coloneqq \int_{[a,b]} \alpha_{c(t)} \left(c'(t) \right) \, dt.$$

Here, $c'(t)$ is shorthand for $(Dc)_t(1)$. It represents the *tangent vector* to the curve c at the point $p = c(t)$, at time t. (Here we are taking advantage of the fact that $[a,b]$ is one-dimensional.)

Now that definition was a pain to write, so we will define a differential 1-form $c^*\alpha$ on $[a,b]$ to swallow that entire thing: specifically, in this case we define $c^*\alpha$ to be

$$\left(c^*\alpha\right)_t (\varepsilon) = \alpha_{c(t)} \cdot (Dc)_t(\varepsilon)$$

(here ε is some displacement in time). Thus, we can more succinctly write

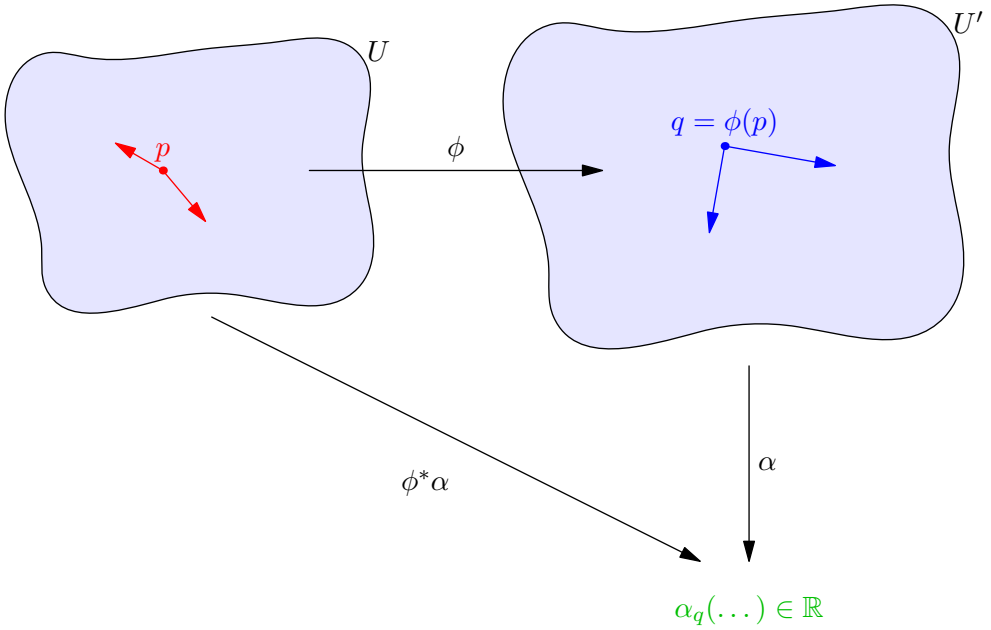
$$\int_c \alpha \coloneqq \int_{[a,b]} c^*\alpha.$$

This is a special case of a *pullback*: roughly, if $\phi \colon U \to U'$ (where $U \subseteq V$, $U' \subseteq V'$), we can change any differential k-form α on U' to a k-form on U. In particular, if $U = [a,b]$,[1] we can resort to our old definition of an integral. Let's now do this in full generality.

§45.2 Pullbacks

Let V and V' be finite dimensional real vector spaces (possibly different dimensions) and suppose U and U' are open subsets of each; next, consider a k-form α on U'.

Given a map $\phi \colon U \to U'$ we now want to define a pullback in much the same way as before. Picture:



Well, there's a total of about one thing we can do. Specifically: α accepts a point in U' and k tangent vectors in V', and returns a real number. We want $\phi^*\alpha$ to accept a point in $p \in U$ and k tangent vectors $v_1, \dots, v_k$ in V, and feed the corresponding information to α.

[1] OK, so $[a,b]$ isn't actually open, sorry. I ought to write $(a-\varepsilon, b+\varepsilon)$, or something.

Clearly we give the point $q = \phi(p)$. As for the tangent vectors, since we are interested in volume, we take the derivative of ϕ at p, $(D\phi)_p$, which will scale each of our vectors v_i into some vector in the target V'. To cut a long story short:

Definition 45.2.1. Given $\phi\colon U \to U'$ and α a k-form, we define the **pullback**

$$(\phi^*\alpha)_p(v_1, \dots, v_k) \coloneqq \alpha_{\phi(p)}\left((D\phi)_p(v_1), \dots, (D\phi)_p(v_k)\right).$$

There is a more concrete way to define the pullback using bases. Suppose $w_1, \dots, w_n$ is a basis of V' and $e_1, \dots, e_m$ is a basis of V. Thus, by the projection principle (Theorem 43.2.1) the map $\phi\colon V \to V'$ can be thought of as

$$\phi(v) = \phi_1(v)w_1 + \cdots + \phi_n(v)w_n$$

where each ϕ_i takes in a $v \in V$ and returns a real number. We know also that α can be written concretely as

$$\alpha = \sum_{I \subseteq \{1,\dots,n\}} f_J \, dw_J.$$

Then, we define

$$\phi^*\alpha = \sum_{I \subseteq \{1,\dots,n\}} (f_I \circ \phi)(D\phi_{i_1} \wedge \cdots \wedge D\phi_{i_k}).$$

A diligent reader can check these definitions are equivalent.

Example 45.2.2 (Computation of a pullback)
Let $V = \mathbb{R}^2$ with basis $\mathbf{e}_1$ and $\mathbf{e}_2$, and suppose $\phi\colon V \to V'$ is given by sending

$$\phi(a\mathbf{e}_1 + b\mathbf{e}_2) = (a^2 + b^2)w_1 + \log(a^2+1)w_2 + b^3 w_3$$

where w_1, w_2, w_3 is a basis for V'. Consider the form $\alpha_q = f(q)w_1^\vee \wedge w_3^\vee$, where $f\colon V' \to \mathbb{R}$. Then

$$(\phi^*\alpha)_p = f(\phi(p)) \cdot (2a\mathbf{e}_1^\vee + 2b\mathbf{e}_2^\vee) \wedge (3b^2\mathbf{e}_2^\vee) = f(\phi(p)) \cdot 6ab^2 \cdot \mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee.$$

It turns out that the pullback basically behaves nicely as possible, e.g.

- $\phi^*(c\alpha + \beta) = c\phi^*\alpha + \phi^*\beta$ (linearity)
- $\phi^*(\alpha \wedge \beta) = (\phi^*\alpha) \wedge (\phi^*\beta)$
- $\phi_1^*(\phi_2^*(\alpha)) = (\phi_2 \circ \phi_1)^*(\alpha)$ (naturality)

but I won't take the time to check these here (one can verify them all by expanding with a basis).

§45.3 Cells

Prototypical example for this section: A disk in $\mathbb{R}^2$ can be thought of as the cell $[0, R] \times [0, 2\pi] \to \mathbb{R}^2$ by $(r, \theta) \mapsto (r\cos\theta)\mathbf{e}_1 + (r\sin\theta)\mathbf{e}_2$.

Now that we have the notion of a pullback, we can define the notion of an integral for more general spaces. Specifically, to generalize the notion of integrals we had before:

Definition 45.3.1. A **k-cell** is a smooth function $c\colon [a_1, b_1] \times [a_2, b_2] \times \dots [a_k, b_k] \to V$.

Example 45.3.2 (Examples of cells)
Let $V = \mathbb{R}^2$ for convenience.

(a) A 0-cell consists of a single point.

(b) As we saw, a 1-cell is an arbitrary curve.

(c) A 2-cell corresponds to a 2-dimensional surface. For example, the map $c\colon [0,R]\times[0,2\pi] \to V$ by
$$c\colon (r,\theta) \mapsto (r\cos\theta, r\sin\theta)$$
can be thought of as a disk of radius R.

So we can now give the definition.

Definition 45.3.3 (How to integrate differential k-forms)**.** Take a differential k-form α and a k-cell $c\colon [0,1]^k \to V$. Define the integral $\int_c \alpha$ using the pullback
$$\int_c \alpha \coloneqq \int_{[0,1]^k} c^*\alpha.$$

Since $c^*\alpha$ is a k-form on the k-dimensional unit box, it can be written as $f(x_1,\dots,x_n)\,dx_1\wedge\cdots\wedge dx_n$, So the above integral could also be written as
$$\int_0^1 \cdots \int_0^1 f(x_1,\dots,x_n)\,dx_1\wedge\cdots\wedge dx_n.$$

Example 45.3.4 (Area of a circle)
Consider $V=\mathbb{R}^2$ and let $c\colon (r,\theta)\mapsto (r\cos\theta)\mathbf{e}_1 + (r\sin\theta)\mathbf{e}_2$ on $[0,R]\times[0,2\pi]$ as before. Take the 2-form α which gives $\alpha_p = \mathbf{e}_1^\vee\wedge\mathbf{e}_2^\vee$ at every point p. Then
$$\begin{aligned} c^*\alpha &= (\cos\theta dr - r\sin\theta d\theta)\wedge(\sin\theta dr + r\cos\theta d\theta) \\ &= r(\cos^2\theta+\sin^2\theta)(dr\wedge d\theta) \\ &= r\,dr\wedge d\theta \end{aligned}$$
Thus,
$$\int_c \alpha = \int_0^R \int_0^{2\pi} r\,dr\wedge d\theta = \pi R^2$$
which is the area of a circle.

Here's some geometric intuition for what's happening. Given a k-cell in V, a differential k-form α accepts a point p and some tangent vectors $v_1,\dots,v_k$ and spits out a number $\alpha_p(v_1,\dots,v_k)$, which as before we view as a signed hypervolume. Then the integral *adds up all these infinitesimals across the entire cell.* In particular, if $V=\mathbb{R}^k$ and we take the form $\alpha\colon p\mapsto \mathbf{e}_1^\vee\wedge\cdots\wedge\mathbf{e}_k^\vee$, then what these α's give is the kth hypervolume of the cell. For this reason, this α is called the **volume form** on $\mathbb{R}^k$.

You'll notice I'm starting to play loose with the term "cell": while the cell $c\colon [0,R]\times[0,2\pi]\to\mathbb{R}^2$ is supposed to be a function I have been telling you to think of it as a unit disk (i.e. in terms of its image). In the same vein, a curve $[0,1]\to V$ should be thought of as a curve in space, rather than a function on time.

This error turns out to be benign. Let α be a k-form on U and $c\colon [a_1,b_1]\times\cdots\times[a_k,b_k]\to U$ a k-cell. Suppose $\phi\colon [a_1',b_1']\times\ldots[a_k',b_k']\to[a_1,b_1]\times\cdots\times[a_k,b_k]$; it is a **reparametrization** if ϕ is bijective and $(D\phi)_p$ is always invertible (think "change of variables"); thus

$$c\circ\phi\colon [a_1',b_1']\times\cdots\times[a_k',b_k']\to U$$

is a k-cell as well. Then it is said to **preserve orientation** if $\det(D\phi)_p>0$ for all p and **reverse orientation** if $\det(D\phi)_p<0$ for all p.

Exercise 45.3.5. Why is it that exactly one of these cases must occur?

Theorem 45.3.6 (Changing variables doesn't affect integrals)
Let c be a k-cell, α a k-form, and ϕ a reparametrization. Then

$$\int_{c\circ\phi}\alpha=\begin{cases}\int_c\alpha & \phi\text{ preserves orientation}\\ -\int_c\alpha & \phi\text{ reverses orientation.}\end{cases}$$

Proof. Use naturality of the pullback to reduce it to the corresponding theorem in normal calculus. □

So for example, if we had parametrized the unit circle as $[0,1]\times[0,1]\to\mathbb{R}^2$ by $(r,t)\mapsto R\cos(2\pi t)\mathbf{e}_1+R\sin(2\pi t)\mathbf{e}_2$, we would have arrived at the same result. So we really can think of a k-cell just in terms of the points it specifies.

§45.4 Boundaries

Prototypical example for this section: The boundary of $[a,b]$ is $\{b\}-\{a\}$. The boundary of a square goes around its edge counterclockwise.

First, I introduce a technical term that lets us consider multiple cells at once.

Definition 45.4.1. A **k-chain** U is a formal linear combination of k-cells over U, i.e. a sum of the form

$$c=a_1c_1+\cdots+a_mc_m$$

where each $a_i\in\mathbb{R}$ and c_i is a k-cell. We define $\int_c\alpha=\sum_i a_i\int c_i$.

In particular, a 0-chain consists of several points, each with a given weight.

Now, how do we define the boundary? For a 1-cell $[a,b]\to U$, as I hinted earlier we want the answer to be the 0-chain $\{c(b)\}-\{c(a)\}$. Here's how we do it in general.

Definition 45.4.2. Suppose $c\colon[0,1]^k\to U$ is a k-cell. Then the **boundary** of c, denoted $\partial c\colon[0,1]^{k-1}\to U$, is the $(k-1)$-chain defined as follows. For each $i=1,\ldots,k$ define $(k-1)$-chains by

$$c_i^{\text{start}}\colon(t_1,\ldots,t_{k-1})\mapsto c(t_1,\ldots,t_{i-1},0,t_i,\ldots,t_{k-1})$$
$$c_i^{\text{stop}}\colon(t_1,\ldots,t_{k-1})\mapsto c(t_1,\ldots,t_{i-1},1,t_i,\ldots,t_{k-1}).$$

Then

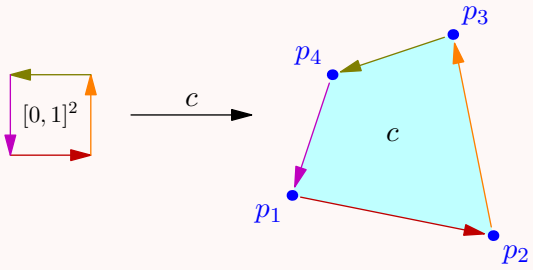
$$\partial c:=\sum_{i=1}^{k}(-1)^{i+1}\left(c_i^{\text{stop}}-c_i^{\text{start}}\right).$$

Finally, the boundary of a chain is the sum of the boundaries of each cell (with the appropriate weights). That is, $\partial(\sum a_ic_i)=\sum a_i\partial c_i$.

Question 45.4.3. Satisfy yourself that one can extend this definition to a k-cell c defined on $c\colon [a_1, b_1] \times \cdots \times [a_k, b_k] \to V$ (rather than from $[0,1]^k \to V$).

Example 45.4.4 (Examples of boundaries)

Consider the 2-cell $c\colon [0,1]^2 \to \mathbb{R}^2$ shown below.



Here p_1, p_2, p_3, p_4 are the images of $(0,0)$, $(0,1)$, $(1,1)$, $(1,0)$, respectively. Formally, ∂c is given by

$$\partial c = (t \mapsto c(1,t)) - (t \mapsto c(0,t)) - (t \mapsto c(t,1)) + (t \mapsto c(t,0)).$$

I apologize for the eyesore notation caused by inline functions. Let's make amends and just write

$$\partial c = [p_2, p_3] - [p_1, p_4] - [p_4, p_3] + [p_1, p_2]$$

where each "interval" represents the 1-cell shown by the reddish arrows on the right, after accounting for the minus signs. We can take the boundary of this as well, and obtain an empty chain as

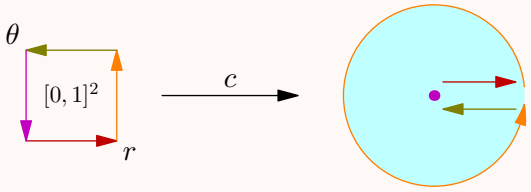
$$\partial(\partial c) = \sum_{i=1}^{4} \{p_{i+1}\} - \{p_i\} = 0.$$

Example 45.4.5 (Boundary of a unit disk)

Consider the unit disk given by

$$c\colon [0,1] \times [0,1] \to \mathbb{R}^2 \quad \text{by} \quad (r,\theta) \mapsto r\cos(2\pi\theta)\mathbf{e}_1 + r\sin(2\pi\theta)\mathbf{e}_2.$$

The four parts of the boundary are shown in the picture below:



Note that two of the arrows more or less cancel each other out when they are integrated. Moreover, we interestingly have a *degenerate* 1-cell at the center of the circle; it is a constant function $[0,1] \to \mathbb{R}^2$ which always gives the origin.

Obligatory theorem, analogous to $d^2 = 0$ and left as a problem.

Theorem 45.4.6 (The boundary of the boundary is empty)
$\partial^2 = 0$, in the sense that for any k-chain c we have $\partial^2(c) = 0$.

§45.5 Stokes' theorem

Prototypical example for this section: $\int_{[a,b]} dg = g(b) - g(a)$.

We now have all the ingredients to state Stokes' theorem for cells.

Theorem 45.5.1 (Stokes' theorem for cells)
Take $U \subseteq V$ as usual, let $c \colon [0,1]^k \to U$ be a k-cell and let $\alpha \colon U \to \bigwedge^{k-1}(V^\vee)$ be a $(k-1)$-form. Then

$$\int_c d\alpha = \int_{\partial c} \alpha.$$

In particular, if $d\alpha = 0$ then the left-hand side vanishes.

For example, if c is the interval $[a,b]$ then $\partial c = \{b\} - \{a\}$, and thus we obtain the fundamental theorem of calculus.

§45.6 Back to Earth: A comparison to what you learned in vector calculus

Now that we've done everything abstractly, let's compare what we've learned to what you might see in $\mathbb{R}^3$ if you're doing a vector calculus course at a typical university.

In Figure 45.1 I've copied a picture I drew in fall 2024 for the 18.02 class at MIT, which is the multivariable calculus class that a lot of first-year students take. For each $0 \le d \le n \le 3$ (besides $d = n = 0$), it shows what kind of integral showed up in the class if you were doing a d-dimensional integral of a function whose domain was $\mathbb{R}^n$. Note that every integral in this picture is real-valued.

I've deliberately used the notation that was actually used at MIT, which I'll refer to as 18.02 notation, because it's similar to what you will see on Wikipedia and other places too. The goal of this section is to provide a translation system from 18.02 notation to Napkin notation. (Throughout the whole section, $\mathbb{R}^n$ is thought of as a normed vector space, so the identification $\mathbf{e}_1 \mapsto \mathbf{e}_1^\vee$ and so on is canonical.)

There is a lot going in Figure 45.1, so let's break it down piece by piece.

0-forms. A 0-form is the same as just a function, so the column of 0-D integrals should be easy to understand: it's just evaluation at point, or a sum of points.

n-forms. The case $n = d$ is also easy: we know it's possible to integrate an n-form in $\mathbb{R}^n$ and get a number. That is:

- A normal integral $\int_a^b \mathrm{d}x$ is the integral across a 1-cell $[a,b]$ across the 1-form $f \cdot \mathbf{e}_1^\vee$.
- An area integral $\int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$ corresponds to integrating the 1-form $f \cdot \mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee$.
- A volume integral $\int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$ corresponds to integrating the 1-form $f \cdot \mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee \wedge \mathbf{e}_3^\vee$.

So this takes care of the green-labeled things on the diagonal.

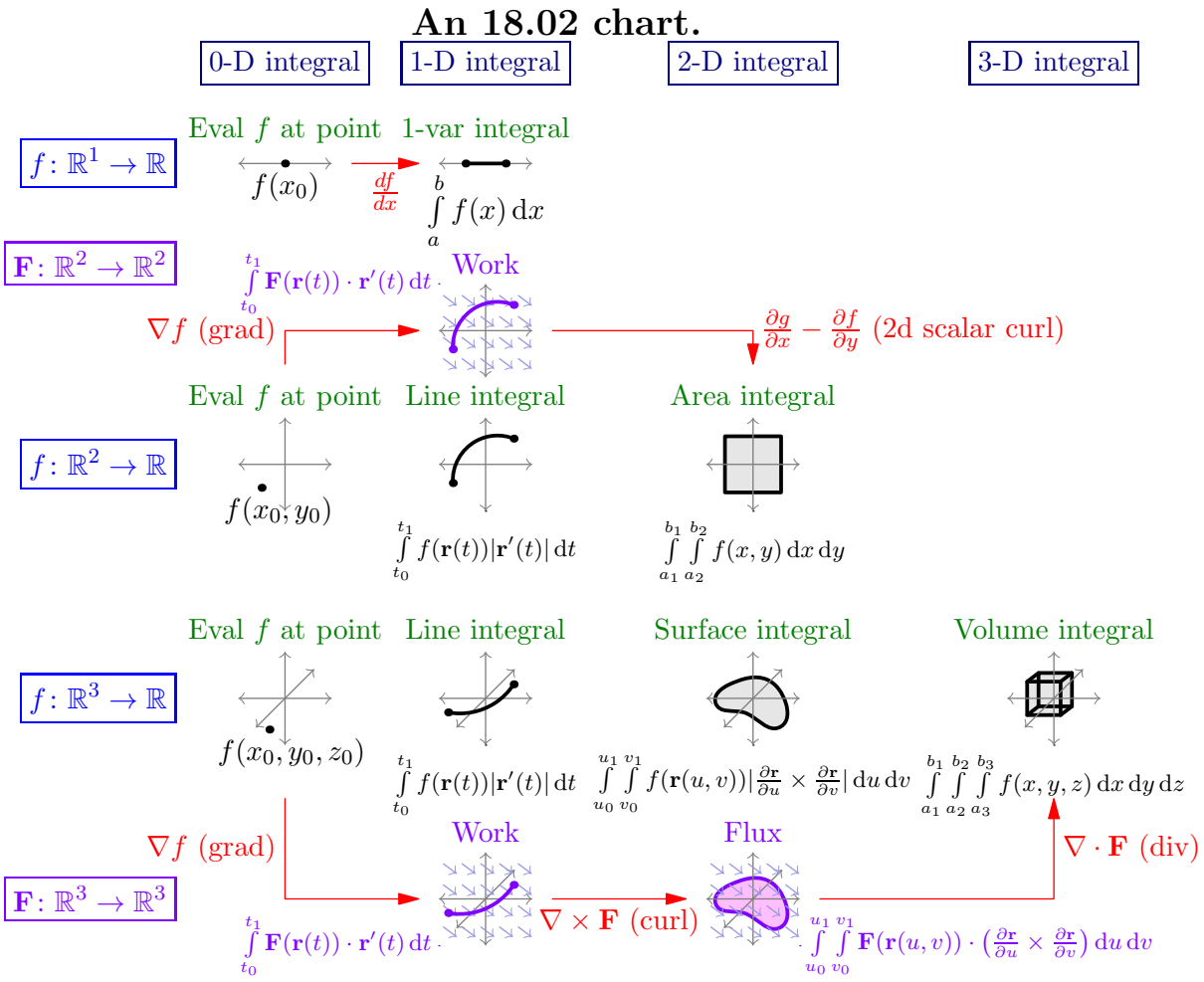


Figure 45.1: Throwback to first year of college? High-resolution version at `https://web.evanchen.cc/upload/1802/integrals-stokes.pdf`.

We can't interpret the remaining three green pictures! The tricky part is the situations where $0 < d < n < 3$. There are three such things, the two line integrals when $d = 1$ and $n \in \{2, 3\}$ and the surface integral when $d = 2$ and $n = 3$.

In fact, these are *not* covered by our theory of differential forms! Indeed, even in the special case where $f = 1$ is a constant function, the line integrals are actually arc length, and as we mentioned in Section 44.6, that integral cannot be viewed as the integral of any differential form. Similarly, surface area isn't a differential form either.

1-forms and 2-forms. However, the three purplish integrals (over vector fields) can be viewed in our framework.

- Consider $d = 1$ and $n = 3$, i.e. the 3-D line integral. We have as input a vector-valued function $\mathbf{F}\colon \mathbb{R}^3 \to \mathbb{R}^3$. By projection principle (Theorem 43.2.1), it's the same as the data of

$$\mathbf{F}(p) = F_1(p)\mathbf{e}_1 + F_2(p)\mathbf{e}_2 + F_3(p)\mathbf{e}_3$$

  for three functions $F_i\colon \mathbb{R}^3 \to \mathbb{R}$ for $i = 1, 2, 3$.

  To convert the 18.02 notation $\mathbf{F}(p)$ into Napkin notation, we identify $\mathbf{F}$ with the

differential form

$$\alpha_p = F_1(p)\mathbf{e}_1^\vee + F_2(p)\mathbf{e}_2^\vee + F_3(p)\mathbf{e}_3^\vee.$$

Meanwhile, the path $\mathbf{r}(t)$ parametrized by time $t \in [t_0, t_1]$ matches the concept of a 1-form $c : [t_0, t_1] \to \mathbb{R}^3$. The "work" in the integral is written as

$$\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t)$$

but that dot product is exactly the pullback $c^*\alpha$.

- The case $d = 1$ and $n = 2$ is exactly the same, with 3 replaced by 2.
- The weirdest case is the flux integral, for $d = 2$ and $n = 3$. The parametrization $\mathbf{r}(u, v)$ is fine, and it corresponds to a 2-cell c. But $\mathbf{F}(p)$ seems to have the wrong type.

  But let's again write

  $$\mathbf{F}(p) = F_1(p)\mathbf{e}_1 + F_2(p)\mathbf{e}_2 + F_3(p)\mathbf{e}_3.$$

  There is a fairly weird hack used to convert this into Napkin notation: the form desired is

  $$\alpha_p = F_1(p)\mathbf{e}_2^\vee \wedge \mathbf{e}_3^\vee + F_2(p)\mathbf{e}_3^\vee \wedge \mathbf{e}_1^\vee + F_3(p)\mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee.$$

  Yes, that's really the identification! For this definition to be possible, we had to exploit the fact that

  $$\binom{3}{1} = \binom{3}{2}.$$

  That is the three-dimensional space $\bigwedge^2(\mathbb{R}^3)$ happens to have the same number of basis elements as $\bigwedge^1(\mathbb{R}^3) \cong \mathbb{R}^3$, so the

  $$\begin{aligned} \star\colon \bigwedge^2(\mathbb{R}^3) &\to \mathbb{R}^3 \\ \mathbf{e}_1 \wedge \mathbf{e}_2 &\mapsto \mathbf{e}_3 \\ \mathbf{e}_2 \wedge \mathbf{e}_3 &\mapsto \mathbf{e}_1 \\ \mathbf{e}_3 \wedge \mathbf{e}_1 &\mapsto \mathbf{e}_2 \end{aligned}$$

  is really an isomorphism, because it maps basis elements to basis elements. We denote this map by $\star$, because it turns out this map generalizes to the so-called Hodge star operator in higher dimensions.

  This is where I talk about cross products, which I've deliberately avoided until now. The cross product is a weird operation that takes two vectors in $\mathbb{R}^3$ and outputs a vector in $\mathbb{R}^3$. Specifically, if $\mathbf{v} = x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3$ and $\mathbf{w} = x'\mathbf{e}_1 + y'\mathbf{e}_2 + z'\mathbf{e}_3$, the definition of cross products taught in school is

  $$\mathbf{v} \times \mathbf{w} \coloneqq (yz' - y'z)\mathbf{e}_1 + (zx' - xz')\mathbf{e}_2 + (xy' - x'y)\mathbf{e}_3.$$

  Where does this come from? The answer is that $\star(\mathbf{v} \vee \mathbf{w})$:

  $$\begin{aligned} \mathbf{v} \wedge \mathbf{w} &= (x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3) \times (x'\mathbf{e}_1 + y'\mathbf{e}_2 + z'\mathbf{e}_3) \\ &= (xy' - x'y)\mathbf{e}_1 \wedge \mathbf{e}_2 + (yz' - y'z)\mathbf{e}_2 \wedge \mathbf{e}_3 + (zx' - xz')\mathbf{e}_3 \wedge \mathbf{e}_1 \\ \star(\mathbf{v} \wedge \mathbf{w}) &\mapsto (xy' - x'y)\mathbf{e}_3 + (yz' - y'z)\mathbf{e}_1 + (zx' - xz')\mathbf{e}_2. \end{aligned}$$

  With that out of the way, the weird dot-cross product

  $$\mathbf{F}(\mathbf{r}(u, v)) \cdot (\mathbf{r}_u \times \mathbf{r}_v)\,\mathrm{d}u\,\mathrm{d}v$$

  is now rigged to correspond to the pullback $c^*\alpha$. So using this Hodge star, we find that flux is actually the integration of a 2-form.

Exterior derivatives Every red arrow in Figure 45.1 corresponds to the exterior derivative of the corresponding form. That is:

- The "grad" operation takes a 0-form f and outputs a vector field corresponding to the 1-form df.
- The "curl" operation takes a 1-form α and outputs a vector field corresponding to the 2-form $d\alpha$. When $n = 3$, this checks out because the space of 1-forms is $\binom{3}{1}$ dimensional, and the space of 2-forms is $\binom{3}{2}$, and thankfully $\binom{3}{1} = 3 = \binom{3}{2}$.

  The weird notation $\nabla \times \mathbf{F}$ can be checked to correspond to the exterior derivative. On the 18.02 side, if we have

  $$\mathbf{F} = F_1\mathbf{e}_1 + F_2\mathbf{e}_2 + F_3\mathbf{e}_3$$

  then the 18.02 definition of curl is that

  $$\operatorname{curl}(\mathbf{F}) \coloneqq \nabla \times \mathbf{F} \coloneqq \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right)\mathbf{e}_1 + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right)\mathbf{e}_2 + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\mathbf{e}_3$$

  The reason for the nonsensical $\nabla\times$ notation is that if you *really* abuse notation you can almost think of this as the cross product of a vector $\nabla = \left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right\rangle$ and the vector $\mathbf{F} = \langle F_1, F_2, F_3 \rangle$.

  Now to convert $\mathbf{F}$ into Napkin notation, remember we identified $\mathbf{F}$ with the differential form

  $$\alpha = F_1\mathbf{e}_1^\vee + F_2\mathbf{e}_2^\vee + F_3\mathbf{e}_3^\vee.$$

  If we follow our formula for exterior derivative in Definition 44.4.4, we get

  $$\begin{aligned} d\alpha &= dF_1 \wedge \mathbf{e}_1^\vee + dF_2 \wedge \mathbf{e}_2^\vee + dF_3 \wedge \mathbf{e}_3^\vee \\ &= \left(\frac{\partial F_1}{\partial x}\mathbf{e}_1^\vee + \frac{\partial F_1}{\partial y}\mathbf{e}_2^\vee + \frac{\partial F_1}{\partial z}\mathbf{e}_3^\vee\right) \wedge \mathbf{e}_1^\vee \\ &+ \left(\frac{\partial F_2}{\partial x}\mathbf{e}_1^\vee + \frac{\partial F_2}{\partial y}\mathbf{e}_2^\vee + \frac{\partial F_2}{\partial z}\mathbf{e}_3^\vee\right) \wedge \mathbf{e}_2^\vee \\ &+ \left(\frac{\partial F_3}{\partial x}\mathbf{e}_1^\vee + \frac{\partial F_3}{\partial y}\mathbf{e}_2^\vee + \frac{\partial F_3}{\partial z}\mathbf{e}_3^\vee\right) \wedge \mathbf{e}_3^\vee \\ &= \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right)\mathbf{e}_2^\vee \wedge \mathbf{e}_3^\vee + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right)\mathbf{e}_3^\vee \wedge \mathbf{e}_1^\vee + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\mathbf{e}_1^\vee \wedge \mathbf{e}_2^\vee. \end{aligned}$$

  Taking the Hodge star and then dropping all the $\vee$'s gives the same thing as $\nabla \times \mathbf{F}$, so this completes the correspondence between the 18.02 notation and the Napkin notation.
- In 18.02 terminology, the divergence div is defined by

  $$\operatorname{div}(\mathbf{F}) \coloneqq \nabla \cdot \mathbf{F} \coloneqq \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}$$

  which is a scalar-valued function for input points $p \in \mathbb{R}^3$. We let you do this one in Problem 45C.

  The reason for the nonsensical $\nabla\cdot$ notation is that if you *really* abuse notation you can almost think of this as the dot product of a vector $\nabla = \left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right\rangle$ and the vector $\mathbf{F} = \langle F_1, F_2, F_3 \rangle$.

Double derivative We know that $d^2 = 0$, which in Figure 45.1 means composing two arrows gives zero. You'll see this in 18.02 as

- The curl of a gradient is zero.
- The flux of a curl is zero.

but really they're the same theorem.

Stokes' theorem Each red arrow also gives an instance of Stokes' theorem for cells. So Stokes' theorem even for cells is really great, because we get six 18.02 theorems as special cases!

- The three arrows from 0-D integrals to 1-D integrals are all called "Fundamental Theorem of Calculus". Some authors will say "Fundamental Theorem of Calculus for line integrals" instead for $n > 1$.
- For $n = 2$, the other red arrow is called "Green's theorem"; we let you work it out as Problem 45A†.
- For $n = 3$, the arrow from work to flux is confusingly also called "Stokes' theorem"; it says the flux of a 2-D surface equals the work on the 1-D boundary.
- The rightmost red arrow for $n = 3$ is called the "divergence theorem"; it says the divergence of a 3-D volume equals the flux of the 2-D boundary surface.

§45.7 A few harder problems to think about

Problem 45A† (Green's theorem)**.** Let $f, g \colon \mathbb{R}^2 \to \mathbb{R}$ be smooth functions and c a 2-cell. Prove that

$$\int_c \left(\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y} \right) dx \wedge dy = \int_{\partial c} (f\, dx + g\, dy).$$

Problem 45B. Show that $\partial^2 = 0$.

Problem 45C. Finish the correspondence of the 18.02 notation with Napkin notation. That is, let $\mathbf{F} \colon \mathbb{R}^3 \to \mathbb{R}^3$ be a vector field, and let α be the 2-form corresponding to it in Napkin version. Show that the scalar-valued function defined by

$$\operatorname{div}(\mathbf{F}) \coloneqq \nabla \cdot \mathbf{F} \coloneqq \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}$$

coincides with evaluation at the 3-form $d\alpha$.

Problem 45D (Pullback and d commute)**.** Let U and U' be open sets of vector spaces V and V' and let $\phi \colon U \to U'$ be a smooth map between them. Prove that for any differential form α on U' we have

$$\phi^*(d\alpha) = d(\phi^* \alpha).$$

Problem 45E (Arc length isn't a form)**.** Show that there does *not* exist a 1-form α on $\mathbb{R}^2$ such that for a curve $c \colon [0,1] \to \mathbb{R}^2$, the integral $\int_c \alpha$ gives the arc length of c.

Problem 45F. An **exact** k-form α is one satisfying $\alpha = d\beta$ for some β. Prove that

$$\int_{C_1} \alpha = \int_{C_2} \alpha$$

where C_1 and C_2 are any concentric circles in the plane and α is some exact 1-form.

46 A bit of manifolds

Last chapter, we stated Stokes' theorem for cells. It turns out there is a much larger class of spaces, the so-called *smooth manifolds*, for which this makes sense.

Unfortunately, the definition of a smooth manifold is *complete garbage*, and so by the time I am done defining differential forms and orientations, I will be too lazy to actually define what the integral on it is, and just wave my hands and state Stokes' theorem.

§46.1 Topological manifolds

Prototypical example for this section: S^2*: "the Earth looks flat".*

Long ago, people thought the Earth was flat, i.e. homeomorphic to a plane, and in particular they thought that $\pi_2(\text{Earth}) = 0$. But in fact, as most of us know, the Earth is actually a sphere, which is not contractible and in particular $\pi_2(\text{Earth}) \cong \mathbb{Z}$. This observation underlies the definition of a manifold:

An n-manifold is a space which locally looks like $\mathbb{R}^n$.

Actually there are two ways to think about a topological manifold M:

- "Locally": at every point $p \in M$, some open neighborhood of p looks like an open set of $\mathbb{R}^n$. For example, to someone standing on the surface of the Earth, the Earth looks much like $\mathbb{R}^2$.
- "Globally": there exists an open cover of M by open sets $\{U_i\}_i$ (possibly infinite) such that each U_i is homeomorphic to some open subset of $\mathbb{R}^n$. For example, from outer space, the Earth can be covered by two hemispherical pancakes.

Question 46.1.1. Check that these are equivalent.

While the first one is the best motivation for examples, the second one is easier to use formally.

Definition 46.1.2. A **topological n-manifold** M is a Hausdorff space with an open cover $\{U_i\}$ of sets homeomorphic to subsets of $\mathbb{R}^n$, say by homeomorphisms

$$\phi_i \colon U_i \xrightarrow{\cong} E_i \subseteq \mathbb{R}^n$$

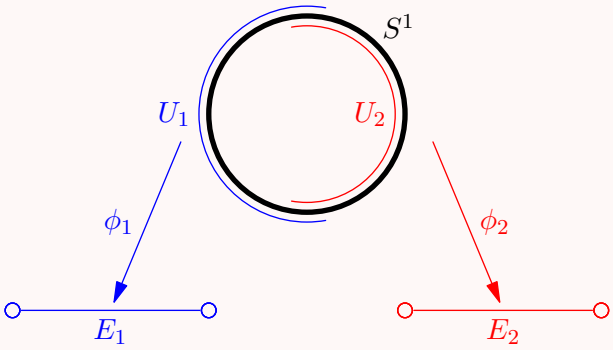
where each E_i is an open subset of $\mathbb{R}^n$. Each $\phi_i \colon U_i \to E_i$ is called a **chart**, and together they form a so-called **atlas**.

Remark 46.1.3 — Here "E" stands for "Euclidean". I think this notation is not standard; usually people just write $\phi_i(U_i)$ instead.

Remark 46.1.4 — This definition is nice because it doesn't depend on embeddings: a manifold is an *intrinsic* space M, rather than a subset of $\mathbb{R}^N$ for some N. Analogy:

an abstract group G is an intrinsic object rather than a subgroup of S_n.

Example 46.1.5 (An atlas on S^1)
Here is a picture of an atlas for S^1, with two open sets.



Question 46.1.6. Where do you think the words "chart" and "atlas" come from?

Example 46.1.7 (Some examples of topological manifolds)
(a) As discussed at length, the sphere S^2 is a 2-manifold: every point in the sphere has a small open neighborhood that looks like D^2. One can cover the Earth with just two hemispheres, and each hemisphere is homeomorphic to a disk.

(b) The circle S^1 is a 1-manifold; every point has an open neighborhood that looks like an open interval.

(c) The torus, Klein bottle, $\mathbb{RP}^2$ are all 2-manifolds.

(d) $\mathbb{R}^n$ is trivially a manifold, as are its open sets.

All these spaces are compact except $\mathbb{R}^n$.
A non-example of a manifold is D^n, because it has a *boundary*; points on the boundary do not have open neighborhoods that look Euclidean.

§46.2 Smooth manifolds

Prototypical example for this section: All the topological manifolds.

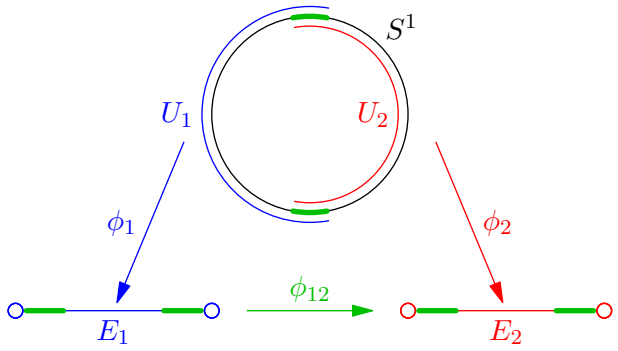
Let M be a topological n-manifold with atlas $\{U_i \xrightarrow{\phi_i} E_i\}$.

Definition 46.2.1. For any i, j such that $U_i \cap U_j \neq \varnothing$, the **transition map** ϕ_{ij} is the composed map

$$\phi_{ij} \colon E_i \cap \phi_i^{\text{img}}(U_i \cap U_j) \xrightarrow{\phi_i^{-1}} U_i \cap U_j \xrightarrow{\phi_j} E_j \cap \phi_j^{\text{img}}(U_i \cap U_j).$$

Sorry for the dense notation, let me explain. The intersection with the image $\phi_i^{\text{img}}(U_i \cap U_j)$ and the image $\phi_j^{\text{img}}(U_i \cap U_j)$ is a notational annoyance to make the map well-defined and a homeomorphism. The transition map is just the natural way to go from $E_i \to E_j$,

restricted to overlaps. Picture below, where the intersections are just the green portions of each E_1 and E_2:



We want to add enough structure so that we can use differential forms.

Definition 46.2.2. We say M is a **smooth manifold** if all its transition maps are smooth.

This definition makes sense, because we know what it means for a map between two open sets of $\mathbb{R}^n$ to be differentiable.

With smooth manifolds we can try to port over definitions that we built for $\mathbb{R}^n$ onto our manifolds. So in general, all definitions involving smooth manifolds will reduce to something on each of the coordinate charts, with a compatibility condition.

As an example, here is the definition of a "smooth map":

Definition 46.2.3. (a) Let M be a smooth manifold. A continuous function $f\colon M \to \mathbb{R}$ is called **smooth** if the composition

$$E_i \xrightarrow{\phi_i^{-1}} U_i \hookrightarrow M \xrightarrow{f} \mathbb{R}$$

is smooth as a function $E_i \to \mathbb{R}$.

(b) Let M and N be smooth with atlases $\{U_i^M \xrightarrow{\phi_i} E_i^M\}_i$ and $\{U_j^N \xrightarrow{\phi_j} E_j^N\}_j$. A map $f\colon M \to N$ is **smooth** if for every i and j, the composed map

$$E_i \xrightarrow{\phi_i^{-1}} U_i \hookrightarrow M \xrightarrow{f} N \twoheadrightarrow U_j \xrightarrow{\phi_j} E_j$$

is smooth, as a function $E_i \to E_j$.

§46.3 Regular value theorem

Prototypical example for this section: $x^2 + y^2 = 1$ *is a circle!*

Despite all that I've written about general manifolds, it would be sort of mean if I left you here because I have not really told you how to actually construct manifolds in practice, even though we know the circle $x^2 + y^2 = 1$ is a great example of a one-dimensional manifold embedded in $\mathbb{R}^2$.

Theorem 46.3.1 (Regular value theorem)

Let V be an n-dimensional real normed vector space, let $U \subseteq V$ be open and let $f_1, \dots, f_m \colon U \to \mathbb{R}$ be smooth functions. Let M be the set of points $p \in U$ such that $f_1(p) = \dots = f_m(p) = 0$.
Assume M is nonempty and that the map

$$V \to \mathbb{R}^m \quad \text{by} \quad v \mapsto ((Df_1)_p(v), \dots, (Df_m)_p(v))$$

has rank m, for every point $p \in M$. Then M is a manifold of dimension $n - m$.

For a proof, see [Sj05, Theorem 6.3].
One very common special case is to take $m = 1$ above.

Corollary 46.3.2 (Level hypersurfaces)

Let V be a finite-dimensional real normed vector space, let $U \subseteq V$ be open and let $f \colon U \to \mathbb{R}$ be smooth. Let M be the set of points $p \in U$ such that $f(p) = 0$. If $M \neq \varnothing$ and $(Df)_p$ is not the zero map for any $p \in M$, then M is a manifold of dimension $\dim V - 1$.

Example 46.3.3 (The circle $x^2 + y^2 - c = 0$)

Let $f(x, y) = x^2 + y^2 - c$, $f \colon \mathbb{R}^2 \to \mathbb{R}$, where c is a positive real number. Note that

$$Df = 2x \cdot dx + 2y \cdot dy$$

which in particular is nonzero as long as $(x, y) \neq (0, 0)$, i.e. as long as $c \neq 0$. Thus:

- When $c > 0$, the resulting curve — a circle with radius $\sqrt{c}$ — is a one-dimensional manifold, as we knew.
- When $c = 0$, the result fails. Indeed, M is a single point, which is actually a zero-dimensional manifold!

We won't give further examples since I'm only mentioning this in passing in order to increase your capacity to write real concrete examples. (But [Sj05, Chapter 6.2] has some more examples, beautifully illustrated.)

§46.4 Differential forms on manifolds

We already know what a differential form is on an open set $U \subseteq \mathbb{R}^n$. So, we naturally try to port over the definition of differentiable form on each subset, plus a compatibility condition.

Let M be a smooth manifold with atlas $\{U_i \xrightarrow{\phi_i} E_i\}_i$.

Definition 46.4.1. A **differential k-form** α on a smooth manifold M is a collection $\{\alpha_i\}_i$ of differential k-forms on each E_i, such that for any j and i we have that

$$\alpha_j = \phi_{ij}^*(\alpha_i).$$

In English: we specify a differential form on each chart, which is compatible under pullbacks of the transition maps.

§46.5 Orientations

Prototypical example for this section: Left versus right, clockwise vs. counterclockwise.

This still isn't enough to integrate on manifolds. We need one more definition: that of an orientation.

The main issue is the observation from standard calculus that

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$$

Consider then a space M which is homeomorphic to an interval. If we have a 1-form α, how do we integrate it over M? Since M is just a topological space (rather than a subset of $\mathbb{R}$), there is no default "left" or "right" that we can pick. As another example, if $M = S^1$ is a circle, there is no default "clockwise" or "counterclockwise" unless we decide to embed M into $\mathbb{R}^2$.

To work around this we have to actually have to make additional assumptions about our manifold.

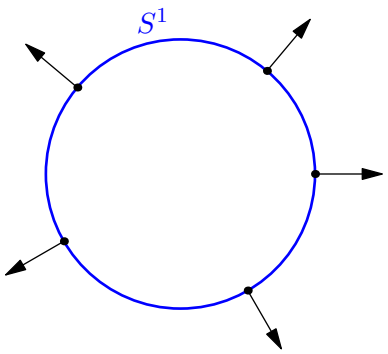
Definition 46.5.1. A smooth n-manifold is **orientable** if there exists a differential n-form ω on M such that for every $p \in M$,

$$\omega_p \neq 0.$$

Recall here that ω_p is an element of $\bigwedge^n(V^\vee)$. In that case we say ω is a **volume form** of M.

How do we picture this definition? If we recall that an differential form is supposed to take tangent vectors of M and return real numbers. To this end, we can think of each point $p \in M$ as having a **tangent plane** $T_p(M)$ which is n-dimensional. Now since the volume form ω is n-dimensional, it takes an entire basis of the $T_p(M)$ and gives a real number. So a manifold is orientable if there exists a consistent choice of sign for the basis of tangent vectors at every point of the manifold.

For "embedded manifolds", this just amounts to being able to pick a nonzero field of normal vectors to each point $p \in M$. For example, S^1 is orientable in this way.



Similarly, one can orient a sphere S^2 by having a field of vectors pointing away (or towards) the center. This is all non-rigorous, because I haven't defined the tangent plane $T_p(M)$; since M is in general an intrinsic object one has to be quite roundabout to define $T_p(M)$ (although I do so in an optional section later). In any event, the point is that guesses about the orientability of spaces are likely to be correct.

Example 46.5.2 (Orientable surfaces)

(a) Spheres S^n, planes, and the torus $S^1 \times S^1$ are orientable.

(b) The Möbius strip and Klein bottle are *not* orientable: they are "one-sided".

(c) $\mathbb{CP}^n$ is orientable for any n.

(d) $\mathbb{RP}^n$ is orientable only for odd n.

§46.6 Stokes' theorem for manifolds

Stokes' theorem in the general case is based on the idea of a **manifold with boundary** M, which I won't define, other than to say its boundary ∂M is an $n-1$ dimensional manifold, and that it is oriented if M is oriented. An example is $M = D^2$, which has boundary $\partial M = S^1$.

Next,

Definition 46.6.1. The **support** of a differential form α on M is the closure of the set

$$\{p \in M \mid \alpha_p \neq 0\}.$$

If this support is compact as a topological space, we say α is **compactly supported**.

Remark 46.6.2 — For example, volume forms are supported on all of M.

Now, one can define integration on oriented manifolds, but I won't define this because the definition is truly awful. Then Stokes' theorem says

Theorem 46.6.3 (Stokes' theorem for manifolds)

Let M be a smooth oriented n-manifold with boundary and let α be a compactly supported $(n-1)$-form. Then

$$\int_M d\alpha = \int_{\partial M} \alpha.$$

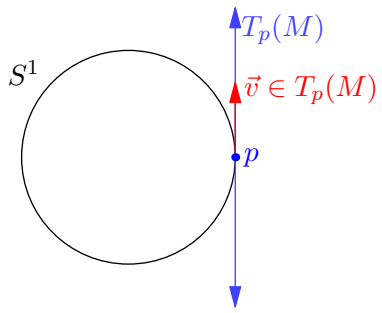
All the omitted details are developed in full in [Sj05].

§46.7 (Optional) The tangent and cotangent space

Prototypical example for this section: Draw a line tangent to a circle, or a plane tangent to a sphere.

Let M be a smooth manifold and $p \in M$ a point. I omitted the definition of $T_p(M)$ earlier, but want to actually define it now.

As I said, geometrically we know what this *should* look like for our usual examples. For example, if $M = S^1$ is a circle embedded in $\mathbb{R}^2$, then the tangent vector at a point p should just look like a vector running off tangent to the circle. Similarly, given a sphere $M = S^2$, the tangent space at a point p along the sphere would look like plane tangent to M at p.



However, one of the points of all this manifold stuff is that we really want to see the manifold as an *intrinsic object*, in its own right, rather than as embedded in $\mathbb{R}^n$.[1] So, we would like our notion of a tangent vector to not refer to an ambient space, but only to intrinsic properties of the manifold M in question.

§46.7.i Tangent space

To motivate this construction, let us start with an embedded case for which we know the answer already: a sphere.

Suppose $f\colon S^2 \to \mathbb{R}$ is a function on a sphere, and take a point p. Near the point p, f looks like a function on some open neighborhood of the origin. Thus we can think of taking a *directional derivative* along a vector $\vec{v}$ in the imagined tangent plane (i.e. some partial derivative). For a fixed $\vec{v}$ this partial derivative is a linear map

$$D_{\vec{v}}\colon C^\infty(M) \to \mathbb{R}.$$

It turns out this goes the other way: if you know what $D_{\vec{v}}$ does to every smooth function, then you can recover v. This is the trick we use in order to create the tangent space. Rather than trying to specify a vector $\vec{v}$ directly (which we can't do because we don't have an ambient space),

The vectors *are* partial-derivative-like maps.

More formally, we have the following.

Definition 46.7.1. A **derivation** D at p is a linear map $D\colon C^\infty(M) \to \mathbb{R}$ (i.e. assigning a real number to every smooth f) satisfying the following Leibniz rule: for any f, g we have the equality

$$D(fg) = f(p) \cdot D(g) + g(p) \cdot D(f) \in \mathbb{R}.$$

This is just a "product rule". Then the tangent space is easy to define:

Definition 46.7.2. A **tangent vector** is just a derivation at p, and the **tangent space** $T_p(M)$ is simply the set of all these tangent vectors.

In this way we have constructed the tangent space.

[1] This can be thought of as analogous to the way that we think of a group as an abstract object in its own right, even though Cayley's Theorem tells us that any group is a subgroup of the permutation group.

Note this wasn't always the case! During the 19th century, a group was literally defined as a subset of $\mathrm{GL}(n)$ or of S_n. In fact Sylow developed his theorems without the word "group". Only much later did the abstract definition of a group was given, an abstract set G which was independent of any *embedding* into S_n, and an object in its own right.

§46.7.ii The cotangent space

In fact, one can show that the product rule for D is equivalent to the following three conditions:

1. D is linear, meaning $D(af+bg) = aD(f) + bD(g)$.
2. $D(1_M) = 0$, where 1_M is the constant function on M.
3. $D(fg) = 0$ whenever $f(p) = g(p) = 0$. Intuitively, this means that if a function $h = fg$ vanishes to second order at p, then its derivative along D should be zero.

This suggests a third equivalent definition: suppose we define

$$\mathfrak{m}_p \coloneqq \{f \in C^\infty M \mid f(p) = 0\}$$

to be the set of functions which vanish at p (this is called the *maximal ideal* at p). In that case,

$$\mathfrak{m}_p^2 = \left\{ \sum_i f_i \cdot g_i \mid f_i(p) = g_i(p) = 0 \right\}$$

is the set of functions vanishing to second order at p. Thus, a tangent vector is really just a linear map

$$\mathfrak{m}_p/\mathfrak{m}_p^2 \to \mathbb{R}.$$

In other words, the tangent space is actually the dual space of $\mathfrak{m}_p/\mathfrak{m}_p^2$; for this reason, the space $\mathfrak{m}_p/\mathfrak{m}_p^2$ is defined as the **cotangent space** (the dual of the tangent space). This definition is even more abstract than the one with derivations above, but has some nice properties:

- it is coordinate-free, and
- it's defined only in terms of the smooth functions $M \to \mathbb{R}$, which will be really helpful later on in algebraic geometry when we have varieties or schemes and can repeat this definition.

§46.7.iii Sanity check

With all these equivalent definitions, the last thing I should do is check that this definition of tangent space actually gives a vector space of dimension n. To do this it suffices to show verify this for open subsets of $\mathbb{R}^n$, which will imply the result for general manifolds M (which are locally open subsets of $\mathbb{R}^n$). Using some real analysis, one can prove the following result:

Theorem 46.7.3

Suppose $M \subset \mathbb{R}^n$ is open and $0 \in M$. Then

$$\begin{aligned} \mathfrak{m}_0 &= \{\text{smooth functions } f \mid f(0) = 0\} \\ \mathfrak{m}_0^2 &= \{\text{smooth functions } f \mid f(0) = 0, (\nabla f)_0 = 0\}. \end{aligned}$$

In other words $\mathfrak{m}_0^2$ is the set of functions which vanish at 0 and such that all first derivatives of f vanish at zero.

Thus, it follows that there is an isomorphism

$$\mathfrak{m}_0/\mathfrak{m}_0^2 \cong \mathbb{R}^n \quad \text{by} \quad f \mapsto \left[\frac{\partial f}{\partial x_1}(0), \dots, \frac{\partial f}{\partial x_n}(0)\right]$$

and so the cotangent space, hence tangent space, indeed has dimension n.

§46.8 A few harder problems to think about

Problem 46A. Show that a differential 0-form on a smooth manifold M is the same thing as a smooth function $M \to \mathbb{R}$.

some applications of regular value theorem here

XIII

Riemann Surfaces

Part XIII: Contents

47 Basic definitions of Riemann surfaces

Roughly speaking, the theory of Riemann surfaces is just the generalization of complex analysis using ideas from differential geometry: Just like how a 2-manifold can be viewed as a collection of patches of the real plane $\mathbb{R}^2$ smoothly welded together to form a more complicated object, we take "pieces" of the complex plane $\mathbb{C}$, *analytically* welded together .

We already know that the theory of holomorphic function is very nice — they're all analytic! The same amount of rigidity is to be expected here.

In fact, on *compact* Riemann surfaces, the theories are even nicer than the case of holomorphic functions! For example:

- For two Riemann surfaces X and Y where Y is compact, any meromorphic function $f\colon X \to Y$ must in fact be holomorphic i.e. defined everywhere.
- If X is a compact Riemann surface, then a holomorphic function $f\colon X \to \mathbb{C}$ is constant.
- In the same setting as above, furthermore we have that if $g\colon X \to \mathbb{C}$ is meromorphic, then the number of zeros of g is equal to the number of poles of g, with multiplicity.

Remark 47.0.1 (Why do we have these nice properties?) — Roughly speaking, $\mathbb{C}$ is not compact — it is isomorphic to the Riemann sphere with a hole removed. By filling in the hole, we allow meromorphic functions to be extended taking value ∞ at places that previously was a pole.

As an orientable 2-manifold, we can define the **genus** of a Riemann surface — it is a purely topological concept, yet it is crucially linked to several algebraic invariants in very surprising ways. You may have heard of the *elliptic curve* in cryptography — they also present as a Riemann surface, and a generalization, **hyperelliptic curve**, form a family of Riemann surfaces of arbitrary genus ≥ 2!

§47.1 Complex structures

Recall the definitions in the previous chapters:

- A topological n-manifold is a Hausdorff space, with a covering $\{U_i\}$, each being homeomorphic to $\mathbb{R}^n$.
- A smooth n-manifold is a topological n-manifold, where all the transition maps are smooth.

What do they have in common? Seemingly not too much. But essentially, they're all describing the same philosophy:

We take countably many patches $\{U_i\}$, and weld them together while keeping the underlying structure.

Here, a topological manifold has a *topological structure*, and a smooth manifold has a *smooth structure*. In a similar manner, a complex manifold has a *complex structure*.

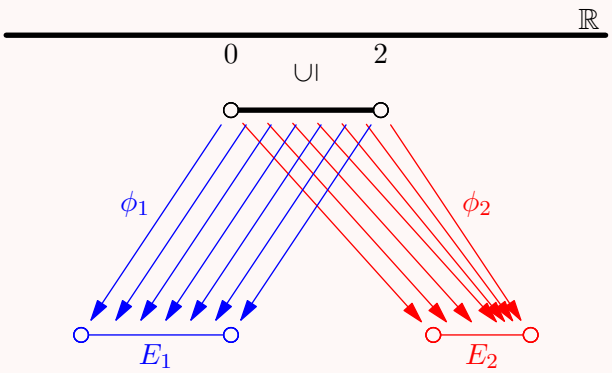
What do we mean by "structure" here?

First, a topological structure is familiar to you — it's just a topology. Formally, the topology is defined by the collection of open sets, but the actual *meaning* of a topological structure dictates:

- whether a set is considered open or closed,
- whether a sequence of points converge to a given point,
- whether a map $X \to Y$ or $Y \to X$ is continuous (given Y is another topological space),
- etc.

Given a topological n-manifold with an existing (Hausdorff) topology on it, we can tell whether a local chart *respects the topological structure*; in other words, is a homeomorphism.

Example 47.1.1 ($(0,2)$ is a topological 1-manifold)
The open interval $(0,2)$ included in $\mathbb{R}$ can be considered a topological manifold.



Two possible charts for the space, ϕ_1 and ϕ_2, are shown.
Their formulas are $\phi_1 \colon (0,2) \to (0,2), \phi_1(x) = x$ and $\phi_2 \colon (0,2) \to (0,1.3), \phi_2(x) = x + 0.35 \cdot (1 - x - |1 - x|)$.

In the example above, you may notice that, even though the chart ϕ_2 is a homeomorphism, it doesn't look *smooth*. So, you want to define a smooth 2-manifold as something like:

> A surface $S \subseteq \mathbb{R}^3$ is a smooth 2-manifold if, for each $p \in S$, there exists an open neighborhood $V \subseteq S$ that is diffeomorphic to $E \subseteq \mathbb{R}^2$.

In fact, this is the actual definition in classical differential geometry — of course, this isn't completely general, for instance, we know that the Klein bottle cannot be embedded into $\mathbb{R}^3$.

So, why didn't we define something like this in Definition 46.2.2? The problem is, the concept of a diffeomorphism isn't defined on a Hausdorff topological space — in fact it can't be defined, right in the example above, you can see a homeomorphism that is not a

diffeomorphism — in other words, a topological space can be assigned different *smooth structures*.

So, the essence of what the definition Definition 46.2.2 is doing is, it implicitly defines what a *smooth structure* mean, by inducing the smooth structure from each patch $E_i \subseteq \mathbb{R}^n$ to the topological space M. The condition that the transition functions need to be smooth is, of course, to ensure that the smooth structures on M induced by different ϕ_i are the same.

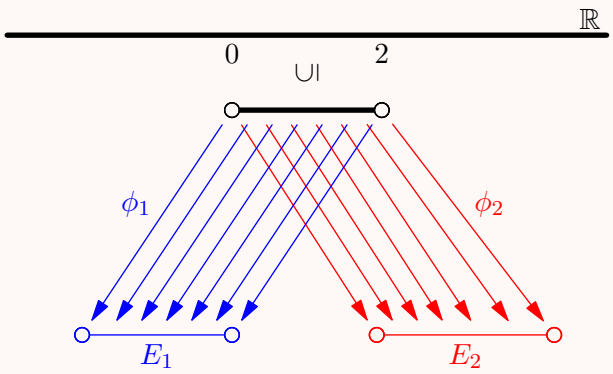
In completely the same way, we could have replaced Definition 46.1.2 by:

> A topological n-manifold M is a set with a collection of subsets $\{U_i\}$ that covers M, for each U_i there is a bijective map from it to a subset of $\mathbb{R}^n$, say
>
> $$\phi_i \colon U_i \to E_i \subseteq \mathbb{R}^n$$
>
> where each E_i is an open subset of $\mathbb{R}^n$, satisfying that all the transition maps are topological homeomorphisms.

Here, the ϕ_i are "set isomorphisms" and plays a similar role as the homeomorphisms in Definition 46.2.2, and the topological space structure is similarly induced from the patches E_i.

Example 47.1.2 ($(0,2)$ is a smooth 1-manifold)

Just as above, the open interval $(0,2)$ included in $\mathbb{R}$ can also be considered a smooth manifold.



This time around, ϕ_1 is the same as above, but $\phi_2 \colon (0,2) \to (0, 2+\frac{1}{e})$ is defined by

$$\phi_2(x) = \begin{cases} x & \text{if } x \le 1 \\ x + e^{-1/(x-1)} & \text{otherwise.} \end{cases}$$

Because all of ϕ_1, ϕ_2, and their inverses are smooth functions, the transition maps $\phi_1 \circ \phi_2^{-1}$ and $\phi_2 \circ \phi_1^{-1}$ are thus smooth, satisfying the hypothesis of Definition 46.2.2.

You should take a moment to think through this idea — because smooth functions on $\mathbb{R}^n$ are so natural, it's easy to forget that a smooth manifold carries more structure than just the topology.

Once again, as we have seen in the example above, $\mathbb{R}^n$ has more structure than just being smooth — it has an *analytic structure*. The chart ϕ_2 does not preserve this structure.

So, for Riemann surface, we will just have:

A Riemann surface is a smooth (real) 2-manifold which locally looks like $\mathbb{C}$, and carries an *complex-smooth structure*.

Of course, by the miracle of complex analysis — holomorphic functions are analytic! — this is equivalent to stating that a Riemann surface carries a complex-analytic structure.

§47.2 Riemann surface

Prototypical example for this section: The Riemann sphere, or any open subset of $\mathbb{C}$ such as $\{z \in \mathbb{C} \mid |z| < 1\}$.

From the motivation above, the definition of a Riemann surface naturally falls out:

Definition 47.2.1 (Riemann surface)**.** A **Riemann surface** X is a second countable connected Hausdorff space with an open cover $\{U_i\}$ of countably many sets homeomorphic to open subsets of $\mathbb{C}$, say by homeomorphisms

$$\phi_i \colon U_i \xrightarrow{\cong} E_i \subseteq \mathbb{C}$$

such that the **transition maps** ϕ_{ij} defined by

$$\phi_{ij} \colon E_i \cap \phi_i^{\text{img}}(U_i \cap U_j) \xrightarrow{\phi_i^{-1}} U_i \cap U_j \xrightarrow{\phi_j} E_j \cap \phi_j^{\text{img}}(U_i \cap U_j).$$

are analytic functions. Each ϕ_i is called a **complex chart**, and together they form a **complex atlas**.

We say that the complex atlas gives the Hausdorff space a **complex structure**. Thus, in other words, a Riemann surface is a (second countable, connected, Hausdorff) topological space with a complex structure.

[Mi95] has an alternative definition, by a maximal complex atlas. Both definitions are the same, but in practice, it's easier to specify finitely many complex charts than specifying infinitely many ones.

A complex chart $U_i \to E_i$ should be think of as giving a **local coordinate** on U_i. Formally:

Definition 47.2.2. For a point $p \in X$, open set $U \subseteq X$ and complex chart $\phi \colon U \to \mathbb{C}$, let $z = \phi(x)$ for each $x \in U$, we call z a **local coordinate**. We say that the local coordinate is **centered** at p if $\phi(p) = 0$.

§47.3 Complex manifold

Analogously to the definition of a real n-manifold, we can define a complex manifold. Just as above, the structure has much more rigidity than a smooth surface.

Definition 47.3.1 (Complex n-manifold)**.** A **complex n-manifold** is a Hausdorff space with an open cover $\{U_i\}$ of countably many sets homeomorphic to open subsets of $\mathbb{C}^n$, say by homeomorphisms

$$\phi_i \colon U_i \xrightarrow{\cong} E_i \subseteq \mathbb{C}^n$$

such that the **transition maps** ϕ_{ij} are analytic functions.

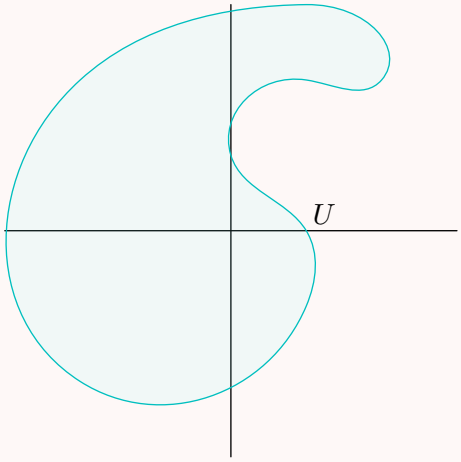
Of course, a complex n-manifold is naturally a smooth (real) $2n$-manifold.

§47.4 Examples of Riemann surfaces

In this chapter, several examples will be given.

Example 47.4.1 (Open subsets of $\mathbb{C}$)

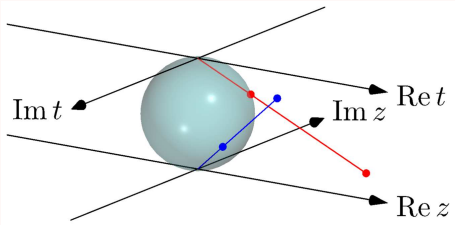
Any connected open subset $U \subseteq \mathbb{C}$ is a Riemann surface.



This is a boring example (the whole thing can be defined without any welding), but let's go on.

Example 47.4.2 (The Riemann sphere)

The Riemann sphere $\mathbb{C}_\infty$, as a smooth 2-manifold, is just a sphere.



Its complex structure is defined as follows:
Embed the sphere in $\mathbb{R}^3$ such that $N = (0,0,1)$ and $S = (0,0,0)$ are two antipodal points.
Let E_1 be the xy-plane, and let E_2 be the set of points with $z = 1$.
Then, let $\phi_1 \colon \mathbb{C}_\infty \setminus \{N\} \to E_1$ be the stereographic projection from the sphere (except the point N) to E_1 through the point N, and let $\phi_2 \colon \mathbb{C}_\infty \setminus \{S\} \to E_2$ be the stereographic projection from the sphere (except the point S) to E_2 through the point S.
We think of E_1 and E_2 as copies of the complex plane embedded in $\mathbb{R}^3$ by $z \mapsto (\operatorname{Re} z, \operatorname{Im} z, 0) \in E_1$ and $t \mapsto (\operatorname{Re} t, -\operatorname{Im} t, 1) \in E_2$. Then ϕ_1 and ϕ_2 are complex charts for $\mathbb{C}_\infty$.
The domain of ϕ_1 and ϕ_2 covers $\mathbb{C}_\infty$. To make $\mathbb{C}_\infty$ into a complex manifold, we must ensure that the complex structure induced by ϕ_1 and ϕ_2 are the same — indeed, over any open set U that contains neither N nor S, the projections are related by $\phi_1(p) = \frac{1}{\phi_2(p)}$ for all $p \in U$.

This also explains why the minus sign is needed in $t \mapsto (\operatorname{Re} t, -\operatorname{Im} t, 1)$ — otherwise, the projections will be related by $\phi_1(p) = \frac{1}{(\phi_2(p))}$, which is not analytic.

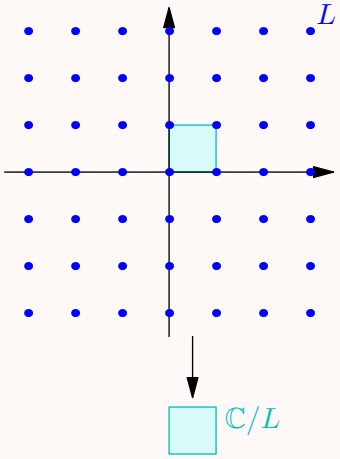
We can think of the Riemann sphere as the result of welding two copies of $\mathbb{C}$ together in order to "fill in" the missing point ∞.

In the example above, the local coordinate given by ϕ_1 is centered at S, and the local coordinate given by ϕ_2 is centered at N. The point $\phi_1^{-1}(4)$ would have local coordinate $z = 4$ under the chart ϕ_1, and local coordinate $t = \frac{1}{4}$ under the chart ϕ_2.

Example 47.4.3 (The complex torus)

Let L be the set $\mathbb{Z}[i]$ of complex numbers with both real and imaginary parts of $\mathbb{C}$. Then L forms an additive subgroup of $\mathbb{C}$.

Consider the quotient $\mathbb{C}/L$. The quotient map $\mathbb{C} \to \mathbb{C}/L$ induces a natural complex structure on $\mathbb{C}/L$.



Here we draw $\mathbb{C}/L$ as a square, but you should imagine that the top and bottom edge, as well as the left and right edges, are smoothly welded together.

For each small patch of the torus, we can isomorphically map it to $\mathbb{C}$ by taking a suitable component of the preimage of the quotient map — the different choices of the projection are related by transition functions $\phi_{ij}(x) = x + a$ for $a \in L$, this is analytic.

The complex torus is compact, thus any holomorphic function on $\mathbb{C}/L$ is constant. Meromorphic functions are more interesting, and also difficult to construct.

And some non-examples.

Example 47.4.4

The disjoint union of two Riemann spheres is not a Riemann surface, because it is not connected.

The condition that a Riemann surface must be connected is merely a technical condition such that theorems are nice — we don't lose much by requiring this, because any topological space with a complex structure can be broken down into disjoint union of

Riemann surfaces, one for each connected component.

48 Morphisms between Riemann surfaces

§48.1 Definition

The definition is what we would expect — since a Riemann surface's main feature is a complex structure, a map $f\colon \mathbb{C} \to \mathbb{C}$ is a morphism between Riemann surfaces if and only if it is holomorphic.

Definition 48.1.1. Let X and Y be Riemann surfaces. A mapping $f\colon X \to Y$ is holomorphic at $p \in X$ if and only if there exists charts $\phi_1\colon U_1 \to E_1$ on X with $p \in U_1$ and $\phi_2\colon U_2 \to E_2$ on Y with $f(p) \in U_2$ such that the composition $\phi_2 \circ f \circ \phi_1^{-1}$ is holomorphic at $\phi_1(p)$. We say f is a **morphism between Riemann surfaces** if and only if it is holomorphic at all points of X.

In other words: f is holomorphic if and only if it is holomorphic as function mapping between local coordinates.

Example 48.1.2

Some examples follows.

- The function $f\colon \mathbb{C} \to \mathbb{C}$ by $f(x) = x^3$ is a morphism.

  Note that this function is not bijective. At each point $p \neq 0$, there is an open neighborhood on which f has an inverse, but f has no inverse at 0.

- The embedding of the complex plane into the Riemann sphere, $\mathbb{C} \hookrightarrow \mathbb{C}_\infty$, is a morphism.

§48.2 Functions to the Riemann sphere

Prototypical example for this section: The meromorphic function $\frac{1}{z}$ can be made into a holomorphic $\mathbb{C} \to \mathbb{C}_\infty$ function.

In this section, we will see that the Riemann sphere $\mathbb{C}_\infty$ can be viewed as "$\mathbb{C}$ with a point at infinity added". This interpretation allows us to interpret meromorphic functions $f\colon X \to \mathbb{C}$ as holomorphic maps $g\colon X \to \mathbb{C}_\infty$, which allows a much better handling of meromorphic functions — there's no longer any singularity, the resulting function g is holomorphic everywhere!

First, we see that $\mathbb{C}_\infty$ can be naturally interpreted as $\mathbb{C}$ with a single point added: With notation as in Example 47.4.2, identify $\mathbb{C}_\infty \setminus \{N\}$ with E_1 (and thus with $\mathbb{C}$) through the map ϕ_1, and we let ∞ be the point N.

Question 48.2.1. Convince yourself that it makes sense to call the point ∞ — for every sequence of points $\{z_i\}$ on $\mathbb{C}$ such that $|z_i| \to +\infty$, then $\phi_1^{-1}(z_i) \to \infty$ on $\mathbb{C}_\infty$ as a topological space.

So, let X be a Riemann surface, and $f\colon X \to \mathbb{C}$ be a meromorphic function on X. Naturally, g can be defined by

$$g(z) = \begin{cases} f(z) & \text{if } f(z) \neq \infty \\ \infty & \text{if } f(z) = \infty. \end{cases}$$

Then g is continuous — but furthermore, it's analytic.

Question 48.2.2. Clearly, at points $z \in X$ where $g(z) \neq \infty$, then g is analytic. Convince yourself that g is also analytic at $z \in X$ where $g(z) = \infty$. (With notation as in Example 47.4.2, take a small open set $U \subseteq X$, and re-parametrize $g^{\text{img}}(U) \subseteq \mathbb{C}_\infty$ by $t = 1/z$.)

Therefore,

Proposition 48.2.3

There is a one-to-one correspondence between meromorphic functions $f\colon X \to \mathbb{C}$ and holomorphic maps $g\colon X \to \mathbb{C}_\infty$ such that g is not identically ∞.

Or, more informally,

Plugging in the hole at ∞ of $\mathbb{C}$ allows us to analytically extend meromorphic functions to $\mathbb{C} \cup \infty$ maps which is holomorphic everywhere.

§48.3 Some other nice properties

We have just seen in the last section that the Riemann sphere $\mathbb{C}_\infty$ allows us to remove the singularities of a meromorphic functions.

Informally speaking, this is because $\mathbb{C}_\infty$ is a "compactfication" of $\mathbb{C}$ — adding a point to make it compact — and compact Riemann surfaces enjoy many nice properties.

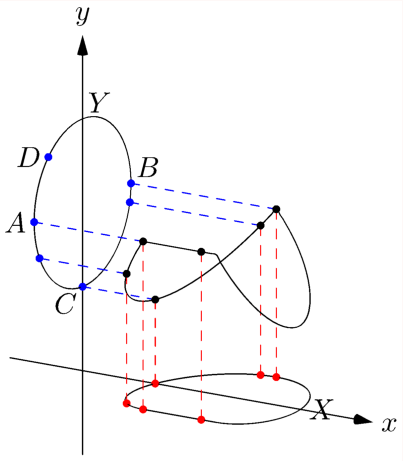
Proposition 48.3.1

Let X and Y be compact, $f\colon X \to Y$ be holomorphic and not constant. For each point $y \in Y$, define d_y be the total multiplicity of the points in the preimage of y. Then, d_y is well-defined and constant.

You can see why this proposition is surprising:

Example 48.3.2 (The proposition does not hold for smooth compact manifolds)

Consider the following function $f\colon X \to Y$ between compact smooth real 1-manifold, depicted as a plot with x and y-axis. (Note that a compact 1-manifold cannot be embedded into $\mathbb{R}$, because compact subsets of $\mathbb{R}$ are closed and bounded, thus necessarily have a boundary. A proper graph would live in a 4-dimensional space, which is rather difficult to visualize, so we settle with an approximate representation.)



Here, X and Y are both isomorphic to the unit circle.
We count the number of points in the fiber above each point in Y:

- Above point A, there are infinitely many points.
- Above point B, there is only one point. (You can argue that this point has "multiplicity 2" however)
- Above point C, there are two points.
- Above point D, the fiber is empty.

Definition 48.3.3. The value d_y above is called the **degree** of the map f, written $\deg(f)$.

Example 48.3.4

The map $z \mapsto z^k$, when extended to a $\mathbb{C}_\infty \to \mathbb{C}_\infty$ map, has degree k.

If $t \neq 0$, then we know that t has k distinct k-th roots. But if $t = 0$, its preimage only consist of the point 0 — in this case, we wish to say $z = 0$ is a "multiple point" — we will formalize it next section when we defines the multiplicity of a map.

If you have read Section 74.1, this is in fact the same as the concept of a degree in homology when X and Y are both Riemann spheres — it counts how many spherical bags that im f consists of. But, in this case, the theory is extra nice — not only that the graph is homotopy equivalent to one that covers each point d times, but each point is in fact covered *exactly* d times!

This theme will be recurrent in complex analysis and Riemann surfaces. Basically:

If the "things" are counted properly, the formula is very nice.

The proof of the proposition is not difficult — the main observation is that the theorem is true for functions of the form $f(z) = z^n$, and locally around each point $p \in X$, f is either an isomorphism or has the form above. So d_y is locally constant, and thus constant because Y is connected.

§48.4 Multiplicity of a map

Prototypical example for this section: $f(x) = (x-3)^2$ *has* $\text{mult}_3(f) = 2$.

In the previous section, we informally talk about the multiplicity of a map at a point. We will rigorously define it in this section.

Example 48.4.1

Consider the map $f\colon \mathbb{C} \to \mathbb{C}$ given by $f(z) = z^5 + 1$.
Above each point $y \in \mathbb{C}$, the fiber $f^{\text{pre}}(y)$ has 5 points — except when $y = 1$, then $f^{\text{pre}}(1) = \{0\}$ has only 1 point.

This behavior is *undesirable*, and we would like to say that the function f maps 5 "identical copies" of the point 0 to the point 1. (Another way you could see it is that, for each sequences $\{y_i\}$ converging to 1, there are 5 different sequences $\{x_i\}$ converging to 0 such that $f(x_i) = y_i$ for each i.)

Inspired by this, we will define multiplicity in a way such that:

- $z \mapsto z^m$ has multiplicity m, for integer $m \geq 1$.
- If we perform an analytic reparametrization of the source or the target, then the degree does not change.

Turns out these two properties completely defines the degree! We have the following.

Proposition 48.4.2

Let $f\colon X \to Y$ be a nonconstant holomorphic map defined at $p \in X$. Then there is a unique integer $m \geq 1$ such that, for every chart $\phi_2\colon U_2 \to V_2$ on Y centered at $f(p)$ (that is, $\phi_2(f(p)) = 0$), there is a chart $\phi_1\colon U_1 \to V_1$ on X centered at p such that the induced map $\phi_2 \circ F \circ \phi_1^{-1}\colon V_1 \to V_2$ has the form $z \mapsto z^m$.

In other words, once we fix a chart of Y, there exist a chart of (an open subset of) X such that the induced map between open subsets of $\mathbb{C}$ is a power map; furthermore, the exponent is independent of the selection.

Every map looks locally like $z \mapsto z^m$.

Proof. Essentially, use the Taylor expansion to determine m, then the selection of ϕ_1 is pretty much fixed by the restrictions. □

Definition 48.4.3. The value m above is the **multiplicity** of f at point p, written $\text{mult}_p(f)$.

Example 48.4.4 (More examples of multiplicity of a map at a point)

We consider some examples.

- The function $z \mapsto z^{-2}$, extended to a $\mathbb{C} \to \mathbb{C}_\infty$ map, has multiplicity 2 at point 0 — "two copies" of the point 0 is mapped to the point ∞.
- The function $f(z) = (z-1)(z-2)^5$ has $\text{mult}_2(f) = 5$ — more generally, if p

is a root of f, then $\text{mult}_p(f)$ is the multiplicity of the root.

- The function $z \mapsto z+1$ has multiplicity 1 everywhere — in fact, the multiplicity of a nonconstant map at "most" points will be 1.

These are the official terms:

Definition 48.4.5. A point p such that $\text{mult}_p(f) > 1$ is called a **ramification point**. In that case, the point $f(p)$ is called a **branch point**.

§48.5 The sum of the orders of a meromorphic function

Yet another case where we get a nice formula.

Example 48.5.1

Let us consider some meromorphic $\mathbb{C}_\infty \to \mathbb{C}_\infty$ functions (defined by extending a $\mathbb{C} \to \mathbb{C}$ function the obvious way), and list the zeros and poles of it (with multiplicity).

Function	Zeros	Poles
5	None	None
$(x+1)^2$	$-1, -1$	∞, ∞
$\frac{1}{x^2+1}$	∞, ∞	$i, -i$
$\frac{x+1}{x+2}$	-1	-2

Every time, the number of zeros equals the number of poles. This is not a coincidence!

Proposition 48.5.2

Let $f\colon X \to \mathbb{C}$ be a nonconstant meromorphic function on a compact Riemann surface X. Then

$$\sum_p \text{ord}_p(f) = 0.$$

Of course, we need X to be compact — there certainly are $\mathbb{C} \to \mathbb{C}$ functions that has several zeros, but no poles.

Proof. We extend f to a $X \to \mathbb{C}_\infty$ function, then the sum of multiplicities of points in the fiber of 0 is equal to that in the fiber of ∞. □

§48.6 The Hurwitz formula

write this one. It's quite nice actually

§48.7 The identity theorem

The following propositions are expected — the same behavior is seen in complex analysis with holomorphic functions.

Theorem 48.7.1

Let $f, g \colon X \to Y$ be holomorphic maps between Riemann surfaces. If $f = g$ on a nonempty open subset of X, then $f = g$.

This is the analog of Problem 31C*. Note that here the assumption that X is connected is used — the disjoint union of two copies of $\mathbb{C}$ is a smooth 2-manifold, but not a Riemann manifold.

That is,

Holomorphic maps are *rigid* — the value of a function on a tiny subset determines its value everywhere.

49 Affine and projective plane curves

In this chapter, we will define affine and projective plane curves. This has two purposes:

- Many interesting curves in $\mathbb{R}^2$ can be defined as the set of roots of a polynomial. This is just a natural generalization.
- We will see that, in fact, *every* compact Riemann surfaces can be written as a projective curve! Thus, by studying the projective curves, we have in fact studied all compact Riemann surfaces.

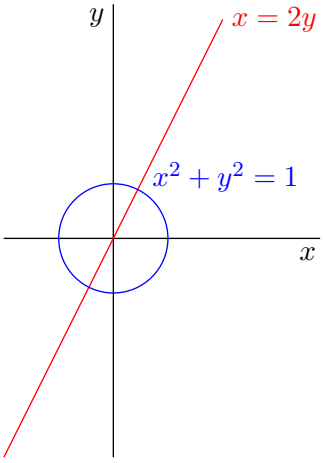
We will see what these means in the following sections.

§49.1 Affine plane curves

Consider some familiar curves on the plane.

- A line can be represented by an equation $y = ax + b$, or $x = c$.
- A circle can be written as the set of $y = \pm\sqrt{1 - x^2}$ for $-1 \le x \le 1$.

There is not much going on so far, but here is a picture.



As you can see, the definitions above are actually quite clumsy. We can do better by defining the points on the curve *implicitly*:

- A line can be represented as the set of (x, y) such that $ax + by + c = 0$.
- A circle can be represented as the set of (x, y) such that $x^2 + y^2 = 1$.

Of course, this way it is harder computationally to compute the coordinate of a point, but the definition is nicer.

The point is:

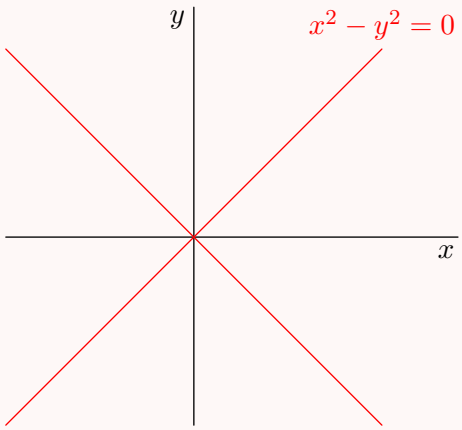
Many of the interesting curves can be written as the set of roots of a polynomial.

So we will try to do the same here — intuitively, if we start with complex dimension 2 and specify one polynomial, then the remaining part has complex dimension 1 i.e. a Riemann surface.

First, there is a technical detail we need to sort out — the set of roots of a polynomial need not be a smooth curve.

Example 49.1.1

The set of roots of $x^2 - y^2 = 0$ in $\mathbb{R}^2$ is not a curve near the origin — there are two intersecting curves.



This can be easily handled by placing a restriction on the polynomial. Let $f(x, y)$ a polynomial, and $X = \{(x, y) \in \mathbb{R}^2 \mid f(x, y) = 0\}$. Then:

Theorem 49.1.2

For a point $(x, y) \in X$ such that not both $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ vanishes, then X is smooth near (x, y).

If at a point $(x, y) \in X$ such that $\frac{\partial f}{\partial x} \neq 0$ or $\frac{\partial f}{\partial y} \neq 0$, we say X is **smooth** or **nonsingular** at (x, y).

In fact, we have something more. With notation as above, let $(x, y) \in X$, then:

- Suppose $\frac{\partial f}{\partial x} \neq 0$, then near the point (x, y), X can be parametrized by $x = g(y)$ for some analytic function g.
- Suppose $\frac{\partial f}{\partial y} \neq 0$, then near the point (x, y), X can be parametrized by $y = h(x)$ for some analytic function h.

All these are just the implicit function theorem.

Exercise 49.1.3. Check the statement above on the circle $x^2 + y^2 = 1$, at the points $(0, 1)$ and $(1, 0)$.

The exact same statement holds if we replace $\mathbb{R}^2$ with $\mathbb{C}^2$.

Next, we want the set of roots X to actually be a *Riemann surface*, not just a set of points in $\mathbb{C}^2$. So, we would need to find a suitable analytic structure on X.

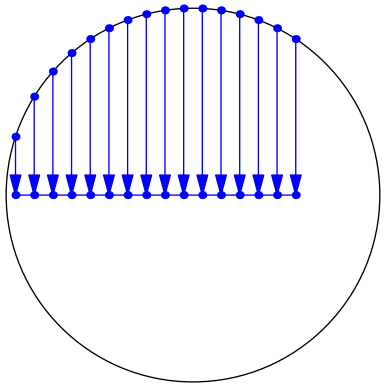
In the circle above, what would be a suitable analytic structure? One possible thought is to unroll the circle by arc-length and map it onto $\mathbb{R}$, but for a Riemann surface this

isn't even well-defined — how would you unroll, let's say a sphere onto a plane?

Another possibility is, given the statement of the implicit function above, we declare:

- On an open set $U \subseteq X$ where $\frac{\partial f}{\partial x} \neq 0$ for all points in U, suppose U is small enough such that X can be parametrized by $x = g(y)$ for some analytic function g, then the map ϕ such that $\phi(x, y) = \phi(g(y), y) = y$ is a complex chart.
- Similar for the $\frac{\partial f}{\partial y} \neq 0$ case.

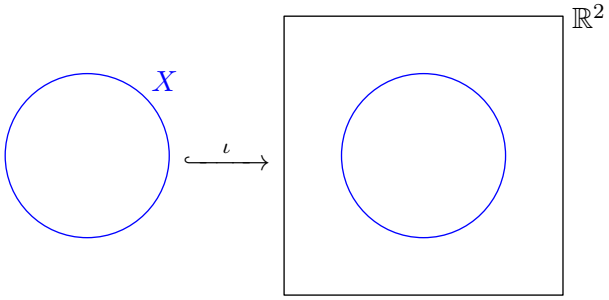
A possible complex chart is depicted below. Intuitively, the fact that $\frac{\partial f}{\partial y} = 0$ at the two points $(1, 0)$ and $(-1, 0)$ reflects that this "project-to-x" complex chart cannot be used at these points.



Actually, in the real analytic case, the two definitions above are equivalent. You can optionally do the exercise below.

Exercise 49.1.4. Show this for the circle above. (One possibility is to write down an explicit formula for the arc length and show it is analytic)

While this definition is already somewhat natural, there is something more to this. In category theory, we study properties of objects by studying the maps between them. The set X above has a natural map — the inclusion map into $\mathbb{R}^2$, and $\mathbb{R}^2$ has an obvious existing analytic structure.



The analytic structure defined above is natural in the following sense:

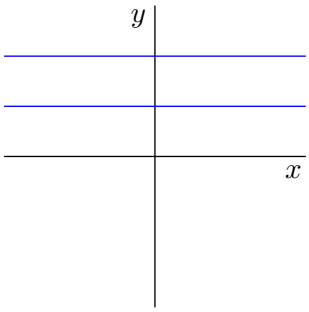
- For a function g such that $Y \xrightarrow{g} X \xrightarrow{\iota} \mathbb{R}^2$, then g is analytic if and only if $\iota \circ g$ is analytic.
- For a function $X \xrightarrow{\iota} \mathbb{R}^2 \xrightarrow{g} Y$, then g is analytic if and only if $g \circ \iota$ is analytic.
- $X \xrightarrow{\iota} \mathbb{R}^2$, then ι is analytic, and for any other complex structure $X' \xrightarrow{\iota'} \mathbb{R}^2$ such that ι' is analytic, there exists an unique analytic map $X' \to X$.

In fact, each of the bullet point uniquely determines the complex structure on X.

In some sense, this is like a universal property for our natural analytic structure.

Of course, we haven't defined what an analytic real manifold is. Brave readers may try to rigorously formalize all these concepts and prove the statement above.

There is another technical detail that needs to be sorted out. The set of zeros of $f(x,y) = (y-1)(y-2)$ is:



This is certainly smooth — but it's not connected. We required a Riemann surface to be connected.

Apart from these two issues, our final statement is:

Definition 49.1.5. Given a polynomial $f(z,w) \in \mathbb{C}[z,w]$, let $X = \{(z,w) \in \mathbb{C}^2 \mid f(z,w) = 0\}$ be the set of roots of f. Suppose that X is connected, and for all $(z,w) \in X$, f is smooth at (z,w) (that is, either $\frac{\partial f}{\partial z} \neq 0$ or $\frac{\partial f}{\partial w} \neq 0$). Then, X is a Riemann surface — we call X an **(smooth) affine plane curve**, with complex charts defined by:

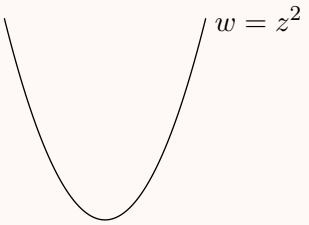
- On an open set U such that $\frac{\partial f}{\partial z} \neq 0$ everywhere on U, then $\phi \colon U \to \mathbb{C}$, $\phi(z,w) = w$ is a complex chart.
- On an open set U such that $\frac{\partial f}{\partial w} \neq 0$ everywhere on U, then $\phi \colon U \to \mathbb{C}$, $\phi(z,w) = z$ is a complex chart.

We call them affine because the plane is "flat", unlike the projective plane $\mathbb{CP}^2$ which is more "curved" in some sense.

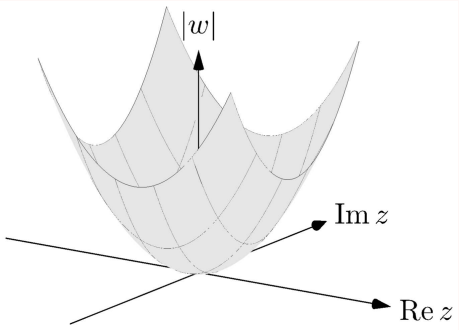
Of course, we should have some examples — with these tools, we are in a position to define an (affine) elliptic curve, and other affine curves.

Example 49.1.6 (A parabola)

Consider the Riemann surface cut out by $w = z^2$. Its real part looks like a parabola:



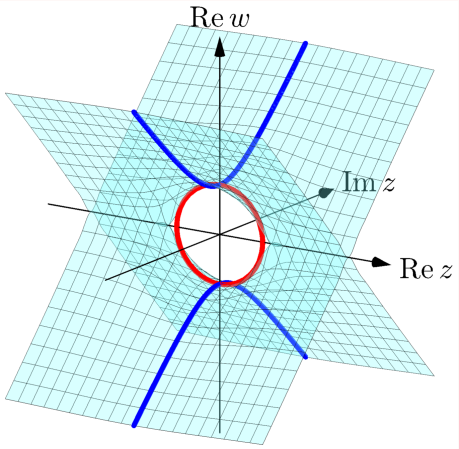
Since drawing a graph in 4 dimensions is difficult, we will project the Riemann surface onto 3 dimensions. The result is:



This Riemann surface is in fact isomorphic to the complex plane $\mathbb{C}$ by $(z, w) \mapsto z$.

Example 49.1.7 (The circle)

We all know what the real part of the circle looks like. Visualizing the whole Riemann surface is a bit more difficult, however.

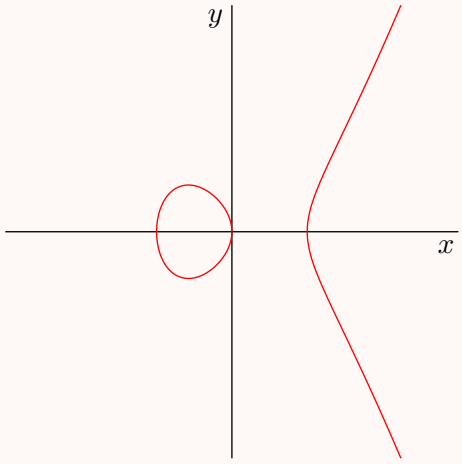


The highlighted red circle is the real part. Note that the fact that the plane is shown to be self-intersecting is merely an artifact of the projection.
Although the circle is not isomorphic to the complex plane $\mathbb{C}$ (we won't be able to prove this any time soon[a]), it is in fact isomorphic to the hyperbola $x^2 - y^2 = 1$ given by the transformation $y \mapsto y \cdot i$. With another rotation and multiplication by a constant, it is in turn isomorphic to the hyperbola $xy = 1$, which is "almost" isomorphic to the line $x = y$, missing one point $(0, 0)$.

[a] If you have read the homotopy chapter, this Riemann surface has a deformation retract to its real part — the circle, thus is homotopic to it. We know the complex plane $\mathbb{C}$ is nulhomotopic instead.

Example 49.1.8 (The elliptic curve $y^2 = x^3 - x$)

The real part looks like this. (The complex part is not drawn this time.)



While we won't be able to prove this any time soon, turns out this Riemann surface is not isomorphic to $\mathbb{C}$ — even if we allow deleting finitely many points.

§49.2 The projective line $\mathbb{CP}^1$

We will define the projective line — as it will turn out, it is isomorphic to the Riemann sphere $\mathbb{C}_\infty$ which we have already defined. So this section is only to show how our tools work.

As you might have guessed by the name: as a set of points, $\mathbb{CP}^1$ is the quotient of the set of points $\mathbb{C}^2 \setminus \{0\}$, modulo the relation $(x,y) \sim (\lambda x, \lambda y)$ for any $\lambda \in \mathbb{C} \setminus \{0\}$.

As a topological complex manifold, fortunately, it is still easy — $\mathbb{C}^2 \setminus \{0\}$ has a natural topology, and $\mathbb{CP}^1$ gets the quotient topology.

Exercise 49.2.1. Define the topology on the space $\mathbb{RP}^1$ analogously.

Exercise 49.2.2. Let $X \subseteq \mathbb{R}^2$ be a line that does not pass through the point $(0,0)$. Show that $X \xrightarrow{f} \mathbb{R}^2 \xrightarrow{q} \mathbb{RP}^1$ is an embedding i.e. $X \xrightarrow{q \circ f} \operatorname{im}(q \circ f) \subseteq \mathbb{RP}^1$ is a homeomorphism.

As a Riemann surface, the usual textbook definition goes:

Definition 49.2.3 (Complex structure of $\mathbb{CP}^1$)**.** Cover $\mathbb{CP}^1$ by two open sets, U_1 consisting of points with nonzero x coordinate, and U_2 consisting of points with nonzero y-coordinate. Then the two complex charts $\phi_1 \colon U_1 \to \mathbb{C}$ given by $\phi_1(x,y) = y/x$ and $\phi_2 \colon U_2 \to \mathbb{C}$ given by $\phi_2(x,y) = x/y$ determines a complex structure.

And goes on to prove that the two open sets indeed cover the whole of $\mathbb{CP}^1$, the value y/x is well-defined, transition maps are holomorphic, etc.

The definition above is elementary, but uninstructive. Where does the complex charts come from?

Given what we have done in the previous chapter, it should be obvious where we should go from here. There are two things to try:

- Let X be an affine plane curve in $\mathbb{C}^2$ that does not contain the point 0. Then the map $X \hookrightarrow \mathbb{C}^2 \twoheadrightarrow \mathbb{CP}^1$ should be an isomorphism whenever some certain derivative does not vanish.

- We can also use maps: the complex structure is such that whenever we have $Y \xrightarrow{f} \mathbb{C}^2 \xrightarrow{q} \mathbb{CP}^1$ or $\mathbb{C}^2 \xrightarrow{q} \mathbb{CP}^1 \xrightarrow{g} Y$, then f is analytic if and only if $q \circ f$ is analytic; and g is analytic if and only if $g \circ q$ is analytic.

Both are equivalent to the definition above — in fact, the definition is merely a special case of the first bullet point, where X is taken to be the line $x = 1$ and $y = 1$ respectively. Coincidentally, the 2 resulting complex charts is the simplest one to write down algebraically, and they already cover the whole $\mathbb{CP}^1$, so it is often taken to be the definition. There is no reason why it must be these 2 lines however — you might as well use $x + y = 1$ and $x - y = 1$.

§49.3 Projective plane curves

Instead of using affine plane curves $X \subseteq \mathbb{C}^2$, this time around, we will define projective plane curves $X \subseteq \mathbb{CP}^2$.

Apart from "just another source of example", projective plane curves have a distinctive advantage — *they're compact*! This allows many nice properties to hold — we have seen a few in the last chapter.

We start with defining the **projective plane** $\mathbb{CP}^2$. Of course it is $\mathbb{C}^3 \setminus \{0\}$ quotient by the relation $(x, y, z) \sim (\lambda x, \lambda y, \lambda z)$. It has a natural 2-dimensional complex structure induced from $\mathbb{C}^3$ by the quotient map.

The above definition is natural, but abstract. Concretely, we can write:

Question 49.3.1. Define the three complex-manifold charts (on the open set where they're well-defined) by:

$$\begin{aligned}\phi_0(x, y, z) &= (y/x, z/x)\\ \phi_1(x, y, z) &= (x/y, z/y)\\ \phi_2(x, y, z) &= (x/z, y/z).\end{aligned}$$

Convince yourself that this complex manifold structure is the correct one.

Then, a projective plane curve X is defined to be the set of points (x, y, z) such that $f(x, y, z) = 0$ — again, satisfying certain smoothness and connectedness conditions. Unfortunately, if the polynomial were e.g. $f(x, y, z) = x - 1$, it will not be well-defined, as $f(1, 0, 0) = 0$ but $f(2, 0, 0) = 1$. So we require that f is homogeneous — that way, $f(x, y, z)$ is still not well-defined, but at least we know whether $f(x, y, z) = 0$.

The complex structure on a projective plane curve is similarly defined by the universal property.

The definition is short and natural, but abstract. A more concrete definition is given below.

Question 49.3.2. With notation as above, define U_0, U_1 and U_2 to be the domain of ϕ_0, ϕ_1 and ϕ_2 respectively. Note that $U_i \xrightarrow{\phi_i} \mathbb{C}^2$ gives an isomorphism between U_i and the affine plane $\mathbb{C}^2$.
Convince yourself that the intersection of a projective plane curve X with one of the U_i is a (possibly empty) affine plane curve, when mapped to $\mathbb{C}^2$, and all the mappings are isomorphisms.

We need some examples.

Example 49.3.3 (The Riemann sphere, again)

The Riemann sphere can alternatively be defined as the set of points where $z = 0$ in $\mathbb{CP}^2$.

There's nothing interesting about this — we already know how the Riemann surface looks like. It just serves as a trivial example.

Example 49.3.4 (An elliptic curve, again)

Let $f(x, y) = x^3 - x - y^2$. We know that the set of roots of f in the affine plane $\mathbb{C}^2$ is the elliptic curve.

Identifying $\mathbb{C}^2$ with U_2, most points in $\mathbb{CP}^2$ can be written as $(x, y, 1)$. We want to find a polynomial $g(x, y, z)$ such that its set of roots in $\mathbb{CP}^2$, restricted to U_2, equals to the elliptic curve.

Intuitively, by the identity theorem, this should suffices to uniquely determines the Riemann surface. Indeed, our target polynomial g is:

$$g(x, y, z) = x^3 - xz^2 - y^2z.$$

This is just the laziest way to homogenize the polynomial f, multiplying the least power of z to make the result a homogeneous polynomial, and that $g(x, y, 1) = f(x, y)$.

We have that $\mathbb{CP}^2$ is compact, and the set of roots of g is closed, therefore the resulting Riemann surface is *compact*! As promised.

visualize this

As it turns out, unlike the Riemann sphere, the Riemann surface defined by the elliptic curve above has *genus 1*! We have the first example that is definitely distinct from the Riemann sphere.

Exercise 49.3.5. In the example above, what if we multiply a larger power of z? For instance

$$g(x, y, z) = x^3z - xz^2 - y^2z.$$

Example 49.3.6 (A hyperelliptic curve)

Let $f(x, y) = (x - x_1)(x - x_2) \cdots (x - x_{2k+1}) - y^2$, where all of $x_1, \ldots, x_{2k+1}$ are distinct complex numbers.

We can homogenize f to get $g(x, y, z) = (x - x_1z)(x - x_2z) \cdots (x - x_{2k+1}z) - y^2z^{2k-1}$.

As above, the set of roots of g in $\mathbb{CP}^2$ cuts out a Riemann surface — once again, this has *genus* k!

Therefore, we have seen examples of compact Riemann surfaces of all the genera simply by picking different values of k.

Saying that we have "seen" the surfaces themselves is not quite accurate — but you can try to visualize these hyperelliptic curves the same way the elliptic curve is visualized.

§49.4 Filling in the holes

Prototypical example for this section: The Riemann sphere is formed by filling in a single hole in the complex plane $\mathbb{C}$.

write this

§49.5 Nodes of a plane curve

Prototypical example for this section: The set defined by the equation $x^2 - y^2 = 0$ has a simple node.

50 Differential forms

§50.1 Differential form on $\mathbb{C}$

In this section, we will generalize the definition of what can be contour integrated.

Definition 50.1.1 (Differential 1-forms on $\mathbb{C}$)**.** A 1-form ω on an open set $U \subseteq \mathbb{C}$ is an expression of the form $f(z)d\operatorname{Re} + g(z)d\operatorname{Im}$, where $f(z)$ and $g(z)$ are smooth $U \to \mathbb{C}$ functions.

Here, smooth means being infinitely differentiable when interpreted as $\mathbb{R}^2 \to \mathbb{R}^2$ functions.

This is almost the same as the definition of a 1-form on $\mathbb{R}^2$! Here, Re and Im takes the role of $\mathbf{e}_1^\vee$ and $\mathbf{e}_2^\vee$ the obvious way.

The only difference is, as you can observe, $f(z)$ and $g(z)$ returns complex numbers instead of real numbers — but this is mostly inconsequential, by the projection principle (Theorem 43.2.1), the 1-form ω is equivalent to a pair of real-valued 1-forms $(\operatorname{Re}\omega, \operatorname{Im}\omega)$.

The reason why we want to do what we did is simply for convenience — by abuse of notation, let z be the function $z \mapsto z$, then we want dz to be a 1-form that returns the change in z.

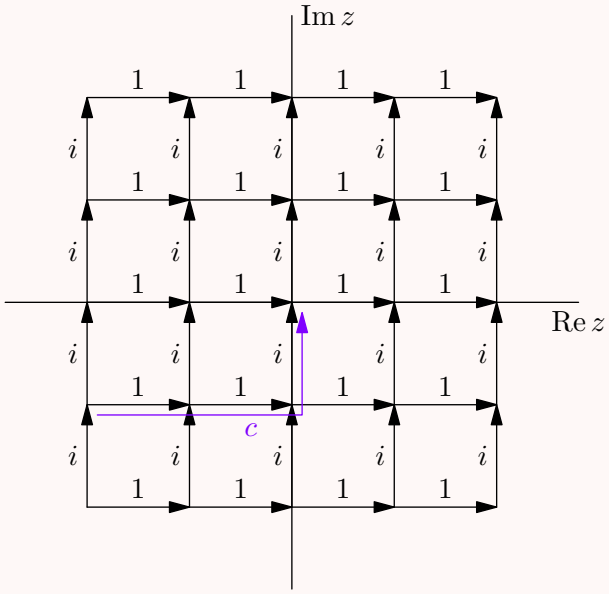
§50.2 Visualization of differential forms

Because ω takes in a point and returns a $\mathbb{R}$-linear map from the tangent space, the obvious way to visualize it is to draw a quiver diagram — for each point, the value of $\omega_p(v)$ is plotted for vectors v, which we interpret as "if we integrate a curve c in the direction of v, with length approximately the length of v, close to the point p, then the result is approximately the labeled value.

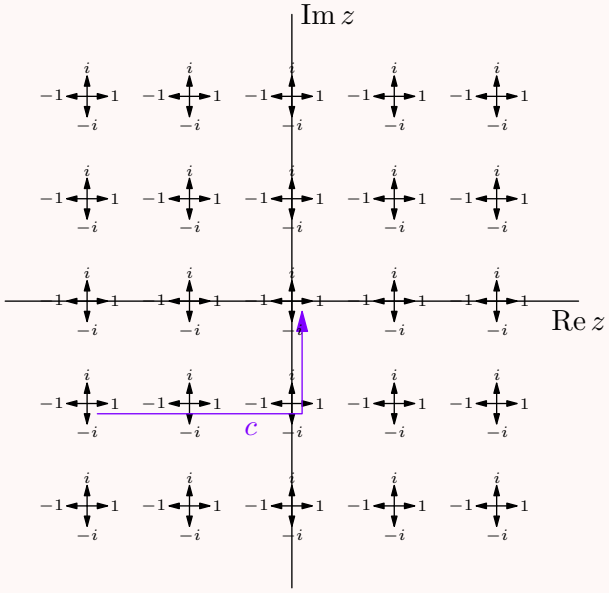
To integrate a differential form ω over a curve c, simply add up the numbers that corresponds to the direction of movement of c that appears in the diagram.

With that visualization, here are some examples.

Example 50.2.1 (The 1-form dz)
We may visualize dz, which is just the change in z, as follows.

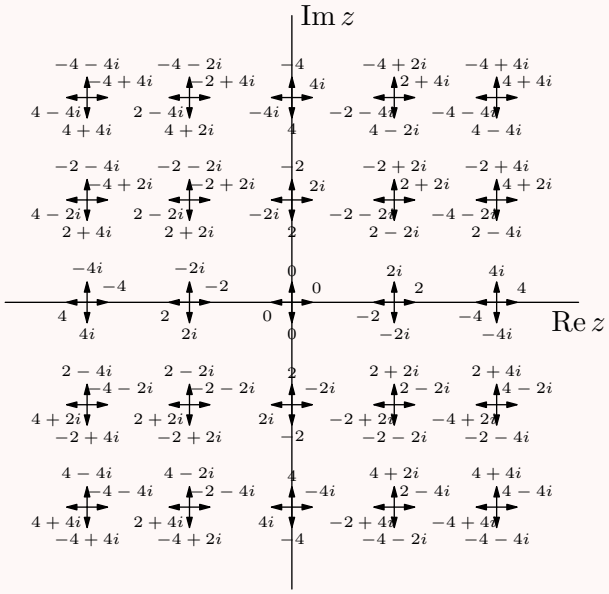


The integration $\int_c dz$ can be computed by adding up the value of the vectors together, so we get $2+i$ — this is indeed equal to the change of z as we travel along the curve, $0-(-2-i)=2+i$.
Since having the arrows extending the whole length can be confusing, we will shorten the arrow like the following.



Example 50.2.2 (Another holomorphic 1-form: $d(z^2) = 2z \cdot dz$)

While we have never defined what a holomorphic 1-form is, it makes intuitive sense for the definition to satisfy that: if $f(z)$ is holomorphic, then $df(z)$ should be a holomorphic 1-form.
In any case, if you perform the same procedure as above and compute the differential change of z^2 along the tangent vectors, you will get the following.



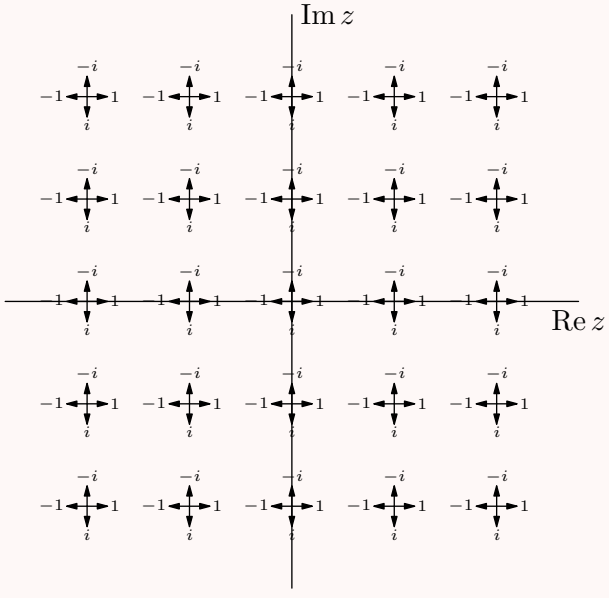
Unfortunately, it gets a bit cluttered regardless. Anyway, as you can see, at each point z and along each direction, the value of the 1-form $d(z^2)$ is $2z$ times the corresponding value of the 1-form dz, thus it makes sense for us to define multiplication such that $d(z^2) = 2z \cdot dz$.

Example 50.2.3 (A non-holomorphic form: $d\bar{z}$)

In both of the examples above, we see that, at each point z, $\omega_z(\mathbf{e}_2) = i \cdot \omega_z(\mathbf{e}_1)$, where $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$ — in other words, rotating the vector counterclockwise by 90 deg multiplies the value of the differential form by i.

The natural question would be: Is this the property of all 1-forms? Turns out it isn't. (Later on, we will see that this is a common property of all holomorphic 1-forms, or more generally, all type $(1, 0)$ 1-forms.)

Consider the following example: let $\omega = d\bar{z}$ — this is just the change in value of $\bar{z}$.



Example 50.2.4 (Another non-holomorphic form: $\overline{z} \cdot dz$)
Just as how a smooth function $f(z)$ is holomorphic if and only if it is complex-differentiable, we should define a holomorphic 1-form such that a smooth 1-form ω is holomorphic if and only if we can compute its differential $d\omega$.
We certainly haven't defined a 2-form yet, nor have we defined what it means to differentiate a 1-form ω to a 2-form.

§50.2.i Holomorphic forms

With the above examples in mind, we defines:

Definition 50.2.5 (Holomorphic 1-forms on the complex plane)**.** A 1-form ω is holomorphic if and only if it can be written as $f(z) \cdot dz$ for some holomorphic function f.

And also a few other types.

Definition 50.2.6 (Type $(1,0)$ and type $(0,1)$ 1-forms)**.** A 1-form is of **type** $(1,0)$ if it is locally of the form $f(z)dz$ for smooth f. Similarly, a 1-form is of **type** $(0,1)$ if it is locally of the form $f(z)d\overline{z}$ for smooth f.

Example 50.2.7 (Some type $(1,0)$ 1-forms)
Looking at the examples above:

- The holomorphic forms, dz or $2z \cdot dz$, are of course type $(1,0)$.
- $\overline{z} \cdot dz$ is still a type $(1,0)$ form, even though it is not holomorphic.
- $d\overline{z}$, however, is not a type $(1,0)$ form.

Why do we care? Note that it is nontrivial that the definition above is well-defined — it only makes sense because a holomorphic function scales every direction the same amount!

Intuitively,

A $(1,0)$ form ω is a form such that $\omega_p(\mathbf{e}_2) = \omega_p(\mathbf{e}_1) \cdot i$.

§50.2.ii Putting the pieces together: 1-forms on a Riemann surface

write this

Unsurprisingly, now we can define a 1-form on a Riemann surface.

51 The Riemann-Roch theorem

§51.1 Motivation

Recall a basic fact in complex analysis:

> A holomorphic $\mathbb{C} \to \mathbb{C}$ function is uniquely determined by its Taylor series expansion at the origin.

Compared to the case of real smooth function, this is already very rigid — the value of the function in a small neighborhood of the origin determines the value of the function everywhere — but, in order to specify a function, you still need infinitely many coordinates!

Meanwhile, we have Liouville's theorem:

> A bounded holomorphic $\mathbb{C} \to \mathbb{C}$ function is constant.

As we have learnt earlier, this theorem, when phrased in terms of Riemann surfaces, can be more elegantly rephrased to the following:

> A holomorphic $\mathbb{C}_\infty \to \mathbb{C}$ function is constant.

In other words, in order to specify a holomorphic $\mathbb{C}_\infty \to \mathbb{C}$, you only need a *single complex number*! That is, the $\mathbb{C}$-vector space $\operatorname{Hom}(\mathbb{C}_\infty, \mathbb{C})$ has dimension 1.

Naturally, you may ask, "is there anything inbetween"? There is! And the Riemann-Roch theorem is the main ingredient to understand how these things work.

So, how are we going to define this? If you compare the two situations above, a holomorphic $\mathbb{C} \to \mathbb{C}$ function is a meromorphic $\mathbb{C}_\infty \to \mathbb{C}$ function, which is allowed to have a pole at ∞, and nowhere else.

So,

By smoothly interpolate between "allow pole of arbitrary order" and "must be holomorphic", we can produce many interesting spaces of functions.

Conveniently, back in chapter 32, we have defined the **multiplicity** of a zero and the **order** of a pole of a meromorphic function. So, the natural point between these two extremes is to allow a pole of order at most d.

For notational convenience, we defines:

Definition 51.1.1 (Order of a meromorphic function)**.** Let f be meromorphic at p. We define $\operatorname{ord}_p(f)$ to be:

- d, if f has a zero of multiplicity d at p;
- $-d$, if f has a pole of order d at p;
- 0, otherwise.

Example 51.1.2 (The space of functions with pole of order at most 4 on $\mathbb{C}_\infty$)
Let $L(4 \cdot \infty)$ be the set of meromorphic $\mathbb{C}_\infty \to \mathbb{C}$ function, being holomorphic everywhere except ∞, and has a pole of order at most 4 at ∞ — in other words,

$$L(4 \cdot \infty) = \{f \text{ meromorphic on } \mathbb{C}_\infty \mid f \text{ defined on } \mathbb{C}_\infty \setminus \{\infty\}, \operatorname{ord}_\infty(f) \geq -4\}.$$

(The notation $L(-)$ will be explained later.)
Obviously, this forms a natural $\mathbb{C}$-vector space.
Consider the Taylor series of any $f \in L(4 \cdot \infty)$ at the origin:

$$f(z) = \frac{c_{-m}}{z^m} + \frac{c_{-m+1}}{z^{m-1}} + \cdots + \frac{c_{-1}}{z} + c_0 + c_1 z + \cdots$$

Obviously, because f is defined at the origin, it cannot have any nonzero coefficient c_{-m} for $m > 0$. But more importantly, it cannot have any nonzero coefficient c_m for $m > 4$ either![a]
Did you see what happened here? We start with requiring the function to be analytic and does not blow up too badly, and we end up with just the *algebraic* function — the polynomials!
In particular, $L(4 \cdot \infty)$ consists of the polynomials of degree ≤ 4, and

$$\dim L(4 \cdot \infty) = 5$$

as a $\mathbb{C}$-vector space.

[a] The reason is actually not very straightforward, but you can see for yourself why it is true: if there are only finitely many nonzero terms, then the order of the pole at ∞, $(-\operatorname{ord}_\infty(f))$, is precisely the degree of the highest nonzero coefficient.

Example 51.1.3 (More complicated $L(-)$ spaces)
There's no reason why we should restrict ourselves to considering only the functions that blow up at ∞ — as we will see, more general meromorphic functions can be considered, as long as we restrict the order of the poles.
Let $L(-1 \cdot 3 + 4 \cdot i + 5 \cdot \infty)$ be the set of meromorphic functions $f \colon \mathbb{C}_\infty \to \mathbb{C}$ that are:

- holomorphic everywhere in $\mathbb{C}_\infty$, possibly with the exception of the points 3, i, and ∞;
- at 3, it must have a root of order ≥ 1;
- at i, it cannot have a pole of order more than 4;
- at ∞, it cannot have a pole of order more than 5.

So, for example, $(z \mapsto z - 3)$, $(z \mapsto \frac{(z-3)^3}{(z-i)^2})$, or $(z \mapsto (z-3)^4)$ are functions in the set, but not $(z \mapsto (z-3)^2 + 1)$ or $(z \mapsto (z-3)^7)$.
As before, this is a $\mathbb{C}$-vector space, and furthermore, it is also finite-dimensional! What should its dimension be?
Well, note that there is a 1-1 bijection between functions $f \in L(-1 \cdot 3 + 4 \cdot i + 3 \cdot \infty)$ and functions $g \in L(-1 \cdot 3 + 7 \cdot \infty)$ by

$$g = \Phi(f) = (z \mapsto f(z) \cdot (z - i)^4),$$

where, as you could probably have guessed by now, $L(-1 \cdot 3 + 7 \cdot \infty)$ is the space of meromorphic functions that has at least a zero at 3 and at most a pole of order 7 at ∞.
Using that information, it shouldn't be too hard for you to see that the dimension should be 7.

For another piece of motivation: Later on, we will also define the concept of **divisors** and **line bundles**. If you have learned about these concepts in algebraic geometry context, you might be interested to learn what they are actually about; otherwise, it is still very surprising that these theorems can be naturally generalized to completely algebraic settings, and *your intuition from the case of analytic manifold will mostly work verbatim* — in fact, you can even define the genus of a number field, like $\mathbb{Q}[\sqrt{2}]$!

§51.2 Divisors

Prototypical example for this section: $(-3) \cdot i + (-4) \cdot \infty$ *is a divisor on* $\mathbb{C}_\infty$.

We start with defining a convenient notation for the above concepts.

First, observe that the condition "f must have a zero of order at most 4 at the origin" can be conveniently written as

$$z^4 \mid f.$$

In other words, z^4 must be a **divisor** of f.

This notation works if f is a polynomial, since we already know what it means for two polynomials to divide each other.

Generalizing, we could say "f cannot have a pole of order more than 3 at the point i, and f cannot have a pole of order more than 4 at the point ∞" by

$$(z-i)^{-3} \cdot (z-\infty)^{-4} \mid f.$$

Of course, at this point, the notation is purely formal — there is no interpretation as "functions" that could be assigned to the expression $(z-\infty)$, for instance.

Those objects are, appropriately enough, called **divisors**. So we come to the formal definition:

Definition 51.2.1 (Divisors)**.** Let X be a Riemann manifold, then a divisor D on X is a function $D\colon X \to \mathbb{Z}$, which is nonzero on a discrete set of points.

The formal objects $(z-i)^{-3} \cdot (z-\infty)^{-4}$ above, from now on, we will consider it as a function $D\colon \mathbb{C}_\infty \to \mathbb{Z}$ by

$$D(z) = \begin{cases} -3 & z = i \\ -4 & z = \infty \\ 0 & \text{otherwise.} \end{cases}$$

Abuse of Notation 51.2.2. For a point $p \in X$, we write p to mean the divisor that takes value 1 at p, and 0 elsewhere.

Because divisors are integer-valued functions, we can add two divisors together or multiply a divisor with an integer, the result is an integer. So,

Example 51.2.3 ($(z-i)^{-3} \cdot (z-\infty)^{-4}$ as a divisor)
The divisor D that corresponds to the formal object $(z-i)^{-3} \cdot (z-\infty)^{-4}$ above can be written as $(-3) \cdot i + (-4) \cdot \infty$.

§51.3 Degree of a divisor

Prototypical example for this section: $\deg((-3) \cdot i + (-4) \cdot \infty) = -7$.

If the surface X is compact, any discrete set of points is finite. Thus, a divisor D on X has finite support.

This allows us to define the degree of a divisor:

Definition 51.3.1 (Degree of a divisor)**.** For a divisor D on a compact surface X, its degree is $\sum_{p \in X} D(p)$.

Of course, the sum is well-defined because only finitely many terms are nonzero.

§51.4 The principal divisor of a meromorphic function

Prototypical example for this section: $\operatorname{div} \frac{(z-3)^2}{z-i} = 2 \cdot 3 + (-1) \cdot i + (-1) \cdot \infty$ *has degree* 0.

After defining a divisor, we want a convenient notation to formalize our fuzzy notation earlier of a divisor "divides" a function.

Definition 51.4.1 (Divisor of a meromorphic function)**.** Let f be meromorphic on a Riemann surface X. Then the divisor of f, $\operatorname{div}(f)$, is such that

$$\operatorname{div}(f)(p) = \operatorname{ord}_p(f).$$

We can also write it as a formal sum: $\operatorname{div}(f) = \sum_p \operatorname{ord}_p(f) \cdot p$ — by the abuse of notation above, this would be an actual sum if f has finitely many roots and poles.

If a divisor D is equal to $\operatorname{div}(f)$ for some f, we call D a **principal divisor**. (Compare this with a principal ideal, being an ideal generated by one element!)

Looking at the prototype example of this section, you might have guessed the following for the Riemann sphere. In fact, much more is true:

Proposition 51.4.2
If a divisor D on a compact surface X is principal, then $\deg D = 0$.

Let us not forget our goal of defining a convenient notation to talk about the space of functions with bounded poles. With the notation defined above, if $f = z$, $\operatorname{div} f = 1 \cdot 0 + (-1) \cdot \infty$, and we want to say f "divides" the divisor $(-1) \cdot \infty$. The natural definition would be:

Definition 51.4.3 (The partial ordering of divisors)**.** We write $D \geq 0$ if $D(p) \geq 0$ for all $p \in X$.

For two divisors D_1 and D_2, we write $D_1 \geq D_2$ if $D_1 - D_2 \geq 0$.

And finally,

Definition 51.4.4. Let D be a divisor on a Riemann surface X. Then the space of meromorphic functions with poles bounded by D is

$$L(D) = \{f \text{ meromorphic on } X \mid \operatorname{div}(f) \geq -D\}.$$

Exercise 51.4.5. This exercise is just for you to get familiar with the notation. Show the following:

- For two divisors $D_1 \leq D_2$, then $L(D_1) \subseteq L(D_2)$.
- If X is compact, then $L(0) \cong \mathbb{C}$.
- If X is compact and $\deg D < 0$, then $L(D) = \{0\}$.

§51.5 The Riemann-Roch theorem

Explain canonical divisor

Theorem 51.5.1 (The Riemann-Roch theorem)

Let D be a divisor on an algebraic curve X of genus g, and K be a canonical divisor on X. Then

$$\dim L(D) - \dim L(K - D) = \deg(D) + 1 - g.$$

List some applications

52 Line bundles

§52.1 Overview

You might have heard about line bundles, which is somehow "a set L with a map $\pi\colon L \to X$ where the preimage of each point is a line". And then, in the algebraic geometry section, you come across the concept of "section" which appears to be just a function.

That sounds reasonable, but you may ask, "so what? Isn't it then just another complex manifold which has one more dimension than X? Why not just study complex manifold?"

It's true, but there are more structures on a line bundle:

- You can take the product of two line bundles, which somehow "add up the twists" of both line bundles.
- A section is not just a function — you can think of a section as the graph of a function in the special case that the "graph paper" itself is flat, but if it is curved like a Möbius strip, you will see that there is no way to assign a "function value" to each point of the "graph paper" — a situation which we will call "the line bundle is not trivial".

In other words, a line bundle vastly generalizes the "space of the graph of a function".

Later on, you will see a deep hidden connection between line bundles and linearly equivalent classes of divisors, and how they are all linked by the so-called Picard group.

§52.2 Definition

Let X be a Riemann surface.

In this section, we will view X as just a curve — that is, a 1-dimensional object instead of a 2-dimensional object — because:

- It is easier to visualize things when they can be embedded in 3-dimensional space. (Try to draw the graph of ... with both real and complex part, and you will see what I mean!)
- Since all of our functions of interest are analytic, the behavior of a function elsewhere is determined by its value on the real axis.

Looking only at the real part can makes some intuition slipped however — for example, it is possible to overlook that the circle $x^2 + y^2 = 1$ and the hyperbola $x^2 = 1 + y^2$ cuts out Riemann surfaces in $\mathbb{C}^2$ of the same shape, or that the function $\frac{1}{x^2+1}$ has a pole at $x = \pm i$. So, be careful.

Definition 52.2.1. A **line bundle** L is a set, together with:

- A projection map $\pi\colon L \to X$,
- An open cover $\{U_i\}$ of X,
- For each U_i, a **line bundle chart** $\phi\colon \pi^{-1}(U) \to \mathbb{C} \times U$ that bijectively maps each point in $\pi^{-1}(p)$ to a point in $\mathbb{C} \times p$,

- For two open sets U_1 and U_2, the **transition function** $\phi_2 \circ \phi_1^{-1} \colon \mathbb{C} \times U_1 \to \mathbb{C} \times U_2$ must be a $\mathbb{C}$-*vector space isomorphism* restricted to $\mathbb{C} \times p \to \mathbb{C} \times p$ for each point $p \in U_1 \times U_2$, and the scaling factor must be an analytic function on U.

Remark 52.2.2 (Warning) — Typically, we draw a graph of the function $f(x)$ by the set of points (x, y) where $y = f(x)$.
This time, we use the notation in [Mi95] — the target of a line bundle chart is $\mathbb{C} \times U$ instead of $U \times \mathbb{C}$ — so if we consider a section the generalization of a function, the coordinate would look like (y, x) instead.

The definition is dense, but essentially:

A line bundle is a set with a line bundle structure, consisting of an analytic structure and a 1-dimensional vector space structure.

The transition maps is simply to weld the pieces of the line bundle together, just like how they welded pieces of a Riemann surface in Chapter 47.

Another definition, we will explain this one later.

Definition 52.2.3 (Sections of a line bundle)**.** Let L be a line bundle. A **section** on an open set U is a map $f \colon U \to L$ such that $\pi \circ f$ is the identity map on U.

We call a section $f \colon X \to L$ a **global section**.

The section $f \colon U \to L$ is an **analytic section** if for every $U_1 \subseteq U$ such that there is a line bundle chart $\phi \colon \pi^{-1}(U_1) \to \mathbb{C} \times U_1$, then $\phi \circ f|_{U_1} \colon U_1 \to \mathbb{C} \times U_1$ is analytic.

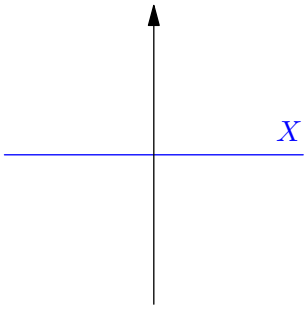
We will see this definition later on in algebraic geometry, Definition 82.2.2.

Remark 52.2.4 — In most books, they will first define what a sheaf is, then instead of "analytic section", they say "a section of the sheaf of analytic functions" (or regular functions, etc.)

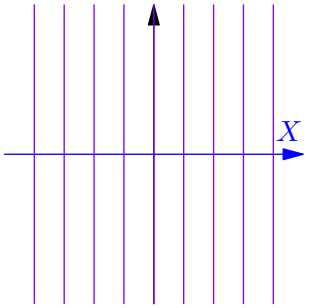
§52.3 Visualizing a line bundle

Just as how you can keep all the information of the Riemann sphere $\mathbb{C}_\infty$ in your head at once just by visualizing a sphere (with the analytic structure viewed as some "compatible grids" on the surface), you should also be able to keep all the information of a line bundle in your head at once — at least in the simplest cases.

First, we visualize $\mathbb{C} \times X$ where $X = \mathbb{C}$. Looking at only the real parts, it looks like a plane.

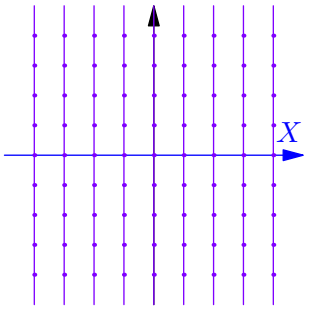


As a line bundle, the preimage of each point is a line.

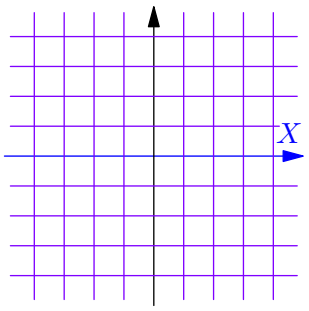


Question 52.3.1. In symbols, what subset of $\mathbb{C} \times X$ does a vertical line correspond to?

They are not just disparate lines however — there are two more structures. First one is a vector space structure — of course the dimension of $\mathbb{C}$ as a $\mathbb{C}$-vector space is 1. We can visualize it by marking the points $1, 2, 3, \ldots$ on the line.

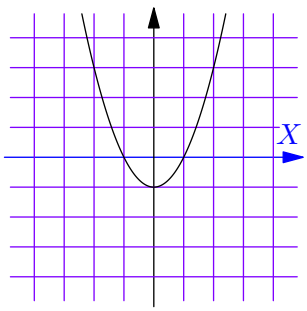


The other structure is that the lines must "smoothly varies" as p varies over X. We visualize this by drawing, well, a grid.



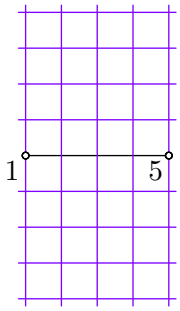
Question 52.3.2. How does the picture of the grid correspond to the formal definition of a line bundle chart? (Hint: take the preimage of the vertical lines $x = c$ and the horizontal lines $y = c$ with respect to the line bundle chart ϕ: $L \to \mathbb{C} \times U$, where $U \subseteq X$, then use the analytic structure on X to identify open subsets of U with open subsets of $\mathbb{C}$.)

So far, nothing surprising — this is just the usual grid graph, where we can draw functions on it like $y = x^2 - 1$, and a function is analytic if it is analytic with respect to the grid.

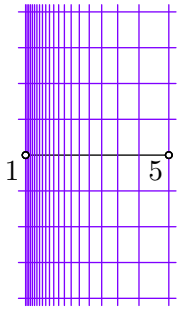


Of course, instead of a function, we call this a *section*. This particular section is in fact analytic, as you would expect.

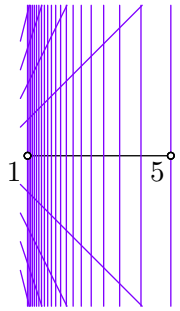
Let us take a look at which "grids" represent the same line bundle structure. For this part, we will look at $\{z \in \mathbb{C} \mid |z-3| < 2\}$, its real part being $(1,5)$.



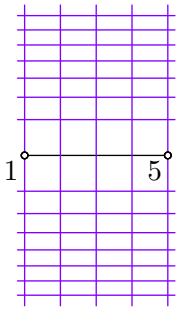
If we apply an analytic reparametrization on the segment $(1,5)$ — for example let $t = 25/x$, then the grid becomes like the following. It still represents the same line bundle structure — in other words, the two charts are compatible.



If we rescale the vertical direction by an analytically-varying function, it still represents the same line bundle structure.



However, if we rescale the vertical direction by something that is not linear, the vector space structure will be changed. The following grid *does not* represent the same line bundle structure:



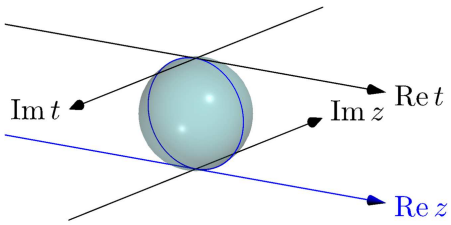
Intuitively, this makes sense — in a *vector space*, you can add two elements together and get another element — in our case, if $a, b \in L$ such that $\pi(a) = \pi(b)$, we can compute $c = a + b$ and get another element $c \in L$ with $\pi(c) = \pi(a) = \pi(b)$. If we rescale the vertical direction non-linearly, the element c will be changed.

Finally, don't forget that L still has an analytic structure — even though a section isn't necessarily a function, we are still able to say when a section is analytic.

Question 52.3.3. Verify that everything explained above matches the formal definition. (This is important! Fuzzy pictures won't help you to understand the concepts; and if your intuition is incomplete or inaccurate, you will have a lot of trouble understanding the subsequent parts.)

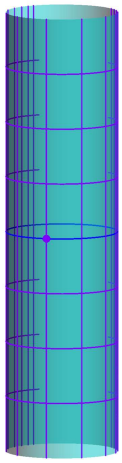
So far, everything just looks like a graph paper, on which a section looks just like a function.[1] Let us consider a more complicated space X — the Riemann sphere.

Because we are looking at the real part only, so once again, $\mathbb{C}_\infty$ looks like just a circle.



As before, we let z and t parametrize the points on the surface, with $t = \frac{1}{z}$ wherever both are defined.

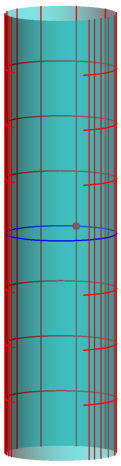
Still, we need two dimensions to embed a circle. So, the real part of $\mathbb{C} \times \mathbb{C}_\infty$ may looks something like the following:



[1] As warned above, "graph coordinate" is written (y, x).

The grid lines are drawn, and the origin $z = 0$, is marked with a dot. The vertical lines mark the position $z = 0$, $z = \frac{1}{2}$, $z = 1$, $z = \frac{3}{2}$, ...

On the opposite side, we may have something like the following. The vertical lines mark the position $t = 0$, $t = \frac{1}{2}$, $t = 1$, $t = \frac{3}{2}$, ...

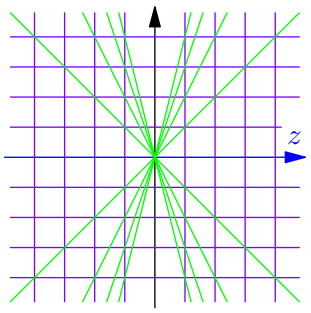


Question 52.3.4. Check that, on $\mathbb{C} \times U$ for any open set $U \subseteq \mathbb{C}_\infty$ that contains neither 0 nor ∞, the two line bundle charts above define the same line bundle structure. (What are the transition functions?)

So, we have the so-called trivial line bundle $\mathbb{C} \times \mathbb{C}_\infty$.

As promised, there are also nontrivial line bundles here.

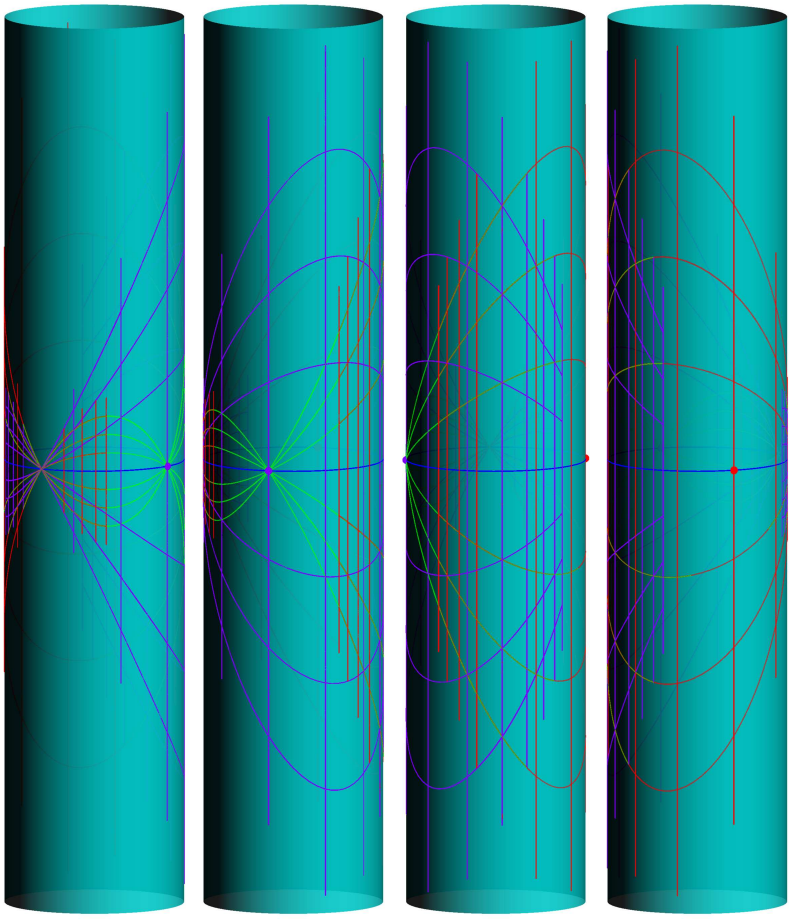
First, recall from the section above: over any open set U that contains neither 0 nor ∞, we can consider another line bundle chart that scales the vertical direction by a factor of z, this induces the same line bundle structure on U.



Question 52.3.5. Let ϕ_p and ϕ_g be the line bundle charts corresponding to the purple and green grid, respectively. Verify that if a point $q \in L$ satisfies $\phi_p(q) = (y, z)$ for $z \notin \{0, \infty\}$, then $\phi_g(q) = (\frac{y}{z}, z)$.

Now — note that the trivial line bundle $\mathbb{C} \times \mathbb{C}_\infty$ above can be seen as welding the two pieces together, such that the purple line (y, z) gets welded to the red line (y, t) for each y. There is nothing that restricts us to that specific welding method, however — this time around, we will try to weld the green line $(\frac{y}{z}, z)$ to the red line (y, t) for each y.

The thing will look like this. It looks quite complicated, so this time 4 views are shown.



The cylinder this time is only for illustrative purpose. Let us see what is going on.

- First, near the purple and the red point, the graph lines looks like our usual situation.

  Note that because $t = \frac{1}{z}$, looking from outside, the red coordinate lines will looks flipped.

- On the positive side ($z > 0$ and $t > 0$), no problem — we just need to squeeze the purple lines closer together — as depicted in the figure.

- On the negative side, however — note that the green line $(\frac{y}{z}, z)$ moves *downwards* when y increases, so we will need to "twist the graph paper" for it to go up.

  In the figure, this is depicted as a singularity where all the horizontal lines intersect, but in reality, you should think of it as we twisting the "graph paper" by 180 degrees and weld it to the other part.

This is a Möbius strip!

Thus, it appears to be obvious that this line bundle is not isomorphic to the trivial one, whatever "isomorphic" might mean.

Question 52.3.6. Check that what we did above makes sense when y, z and t are not real — in particular, $\mathbb{C} \setminus \{0\}$ is a connected set, unlike $\mathbb{R} \setminus \{0\}$. You probably won't be able to visualize the "graph paper" this time (it is 4-dimensional!), so you will have to keep your intuition confined in the real part and use algebra for the rest.

§52.4 Morphisms between line bundles

In order to formally define what it means for two line bundles to be isomorphic, we need to be able to define morphisms. It is exactly what you expect — it must respect the line bundle structure (that is, the vector space structure and the analytic structure) on L_1 and L_2.

Definition 52.4.1. Let $\pi_1 \colon L_1 \to X$ and $\pi_2 \colon L_2 \to X$ be line bundles. A line bundle morphism $\alpha \colon L_1 \to L_2$ is a set morphism such that:

- $\pi_2 \circ \alpha = \pi_1$, and
- if $\phi_1 \colon \pi_1^{-1}(U_1) \to \mathbb{C} \times U_1$ and $\phi_2 \colon \pi_2^{-1}(U_2) \to \mathbb{C} \times U_2$ are line bundle charts, then the composition

$$\phi_2 \circ \alpha \circ \phi_1^{-1} \colon \mathbb{C} \times (U_1 \cap U_2) \to \mathbb{C} \times (U_1 \cap U_2)$$

has the form $(s, p) \mapsto (f(p) \cdot s, p)$ where f is analytic on $U_1 \cap U_2$.

Exercise 52.4.2. The function $f(p)$ above must be nonzero for all $p \in U_1 \cap U_2$. Why? (Hint: invert the function by swapping the role of ϕ_1 and ϕ_2.)

Question 52.4.3. Check that the above definition is the equivalent to the following: α is a line bundle morphism if and only if

- it maps a point $q \in \pi_1^{-1}(x)$ to some point $\alpha(q) \in \pi_2^{-1}(x)$ (that is, each fiber gets mapped to the corresponding fiber), and
- for every analytic section $s \colon U \to L_1$ on open set $U \subseteq X$, then $\alpha(s)$ is an analytic section $s \colon U \to L_2$.

Example 52.4.4

Let $X = \mathbb{C}$. Then $\alpha \colon \mathbb{C} \times X \to \mathbb{C} \times X$ by $\alpha(y, x) = (y \cdot x^2, x)$ is a line bundle homomorphism.
This line bundle homomorphism is not an isomorphism, because every point $(y, 0)$ gets mapped to $(0, 0)$.

The definition of line bundle isomorphism is what you would expect.

Definition 52.4.5 (Isomorphism of line bundles)**.** Two line bundles L_1 and L_2 are **isomorphic** if there are line bundle isomorphisms $\alpha \colon L_1 \to L_2$ and $\beta \colon L_2 \to L_1$ that are inverse of each other.

§52.5 Relation to invertible sheaves

write (actually we haven't defined sheaf yet either)

XIV

Algebraic NT I: Rings of Integers

Part XIV: Contents

53 Algebraic integers

Here's a first taste of algebraic number theory.

This is really close to the border between olympiads and higher math. You've always known that $a+\sqrt{2}b$ had a "norm" a^2-2b^2, and that somehow this norm was multiplicative. You've also always known that roots come in conjugate pairs. You might have heard of minimal polynomials but not know much about them.

This chapter and the next one will make all these vague notions precise. It's drawn largely from the first chapter of [Og10].

§53.1 Motivation from high school algebra

This is adapted from my blog, *Power Overwhelming*[1].

In high school precalculus, you'll often be asked to find the roots of some polynomial with integer coefficients. For instance,

$$x^3 - x^2 - x - 15 = (x-3)(x^2+2x+5)$$

has roots 3, $-1+2i$, $-1-2i$. Or as another example,

$$x^3 - 3x^2 - 2x + 2 = (x+1)(x^2-4x+2)$$

has roots -1, $2+\sqrt{2}$, $2-\sqrt{2}$. You'll notice that the irrational roots, like $-1 \pm 2i$ and $2 \pm \sqrt{2}$, are coming up in pairs. In fact, I think precalculus explicitly tells you that the complex roots come in conjugate pairs. More generally, it seems like all the roots of the form $a + b\sqrt{c}$ come in "conjugate pairs". And you can see why.

But a polynomial like

$$x^3 - 8x + 4$$

has no rational roots. (The roots of this are approximately -3.0514, 0.51730, 2.5341.) Or even simpler,

$$x^3 - 2$$

has only one real root, $\sqrt[3]{2}$. These roots, even though they are irrational, have no "conjugate" pairs. Or do they?

Let's try and figure out exactly what's happening. Let α be any complex number. We define a **minimal polynomial** of α over $\mathbb{Q}$ to be a polynomial such that

- $P(x)$ has rational coefficients, and leading coefficient 1,
- $P(\alpha) = 0$.
- The degree of P is as small as possible. We call $\deg P$ the **degree** of α.

Example 53.1.1 (Examples of minimal polynomials)

(a) $\sqrt{2}$ has minimal polynomial $x^2 - 2$.

(b) The imaginary unit $i = \sqrt{-1}$ has minimal polynomial $x^2 + 1$.

[1]URL: https://blog.evanchen.cc/2014/10/19/why-do-roots-come-in-conjugate-pairs/

(c) A primitive pth root of unity, $\zeta_p = e^{\frac{2\pi i}{p}}$, has minimal polynomial $x^{p-1} + x^{p-2} + \cdots + 1$, where p is a prime.

Note that $100x^2 - 200$ is also a polynomial of the same degree which has $\sqrt{2}$ as a root; that's why we want to require the polynomial to be monic. That's also why we choose to work in the rational numbers; that way, we can divide by leading coefficients without worrying if we get non-integers.

Why do we care? The point is as follows: suppose we have another polynomial $A(x)$ such that $A(\alpha) = 0$. Then we claim that $P(x)$ actually divides $A(x)$! That means that all the other roots of P will also be roots of A.

The proof is by contradiction: if not, by polynomial long division we can find a quotient and remainder $Q(x)$, $R(x)$ such that

$$A(x) = Q(x)P(x) + R(x)$$

and $R(x) \not\equiv 0$. Notice that by plugging in $x = \alpha$, we find that $R(\alpha) = 0$. But $\deg R < \deg P$, and $P(x)$ was supposed to be the minimal polynomial. That's impossible!

It follows from this and the monicity of the minimal polynomial that it is unique (when it exists), so actually it is better to refer to *the* minimal polynomial.

Exercise 53.1.2. Can you find an element in $\mathbb{C}$ that has no minimal polynomial?

Let's look at a more concrete example. Consider $A(x) = x^3 - 3x^2 - 2x + 2$ from the beginning. The minimal polynomial of $2 + \sqrt{2}$ is $P(x) = x^2 - 4x + 2$ (why?). Now we know that if $2 + \sqrt{2}$ is a root, then $A(x)$ is divisible by $P(x)$. And that's how we know that if $2 + \sqrt{2}$ is a root of A, then $2 - \sqrt{2}$ must be a root too.

As another example, the minimal polynomial of $\sqrt[3]{2}$ is $x^3 - 2$. So $\sqrt[3]{2}$ actually has **two** conjugates, namely, $\alpha = \sqrt[3]{2}(\cos 120^\circ + i \sin 120^\circ)$ and $\beta = \sqrt[3]{2}(\cos 240^\circ + i \sin 240^\circ)$. Thus any polynomial which vanishes at $\sqrt[3]{2}$ also has α and β as roots!

Question 53.1.3 (Important but tautological: irreducible $\iff$ minimal)**.** Let α be a root of the polynomial $P(x)$. Show that $P(x)$ is the minimal polynomial if and only if it is irreducible.

§53.2 Algebraic numbers and algebraic integers

Prototypical example for this section: $\sqrt{2}$ is an algebraic integer (root of $x^2 - 2$), $\frac{1}{2}$ is an algebraic number but not an algebraic integer (root of $x - \frac{1}{2}$).

Let's now work in much vaster generality. First, let's give names to the new numbers we've discussed above.

Definition 53.2.1. An **algebraic number** is any $\alpha \in \mathbb{C}$ which is the root of *some* polynomial with coefficients in $\mathbb{Q}$. The set of algebraic numbers is denoted $\overline{\mathbb{Q}}$.

Remark 53.2.2 — One can equally well say algebraic numbers are those that are roots of some polynomial with coefficients in $\mathbb{Z}$ (rather than $\mathbb{Q}$), since any polynomial in $\mathbb{Q}[x]$ can be scaled to one in $\mathbb{Z}[x]$.

Definition 53.2.3. Consider an algebraic number α and its minimal polynomial P (which is monic and has rational coefficients). If it turns out the coefficients of P are integers, then we say α is an **algebraic integer**.

The set of algebraic integers is denoted $\overline{\mathbb{Z}}$.

Remark 53.2.4 — One can show, using *Gauss's Lemma*, that if α is the root of *any* monic polynomial with integer coefficients, then α is an algebraic integer. So in practice, if I want to prove that $\sqrt{2}+\sqrt{3}$ is an algebraic integer, then I only have to say "the polynomial $(x^2-5)^2-24$ works" without checking that it's minimal.

Sometimes for clarity, we refer to elements of $\mathbb{Z}$ as **rational integers**.

Example 53.2.5 (Examples of algebraic integers)

The numbers

$$4,\ i=\sqrt{-1},\ \sqrt[3]{2},\ \sqrt{2}+\sqrt{3}$$

are all algebraic integers, since they are the roots of the monic polynomials $x-4$, x^2+1, x^3-2 and $(x^2-5)^2-24$.

The number $\frac{1}{2}$ has minimal polynomial $x-\frac{1}{2}$, so it's an algebraic number but not an algebraic integer. (In fact, the rational root theorem also directly implies that any monic integer polynomial does not have $\frac{1}{2}$ as a root!)

There are two properties I want to state off the bat, because they'll be used extensively in the tricky (but nice) problems at the end of the section. The first we prove now, since it's very easy:

Proposition 53.2.6 (Rational algebraic integers are rational integers)

An algebraic integer is rational if and only if it is a rational integer. In symbols,

$$\overline{\mathbb{Z}} \cap \mathbb{Q} = \mathbb{Z}.$$

Proof. Let α be a rational number. If α is an integer, it is the root of $x-\alpha$, hence an algebraic integer too.

Conversely, if P is a monic polynomial with integer coefficients such that $P(\alpha)=0$ then (by the rational root theorem, say) it follows α must be an integer. □

The other is that:

Proposition 53.2.7 ($\overline{\mathbb{Z}}$ is a ring and $\overline{\mathbb{Q}}$ is a field)

The algebraic integers $\overline{\mathbb{Z}}$ form a ring. The algebraic numbers $\overline{\mathbb{Q}}$ form a field.

We could prove this now if we wanted to, but the results in the next chapter will more or less do it for us, and so we take this on faith temporarily.

Remark 53.2.8 — For α an algebraic integer with minimal polynomial P, it's clear by definition that all other roots of P are also algebraic integers. One can check that this property (along with the two properties above) characterize the set

of algebraic integers. From this point of view, the algebraic integers can be thought of as an intrinsically-defined generalization of the ring of integers $\mathbb{Z} \subseteq \mathbb{Q}$ to the field of all algebraic numbers $\overline{\mathbb{Q}}$.

§53.3 Number fields

Prototypical example for this section: $\mathbb{Q}(\sqrt{2})$ *is a typical number field.*

Given any algebraic number α, we're able to consider fields of the form $\mathbb{Q}(\alpha)$. Let us write down the more full version.

Definition 53.3.1. A **number field** K is a field containing $\mathbb{Q}$ as a subfield which is a *finite-dimensional* $\mathbb{Q}$-vector space. The **degree** of K is its dimension.

Example 53.3.2 (Prototypical example)
Consider the field
$$K = \mathbb{Q}(\sqrt{2}) = \left\{ a + b\sqrt{2} \mid a, b \in \mathbb{Q} \right\}.$$
This is a field extension of $\mathbb{Q}$, and has degree 2 (the basis being 1 and $\sqrt{2}$).

You might be confused that I wrote $\mathbb{Q}(\sqrt{2})$ (which should permit denominators) instead of $\mathbb{Q}[\sqrt{2}]$, say. But if you read through Example 5.5.4, you should see that the denominators don't really matter: $\frac{1}{3-\sqrt{2}} = \frac{1}{7}(3+\sqrt{2})$ anyways, for example. You can either check this now in general, or just ignore the distinction and pretend I wrote square brackets everywhere.

Exercise 53.3.3 (Unimportant)**.** Show that if α is an algebraic number, then $\mathbb{Q}(\alpha) \cong \mathbb{Q}[\alpha]$.

Example 53.3.4 (Adjoining an algebraic number)
Let α be the root of some irreducible polynomial $P(x) \in \mathbb{Q}[x]$. The field $\mathbb{Q}(\alpha)$ is a field extension as well, and the basis is $1, \alpha, \alpha^2, \ldots, \alpha^{m-1}$, where $m = \deg P$. In particular, the degree of $\mathbb{Q}(\alpha)$ is the degree of P.

Example 53.3.5 (Non-examples of number fields)
$\mathbb{R}$ and $\mathbb{C}$ are not number fields since there is no *finite* $\mathbb{Q}$-basis of them.

§53.4 Primitive element theorem, and monogenic extensions

Prototypical example for this section: $\mathbb{Q}(\sqrt{3}, \sqrt{5}) \cong \mathbb{Q}(\sqrt{3}+\sqrt{5})$*. Can you see why?*

I'm only putting this theorem here because I was upset that no one told me it was true (it's a very natural conjecture), and I hope to not do the same to the reader. However, I'm not going to use it in anything that follows.

Theorem 53.4.1 (Artin's primitive element theorem)
Every number field K is isomorphic to $\mathbb{Q}(\alpha)$ for some algebraic number α.

The proof is left as Problem 59F, since to prove it I need to talk about field extensions first.

The prototypical example

$$\mathbb{Q}(\sqrt{3}, \sqrt{5}) \cong \mathbb{Q}(\sqrt{3} + \sqrt{5})$$

makes it clear why this theorem should not be too surprising.

§53.5 A few harder problems to think about

Problem 53A. Find a polynomial with integer coefficients which has $\sqrt{2} + \sqrt[3]{3}$ as a root.

Problem 53B (Brazil 2006)**.** Let p be an irreducible polynomial in $\mathbb{Q}[x]$ and degree larger than 1. Prove that if p has two roots r and s whose product is 1 then the degree of p is even.

Problem 53C*. Consider n roots of unity $\varepsilon_1, \ldots, \varepsilon_n$. Assume the average $\frac{1}{n}(\varepsilon_1 + \cdots + \varepsilon_n)$ is an algebraic integer. Prove that either the average is zero or $\varepsilon_1 = \cdots = \varepsilon_n$. (Used in Lemma 22.2.2.)

Problem 53D†. Which rational numbers q satisfy $\cos(q\pi) \in \mathbb{Q}$?

Problem 53E (MOP 2010)**.** There are $n > 2$ lamps arranged in a circle; initially one is on and the others are off. We may select any regular polygon whose vertices are among the lamps and toggle the states of all the lamps simultaneously. Show it is impossible to turn all lamps off.

Problem 53F (Kronecker's theorem)**.** Let α be an algebraic integer. Suppose all its Galois conjugates have absolute value one. Prove that $\alpha^N = 1$ for some positive integer N.

Problem 53G. Is there an algebraic integer with absolute value one which is not a root of unity?

Problem 53H. Is the ring of algebraic integers Noetherian?

54 The ring of integers

§54.1 Norms and traces

Prototypical example for this section: $a+b\sqrt{2}$ *as an element of* $\mathbb{Q}(\sqrt{2})$ *has norm* a^2-2b^2 *and trace* $2a$.

Remember when you did olympiads and we had like a^2+b^2 was the "norm" of $a+bi$? Cool, let me tell you what's actually happening.

First, let me make precise the notion of a conjugate.

Definition 54.1.1. Let α be an algebraic number, and let $P(x)$ be its minimal polynomial, of degree m. Then the m roots of P are the (Galois) **conjugates** of α.

It's worth showing at the moment that there are no repeated conjugates.

Lemma 54.1.2 (Irreducible polynomials have distinct roots)
An irreducible polynomial in $\mathbb{Q}[x]$ cannot have a complex double root.

Proof. Let $f(x) \in \mathbb{Q}[x]$ be the irreducible polynomial and assume it has a double root α. **Take the derivative** $f'(x)$**.** This derivative has three interesting properties.

- The degree of f' is one less than the degree of f.
- The polynomials f and f' are not relatively prime because they share a factor $x-\alpha$.
- The coefficients of f' are also in $\mathbb{Q}$.

Consider $g = \gcd(f, f')$. We must have $g \in \mathbb{Q}[x]$ by Euclidean algorithm. But the first two facts about f' ensure that g is nonconstant and $\deg g < \deg f$. Yet g divides f, contradiction to the fact that f should be a minimal polynomial. □

Hence α has exactly as many conjugates as the degree of α.

Now, we would *like* to define the *norm* of an element $\mathrm{N}(\alpha)$ as the product of its conjugates. For example, we want $2+i$ to have norm $(2+i)(2-i)=5$, and in general for $a+bi$ to have norm a^2+b^2. It would be *really cool* if the norm was multiplicative; we already know this is true for complex numbers!

Unfortunately, this doesn't quite work: consider

$$\mathrm{N}(2+i) = 5 \text{ and } \mathrm{N}(2-i) = 5.$$

But $(2+i)(2-i)=5$, which doesn't have norm 25 like we want, since 5 is degree 1 and has no conjugates at all. The reason this "bad" thing is happening is that we're trying to define the norm of an *element*, when we really ought to be defining the norm of an element *with respect to a particular* K.

What I'm driving at is that the norm should have different meanings depending on which field you're in. If we think of 5 as an element of $\mathbb{Q}$, then its norm is 5. But thought of as an element of $\mathbb{Q}(i)$, its norm really ought to be 25. Let's make this happen: for K a number field, we will now define $\mathrm{N}_{K/\mathbb{Q}}(\alpha)$ to be the norm of α *with respect to* K as follows.

Definition 54.1.3. Let $\alpha \in K$ have degree n, so $\mathbb{Q}(\alpha) \subseteq K$, and set $k = (\deg K)/n$. The **norm** of α is defined as

$$\mathrm{N}_{K/\mathbb{Q}}(\alpha) \coloneqq \left(\prod \text{Galois conj of } \alpha\right)^k.$$

The **trace** is defined as

$$\mathrm{Tr}_{K/\mathbb{Q}}(\alpha) \coloneqq k \cdot \left(\sum \text{Galois conj of } \alpha\right).$$

The exponent of k is a "correction factor" that makes the norm of 5 into $5^2 = 25$ when we view 5 as an element of $\mathbb{Q}(i)$ rather than an element of $\mathbb{Q}$. For a "generic" element of K, we expect $k = 1$.

Exercise 54.1.4. Use what you know about nested vector spaces to convince yourself that k is actually an integer.

Example 54.1.5 (Norm of $a + b\sqrt{2}$)
Let $\alpha = a + b\sqrt{2} \in \mathbb{Q}(\sqrt{2}) = K$. If $b \neq 0$, then α and K have the degree 2. Thus the only conjugates of α are $a \pm b\sqrt{2}$, which gives the norm

$$(a + b\sqrt{2})(a - b\sqrt{2}) = a^2 - 2b^2,$$

The trace is $(a - b\sqrt{2}) + (a + b\sqrt{2}) = 2a$.
Nicely, the formula $a^2 - 2b^2$ and $2a$ also works when $b = 0$.

Example 54.1.6 (Norm of $a + b\sqrt[3]{2} + c\sqrt[3]{4}$)
Let $\alpha = a + b\sqrt[3]{2} + c\sqrt[3]{4} \in \mathbb{Q}(\sqrt[3]{2}) = K$. As above, if $b \neq 0$ or $c \neq 0$, then α and K have the same degree 3. The conjugates of α are $a + b\sqrt[3]{2}\omega + c\sqrt[3]{4}\omega^2$ and $a + b\sqrt[3]{2}\omega^2 + c\sqrt[3]{4}\omega$, and we can compute $\mathrm{N}_{K/\mathbb{Q}}(\alpha) = a^3 + 2b^3 + 4c^3 - 6abc$ and $\mathrm{Tr}_{K/\mathbb{Q}}(\alpha) = 3a$.
Note that in this case the conjugates of α does not lie in the field K!

Of importance is:

Proposition 54.1.7 (Norms and traces are rational integers)
If α is an algebraic integer, its norm and trace are rational integers.

Question 54.1.8. Prove it. (Vieta formula.)

That's great, but it leaves a question unanswered: why is the norm multiplicative? To do this, I have to give a new definition of norm and trace.

Remark 54.1.9 — Another way to automatically add the "corrective factor" is to use the embeddings of K into $\mathbb{C}$.
As we will see later, in Theorem 59.3.1, there are exactly $d = \deg K$ embeddings of K into $\mathbb{C}$, say $\sigma_1, \dots, \sigma_d$. Then, $\mathrm{Tr}_{K/\mathbb{Q}}(\alpha) = \sum_{i=1}^{d} \sigma_i(\alpha)$ and $\mathrm{N}_{K/\mathbb{Q}}(\alpha) = \prod_{i=1}^{d} \sigma_i(\alpha)$.

Theorem 54.1.10 (Morally correct definition of norm and trace)
Let K be a number field of degree n, and let $\alpha \in K$. Let $\mu_\alpha \colon K \to K$ denote the map

$$x \mapsto \alpha x$$

viewed as a linear map of $\mathbb{Q}$-vector spaces. Then,

- the norm of α equals the determinant $\det \mu_\alpha$, and
- the trace of α equals the trace $\operatorname{Tr} \mu_\alpha$.

The definition of the determinant has an obvious geometrical interpretation: viewing $K \cong \mathbb{Q}^n$ as a vector space, the determinant measures how much $\mathbb{Q}^n$ is stretched when multiplied by α. That is, given a parallelepiped with volume v in $\mathbb{Q}^n$, it will be transformed to one with volume $|\operatorname{N}(\alpha)|v$ under the transformation μ_α.

Since the trace and determinant don't depend on the choice of basis, you can pick whatever basis you want and use whatever definition you got in high school. Fantastic, right?

Example 54.1.11 (Explicit computation of matrices for $a + b\sqrt{2}$)
Let $K = \mathbb{Q}(\sqrt{2})$, and let 1, $\sqrt{2}$ be the basis of K. Let

$$\alpha = a + b\sqrt{2}$$

(possibly even $b = 0$), and notice that

$$\left(a + b\sqrt{2}\right)\left(x + y\sqrt{2}\right) = (ax + 2yb) + (bx + ay)\sqrt{2}.$$

We can rewrite this in matrix form as

$$\begin{bmatrix} a & 2b \\ b & a \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + 2yb \\ bx + ay \end{bmatrix}.$$

Consequently, we can interpret μ_α as the matrix

$$\mu_\alpha = \begin{bmatrix} a & 2b \\ b & a \end{bmatrix}.$$

Of course, the matrix will change if we pick a different basis, but the determinant and trace do not: they are always given by

$$\det \mu_\alpha = a^2 - 2b^2 \text{ and } \operatorname{Tr} \mu_\alpha = 2a.$$

This interpretation explains why the same formula should work for $a + b\sqrt{2}$ even in the case $b = 0$.

Proof. I'll prove the result for just the norm; the trace falls out similarly. Set

$$n = \deg \alpha, \qquad kn = \deg K.$$

The proof is split into two parts, depending on whether or not $k = 1$.

Proof if $k = 1$. Set $n = \deg \alpha = \deg K$. Thus the norm actually *is* the product of the Galois conjugates. Also,

$$\{1, \alpha, \dots, \alpha^{n-1}\}$$

is linearly independent in K, and hence a basis (as $\dim K = n$). Let's use this as the basis for μ_α.

Let

$$x^n + c_{n-1}x^{n-1} + \cdots + c_0$$

be the minimal polynomial of α. Thus $\mu_\alpha(1) = \alpha$, $\mu_\alpha(\alpha) = \alpha^2$, and so on, but $\mu_\alpha(\alpha^{n-1}) = -c_{n-1}\alpha^{n-1} - \cdots - c_0$. Therefore, μ_α is given by the matrix

$$M = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & -c_0 \\ 1 & 0 & 0 & \dots & 0 & -c_1 \\ 0 & 1 & 0 & \dots & 0 & -c_2 \\ \vdots & \vdots & \vdots & \ddots & 0 & -c_{n-2} \\ 0 & 0 & 0 & \dots & 1 & -c_{n-1} \end{bmatrix}.$$

Thus

$$\det M = (-1)^n c_0$$

and we're done by Vieta's formulas. ∎

Proof if $k > 1$. We have nested vector spaces

$$\mathbb{Q} \subseteq \mathbb{Q}(\alpha) \subseteq K.$$

Let $e_1, \dots, e_k$ be a $\mathbb{Q}(\alpha)$-basis for K (meaning: interpret K as a vector space over $\mathbb{Q}(\alpha)$, and pick that basis). Since $\{1, \alpha, \dots, \alpha^{n-1}\}$ is a $\mathbb{Q}$ basis for $\mathbb{Q}(\alpha)$, the elements

$$\begin{array}{cccc} e_1, & e_1\alpha, & \dots, & e_1\alpha^{n-1} \\ e_2, & e_2\alpha, & \dots, & e_2\alpha^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ e_k, & e_k\alpha, & \dots, & e_k\alpha^{n-1} \end{array}$$

constitute a $\mathbb{Q}$-basis of K. Using *this* basis, the map μ_α looks like

$$\underbrace{\begin{bmatrix} M & & & \\ & M & & \\ & & \ddots & \\ & & & M \end{bmatrix}}_{k \text{ times}}$$

where M is the same matrix as above: we just end up with one copy of our old matrix for each e_i. Thus $\det \mu_\alpha = (\det M)^k$, as needed. ∎

Question 54.1.12. Verify the result for traces as well. □

From this it follows immediately that

$$\mathrm{N}_{K/\mathbb{Q}}(\alpha\beta) = \mathrm{N}_{K/\mathbb{Q}}(\alpha)\,\mathrm{N}_{K/\mathbb{Q}}(\beta)$$

because by definition we have

$$\mu_{\alpha\beta} = \mu_\alpha \circ \mu_\beta,$$

and that the determinant is multiplicative. In the same way, the trace is additive.

§54.2 The ring of integers

Prototypical example for this section: *If $K = \mathbb{Q}(\sqrt{2})$, then $\mathcal{O}_K = \mathbb{Z}[\sqrt{2}]$. But if $K = \mathbb{Q}(\sqrt{5})$, then $\mathcal{O}_K = \mathbb{Z}[\frac{1+\sqrt{5}}{2}]$.*

$\mathbb{Z}$ makes for better number theory than $\mathbb{Q}$. In the same way, focusing on the *algebraic integers* of K gives us some really nice structure, and we'll do that here.

Definition 54.2.1. Given a number field K, we define

$$\mathcal{O}_K \coloneqq K \cap \overline{\mathbb{Z}}$$

to be the **ring of integers** of K; in other words $\mathcal{O}_K$ consists of the algebraic integers of K.

We do the classical example of a quadratic field now. Before proceeding, I need to write a silly number theory fact.

Exercise 54.2.2 (Annoying but straightforward)**.** Let a and b be rational numbers, and d a squarefree positive integer.

- If $d \equiv 2, 3 \pmod 4$, prove that $2a, a^2 - db^2 \in \mathbb{Z}$ if and only if $a, b \in \mathbb{Z}$.
- For $d \equiv 1 \pmod 4$, prove that $2a, a^2 - db^2 \in \mathbb{Z}$ if and only if $a, b \in \mathbb{Z}$ OR if $a - \frac{1}{2}, b - \frac{1}{2} \in \mathbb{Z}$.

You'll need to take mod 4.

Example 54.2.3 (Ring of integers of $K = \mathbb{Q}(\sqrt{3})$)

Let K be as above. We claim that

$$\mathcal{O}_K = \mathbb{Z}[\sqrt{3}] = \left\{ m + n\sqrt{3} \mid m, n \in \mathbb{Z} \right\}.$$

We set $\alpha = a + b\sqrt{3}$. Then $\alpha \in \mathcal{O}_K$ when the minimal polynomial has integer coefficients.
If $b = 0$, then the minimal polynomial is $x - \alpha = x - a$, and thus α works if and only if it's an integer. If $b \neq 0$, then the minimal polynomial is

$$(x-a)^2 - 3b^2 = x^2 - 2a \cdot x + (a^2 - 3b^2).$$

From the exercise, this occurs exactly for $a, b \in \mathbb{Z}$.

Example 54.2.4 (Ring of integers of $K = \mathbb{Q}(\sqrt{5})$)

We claim that in this case

$$\mathcal{O}_K = \mathbb{Z}\left[\frac{1+\sqrt{5}}{2}\right] = \left\{ m + n \cdot \frac{1+\sqrt{5}}{2} \mid m, n \in \mathbb{Z} \right\}.$$

The proof is exactly the same, except the exercise tells us instead that for $b \neq 0$, we have both the possibility that $a, b \in \mathbb{Z}$ or that $a, b \in \mathbb{Z} - \frac{1}{2}$. This reflects the fact that $\frac{1+\sqrt{5}}{2}$ is the root of $x^2 - x - 1 = 0$; no such thing is possible with $\sqrt{3}$.

In general, the ring of integers of $K = \mathbb{Q}(\sqrt{d})$ is

$$\mathcal{O}_K = \begin{cases} \mathbb{Z}[\sqrt{d}] & d \equiv 2,3 \pmod 4 \\ \mathbb{Z}\left[\frac{1+\sqrt{d}}{2}\right] & d \equiv 1 \pmod 4. \end{cases}$$

What we're going to show is that $\mathcal{O}_K$ behaves in K a lot like the integers do in $\mathbb{Q}$. First we show K consists of quotients of numbers in $\mathcal{O}_K$. In fact, we can do better:

Example 54.2.5 (Rationalizing the denominator)
For example, consider $K = \mathbb{Q}(\sqrt{3})$. The number $x = \frac{1}{4+\sqrt{3}}$ is an element of K, but by "rationalizing the denominator" we can write

$$\frac{1}{4+\sqrt{3}} = \frac{4-\sqrt{3}}{13}.$$

So we see that in fact, x is $\frac{1}{13}$ of an integer in $\mathcal{O}_K$.

The theorem holds true more generally.

Theorem 54.2.6 ($K = \mathbb{Q} \cdot \mathcal{O}_K$)
Let K be a number field, and let $x \in K$ be any element. Then there exists an integer n such that $nx \in \mathcal{O}_K$; in other words,

$$x = \frac{1}{n}\alpha$$

for some $\alpha \in \mathcal{O}_K$.

Exercise 54.2.7. Prove this yourself. (Start by using the fact that x has a minimal polynomial with rational coefficients. Alternatively, take the norm.)

Now we are going to show $\mathcal{O}_K$ is a ring; we'll check it is closed under addition and multiplication. To do so, the easiest route is:

Lemma 54.2.8 ($\alpha \in \overline{\mathbb{Z}} \iff \mathbb{Z}[\alpha]$ finitely generated)
Let $\alpha \in \overline{\mathbb{Q}}$. Then α is an algebraic integer if and only if the abelian group $\mathbb{Z}[\alpha]$ is finitely generated.

Proof. Note that α is an algebraic integer if and only if it's the root of some nonzero, monic polynomial with integer coefficients. Suppose first that

$$\alpha^N = c_{N-1}\alpha^{N-1} + c_{N-2}\alpha^{N-2} + \cdots + c_0.$$

Then the set $1, \alpha, \dots, \alpha^{N-1}$ generates $\mathbb{Z}[\alpha]$, since we can repeatedly replace α^N until all powers of α are less than N.

Conversely, suppose that $\mathbb{Z}[\alpha]$ is finitely generated by some $b_1, \dots, b_m$. Viewing the b_i as polynomials in α, we can select a large integer N (say $N = \deg b_1 + \cdots + \deg b_m + 2015$)

and express α^N in the b_i's to get

$$\alpha^N = c_1 b_1(\alpha) + \cdots + c_m b_m(\alpha).$$

The above gives us a monic polynomial in α, and the choice of N guarantees it is not zero. So α is an algebraic integer. □

Example 54.2.9 ($\frac{1}{2}$ isn't an algebraic integer)
We already know $\frac{1}{2}$ isn't an algebraic integer. So we expect

$$\mathbb{Z}\left[\frac{1}{2}\right] = \left\{ \frac{a}{2^m} \mid a, m \in \mathbb{Z} \text{ and } m \geq 0 \right\}$$

to not be finitely generated, and this is the case.

Question 54.2.10. To make the last example concrete: name all the elements of $\mathbb{Z}[\frac{1}{2}]$ that cannot be written as an integer combination of

$$\left\{ \frac{1}{2}, \frac{7}{8}, \frac{13}{64}, \frac{2015}{4096}, \frac{1}{1048576} \right\}$$

Now we can state the theorem.

Theorem 54.2.11 (Algebraic integers are closed under $+$ and $\times$)
The set $\overline{\mathbb{Z}}$ is closed under addition and multiplication; i.e. it is a ring. In particular, $\mathcal{O}_K$ is also a ring for any number field K.

Proof. Let $\alpha, \beta \in \overline{\mathbb{Z}}$. Then $\mathbb{Z}[\alpha]$ and $\mathbb{Z}[\beta]$ are finitely generated. Hence so is $\mathbb{Z}[\alpha, \beta]$. (Details: if $\mathbb{Z}[\alpha]$ has $\mathbb{Z}$-basis $a_1, \ldots, a_m$ and $\mathbb{Z}[\beta]$ has $\mathbb{Z}$-basis $b_1, \ldots, b_n$, then take the mn elements $a_i b_j$.)

Now $\mathbb{Z}[\alpha \pm \beta]$ and $\mathbb{Z}[\alpha\beta]$ are subsets of $\mathbb{Z}[\alpha, \beta]$ and so they are also finitely generated. Hence $\alpha \pm \beta$ and $\alpha\beta$ are algebraic integers. □

In fact, something even better is true. As you saw, for $\mathbb{Q}(\sqrt{3})$ we had $\mathcal{O}_K = \mathbb{Z}[\sqrt{3}]$; in other words, $\mathcal{O}_K$ was generated by 1 and $\sqrt{3}$. Something similar was true for $\mathbb{Q}(\sqrt{5})$. We claim that in fact, the general picture looks exactly like this.

Theorem 54.2.12 ($\mathcal{O}_K$ is a free $\mathbb{Z}$-module of rank n)
Let K be a number field of degree n. Then $\mathcal{O}_K$ is a free $\mathbb{Z}$-module of rank n, i.e. $\mathcal{O}_K \cong \mathbb{Z}^{\oplus n}$ as an abelian group. In other words, $\mathcal{O}_K$ has a $\mathbb{Z}$-basis of n elements as

$$\mathcal{O}_K = \{c_1\alpha_1 + \cdots + c_{n-1}\alpha_{n-1} + c_n\alpha_n \mid c_i \in \mathbb{Z}\}$$

where α_i are algebraic integers in $\mathcal{O}_K$.

The proof will be postponed to a later chapter.

This last theorem shows that in many ways $\mathcal{O}_K$ is a "lattice" in K. That is, for a

number field K we can find $\alpha_1, \dots, \alpha_n$ in $\mathcal{O}_K$ such that

$$\mathcal{O}_K \cong \alpha_1\mathbb{Z} \oplus \alpha_2\mathbb{Z} \oplus \cdots \oplus \alpha_n\mathbb{Z}$$
$$K \cong \alpha_1\mathbb{Q} \oplus \alpha_2\mathbb{Q} \oplus \cdots \oplus \alpha_n\mathbb{Q}$$

as abelian groups.

§54.3 On monogenic extensions

Recall that it turned out number fields K could all be expressed as $\mathbb{Q}(\alpha)$ for some α. We might hope that something similar is true of the ring of integers: that we can write

$$\mathcal{O}_K = \mathbb{Z}[\theta]$$

in which case $\{1, \theta, \dots, \theta^{n-1}\}$ serves both as a basis of K and as the $\mathbb{Z}$-basis for $\mathcal{O}_K$ (here $n = [K : \mathbb{Q}]$). In other words, we hope that the basis of $\mathcal{O}_K$ is actually a "power basis".

This is true for the most common examples we use:

- the quadratic field, and
- the cyclotomic field in Problem 54E†.

Unfortunately, it is not true in general: the first counterexample is $K = \mathbb{Q}(\alpha)$ for α a root of $X^3 - X^2 - 2X - 8$.

We call an extension with this nice property **monogenic**. As we'll later see, monogenic extensions have a really nice factoring algorithm, Theorem 55.5.4.

> **Remark 54.3.1** (What went wrong with $\mathcal{O}_K$?) — As we have just mentioned above, as an abelian group, $\mathcal{O}_K \cong \mathbb{Z}^3$, so it's generated by finitely many elements.
> In fact, $\{1, \alpha, \beta\}$ is a basis of $\mathcal{O}_K$, where $\beta = \frac{\alpha+\alpha^2}{2}$. The group generated by $\{1, \alpha, \alpha^2\}$ has index 2 in $\mathcal{O}_K$ – that is, $|\mathcal{O}_K/\langle 1, \alpha, \alpha^2\rangle| = 2$, and we misses β.
> If we try to pick $\{1, \beta, \beta^2\}$ as a basis instead, again we get $|\mathcal{O}_K/\langle 1, \beta, \beta^2\rangle| = 2$, and we misses α. If you explicitly compute it out, you can get $\beta^2 = \frac{3\alpha^2+7\alpha}{2} + 6 = 3\beta + 2\alpha + 6$.
> While this is not a proof that the extension is not monogenic, hopefully it gives you a feeling of the structure of $\mathcal{O}_K$.

§54.4 A few harder problems to think about

Problem 54A*. Show that α is a unit of $\mathcal{O}_K$ (meaning $\alpha^{-1} \in \mathcal{O}_K$) if and only if $\mathrm{N}_{K/\mathbb{Q}}(\alpha) = \pm 1$.

Problem 54B*. Let K be a number field. What is the field of fractions of $\mathcal{O}_K$?

Problem 54C (Russian olympiad 1984)**.** Find all integers m and n such that

$$\left(5 + 3\sqrt{2}\right)^m = \left(3 + 5\sqrt{2}\right)^n.$$

Problem 54D (USA TST 2012)**.** Decide whether there exist $a, b, c > 2010$ satisfying

$$a^3 + 2b^3 + 4c^3 = 6abc + 1.$$

Problem 54E† (Cyclotomic Field)**.** Let p be an odd rational prime and ζ_p a primitive pth root of unity. Let $K = \mathbb{Q}(\zeta_p)$. Prove that $\mathcal{O}_K = \mathbb{Z}[\zeta_p]$. (In fact, the result is true even if p is not a prime.)

55 Unique factorization (finally!)

Took long enough.

§55.1 Motivation

Suppose we're interested in solutions to the Diophantine equation $n = x^2 + 5y^2$ for a given n. The idea is to try and "factor" n in $\mathbb{Z}[\sqrt{-5}]$, for example

$$6 = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

Unfortunately, this is not so simple, because as I've said before we don't have unique factorization of elements:

$$6 = 2 \cdot 3 = \left(1 + \sqrt{-5}\right)\left(1 - \sqrt{-5}\right).$$

One reason this doesn't work is that we don't have a notion of a *greatest common divisor*. We can write $(35, 77) = 7$, but what do we make of $(3, 1 + \sqrt{-5})$?

The trick is to use ideals as a "generalized GCD". Recall that by (a, b) I mean the ideal $\{ax + by \mid x, y \in \mathbb{Z}[\sqrt{-5}]\}$. You can see that $(35, 77) = (7)$, but $(3, 1 + \sqrt{-5})$ will be left "unsimplified" because it doesn't represent an actual value in the ring. Using these *sets* (ideals) as elements, it turns out that we can develop a full theory of prime factorization, and we do so in this chapter.

In other words, we use the ideal $(a_1, \dots, a_m)$ to interpret a "generalized GCD" of a_1, $\dots$, a_m. In particular, if we have a number x we want to represent, we encode it as just (x).

Going back to our example of 6,

$$(6) = (2) \cdot (3) = \left(1 + \sqrt{-5}\right) \cdot \left(1 - \sqrt{-5}\right).$$

Please take my word for it that in fact, the complete prime factorization of (6) into prime ideals is

$$(6) = (2, 1 - \sqrt{-5})^2 (3, 1 + \sqrt{-5})(3, 1 - \sqrt{-5}) = \mathfrak{p}^2 \mathfrak{q}_1 \mathfrak{q}_2.$$

In fact, $(2) = \mathfrak{p}^2$, $(3) = \mathfrak{q}_1\mathfrak{q}_2$, $(1 + \sqrt{-5}) = \mathfrak{p}\mathfrak{q}_1$, $(1 - \sqrt{-5}) = \mathfrak{p}\mathfrak{q}_2$. So 6 indeed factorizes uniquely into ideals, even though it doesn't factor into elements.

As one can see above, ideal factorization is more refined than element factorization. Once you have the factorization into *ideals*, you can from there recover all the factorizations into *elements*. The upshot of this is that if we want to write n as $x^2 + 5y^2$, we just have to factor n into ideals, and from there we can recover all factorizations into elements, and finally all ways to write n as $x^2 + 5y^2$. Since we can already break n into rational prime factors (for example $6 = 2 \cdot 3$ above) we just have to figure out how each rational prime $p \mid n$ breaks down. There's a recipe for this, Theorem 55.5.4! In fact, I'll even tell you what is says in this special case:

- If $t^2 + 5$ factors as $(t + c)(t - c) \pmod p$, then $(p) = (p, c + \sqrt{-5})(p, c - \sqrt{-5})$.
- Otherwise, (p) is a prime ideal.

In this chapter we'll develop this theory of unique factorization in full generality.

Remark 55.1.1 — In this chapter, I'll be using the letters $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{p}$, $\mathfrak{q}$ for ideals of $\mathcal{O}_K$. When fractional ideals arise, I'll use I and J for them.

§55.2 Ideal arithmetic

Prototypical example for this section: $(x)(y) = (xy)$, *and* $(x) + (y) = (\gcd(x, y))$. *In any case, think in terms of generators.*

First, I have to tell you how to add and multiply two ideals $\mathfrak{a}$ and $\mathfrak{b}$.

Definition 55.2.1. Given two ideals $\mathfrak{a}$ and $\mathfrak{b}$ of a ring R, we define

$$\mathfrak{a} + \mathfrak{b} \coloneqq \{a + b \mid a \in \mathfrak{a}, b \in \mathfrak{b}\}$$
$$\mathfrak{a} \cdot \mathfrak{b} \coloneqq \{a_1 b_1 + \dots + a_n b_n \mid a_i \in \mathfrak{a}, b_i \in \mathfrak{b}\}.$$

(Note that infinite sums don't make sense in general rings, which is why in $\mathfrak{a} \cdot \mathfrak{b}$ we cut off the sum after some finite number of terms.) You can readily check these are actually ideals. This definition is more natural if you think about it in terms of the generators of $\mathfrak{a}$ and $\mathfrak{b}$.

Proposition 55.2.2 (Ideal arithmetic via generators)
Suppose $\mathfrak{a} = (a_1, a_2, \dots, a_n)$ and $\mathfrak{b} = (b_1, \dots, b_m)$ are ideals in a ring R. Then

(a) $\mathfrak{a} + \mathfrak{b}$ is the ideal generated by $a_1, \dots, a_n, b_1, \dots, b_m$.

(b) $\mathfrak{a} \cdot \mathfrak{b}$ is the ideal generated by $a_i b_j$, for $1 \le i \le n$ and $1 \le j \le m$.

Proof. Pretty straightforward; just convince yourself that this result is correct. □

In other words, for sums you append the two sets of generators together, and for products you take products of the generators. Note that for principal ideals, this coincides with "normal" multiplication, for example

$$(3) \cdot (5) = (15)$$

in $\mathbb{Z}$.

Remark 55.2.3 — Note that for an ideal $\mathfrak{a}$ and an element c, the set

$$c\mathfrak{a} = \{ca \mid a \in \mathfrak{a}\}$$

is equal to $(c) \cdot \mathfrak{a}$. So "scaling" and "multiplying by principal ideals" are the same thing. This is important, since we'll be using the two notions interchangeably.

Remark 55.2.4 — The addition of two ideals does not correspond to the addition of elements — for example, $(4) + (6) = (4, 6) = (2)$, but $4 + 6 = 10$.
This is the best we can hope for — addition of elements does not make sense for ideals — for example, $1 + 1 = 2$ and $1 + (-1) = 0$, but as ideals, $(1) = (-1)$.
In fact, addition of ideal is the straightforward generalization of gcd of elements — as you can check in the example above, $\gcd(4, 6) = 2$.
Nevertheless, I hope you agree that $\mathfrak{a} + \mathfrak{b}$ is a natural notation, compared to

something like $(\mathfrak{a}, \mathfrak{b})$.
Because factorization involves *multiplying*, instead of adding, the ideals together, we will not need to use the notation $\mathfrak{a} + \mathfrak{b}$ any time soon.

Finally, since we want to do factorization we better have some notion of divisibility. So we define:

Definition 55.2.5. We say $\mathfrak{a}$ divides $\mathfrak{b}$ and write $\mathfrak{a} \mid \mathfrak{b}$ if $\mathfrak{a} \supseteq \mathfrak{b}$.

Note the reversal of inclusions! So (3) divides (15), because (15) is contained in (3); every multiple of 15 is a multiple of 3. And from the example in the previous section: In $\mathbb{Z}[\sqrt{-5}]$, $(3, 1-\sqrt{-5})$ divides (3) and $(1-\sqrt{-5})$.

Finally, the **prime ideals** are defined as in Definition 5.3.1: $\mathfrak{p}$ is prime if $xy \in \mathfrak{p}$ implies $x \in \mathfrak{p}$ or $y \in \mathfrak{p}$. This is compatible with the definition of divisibility:

Exercise 55.2.6. A nonzero proper ideal $\mathfrak{p}$ is prime if and only if whenever $\mathfrak{p}$ divides $\mathfrak{ab}$, $\mathfrak{p}$ divides one of $\mathfrak{a}$ or $\mathfrak{b}$.

As mentioned in Remark 5.3.3, this also lets us ignore multiplication by units: $(-3) = (3)$.

§55.3 Dedekind domains

Prototypical example for this section: Any $\mathcal{O}_K$ is a Dedekind domain.

We now define a Dedekind domain as follows.

Definition 55.3.1. An integral domain $\mathcal{A}$ is a **Dedekind domain** if it is Noetherian, integrally closed, and *every nonzero prime ideal of $\mathcal{A}$ is in fact maximal.* (The last condition is the important one.)

Remark 55.3.2 — Note that $\mathcal{A}$ is a Dedekind domain if and only if $\mathcal{A} = \mathcal{O}_K$ for some field K, as we will prove below. We're just defining this term for historical reasons...

Here there's one new word I have to define for you, but we won't make much use of it.

Definition 55.3.3. Let R be an integral domain and let K be its field of fractions. We say R is **integrally closed** if the only elements $a \in K$ which are roots of *monic* polynomials in R are the elements of R (which are roots of the trivial $x - r$ polynomial).

The *interesting* condition in the definition of a Dedekind domain is the last one: prime ideals and maximal ideals are the same thing. The other conditions are just technicalities, but "primes are maximal" has real substance.

Example 55.3.4 ($\mathbb{Z}$ is a Dedekind domain)
The ring $\mathbb{Z}$ is a Dedekind domain. Note that

- $\mathbb{Z}$ is Noetherian (for obvious reasons).
- $\mathbb{Z}$ has field of fractions $\mathbb{Q}$. If $f(x) \in \mathbb{Z}[x]$ is monic, then by the rational root theorem any rational roots are integers (this is the same as the proof that $\overline{\mathbb{Z}} \cap \mathbb{Q} = \mathbb{Z}$). Hence $\mathbb{Z}$ is integrally closed.

- The nonzero prime ideals of $\mathbb{Z}$ are (p), which also happen to be maximal.

The case of interest is a ring $\mathcal{O}_K$ in which we wish to do factorizing. We're now going to show that for any number field K, the ring $\mathcal{O}_K$ is a Dedekind domain. First, the boring part.

Proposition 55.3.5 ($\mathcal{O}_K$ integrally closed and Noetherian)
For any number field K, the ring $\mathcal{O}_K$ is integrally closed and Noetherian.

Proof. Boring, but here it is anyways for completeness.

Since $\mathcal{O}_K \cong \mathbb{Z}^{\oplus n}$,[1] we get that it's Noetherian.

Now we show that $\mathcal{O}_K$ is integrally closed. Suppose that $\eta \in K$ is the root of some polynomial with coefficients in $\mathcal{O}_K$. Thus

$$\eta^n = \alpha_{n-1} \cdot \eta^{n-1} + \alpha_{n-2} \cdot \eta^{n-2} + \cdots + \alpha_0$$

where $\alpha_i \in \mathcal{O}_K$. We want to show that $\eta \in \mathcal{O}_K$ as well.

Well, from the above, $\mathcal{O}_K[\eta]$ is finitely generated... thus $\mathbb{Z}[\eta] \subseteq \mathcal{O}_K[\eta]$ is finitely generated. So $\eta \in \overline{\mathbb{Z}}$, and hence $\eta \in K \cap \overline{\mathbb{Z}} = \mathcal{O}_K$. □

Now let's do the fun part. We'll prove a stronger result, which will re-appear repeatedly.

Theorem 55.3.6 (Important: prime ideals divide rational primes)
Let $\mathcal{O}_K$ be a ring of integers and $\mathfrak{p}$ a nonzero prime ideal inside it. Then $\mathfrak{p}$ contains a rational prime p. Moreover, $\mathfrak{p}$ is maximal.

For a concrete example, consider $(2+i) \subseteq \mathbb{Z}[i]$. In this case, $p = 5$, and because $p \in (2+i)$, we get that the lattice is periodic in both dimensions with period 5, which implies finitely many points in $\{a+bi \mid 0 \le a, b \le 4, a, b \in \mathbb{Z}\}$ suffices to cover all cosets modulo the ideal.

The proof of the finiteness of the quotient here is closely related to the statement that the mesh of the lattice is finite, which will be covered in the next section.

Proof. Take any $\alpha \neq 0$ in $\mathfrak{p}$. Its Galois conjugates are algebraic integers so their product $\mathrm{N}(\alpha)/\alpha$ is in $\mathcal{O}_K$ (even though each individual conjugate need not be in K). Consequently, $\mathrm{N}(\alpha) \in \mathfrak{p}$, and we conclude $\mathfrak{p}$ contains some integer.

Then take the smallest positive integer in $\mathfrak{p}$, say p. We must have that p is a rational prime, since otherwise $\mathfrak{p} \ni p = xy$ implies one of $x, y \in \mathfrak{p}$. This shows the first part.

We now do something pretty tricky to show $\mathfrak{p}$ is maximal. Look at $\mathcal{O}_K/\mathfrak{p}$; since $\mathfrak{p}$ is prime it's supposed to be an integral domain... but we claim that it's actually finite! To do this, we forget that we can multiply on $\mathcal{O}_K$. Recalling that $\mathcal{O}_K \cong \mathbb{Z}^{\oplus n}$ as an abelian group, we obtain a map

$$\mathbb{F}_p{}^{\oplus n} \cong \mathcal{O}_K/(p) \twoheadrightarrow \mathcal{O}_K/\mathfrak{p}.$$

Hence $|\mathcal{O}_K/\mathfrak{p}| \le p^n$ is *finite*. Since finite integral domains are fields (Problem 5D*) we are done. □

[1] By Theorem 54.2.12.

Since every nonzero prime $\mathfrak{p}$ is maximal, we now know that $\mathcal{O}_K$ is a Dedekind domain. Note that this tricky proof is essentially inspired by the solution to Problem 5G†.

Remark 55.3.7 — An alternative proof for the first part is: because $\mathfrak{p}$ is an ideal, $\alpha \cdot \mathcal{O}_K \subseteq \mathfrak{p}$, but $\alpha \cdot \mathcal{O}_K$ is a free $\mathbb{Z}$-module of rank n, so $\mathfrak{p}$ is squeezed between two free $\mathbb{Z}$-modules of rank n, by Theorem 18.1.5 we must have $\mathfrak{p}$ is also free of rank n. So the quotient is finite, then use Lemma 56.8.8.

§55.4 Unique factorization works

Okay, I'll just say it now!

Unique factorization works perfectly in Dedekind domains!

Remark 55.4.1 (Comparison between Dedekind domain and UFD) — If we temporarily forget about the Noetherian and integrally closed condition, we have:

- An integral domain admits unique factorization of elements if the prime elements and the irreducible elements are the same.
- An integral domain admits unique factorization of ideals if the prime ideals and the maximal ideals are the same.

Notice the similarity — in either case, the Noetherian condition is "merely" to ensure that, if you keep extracting prime factors, you will terminate in a finite time.

Example 55.4.2 (What went wrong if $\mathcal{A}$ is not integrally closed?)
Consider $\mathcal{A} = 2\mathbb{Z}$, which is an ideal of $\mathbb{Z}$. Clearly, every nonzero prime ideal is maximal.
Nevertheless, in $\mathcal{A}$, $(2 \cdot 3 \cdot 5) = (60)$ is not a prime ideal (so of course it isn't a maximal ideal), but we cannot break it down into, for example, $(2 \cdot 3) \cdot (5)$.

Theorem 55.4.3 (Prime factorization works)
Let $\mathfrak{a}$ be a nonzero proper ideal of a Dedekind domain $\mathcal{A}$. Then $\mathfrak{a}$ can be written as a finite product of nonzero prime ideals $\mathfrak{p}_i$, say

$$\mathfrak{a} = \mathfrak{p}_1^{e_1}\mathfrak{p}_2^{e_2}\dots\mathfrak{p}_g^{e_g}$$

and this factorization is unique up to the order of the $\mathfrak{p}_i$.
Moreover, $\mathfrak{a}$ divides $\mathfrak{b}$ if and only if for every prime ideal $\mathfrak{p}$, the exponent of $\mathfrak{p}$ in $\mathfrak{a}$ is less than or equal to the corresponding exponent in $\mathfrak{b}$.

I won't write out the proof, but I'll describe the basic method of attack. Section 3 of [UI08] does a nice job of explaining it. When we proved the fundamental theorem of arithmetic, the basic plot was:

(1) Show that if p is a rational prime[2] then $p \mid bc$ means $p \mid b$ or $p \mid c$. (This is called Euclid's Lemma.)

(2) Use strong induction to show that every $N > 1$ can be written as the product of primes (easy).

(3) Show that if $p_1 \dots p_m = q_1 \dots q_n$ for some primes (not necessarily unique), then $p_1 = q_i$ for some i, say q_1.

(4) Divide both sides by p_1 and use induction.

What happens if we try to repeat the proof here? We get step 1 for free, because we're using a better definition of "prime". We can also do step 3, since it follows from step 1. But step 2 doesn't work, because for abstract Dedekind domains we don't really have a notion of size. And step 4 doesn't work because we don't yet have a notion of what the inverse of a prime ideal is.

Well, it turns out that we *can* define the inverse $\mathfrak{a}^{-1}$ of an ideal, and I'll do so by the end of this chapter. You then need to check that $\mathfrak{a} \cdot \mathfrak{a}^{-1} = (1) = \mathcal{A}$. In fact, even this isn't easy. You have to check it's true for prime ideals $\mathfrak{p}$, *then* prove prime factorization, and then prove that this is true. Moreover, $\mathfrak{a}^{-1}$ is not actually an ideal, so you need to work in the field of fractions K instead of $\mathcal{A}$.

So the main steps in the new situation are as follows:

(1) First, show that every ideal $\mathfrak{a}$ divides $\mathfrak{p}_1 \dots \mathfrak{p}_g$ for some finite collection of primes. (This is an application of Zorn's Lemma.)

(2) Define $\mathfrak{p}^{-1}$ and show that $\mathfrak{p}\mathfrak{p}^{-1} = (1)$.

(3) Show that a factorization exists (again using Zorn's Lemma).

(4) Show that it's unique, using the new inverse we've defined.

Finally, let me comment on how nice this is if $\mathcal{A}$ is a PID (like $\mathbb{Z}$). Thus every element $a \in \mathcal{A}$ is in direct correspondence with an ideal (a). Now suppose (a) factors as a product of ideals $\mathfrak{p}_i = (p_i)$, say,

$$(a) = (p_1)^{e_1}(p_2)^{e_2} \dots (p_n)^{e_n}.$$

This verbatim reads

$$a = up_1^{e_1}p_2^{e_2} \dots p_n^{e_n}$$

where u is some unit (recall Definition 4.4.1). Hence, Dedekind domains which are PID's satisfy unique factorization for *elements*, just like in $\mathbb{Z}$. (In fact, the converse of this is true.)

§55.5 The factoring algorithm

Let's look at some examples from quadratic fields. Recall that if $K = \mathbb{Q}(\sqrt{d})$, then

$$\mathcal{O}_K = \begin{cases} \mathbb{Z}[\sqrt{d}] & d \equiv 2, 3 \pmod 4 \\ \mathbb{Z}\left[\frac{1+\sqrt{d}}{2}\right] & d \equiv 1 \pmod 4. \end{cases}$$

Also, recall that the norm of $a + b\sqrt{-d}$ is given by $a^2 + db^2$.

[2]Note that the kindergarten definition of a prime is that "p isn't the product of two smaller integers". This isn't the correct definition of a prime: the definition of a prime is that $p \mid bc$ means $p \mid b$ or $p \mid c$. The kindergarten definition is something called "irreducible". Fortunately, in $\mathbb{Z}$, primes and irreducibles are the same thing, so no one ever told you that your definition of "prime" was wrong.

Example 55.5.1 (Factoring 6 in the integers of $\mathbb{Q}(\sqrt{-5})$)
Let $\mathcal{O}_K = \mathbb{Z}[\sqrt{-5}]$ arise from $K = \mathbb{Q}(\sqrt{-5})$. We've already seen that

$$(6) = (2) \cdot (3) = \left(1 + \sqrt{-5}\right)\left(1 - \sqrt{-5}\right)$$

and you can't get any further with these principal ideals. But let

$$\mathfrak{p} = \left(1 + \sqrt{-5}, 2\right) = \left(1 - \sqrt{-5}, 2\right) \quad \text{and} \quad \mathfrak{q}_1 = (1 + \sqrt{-5}, 3),\ \mathfrak{q}_2 = (1 - \sqrt{-5}, 3).$$

Then it turns out $(6) = \mathfrak{p}^2\mathfrak{q}_1\mathfrak{q}_2$. More specifically, $(2) = \mathfrak{p}^2$, $(3) = \mathfrak{q}_1\mathfrak{q}_2$, and $(1 + \sqrt{-5}) = \mathfrak{p}\mathfrak{q}_1$ and $(1 - \sqrt{-5}) = \mathfrak{p}\mathfrak{q}_2$. (Proof in just a moment.)

I want to stress that all our ideals are computed relative to $\mathcal{O}_K$. So for example,

$$(2) = \{2x \mid x \in \mathcal{O}_K\}.$$

How do we know in this example that $\mathfrak{p}$ is prime/maximal? (Again, these are the same since we're in a Dedekind domain.) Answer: look at $\mathcal{O}_K/\mathfrak{p}$ and see if it's a field. There is a trick to this: we can express

$$\mathcal{O}_K = \mathbb{Z}[\sqrt{-5}] \cong \mathbb{Z}[x]/(x^2 + 5).$$

So when we take *that* mod $\mathfrak{p}$, we get that

$$\mathcal{O}_K/\mathfrak{p} = \mathbb{Z}[x]/(x^2 + 5, 2, 1 + x) \cong \mathbb{F}_2[x]/(x^2 + 5, x + 1)$$

as rings.

Question 55.5.2. Conclude that $\mathcal{O}_K/\mathfrak{p} \cong \mathbb{F}_2$, and satisfy yourself that $\mathfrak{q}_1$ and $\mathfrak{q}_2$ are also maximal.

I should give an explicit example of an ideal multiplication: let's compute

$$\begin{aligned}
\mathfrak{q}_1\mathfrak{q}_2 &= \left((1 + \sqrt{-5})(1 - \sqrt{-5}), 3(1 + \sqrt{-5}), 3(1 - \sqrt{-5}), 9\right) \\
&= \left(6, 3 + 3\sqrt{-5}, 3 - 3\sqrt{-5}, 9\right) \\
&= \left(6, 3 + 3\sqrt{-5}, 3 - 3\sqrt{-5}, 3\right) \\
&= (3)
\end{aligned}$$

where we first did $9 - 6 = 3$ (think Euclidean algorithm!), then noted that all the other generators don't contribute anything we don't already have with the 3 (again these are ideals computed in $\mathcal{O}_K$). You can do the computation for $\mathfrak{p}^2$, $\mathfrak{p}\mathfrak{q}_1$, $\mathfrak{p}\mathfrak{q}_2$ in the same way.

Finally, it's worth pointing out that we should quickly verify that $\mathfrak{p} \neq (x)$ for some x; in other words, that $\mathfrak{p}$ is not principal. Assume for contradiction that it is. Then x divides both $1 + \sqrt{-5}$ and 2, in the sense that $1 + \sqrt{-5} = \alpha_1 x$ and $2 = \alpha_2 x$ for some $\alpha_1, \alpha_2 \in \mathcal{O}_K$. (Principal ideals are exactly the "multiples" of x, so $(x) = x\mathcal{O}_K$.) Taking the norms, we find that $\mathrm{N}_{K/\mathbb{Q}}(x)$ divides both

$$\mathrm{N}_{K/\mathbb{Q}}(1 + \sqrt{-5}) = 6 \quad \text{and} \quad \mathrm{N}_{K/\mathbb{Q}}(2) = 4.$$

Since $\mathfrak{p} \neq (1)$, x cannot be a unit, so its norm must be 2. But there are no elements of norm $2 = a^2 + 5b^2$ in $\mathcal{O}_K$.

Example 55.5.3 (Factoring 3 in the integers of $\mathbb{Q}(\sqrt{-17})$)
Let $\mathcal{O}_K = \mathbb{Z}[\sqrt{-17}]$ arise from $K = \mathbb{Q}(\sqrt{-17})$. We know $\mathcal{O}_K \cong \mathbb{Z}[x]/(x^2+17)$. Now

$$\mathcal{O}_K/3\mathcal{O}_K \cong \mathbb{Z}[x]/(3, x^2+17) \cong \mathbb{F}_3[x]/(x^2-1).$$

This already shows that (3) cannot be a prime (i.e. maximal) ideal, since otherwise our result should be a field. Anyways, we have a projection

$$\mathcal{O}_K \twoheadrightarrow \mathbb{F}_3[x]/\left((x-1)(x+1)\right).$$

Let $\mathfrak{q}_1$ be the pre-image of $(x-1)$ in the image, that is,

$$\mathfrak{q}_1 = (3, \sqrt{-17}-1).$$

Similarly,

$$\mathfrak{q}_2 = (3, \sqrt{-17}+1).$$

We have $\mathcal{O}_K/\mathfrak{q}_1 \cong \mathbb{F}_3$, so $\mathfrak{q}_1$ is maximal (prime). Similarly $\mathfrak{q}_2$ is prime. Magically, you can check explicitly that

$$\mathfrak{q}_1\mathfrak{q}_2 = (3).$$

Hence this is the factorization of (3) into prime ideals.

The fact that $\mathfrak{q}_1\mathfrak{q}_2 = (3)$ looks magical, but it's really true:

$$\begin{aligned}
\mathfrak{q}_1\mathfrak{q}_2 &= (3, \sqrt{-17}-1)(3, \sqrt{-17}+1) \\
&= (9, 3\sqrt{-17}+3, 3\sqrt{-17}-3, 18) \\
&= (9, 3\sqrt{-17}+3, 6) \\
&= (3, 3\sqrt{-17}+3, 6) \\
&= (3).
\end{aligned}$$

In fact, it turns out this always works in general: given a rational prime p, there is an algorithm to factor p in any $\mathcal{O}_K$ of the form $\mathbb{Z}[\theta]$.

Theorem 55.5.4 (Factoring algorithm / Dedekind-Kummer theorem)
Let K be a number field. Let $\theta \in \mathcal{O}_K$ with $|\mathcal{O}_K/\mathbb{Z}[\theta]| = j < \infty$, and let p be a prime not dividing j. Then $(p) = p\mathcal{O}_K$ is factored as follows:

Let f be the minimal polynomial of θ and factor $\overline{f}$ mod p as

$$\overline{f} \equiv \prod_{i=1}^{g} (\overline{f}_i)^{e_i} \pmod{p}.$$

Then $\mathfrak{p}_i = (f_i(\theta), p)$ is prime for each i and the factorization of (p) is

$$\mathcal{O}_K \supseteq (p) = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i}.$$

In particular, if K is monogenic with $\mathcal{O}_K = \mathbb{Z}[\theta]$ then $j = 1$ and the theorem applies for all primes p.

In almost all our applications in this book, K will be monogenic; i.e. $j = 1$. Here $\overline{\psi}$ denotes the image in $\mathbb{F}_p[x]$ of a polynomial $\psi \in \mathbb{Z}[x]$.

Question 55.5.5. There are many possible pre-images f_i we could have chosen (for example if $\overline{f_i} = x^2 + 1 \pmod 3$, we could pick $f_i = x^2 + 3x + 7$.) Why does this not affect the value of $\mathfrak{p}_i$?

Note that earlier, we could check the factorization worked for any particular case. The proof that this works is much the same, but we need one extra tool, the ideal norm. After that we leave the proposition as Problem 55E.

This algorithm gives us a concrete way to compute prime factorizations of (p) in any monogenic number field with $\mathcal{O}_K = \mathbb{Z}[\theta]$. To summarize the recipe:

1. Find the minimal polynomial of θ, say $f \in \mathbb{Z}[x]$.
2. Factor f mod p into irreducible polynomials $\overline{f}_1^{\,e_1}\overline{f}_2^{\,e_2}\dots\overline{f}_g^{\,e_g}$.
3. Compute $\mathfrak{p}_i = (f_i(\theta), p)$ for each i.

Then your $(p) = \mathfrak{p}_1^{e_1} \dots \mathfrak{p}_g^{e_g}$.

Or even shorter:

In order to factorize p in $\mathbb{Z}[x]/(f(x))$, we can factorize $f(x)$ in $\mathbb{Z}[x]/(p)$ instead.

Both are equivalent to factorizing 0 in $\mathbb{Z}[x]/(f(x), p)$ — in other words, writing $\mathbb{Z}[x]/(f(x), p)$ as a direct sum of $\mathbb{Z}[x]$-modules.

Exercise 55.5.6. Factor (29) in $\mathbb{Z}[i]$ using the above algorithm.

§55.6 Fractional ideals

Prototypical example for this section: Analog to $\mathbb{Q}$ for $\mathbb{Z}$, allowing us to take inverses of ideals. Prime factorization works in the nicest way possible.

We now have a neat theory of factoring ideals of $\mathcal{A}$, just like factoring the integers. Now note that our factorization of $\mathbb{Z}$ naturally gives a way to factor elements of $\mathbb{Q}$; just factor the numerator and denominator separately.

Let's make the analogy clearer. The analogue of a rational number is as follows.

Definition 55.6.1. Let $\mathcal{A}$ be a Dedekind domain with field of fractions K. A **fractional ideal** J of K is a set of the form

$$J = \frac{1}{x} \cdot \mathfrak{a} \quad \text{where } x \in \mathcal{A}, \text{ and } \mathfrak{a} \text{ is an integral ideal.}$$

For emphasis, ideals of $\mathcal{A}$ will be sometimes referred to as **integral ideals**.

You might be a little surprised by this definition: one would expect that a fractional ideal should be of the form $\frac{\mathfrak{a}}{\mathfrak{b}}$ for some integral ideals $\mathfrak{a}$, $\mathfrak{b}$. But in fact, it suffices to just take $x \in \mathcal{A}$ in the denominator. The analogy is that when we looked at $\mathcal{O}_K$, we found that we only needed integer denominators: $\frac{1}{4-\sqrt{3}} = \frac{1}{13}(4 + \sqrt{3})$. Similarly here, it will turn out that we only need to look at $\frac{1}{x} \cdot \mathfrak{a}$ rather than $\frac{\mathfrak{a}}{\mathfrak{b}}$, and so we define it this way from the beginning. See Problem 55D† for a different equivalent definition.

Example 55.6.2 ($\frac{5}{2}\mathbb{Z}$ is a fractional ideal)
The set
$$\frac{5}{2}\mathbb{Z} = \left\{ \frac{5}{2}n \mid n \in \mathbb{Z} \right\} = \frac{1}{2}(5)$$
is a fractional ideal of $\mathbb{Z}$.

Now, as we prescribed, the fractional ideals form a multiplicative group:

Theorem 55.6.3 (Fractional ideals form a group)
Let $\mathcal{A}$ be a Dedekind domain and K its field of fractions. For any integral ideal $\mathfrak{a}$, the set
$$\mathfrak{a}^{-1} = \{x \in K \mid x\mathfrak{a} \subseteq (1) = \mathcal{A}\}$$
is a fractional ideal with $\mathfrak{a}\mathfrak{a}^{-1} = (1)$.

(This result is nontrivial. To prove $\mathfrak{a}\mathfrak{a}^{-1} = (1)$, one approach is to prove it first for prime $\mathfrak{a}$, then consider the factorization of $\mathfrak{a}$ into prime ideals.)

Definition 55.6.4. Thus nonzero fractional ideals of K form a group under multiplication with identity $(1) = \mathcal{A}$. This **ideal group** is denoted J_K.

Example 55.6.5 ($(3)^{-1}$ in $\mathbb{Z}$)
Please check that in $\mathbb{Z}$ we have
$$(3)^{-1} = \left\{ \frac{1}{3}n \mid n \in \mathbb{Z} \right\} = \frac{1}{3}\mathbb{Z}.$$

It follows that every fractional ideal J can be uniquely written as
$$J = \prod_i \mathfrak{p}_i^{n_i} \cdot \prod_i \mathfrak{q}_i^{-m_i}$$
where n_i and m_i are positive integers. In fact, $\mathfrak{a}$ is an integral ideal if and only if all its exponents are nonnegative, just like the case with integers. So, a perhaps better way to think about fractional ideals is as products of prime ideals, possibly with negative exponents.

§55.7 The ideal norm

One last tool is the ideal norm, which gives us a notion of the "size" of an ideal.

Definition 55.7.1. The **ideal norm** (or absolute norm) of a nonzero ideal $\mathfrak{a} \subseteq \mathcal{O}_K$ is defined as $|\mathcal{O}_K/\mathfrak{a}|$ and denoted $\mathrm{N}(\mathfrak{a})$.

Example 55.7.2 (Ideal norm of (5) in the Gaussian integers)
Let $K = \mathbb{Q}(i)$, $\mathcal{O}_K = \mathbb{Z}[i]$. Consider the ideal (5) in $\mathcal{O}_K$. We have that
$$\mathcal{O}_K/(5) \cong \{a + bi \mid a, b \in \mathbb{Z}/5\mathbb{Z}\}$$

so (5) has ideal norm 25, corresponding to the fact that $\mathcal{O}_K/(5)$ has $5^2 = 25$ elements.

Example 55.7.3 (Ideal norm of $(2+i)$ in the Gaussian integers)
You'll notice that

$$\mathcal{O}_K/(2+i) \cong \mathbb{F}_5$$

since mod $2+i$ we have both $5 \equiv 0$ and $i \equiv -2$. (Indeed, since $(2+i)$ is prime we had better get a field!) Thus $\mathrm{N}((2+i)) = 5$; similarly $\mathrm{N}((2-i)) = 5$.

Thus the ideal norm measures how "roomy" the ideal is: that is, (5) is a lot more spaced out in $\mathbb{Z}[i]$ than it is in $\mathbb{Z}$. (This intuition will be important when we will actually view $\mathcal{O}_K$ as a lattice.)

Question 55.7.4. What are the ideals with ideal norm one?

Our example with (5) suggests several properties of the ideal norm which turn out to be true:

Lemma 55.7.5 (Properties of the absolute norm)
Let $\mathfrak{a}$ be a nonzero ideal of $\mathcal{O}_K$.

(a) $\mathrm{N}(\mathfrak{a})$ is finite.

(b) For any other nonzero ideal $\mathfrak{b}$, $\mathrm{N}(\mathfrak{ab}) = \mathrm{N}(\mathfrak{a})\,\mathrm{N}(\mathfrak{b})$.

(c) If $\mathfrak{a} = (a)$ is principal, then $\mathrm{N}(\mathfrak{a}) = |\mathrm{N}_{K/\mathbb{Q}}(a)|$.

I unfortunately won't prove these properties, though we already did (a) in our proof that $\mathcal{O}_K$ was a Dedekind domain.

As with the case of the norm of an element, the ideal norm also has a geometrical interpretation: Recall that if $\mathfrak{a} = (a)$, let μ_a be the multiplication-by-a map, then $\mathrm{N}_{K/\mathbb{Q}}(a) = |\det \mu_a|$ measures how much K is stretched under μ_a when viewed as a $\mathbb{Q}$-vector space.

Exercise 55.7.6. Convince yourself that if $\mu_a(\mathcal{O}_K) \subseteq \mathcal{O}_K$, then $|\mathcal{O}_K/a\mathcal{O}_K|$ is exactly equal to $|\det \mu_a|$.

This explains why $\mathrm{N}(\mathfrak{a}) = |\mathrm{N}_{K/\mathbb{Q}}(a)|$, although note that $\mathfrak{a} = (a) = (-a)$, so there need not be an unique multiplication-by-$\mathfrak{a}$ map.

The fact that N is completely multiplicative lets us also consider the norm of a fractional ideal J by the natural extension

$$J = \prod_i \mathfrak{p}_i^{n_i} \cdot \prod_i \mathfrak{q}_i^{-m_i} \quad \implies \quad \mathrm{N}(J) := \frac{\prod_i \mathrm{N}(\mathfrak{p}_i)^{n_i}}{\prod_i \mathrm{N}(\mathfrak{q}_i)^{m_i}}.$$

Thus N is a natural group homomorphism $J_K \to \mathbb{Q}^\times$.

§55.8 A few harder problems to think about

Problem 55A. Show that there are three different factorizations of 77 in $\mathcal{O}_K$, where $K = \mathbb{Q}(\sqrt{-13})$.

Problem 55B. Let $K = \mathbb{Q}(\sqrt[3]{2})$; take for granted that $\mathcal{O}_K = \mathbb{Z}[\sqrt[3]{2}]$. Find the factorization of (5) in $\mathcal{O}_K$.

Problem 55C (Fermat's little theorem)**.** Let $\mathfrak{p}$ be a prime ideal in some ring of integers $\mathcal{O}_K$. Show that for $\alpha \in \mathcal{O}_K$,

$$\alpha^{\mathrm{N}(\mathfrak{p})} \equiv \alpha \pmod{\mathfrak{p}}.$$

Problem 55D†. Let $\mathcal{A}$ be a Dedekind domain with field of fractions K, and pick $J \subseteq K$. Show that J is a fractional ideal if and only if

(i) J is closed under addition and multiplication by elements of $\mathcal{A}$, and

(ii) J is finitely generated as an abelian group.

More succinctly: J is a fractional ideal $\iff$ J is a finitely generated $\mathcal{A}$-module.

Problem 55E. In the notation of Theorem 55.5.4, let $I = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i}$. Assume for simplicity that K is monogenic, hence $\mathcal{O}_K = \mathbb{Z}[\theta]$.

(a) Prove that each $\mathfrak{p}_i$ is prime.

(b) Show that (p) divides I.

(c) Use the norm to show that $(p) = I$.

56 Minkowski bound and class groups

We now have a neat theory of unique factorization of ideals. In the case of a PID, this in fact gives us a UFD. Sweet.

We'll define, in a moment, something called the *class group* which measures how far $\mathcal{O}_K$ is from being a PID; the bigger the class group, the farther $\mathcal{O}_K$ is from being a PID. In particular, $\mathcal{O}_K$ is a PID if it has trivial class group.

Then we will provide some inequalities which let us put restrictions on the class group; for instance, this will let us show in some cases that the class group must be trivial. Astonishingly, the proof will use Minkowski's theorem, a result from geometry.

§56.1 The class group

Prototypical example for this section: PID's have trivial class group.

Let K be a number field, and let J_K denote the multiplicative group of fractional ideals of $\mathcal{O}_K$. Let P_K denote the multiplicative group of **principal fractional ideals**: those of the form $(x) = x\mathcal{O}_K$ for some $x \in K$.

> **Question 56.1.1.** Check that P_K is also a multiplicative group. (This is really easy: name $x\mathcal{O}_K \cdot y\mathcal{O}_K$ and $(x\mathcal{O}_K)^{-1}$.)

As J_K is abelian, we can now define the **class group** (or **ideal class group**) to be the quotient

$$\mathrm{Cl}_K \coloneqq J_K/P_K.$$

The elements of Cl_K are called **classes**.

Equivalently,

The class group Cl_K is the set of nonzero fractional ideals modulo scaling by a constant in K.

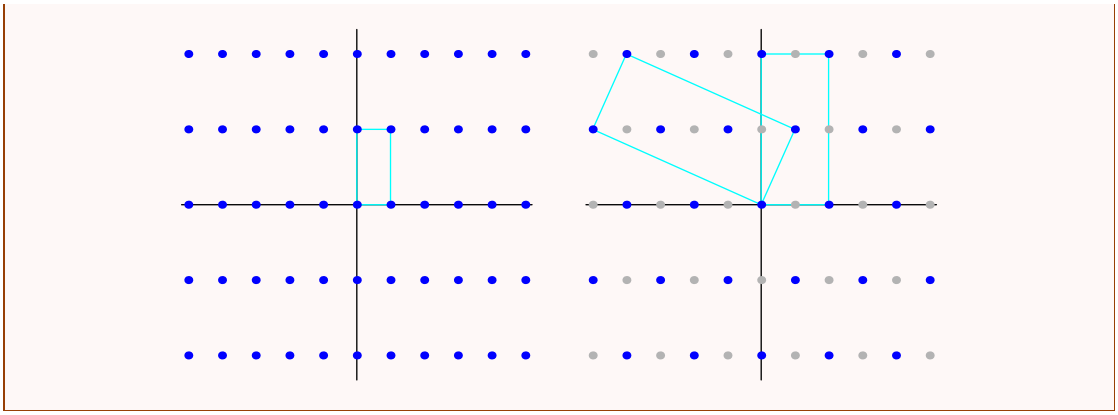
You can also think of the classes as the "shapes" of the ideals, as two ideals belong to the same class if and only if they're isomorphic as $\mathcal{O}_K$-modules.

Example 56.1.2 (Ideal classes in $\mathbb{Q}(\sqrt{-5})$)

If the field is an imaginary quadratic field, visualizing the class of an ideal is really easy: because multiplication by a complex number corresponds to a combination of a scaling and a rotation (i.e. it preserves angles), two ideals belong to the same class if they are similar, that is, you can overlap one onto the another using rotation and scaling.

When the field is $K = \mathbb{Q}(\sqrt{-5})$, the ring of integers is $\mathcal{O}_K = \mathbb{Z}[\sqrt{-5}]$.

The first picture below depicts the ideal $(1) \subseteq \mathcal{O}_K$. The second picture depicts $(2, 1+\sqrt{-5}) \subseteq \mathcal{O}_K$, which is not a principal ideal.



In particular, Cl_K is trivial if all ideals are principal, since the nonzero principal ideals are the same up to scaling.

The size of the class group is called the **class number**. It's a beautiful theorem that the class number is always finite, and the bulk of this chapter will build up to this result. It requires several ingredients.

§56.2 The discriminant of a number field

Prototypical example for this section: *Quadratic fields.*

Let's say I have $K = \mathbb{Q}(\sqrt{2})$. As we've seen before, this means $\mathcal{O}_K = \mathbb{Z}[\sqrt{2}]$, meaning

$$\mathcal{O}_K = \left\{ a + b\sqrt{2} \mid a, b \in \mathbb{Z} \right\}.$$

The key insight now is that you might think of this as a *lattice*: geometrically, we want to think about this the same way we think about $\mathbb{Z}^2$.

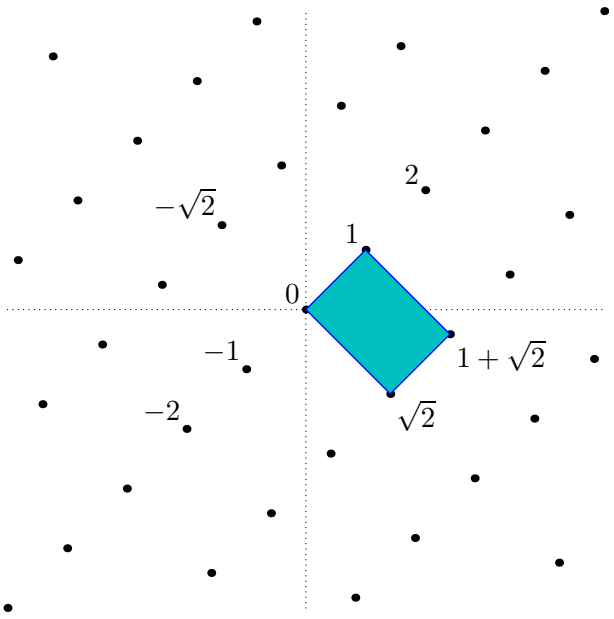
Perversely, we might try to embed this into $\mathbb{Q}^2$ by sending $a + b\sqrt{2}$ to (a, b). But this is a little stupid, since we're rudely making K, which somehow lives inside $\mathbb{R}$ and is "one-dimensional" in that sense, into a two-dimensional space. It also depends on a choice of basis, which we don't like. A better way is to think about the fact that there are two embeddings $\sigma_1 \colon K \to \mathbb{C}$ and $\sigma_2 \colon K \to \mathbb{C}$, namely the identity, and conjugation:

$$\begin{aligned} \sigma_1(a + b\sqrt{2}) &= a + b\sqrt{2} \\ \sigma_2(a + b\sqrt{2}) &= a - b\sqrt{2}. \end{aligned}$$

Fortunately for us, these embeddings both have real image. This leads us to consider the set of points

$$(\sigma_1(\alpha), \sigma_2(\alpha)) \in \mathbb{R}^2 \quad \text{as} \quad \alpha \in K.$$

This lets us visualize what $\mathcal{O}_K$ looks like in $\mathbb{R}^2$. The points of K are dense in $\mathbb{R}^2$, but the points of $\mathcal{O}_K$ cut out a lattice.



To see how big the lattice is, we look at how $\{1, \sqrt{2}\}$, the generators of $\mathcal{O}_K$, behave. The point corresponding to $a + b\sqrt{2}$ in the lattice is

$$a \cdot (1, 1) + b \cdot (\sqrt{2}, -\sqrt{2}).$$

The **mesh** of the lattice[1] is defined as the hypervolume of the "fundamental parallelepiped" I've colored blue above. For this particular case, it ought to be equal to the area of that parallelogram, which is

$$\det \begin{bmatrix} 1 & -\sqrt{2} \\ 1 & \sqrt{2} \end{bmatrix} = 2\sqrt{2}.$$

The definition of the discriminant is precisely this, except with an extra square factor (since permutation of rows could lead to changes in sign in the matrix above). Problem 57B* shows that the squaring makes Δ_K an integer.

To make the next definition, we invoke:

Theorem 56.2.1 (The n embeddings of a number field)

Let K be a number field of degree n. Then there are exactly n field homomorphisms $K \hookrightarrow \mathbb{C}$, say $\sigma_1, \dots, \sigma_n$, which fix $\mathbb{Q}$.

Proof. Deferred to Theorem 59.3.1, once we have the tools of Galois theory. □

In fact, in Theorem 59.3.4 we see that for $\alpha \in K$, we have that $\sigma_i(\alpha)$ runs over the conjugates of α as $i = 1, \dots, n$. It follows that

$$\operatorname{Tr}_{K/\mathbb{Q}}(\alpha) = \sum_{i=1}^{n} \sigma_i(\alpha) \quad \text{and} \quad \operatorname{N}_{K/\mathbb{Q}}(\alpha) = \prod_{i=1}^{n} \sigma_i(\alpha).$$

This allows us to define:

[1] Most authors call this the volume, but I think this is not the right word to use – lattices have "volume" zero since they are just a bunch of points! In contrast, the English word "mesh" really does refer to the width of a "gap".

Definition 56.2.2. Suppose $\alpha_1, \dots, \alpha_n$ is a $\mathbb{Z}$-basis of $\mathcal{O}_K$. The **discriminant** of the number field K is defined by

$$\Delta_K \coloneqq \det \begin{bmatrix} \sigma_1(\alpha_1) & \dots & \sigma_n(\alpha_1) \\ \vdots & \ddots & \vdots \\ \sigma_1(\alpha_n) & \dots & \sigma_n(\alpha_n) \end{bmatrix}^2.$$

This does not depend on the choice of the $\{\alpha_i\}$; we will not prove this here.

Example 56.2.3 (Discriminant of $K = \mathbb{Q}(\sqrt{2})$)
We have $\mathcal{O}_K = \mathbb{Z}[\sqrt{2}]$ and as discussed above the discriminant is

$$\Delta_K = (-2\sqrt{2})^2 = 8.$$

Example 56.2.4 (Discriminant of $\mathbb{Q}(i)$)
Let $K = \mathbb{Q}(i)$. We have $\mathcal{O}_K = \mathbb{Z}[i] = \mathbb{Z} \oplus i\mathbb{Z}$. The embeddings are the identity and complex conjugation which take 1 to $(1,1)$ and i to $(i,-i)$. So

$$\Delta_K = \det \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}^2 = (-2i)^2 = -4.$$

This example illustrates that the discriminant need not be positive for number fields which wander into the complex plane (the lattice picture is a less perfect analogy). But again, as we'll prove in the problems the discriminant is always an integer.

Example 56.2.5 (Discriminant of $\mathbb{Q}(\sqrt{5})$)
Let $K = \mathbb{Q}(\sqrt{5})$. This time, $\mathcal{O}_K = \mathbb{Z} \oplus \frac{1+\sqrt{5}}{2}\mathbb{Z}$, and so the discriminant is going to look a little bit different. The embeddings are still $a + b\sqrt{5} \mapsto a + b\sqrt{5}, a - b\sqrt{5}$. Applying this to the $\mathbb{Z}$-basis $\left\{1, \frac{1+\sqrt{5}}{2}\right\}$, we get

$$\Delta_K = \det \begin{bmatrix} 1 & 1 \\ \frac{1+\sqrt{5}}{2} & \frac{1-\sqrt{5}}{2} \end{bmatrix}^2 = (-\sqrt{5})^2 = 5.$$

Exercise 56.2.6. Extend all this to show that if $K = \mathbb{Q}(\sqrt{d})$ for $d \neq 1$ squarefree, we have

$$\Delta_K = \begin{cases} d & \text{if } d \equiv 1 \pmod 4 \\ 4d & \text{if } d \equiv 2,3 \pmod 4. \end{cases}$$

Actually, let me point out something curious: recall that the polynomial discriminant of $Ax^2 + Bx + C$ is $B^2 - 4AC$. Then:

- In the $d \equiv 1 \pmod 4$ case, Δ_K is the discriminant of $x^2 - x - \frac{d-1}{4}$, which is the minimal polynomial of $\frac{1}{2}(1+\sqrt{d})$. Of course, $\mathcal{O}_K = \mathbb{Z}[\frac{1}{2}(1+\sqrt{d})]$.
- In the $d \equiv 2,3 \pmod 4$ case, Δ_K is the discriminant of $x^2 - d$ which is the minimal polynomial of $\sqrt{d}$. Once again, $\mathcal{O}_K = \mathbb{Z}[\sqrt{d}]$.

This is not a coincidence! Problem 57C⋆ asserts that this is true in general; hence the name "discriminant".

§56.3 The signature of a number field

Prototypical example for this section: $\mathbb{Q}(\sqrt[100]{2})$ *has signature* $(2, 49)$.

In the example of $K = \mathbb{Q}(i)$, we more or less embedded K into the space $\mathbb{C}$. However, K is a degree two extension, so what we'd really like to do is embed it into $\mathbb{R}^2$. To do so, we're going to take advantage of complex conjugation.

Let K be a number field and $\sigma_1, \dots, \sigma_n$ be its embeddings. We distinguish between the **real embeddings** (which map all of K into $\mathbb{R}$) and the **complex embeddings** (which map some part of K outside $\mathbb{R}$). Notice that if σ is a complex embedding, then so is the conjugate $\overline{\sigma} \neq \sigma$; hence complex embeddings come in pairs.

Definition 56.3.1. Let K be a number field of degree n, and set

$$r_1 = \text{number of real embeddings}$$
$$r_2 = \text{number of pairs of complex embeddings.}$$

The **signature** of K is the pair (r_1, r_2). Observe that $r_1 + 2r_2 = n$.

Example 56.3.2 (Basic examples of signatures)

(a) $\mathbb{Q}$ has signature $(1, 0)$.

(b) $\mathbb{Q}(\sqrt{2})$ has signature $(2, 0)$.

(c) $\mathbb{Q}(i)$ has signature $(0, 1)$.

(d) Let $K = \mathbb{Q}(\sqrt[3]{2})$, and let ω be a cube root of unity. The elements of K are

$$K = \left\{ a + b\sqrt[3]{2} + c\sqrt[3]{4} \mid a, b, c \in \mathbb{Q} \right\}.$$

Then the signature is $(1, 1)$, because the three embeddings are

$$\sigma_1 \colon \sqrt[3]{2} \mapsto \sqrt[3]{2}, \quad \sigma_2 \colon \sqrt[3]{2} \mapsto \sqrt[3]{2}\omega, \quad \sigma_3 \colon \sqrt[3]{2} \mapsto \sqrt[3]{2}\omega^2.$$

The first of these is real and the latter two are conjugate pairs.

Example 56.3.3 (Even more signatures)

In the same vein $\mathbb{Q}(\sqrt[99]{2})$ and $\mathbb{Q}(\sqrt[100]{2})$ have signatures $(1, 49)$ and $(2, 49)$.

Question 56.3.4. Verify the signatures of the above two number fields.

From now on, we will number the embeddings of K in such a way that

$$\sigma_1, \sigma_2, \dots, \sigma_{r_1}$$

are the real embeddings, while

$$\sigma_{r_1+1} = \overline{\sigma_{r_1+r_2+1}}, \quad \sigma_{r_1+2} = \overline{\sigma_{r_1+r_2+2}}, \quad \dots, \quad \sigma_{r_1+r_2} = \overline{\sigma_{r_1+2r_2}}.$$

are the r_2 pairs of complex embeddings. We define the **canonical embedding** of K as

$$K \xhookrightarrow{\iota} \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \quad \text{by} \quad \alpha \mapsto (\sigma_1(\alpha), \dots, \sigma_{r_1}(\alpha), \sigma_{r_1+1}(\alpha), \dots, \sigma_{r_1+r_2}(\alpha)).$$

All we've done is omit, for the complex case, the second of the embeddings in each conjugate pair. This is no big deal, since they are just conjugates; the above tuple is all the information we need.

For reasons that will become obvious in a moment, I'll let τ denote the isomorphism

$$\tau \colon \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \xrightarrow{\sim} \mathbb{R}^{r_1+2r_2} = \mathbb{R}^n$$

by breaking each complex number into its real and imaginary part, as

$$\begin{aligned} \alpha \mapsto (&\sigma_1(\alpha), \dots, \sigma_{r_1}(\alpha), \\ &\operatorname{Re}\sigma_{r_1+1}(\alpha), \ \operatorname{Im}\sigma_{r_1+1}(\alpha), \\ &\operatorname{Re}\sigma_{r_1+2}(\alpha), \ \operatorname{Im}\sigma_{r_1+2}(\alpha), \\ &\dots, \\ &\operatorname{Re}\sigma_{r_1+r_2}(\alpha), \ \operatorname{Im}\sigma_{r_1+r_2}(\alpha)). \end{aligned}$$

Example 56.3.5 (Example of canonical embedding)
As before let $K = \mathbb{Q}(\sqrt[3]{2})$ and set

$$\sigma_1 \colon \sqrt[3]{2} \mapsto \sqrt[3]{2}, \quad \sigma_2 \colon \sqrt[3]{2} \mapsto \sqrt[3]{2}\omega, \quad \sigma_3 \colon \sqrt[3]{2} \mapsto \sqrt[3]{2}\omega^2$$

where $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$, noting that we've already arranged indices so $\sigma_1 = \mathrm{id}$ is real while σ_2 and σ_3 are a conjugate pair. So the embeddings $K \xhookrightarrow{\iota} \mathbb{R} \times \mathbb{C} \xrightarrow{\sim} \mathbb{R}^3$ are given by

$$\alpha \xmapsto{\iota} (\sigma_1(\alpha), \sigma_2(\alpha)) \xmapsto{\tau} (\sigma_1(\alpha), \operatorname{Re}\sigma_2(\alpha), \operatorname{Im}\sigma_2(\alpha)).$$

For concreteness, taking $\alpha = 9 + \sqrt[3]{2}$ gives

$$\begin{aligned} 9 + \sqrt[3]{2} &\xmapsto{\iota} \left(9 + \sqrt[3]{2}, \ 9 + \sqrt[3]{2}\omega\right) \\ &= \left(9 + \sqrt[3]{2}, \ 9 - \frac{1}{2}\sqrt[3]{2} + \frac{\sqrt[6]{108}}{2}i\right) \in \mathbb{R} \times \mathbb{C} \\ &\xmapsto{\tau} \left(9 + \sqrt[3]{2}, \ 9 - \frac{1}{2}\sqrt[3]{2}, \ \frac{\sqrt[6]{108}}{2}\right) \in \mathbb{R}^3. \end{aligned}$$

Now, the whole point of this is that we want to consider the resulting lattice when we take $\mathcal{O}_K$. In fact, we have:

Lemma 56.3.6
Consider the composition of the embeddings $K \hookrightarrow \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \xrightarrow{\sim} \mathbb{R}^n$. Then as before, $\mathcal{O}_K$ becomes a lattice L in $\mathbb{R}^n$, with mesh equal to

$$\frac{1}{2^{r_2}}\sqrt{|\Delta_K|}.$$

Proof. Fun linear algebra problem (you just need to manipulate determinants). Left as Problem 56D. □

From this we can deduce:

Lemma 56.3.7

Consider the composition of the embeddings $K \hookrightarrow \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \xrightarrow{\sim} \mathbb{R}^n$. Let $\mathfrak{a}$ be an ideal in $\mathcal{O}_K$. Then the image of $\mathfrak{a}$ is a lattice $L_{\mathfrak{a}}$ in $\mathbb{R}^n$ with mesh equal to

$$\frac{\mathrm{N}(\mathfrak{a})}{2^{r_2}}\sqrt{|\Delta_K|}.$$

Sketch of Proof. Let

$$d = \mathrm{N}(\mathfrak{a}) \coloneqq |\mathcal{O}_K/\mathfrak{a}|.$$

Then in the lattice $L_{\mathfrak{a}}$, we somehow only take $\frac{1}{d}$th of the points which appear in the lattice L, which is why the area increases by a factor of $\mathrm{N}(\mathfrak{a})$. To make this all precise I would need to do a lot more with lattices and geometry than I have space for in this chapter, so I will omit the details. But I hope you can see why this is intuitively true. □

§56.4 Minkowski's theorem

Now I can tell you why I insisted we move from $\mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$ to $\mathbb{R}^n$. In geometry, there's a really cool theorem of Minkowski's that goes as follows.

Theorem 56.4.1 (Minkowski)

Let $S \subseteq \mathbb{R}^n$ be a convex set containing 0 which is centrally symmetric (meaning that $x \in S \iff -x \in S$). Let L be a lattice with mesh d. If either

(a) The volume of S exceeds $2^n d$, or

(b) The volume of S equals $2^n d$ and S is compact,

then S contains a nonzero lattice point of L.

Question 56.4.2. Show that the condition $0 \in S$ is actually extraneous in the sense that any nonempty, convex, centrally symmetric set contains the origin.

Sketch of Proof. Part (a) is surprisingly simple and has a very olympiad-esque solution: it's basically Pigeonhole on areas. We'll prove part (a) in the special case $n = 2$, $L = \mathbb{Z}^2$ for simplicity as the proof can easily be generalized to any lattice and any n. Thus we want to show that any such convex set S with area more than 4 contains a lattice point.

Dissect the plane into 2×2 squares

$$[2a-1, 2a+1] \times [2b-1, 2b+1]$$

and overlay all these squares on top of each other. By the Pigeonhole Principle, we find there exist two points $p \neq q \in S$ which is mapped to the same point. Since S is symmetric, $-q \in S$. Then $\frac{1}{2}(p-q) \in S$ (convexity) and is a nonzero lattice point.

I'll briefly sketch part (b): the idea is to consider $(1+\varepsilon)S$ for $\varepsilon > 0$ (this is "S magnified by a small factor $1+\varepsilon$"). This satisfies condition (a). So for each $\varepsilon > 0$ the set of nonzero lattice points in $(1+\varepsilon)S$, say S_ε, is a *finite nonempty set* of (discrete) points (the "finite" part follows from the fact that $(1+\varepsilon)S$ is bounded). So there has to be some point that's in S_ε for every $\varepsilon > 0$ (why?), which implies it's in S. □

§56.5 The trap box

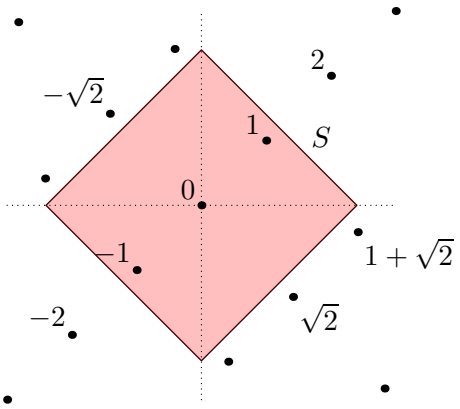
The last ingredient we need is a set to apply Minkowski's theorem to. I propose:

Definition 56.5.1. Let M be a positive real. In $\mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$, define the box S to be the set of points $(x_1, \dots, x_{r_1}, z_1, \dots, z_{r_2})$ such that

$$\sum_{i=1}^{r_1} |x_i| + 2\sum_{j=1}^{r_2} |z_j| \le M.$$

Note that this depends on the value of M.

Think of this box as a *mousetrap*: anything that falls in it is going to have a small norm, and our goal is to use Minkowski to lure some nonzero element into it.



That is, suppose $\alpha \in \mathfrak{a}$ falls into the box I've defined above, which means

$$M \ge \sum_{i=1}^{r_1} |\sigma_i(\alpha)| + 2\sum_{i=r_1+1}^{r_1+r_2} |\sigma_i(\alpha)| = \sum_{i=1}^{n} |\sigma_i(\alpha)|,$$

where we are remembering that the last few σ's come in conjugate pairs. This looks like the trace, but the absolute values are in the way. So instead, we apply AM-GM to obtain:

Lemma 56.5.2 (Effect of the mousetrap)
Let $\alpha \in \mathcal{O}_K$, and suppose $\iota(\alpha)$ is in S (where $\iota \colon K \hookrightarrow \mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$ as usual). Then

$$\mathrm{N}_{K/\mathbb{Q}}(\alpha) = \prod_{i=1}^{n} |\sigma_i(\alpha)| \le \left(\frac{M}{n}\right)^n.$$

The last step we need to do is compute the volume of the box. This is again some geometry I won't do, but take my word for it:

Lemma 56.5.3 (Size of the mousetrap)
Let $\tau \colon \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \xrightarrow{\sim} \mathbb{R}^n$ as before. Then the image of S under τ is a convex, compact, centrally symmetric set with volume

$$2^{r_1} \cdot \left(\frac{\pi}{2}\right)^{r_2} \cdot \frac{M^n}{n!}.$$

Question 56.5.4. (Sanity check) Verify that the above is correct for the signatures $(r_1, r_2) = (2, 0)$ and $(r_1, r_2) = (0, 1)$, which are the possible signatures when $n = 2$.

§56.6 The Minkowski bound

We can now put everything we have together to obtain the great Minkowski bound.

Theorem 56.6.1 (Minkowski bound)
Let $\mathfrak{a} \subseteq \mathcal{O}_K$ be any nonzero ideal. Then there exists $0 \neq \alpha \in \mathfrak{a}$ such that

$$\mathrm{N}_{K/\mathbb{Q}}(\alpha) \le \left(\frac{4}{\pi}\right)^{r_2} \frac{n!}{n^n} \sqrt{|\Delta_K|} \cdot \mathrm{N}(\mathfrak{a}).$$

Proof. This is a matter of putting all our ingredients together. Let's see what things we've defined already:

$$K \xhookrightarrow{\iota} \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \xrightarrow{\tau} \mathbb{R}^n$$

$$\text{box } S \longmapsto \tau^{\text{img}}(S) \qquad \text{with volume } 2^{r_1} \left(\tfrac{\pi}{2}\right)^{r_2} \tfrac{M^n}{n!}$$

$$\mathcal{O}_K \longmapsto \text{Lattice } L \qquad \text{with mesh } 2^{-r_2}\sqrt{|\Delta_K|}$$

$$\mathfrak{a} \longmapsto \text{Lattice } L_{\mathfrak{a}} \qquad \text{with mesh } 2^{-r_2}\sqrt{|\Delta_K|}\,\mathrm{N}(\mathfrak{a})$$

Pick a value of M such that the mesh of $L_{\mathfrak{a}}$ equals 2^{-n} of the volume of the box. Then Minkowski's theorem gives that some $0 \neq \alpha \in \mathfrak{a}$ lands inside the box — the mousetrap is configured to force $\mathrm{N}_{K/\mathbb{Q}}(\alpha) \le \frac{1}{n^n} M^n$. The correct choice of M is

$$M^n = M^n \cdot 2^n \cdot \frac{\text{mesh}}{\text{vol box}} = 2^n \cdot \frac{n!}{2^{r_1} \cdot \left(\frac{\pi}{2}\right)^{r_2}} \cdot 2^{-r_2}\sqrt{|\Delta_K|}\,\mathrm{N}(\mathfrak{a})$$

which gives the bound after some arithmetic. □

§56.7 The class group is finite

Definition 56.7.1. Let $M_K = \left(\frac{4}{\pi}\right)^{r_2} \frac{n!}{n^n}\sqrt{|\Delta_K|}$ for brevity. Note that it is a constant depending on K.

So that's cool and all, but what we really wanted was to show that the class group is finite. How can the Minkowski bound help? The key idea is the following:

The class of $\mathfrak{a}$ is entirely determined by $(\alpha) \cdot \mathfrak{a}^{-1}$.

Question 56.7.2. Verify this. (That is, if $\mathfrak{a}$ and $\mathfrak{b}$ are such that $(\alpha) \cdot \mathfrak{a}^{-1} = (\beta) \cdot \mathfrak{b}^{-1}$ for some $\alpha \in \mathfrak{a}$, $\beta \in \mathfrak{b}$, then prove that $\mathfrak{a} = (\gamma)\mathfrak{b}$ for some $\gamma \in K$.)

Example 56.7.3

Recall this example:

$$(6) = (2, 1-\sqrt{-5})^2(3, 1+\sqrt{-5})(3, 1-\sqrt{-5}) = \mathfrak{p}^2\mathfrak{q}_1\mathfrak{q}_2.$$

Consider $\mathfrak{a} = (2)$ being principal, pick $\alpha = 2$, then $(\alpha) \cdot \mathfrak{a}^{-1} = (1)$. If $(\alpha) \cdot \mathfrak{a}^{-1} = (1)$ (or (2), or anything principal), we know $\mathfrak{a}$ must be principal.
On the other hand, $\mathfrak{q}_1 = (3, 1+\sqrt{-5})$ is not principal, pick $\alpha = 3$. We know $(3) = \mathfrak{q}_1\mathfrak{q}_2$, so $(\alpha) \cdot \mathfrak{q}_1^{-1} = \mathfrak{q}_2 \neq (1)$.

In both examples above, $\mathfrak{a} \mid (\alpha)$, so their quotient should be an "integer". Indeed:

Question 56.7.4. Show that $(\alpha) \cdot \mathfrak{a}^{-1}$ is an integral ideal. (Unwind definitions.)

You might notice that we can rewrite the Minkowski bound to say

$$\mathrm{N}\left((\alpha) \cdot \mathfrak{a}^{-1}\right) \le M_K$$

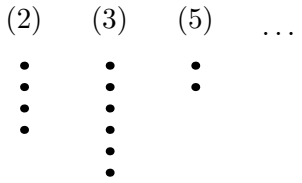
where M_K is some constant depending on K.

This statement is helpful, because in fact, there are only finitely many integral ideals with norm $\le M_K$.

Corollary 56.7.5 (Finiteness of class group)
Class groups are always finite.

Proof. We just have to show there are finitely many integral ideals as above; this will mean there are finitely many classes.

Suppose we want to build such an ideal $\mathfrak{a} = \mathfrak{p}_1^{e_1} \dots \mathfrak{p}_m^{e_m}$. Recall that a prime ideal $\mathfrak{p}_i$ must have some rational prime p inside it, meaning $\mathfrak{p}_i$ divides (p) and p divides $\mathrm{N}(\mathfrak{p}_i)$. So let's group all the $\mathfrak{p}_i$ we want to build $\mathfrak{a}$ with based on which (p) they came from.



To be more dramatic: imagine you have a *cherry tree*; each branch corresponds to a prime (p) and contains as cherries (prime ideals) the factors of (p) (finitely many). Your bucket (the ideal $\mathfrak{a}$ you're building) can only hold a total weight (norm) of M_K. So you can't even touch the branches higher than M_K. You can repeat cherries (oops), but the weight of a cherry on branch (p) is definitely $\ge p$; all this means that the number of ways to build $\mathfrak{a}$ is finite. □

§56.8 Computation of class numbers

Definition 56.8.1. The order of Cl_K is called the **class number** of K.

Remark 56.8.2 — If $\mathrm{Cl}_K = 1$, then $\mathcal{O}_K$ is a PID, hence a UFD.

By computing the actual value of M_K, we can quite literally build the entire "cherry tree" mentioned in the previous proof. Let's give an example how!

Proposition 56.8.3

The field $\mathbb{Q}(\sqrt{-67})$ has class number 1.

Proof. Since $K = \mathbb{Q}(\sqrt{-67})$ has signature $(0,1)$ and discriminant $\Delta_K = -67$ (since $-67 \equiv 1 \pmod 4$) we can compute

$$M_K = \left(\frac{4}{\pi}\right)^1 \cdot \frac{2!}{2^2}\sqrt{67} \approx 5.2.$$

That means we can cut off the cherry tree after (2), (3), (5), since any cherries on these branches will necessarily have norm $\geq M_K$. We now want to factor each of these in $\mathcal{O}_K = \mathbb{Z}[\theta]$, where $\theta = \frac{1+\sqrt{-67}}{2}$ has minimal polynomial $x^2 - x + 17$. But something miraculous happens:

- When we try to reduce $x^2 - x + 17 \pmod 2$, we get an irreducible polynomial $x^2 - x + 1$. By the factoring algorithm (Theorem 55.5.4) this means (2) is prime.
- Similarly, reducing mod 3 gives $x^2 - x + 2$, which is irreducible. This means (3) is prime.
- Finally, for the same reason, (5) is prime.

It's our lucky day; all of the ideals (2), (3), (5) are prime (already principal). To put it another way, each of the three branches has only one (large) cherry on it. That means any time we put together an integral ideal with norm $\leq M_K$, it is actually principal. In fact, these guys have norm 4, 9, 25 respectively... so we can't even touch (3) and (5), and the only ideals we can get are (1) and (2) (with norms 1 and 4).

Now we claim that's all. Suppose $\mathfrak{b}$ is an integral ideal such that $\mathrm{N}(\mathfrak{b}) \leq M_K$. By the above, either $\mathfrak{b} = (1)$ or $\mathfrak{b} = (2)$, both of which are principal, and hence trivial in Cl_K. So J is trivial in Cl_K too, as needed. □

Let's do a couple more.

Theorem 56.8.4 (Gaussian integers $\mathbb{Z}[i]$ form a UFD)

The field $\mathbb{Q}(i)$ has class number 1.

Proof. This is $\mathcal{O}_K$ where $K = \mathbb{Q}(i)$, so we just want Cl_K to be trivial. We have $M_K = \frac{2}{\pi}\sqrt{4} < 2$. So every class has an integral ideal of norm $\mathfrak{b}$ satisfying

$$\mathrm{N}(\mathfrak{b}) \leq \left(\frac{4}{\pi}\right)^1 \cdot \frac{2!}{2^2} \cdot \sqrt{4} = \frac{4}{\pi} < 2.$$

Well, that's silly: we don't have any branches to pick from at all. In other words, we can only have $\mathfrak{b} = (1)$. □

Here's another example of something that still turns out to be unique factorization, but this time our cherry tree will actually have cherries that can be picked.

Proposition 56.8.5 ($\mathbb{Z}[\sqrt{7}]$ is a UFD)
The field $\mathbb{Q}(\sqrt{7})$ has class number 1.

Proof. First we compute the Minkowski bound.

Question 56.8.6. Check that $M_K \approx 2.646$.

So this time, the only branch is (2). Let's factor (2) as usual: the polynomial x^2+7 reduces as $(x-1)(x+1) \pmod 2$, and hence

$$(2) = \left(2, \sqrt{7}-1\right)\left(2, \sqrt{7}+1\right).$$

Oops! We now have two cherries, and they both seem reasonable. But actually, I claim that

$$\left(2, \sqrt{7}-1\right) = \left(3-\sqrt{7}\right).$$

Question 56.8.7. Prove this.

So both the cherries are principal ideals, and as before we conclude that Cl_K is trivial. But note that this time, the prime ideal (2) actually splits; we got lucky that the two cherries were principal but this won't always work. □

How about some nontrivial class groups? First, we use a lemma that will help us with narrowing down the work in our cherry tree.

Lemma 56.8.8 (Ideals divide their norms)
Let $\mathfrak{b}$ be an integral ideal with $\mathrm{N}(\mathfrak{b}) = n$. Then $\mathfrak{b}$ divides the ideal (n).

Proof. By definition, $n = |\mathcal{O}_K/\mathfrak{b}|$. Treating $\mathcal{O}_K/\mathfrak{b}$ as an (additive) abelian group and using Lagrange's theorem, we find

$$0 \equiv \underbrace{\alpha + \cdots + \alpha}_{n \text{ times}} = n\alpha \pmod{\mathfrak{b}} \qquad \text{for all } \alpha \in \mathcal{O}_K.$$

Thus $(n) \subseteq \mathfrak{b}$, done. □

Alternatively, if you have read Chapter 59: If the extension $K/\mathbb{Q}$ is Galois, we can actually prove that, analogous to Remark 54.1.9, $\prod_{\sigma \in \mathrm{Gal}(K/\mathbb{Q})} \sigma(\mathfrak{b}) = (n)$, implying the result $id(\mathfrak{b}) \mid (n)$.

Now we can give such an example.

Proposition 56.8.9 (Class group of $\mathbb{Q}(\sqrt{-17})$)
The number field $K = \mathbb{Q}(\sqrt{-17})$ has class group $\mathbb{Z}/4\mathbb{Z}$.

You are not obliged to read the entire proof in detail, as it is somewhat gory. The idea is just that there are some cherries which are not trivial in the class group.

Proof. Since $\Delta_K = -68$, we compute the Minkowski bound

$$M_K = \frac{4}{\pi}\sqrt{17} < 6.$$

Now, it suffices to factor with (2), (3), (5). The minimal polynomial of $\sqrt{-17}$ is x^2+17, so as usual

$$\begin{aligned}(2) &= (2, \sqrt{-17}+1)^2 \\ (3) &= (3, \sqrt{-17}-1)(3, \sqrt{-17}+1) \\ (5) &= (5)\end{aligned}$$

corresponding to the factorizations of x^2+17 modulo each of 2, 3, 5. Set $\mathfrak{p} = (2, \sqrt{-17}+1)$ and $\mathfrak{q}_1 = (3, \sqrt{-17}-1)$, $\mathfrak{q}_2 = (3, \sqrt{-17}+1)$. We can compute

$$\mathrm{N}(\mathfrak{p}) = 2 \quad \text{and} \quad \mathrm{N}(\mathfrak{q}_1) = \mathrm{N}(\mathfrak{q}_2) = 3.$$

In particular, they are not principal. The ideal (5) is out the window; it has norm 25. Hence, the three cherries are $\mathfrak{p}$, $\mathfrak{q}_1$, $\mathfrak{q}_2$.

The possible ways to arrange these cherries into ideals with norm ≤ 5 are

$$\left\{(1), \mathfrak{p}, \mathfrak{q}_1, \mathfrak{q}_2, \mathfrak{p}^2\right\}.$$

However, you can compute

$$\mathfrak{p}^2 = (2)$$

so $\mathfrak{p}^2$ and (1) are in the same class group; that is, they are trivial. In particular, the class group has order at most 4.

From now on, let $[\mathfrak{a}]$ denote the class (member of the class group) that $\mathfrak{a}$ is in. Since $\mathfrak{p}$ isn't principal (so $[\mathfrak{p}] \neq [(1)]$), it follows that $\mathfrak{p}$ has order two. So Lagrange's theorem says that Cl_K has order either 2 or 4.

Now we claim $[\mathfrak{q}_1]^2 \neq [(1)]$, which implies that $\mathfrak{q}_1$ has order greater than 2. If not, $\mathfrak{q}_1^2$ is principal. We know $\mathrm{N}(\mathfrak{q}_1) = 3$, so this can only occur if $\mathfrak{q}_1^2 = (3)$; this would force $\mathfrak{q}_1 = \mathfrak{q}_2$. This is impossible since $\mathfrak{q}_1 + \mathfrak{q}_2 = (1)$.

Thus, $\mathfrak{q}_1$ has even order greater than 2. So it has to have order 4. From this we deduce

$$\mathrm{Cl}_K \cong \mathbb{Z}/4\mathbb{Z}.$$

□

Remark 56.8.10 — When we did this at Harvard during Math 129, there was a five-minute interruption in which students (jokingly) complained about the difficulty of evaluating $\frac{4}{\pi}\sqrt{17}$. Excerpt:

> "Will we be allowed to bring a small calculator on the exam?" – Student 1
> "What does the size have to do with anything? You could have an Apple Watch" – Professor
> "Just use the fact that $\pi \ge 3$" – me
> "Even [other professor] doesn't know that, how are we supposed to?" – Student 2
> "You have to do this yourself!" – Professor
> "This is an outrage." – Student 1

§56.9 Optional: Proof that $\mathcal{O}_K$ is a free $\mathbb{Z}$-module

We have the suitable tools to prove Theorem 54.2.12 now.

We know $\mathcal{O}_K$ is a ring, so obviously it must be a $\mathbb{Z}$-module. Suppose it is not a free $\mathbb{Z}$-module of degree $n = |K : \mathbb{Q}|$. What could go wrong?

- First, it may happen that it is dense like $\mathbb{Q}$ or the ring extension $\mathbb{Z}[\frac{1}{2}]$ (which makes it not finitely generated and not free).
- Even without that, it may happen that its rank is less than n.

The second possibility is much easier to discard. Let $\alpha_1, \dots, \alpha_n$ be a basis of K. Using Theorem 54.2.6, we have positive integers $d_1, \dots, d_n$ such that $d_1\alpha_1, \dots, d_n\alpha_n \in \mathcal{O}_K$. Since $\alpha_1, \dots, \alpha_n$ are linearly independent, this implies $\operatorname{rank} \mathcal{O}_K \geq n$.

The other direction is harder. We wish to prove $\mathcal{O}_K$ is "discrete" in some sense.

From now on, replace α_i with $d_i\alpha_i$, so they are still a basis of the $\mathbb{Q}$-vector space K, and furthermore they are now in $\mathcal{O}_K$.

Three distinct proofs will be provided.

§56.9.i First proof

We will show that $\mathcal{O}_K$ is contained in some free $\mathbb{Z}$-module of rank n.

Specifically, we will show that $\mathcal{O}_K \subseteq \langle \frac{1}{d}\alpha_1, \dots, \frac{1}{d}\alpha_n \rangle$ for some integer $d \neq 0$.

Let us pretend that we already know $\mathcal{O}_K$ is a free $\mathbb{Z}$-module of rank n. How would we compute d?

> **Exercise 56.9.1.** These are a few naive attempts to compute d; unfortunately, they wouldn't work. Verify that on $\langle 1, 1+2i \rangle \subseteq \mathbb{Z}[i]$.
>
> - Take d to be the first positive integer that belongs to $\langle \alpha_1, \dots, \alpha_n \rangle$. (Attempt inspired by Theorem 55.3.6.)
> - Take d to be the product of the norm of $\alpha_1, \dots, \alpha_n$.

Instead, we compute d by using an idea inspired by how we compute the mesh of the lattice. Let $A = \langle \alpha_1, \dots, \alpha_n \rangle$, then it is a lattice and a free $\mathbb{Z}$-module of rank n.

> **Exercise 56.9.2.** Assume you already know $\mathcal{O}_K$ is free of rank n. Show that $|\mathcal{O}_K/A|$ is finite. Conclude that for every $x \in \mathcal{O}_K$, then $|\mathcal{O}_K/A| \cdot x \in A$.

Using the same argument as Problem 57B*, you can prove that the "discriminant" (squared mesh) of the lattice spanned by $\alpha_1, \dots, \alpha_n$ is an integer. Formally, let

$$d \coloneqq \det \begin{bmatrix} \sigma_1(\alpha_1) & \dots & \sigma_n(\alpha_1) \\ \vdots & \ddots & \vdots \\ \sigma_1(\alpha_n) & \dots & \sigma_n(\alpha_n) \end{bmatrix}^2.$$

> **Exercise 56.9.3.** Convince yourself (at least when all embeddings are real) that the ratio of the meshes of $\mathcal{O}_K$ and A is exactly the size of the quotient abelian group $\mathcal{O}_K/A$, that is $\left|\frac{d}{\Delta_K}\right| = |\mathcal{O}_K/A|^2$. Conclude that for every $x \in \mathcal{O}_K$, then $d \cdot x \in A$.

Which implies $\mathcal{O}_K \subseteq \langle \frac{1}{d}\alpha_1, \dots, \frac{1}{d}\alpha_n \rangle$, which is another free $\mathbb{Z}$-module of rank n.

However, the argument above is circular since it assumes Δ_K exists (it can only serve as a motivation for where the value d comes from). The actual proof is the following.

Similar to Problem 57B*, we define

$$d \coloneqq \det[\mathrm{Tr}_{K/\mathbb{Q}}(\alpha_i \alpha_j)]_{i,j}.$$

Then, because $\{\alpha_i\}_i$ spans K as a $\mathbb{Q}$-vector space, there is some $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{Q}^{\oplus n}$ such that

$$\begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x \end{pmatrix}.$$

Which means

$$\begin{bmatrix} \mathrm{Tr}_{K/\mathbb{Q}}(\alpha_1\alpha_1) & \cdots & \mathrm{Tr}_{K/\mathbb{Q}}(\alpha_1\alpha_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Tr}_{K/\mathbb{Q}}(\alpha_n\alpha_1) & \cdots & \mathrm{Tr}_{K/\mathbb{Q}}(\alpha_n\alpha_n) \end{bmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \mathrm{Tr}_{K/\mathbb{Q}}(\alpha_1 x) \\ \vdots \\ \mathrm{Tr}_{K/\mathbb{Q}}(\alpha_n x) \end{pmatrix}.$$

Or, in short,

$$(\text{some matrix} \in \mathrm{GL}_n(\mathbb{Z}) \text{ with determinant } d) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = (\text{some vector} \in \mathbb{Z}^{\oplus n}).$$

Question 56.9.4. Finish the proof that $d \cdot x \in A$. (Cramer's rule. Or take the adjoint.)

Finally, we have that $\mathcal{O}_K$ is squeezed between two free $\mathbb{Z}$-modules of rank n, so it is also free of rank n.[2]

Question 56.9.5. Finish this. (Theorem 18.1.5 can be used here.)

§56.9.ii Second proof

This time around, instead of dividing by d, we instead take the *dual lattice*.

What is the dual lattice, and why would we think about using it? One way to motivate this proof is to look at the inverse of an ideal: if $(1) \mid \mathfrak{a}$ (i.e. $\mathfrak{a}$ is an integral ideal), then $\mathfrak{a}^{-1} \mid (1)$.

Here, $\mathfrak{a}^{-1} \coloneqq \{x \in K \mid xa \in \mathcal{O}_K \text{ for all } a \in \mathfrak{a}\}$.

We can think of doing something similar, considering

$$\{x \in K \mid xa \in \mathcal{O}_K \text{ for all } a \in \langle \alpha_1, \dots, \alpha_n \rangle\}.$$

This set is actually a lattice of rank n, but this won't work to prove the argument! We're trying to prove $\mathcal{O}_K$ is "discrete" in the first place, if $\mathcal{O}_K = K$ then the set above would equal K as well.

Instead, we must rely on what we already know — the discreteness of $\mathbb{Z}$. Define

$$S \coloneqq \{x \in K \mid \mathrm{Tr}_{K/\mathbb{Q}}(xa) \in \mathbb{Z} \text{ for all } a \in \langle \alpha_1, \dots, \alpha_n \rangle\}.$$

This set is a bit larger than the previous set.

[2] We did something similar in Remark 55.3.7.

Question 56.9.6. Verify that this set is a superset of the previous one. Then conclude that $\mathcal{O}_K \subseteq S$.

Exercise 56.9.7. Consider $\langle 1, 2i \rangle \subseteq \mathbb{Z}[i]$. What would the set S be? (Recall that the trace in $\mathbb{C}$ is just twice the real part.)

S is still a lattice of rank n — and this time, we can actually prove it! Since we already know $\mathbb{Z}$ is discrete.

Now, this is a purely algebraic problem, we will only need to use knowledge of vector space here. Each element $x \in K$ can be written as a vector

$$F(x) \coloneqq \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{Q}^{\oplus n}$$

if we use the basis $\{\alpha_1, \dots, \alpha_n\}$.

Exercise 56.9.8. Show that $(x, y) \mapsto \operatorname{Tr}_{K/\mathbb{Q}}(x \cdot y)$ can then be written as an *invertible* matrix in $\mathrm{GL}_n(\mathbb{Q})$ — specifically, there exists $M \in \mathrm{GL}_n(\mathbb{Q})$ such that $\operatorname{Tr}_{K/\mathbb{Q}}(x \cdot y) = F(x)^\top M F(y)$, where $F(x)^\top$ is the transpose of $F(x)$.

Which makes $(x, y) \mapsto \operatorname{Tr}_{K/\mathbb{Q}}(x \cdot y)$ *almost* an inner product (see Chapter 13), except that it is not positive definite (for example, in $\mathbb{C}$, we have $\operatorname{Tr}((1+i)^2) = 0$). But having the matrix invertible suffices to do the following:

Exercise 56.9.9. Finish the proof. (Hint: Consider the matrix $\begin{bmatrix} F(\alpha_1) & \cdots & F(\alpha_n) \end{bmatrix} \in \mathrm{GL}_n(\mathbb{Q})$. What is the condition on $F(x)$ such that $x \in S$?)

§56.9.iii Third proof

Recall that we wish to prove $\mathcal{O}_K$ is a free $\mathbb{Z}$-module of rank n. To do that, we suppose it cannot be generated by n elements, then show that $\mathcal{O}_K$ is not discrete, and this causes trouble because we know the norm is continuous.

Exercise 56.9.10. Consider $K \cong \mathbb{Q}^{\oplus n}$, this gives a topology on K. Verify that $x \mapsto |\operatorname{N}(x)|$ is indeed continuous.

We have that for every $x \in \mathcal{O}_K$, then $|\operatorname{N}(x)| \in \mathbb{Z}$.

Exercise 56.9.11. Conclude that there is a ball $B(0, r)$ in the topology above that contains no element of $\mathcal{O}_K$, except 0.

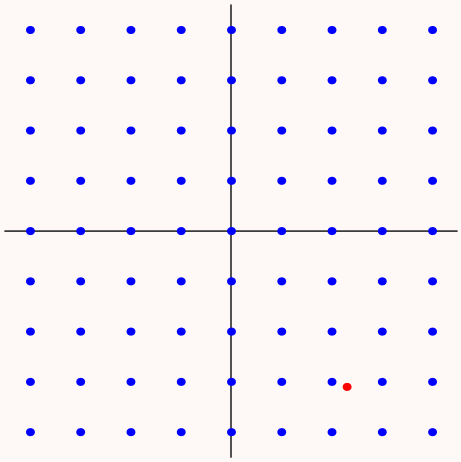
Now, what goes wrong if $\mathcal{O}_K$ cannot be generated by n elements? Things go wrong quickly:

Exercise 56.9.12. Suppose $\langle \alpha_1, \dots, \alpha_n \rangle$ generates K as a $\mathbb{Q}$-vector space. Let $x \in \mathcal{O}_K$ such that x is not a $\mathbb{Z}$-linear combination of $\{\alpha_1, \dots, \alpha_n\}$. Show that there is $z_1, \dots, z_n \in [-\frac{1}{2}, \frac{1}{2}]$ not all zero, and z a $\mathbb{Z}$-linear combination of $\{\alpha_1, \dots, \alpha_n\}$ such that $x + z = \sum_{i=1}^n z_i \alpha_i$.

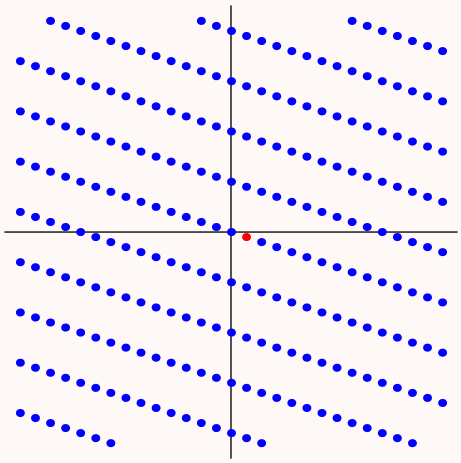
Exercise 56.9.13. With notation as above, suppose $z_1 \neq 0$. Show that if we replace α_1 with z_1, then the new set $\{z_1, \alpha_2, \ldots, \alpha_n\}$ still spans K as a $\mathbb{Q}$-vector space; furthermore, the mesh of the lattice gets decreased by *at least a half*.

Example 56.9.14

Suppose $A \subseteq \mathbb{C}$ is a lattice. We know $1 \in A$ and $i \in A$, so $\mathbb{Z}[i] \subseteq A$. Assume we know in addition that $2.3 - 3.1i \in A$.



Because A is closed under addition, we know that $(2.3 - 3.1i) + (-2 + 3i) = 0.3 - 0.1i \in A$. We replace 1 in the basis with $0.3 - 0.1i$. Then, the lattice $\langle i, 0.3 - 0.1i \rangle \subseteq A$ has smaller mesh than $\langle 1, i \rangle$ as expected.

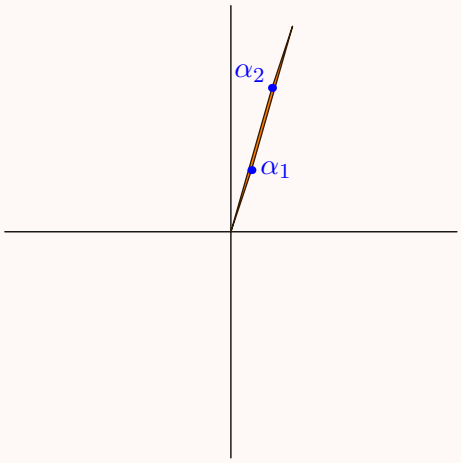


Since we can do this indefinitely ($\mathcal{O}_K$ never gets spanned), the mesh decreases to 0 rapidly.

So, are we done? Since the mesh goes to 0, maybe the smallest distance of some α_i also go to 0? Almost, but not quite:

Example 56.9.15

Consider $\alpha_1 = 1 + 3i$ and $\alpha_2 = 2 + 7i$.



The mesh of the lattice spanned by α_1 and α_2 is 1, however, neither α_1 nor α_2 are particularly close to the origin.

We need to apply Minkowski bound here again: suppose the mesh is d, then the cube centered at origin with volume $2^n d$ contains a nonzero lattice point.[3] This point's distance cannot be more than $\sqrt{n} \cdot \sqrt[n]{d}$ away from the origin.

Remark 56.9.16 — Speaking of which, the LLL lattice basis reduction algorithm can be used to find *in practice* a point on the lattice that is *close enough* to the origin.

For d small enough, $\sqrt{n} \cdot \sqrt[n]{d} < r$ where r is the ball of no lattice point we have shown above, which gives a contradiction. So we're done!

§56.10 A few harder problems to think about

Problem 56A. Show that $K = \mathbb{Q}(\sqrt{-163})$ has trivial class group, and hence $\mathcal{O}_K = \mathbb{Z}\left[\frac{1+\sqrt{-163}}{2}\right]$ is a UFD.[4]

Problem 56B. Determine the class group of $\mathbb{Q}(\sqrt{-31})$.

Problem 56C (China TST 1998)**.** Let n be a positive integer. A polygon in the plane (not necessarily convex) has area greater than n. Prove that one can translate it so that it contains at least $n+1$ lattice points.

Problem 56D (Lemma 56.3.6)**.** Consider the composition of the embeddings $K \hookrightarrow \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \xrightarrow{\sim} \mathbb{R}^n$. Show that the image of $\mathcal{O}_K \subseteq K$ has mesh equal to

$$\frac{1}{2^{r_2}}\sqrt{|\Delta_K|}.$$

Problem 56E. Let $p \equiv 1 \pmod 4$ be a prime. Show that there are unique integers $a > b > 0$ such that $a^2 + b^2 = p$.

[3] Using a sphere would of course make for a better bound, but its volume is a bit harder to calculate.

[4] In fact, $n = 163$ is the largest number for which $\mathbb{Q}(\sqrt{-n})$ has trivial class group. The complete list is $1, 2, 3, 7, 11, 19, 43, 67, 163$, the **Heegner numbers**. You might notice Euler's prime-generating polynomial $t^2 + t + 41$ when doing the above problem. Not a coincidence!

Problem 56F (Korea national olympiad 2014)**.** Let p be an odd prime and k a positive integer such that $p \mid k^2 + 5$. Prove that there exist positive integers m, n such that $p^2 = m^2 + 5n^2$.

57 More properties of the discriminant

I'll remind you that the discriminant of a number field K is given by

$$\Delta_K \coloneqq \det \begin{bmatrix} \sigma_1(\alpha_1) & \dots & \sigma_n(\alpha_1) \\ \vdots & \ddots & \vdots \\ \sigma_1(\alpha_n) & \dots & \sigma_n(\alpha_n) \end{bmatrix}^2$$

where $\alpha_1, \dots, \alpha_n$ is a $\mathbb{Z}$-basis for K, and the σ_i are the n embeddings of K into $\mathbb{C}$.

Several examples, properties, and equivalent definitions follow.

§57.1 A few harder problems to think about

Problem 57A* (Discriminant of cyclotomic field)**.** Let p be an odd rational prime and ζ_p a primitive pth root of unity. Let $K = \mathbb{Q}(\zeta_p)$. Show that

$$\Delta_K = (-1)^{\frac{p-1}{2}} p^{p-2}.$$

Problem 57B* (Trace representation of Δ_K)**.** Let $\alpha_1, \dots, \alpha_n$ be a basis for $\mathcal{O}_K$. Prove that

$$\Delta_K = \det \begin{bmatrix} \operatorname{Tr}_{K/\mathbb{Q}}(\alpha_1^2) & \operatorname{Tr}_{K/\mathbb{Q}}(\alpha_1\alpha_2) & \dots & \operatorname{Tr}_{K/\mathbb{Q}}(\alpha_1\alpha_n) \\ \operatorname{Tr}_{K/\mathbb{Q}}(\alpha_2\alpha_1) & \operatorname{Tr}_{K/\mathbb{Q}}(\alpha_2^2) & \dots & \operatorname{Tr}_{K/\mathbb{Q}}(\alpha_2\alpha_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Tr}_{K/\mathbb{Q}}(\alpha_n\alpha_1) & \operatorname{Tr}_{K/\mathbb{Q}}(\alpha_n\alpha_2) & \dots & \operatorname{Tr}_{K/\mathbb{Q}}(\alpha_n\alpha_n) \end{bmatrix}.$$

In particular, Δ_K is an integer.

Problem 57C* (Root representation of Δ_K)**.** The **discriminant** of a quadratic polynomial $Ax^2 + Bx + C$ is defined as $B^2 - 4AC$. More generally, the polynomial discriminant of a polynomial $f \in \mathbb{Z}[x]$ of degree n is

$$\Delta(f) \coloneqq c^{2n-2} \prod_{1 \le i < j \le n} (z_i - z_j)^2$$

where $z_1, \dots, z_n$ are the roots of f, and c is the leading coefficient of f.

Suppose K is monogenic with $\mathcal{O}_K = \mathbb{Z}[\theta]$. Let f denote the minimal polynomial of θ (hence monic). Show that

$$\Delta_K = \Delta(f).$$

Problem 57D. Show that if $K \neq \mathbb{Q}$ is a number field then $|\Delta_K| > 1$.

Problem 57E (Brill's theorem)**.** For a number field K with signature (r_1, r_2), show that $\Delta_K > 0$ if and only if r_2 is even.

Problem 57F (Stickelberger theorem)**.** Let K be a number field. Prove that

$$\Delta_K \equiv 0 \text{ or } 1 \pmod 4.$$

58 Bonus: Let's solve Pell's equation!

This is an optional aside, and can be safely ignored. (On the other hand, it's pretty short.)

§58.1 Units

Prototypical example for this section: ± 1, *roots of unity,* $3 - 2\sqrt{2}$ *and its powers.*

Recall according to Problem 54A* that $\alpha \in \mathcal{O}_K$ is invertible if and only if

$$\mathrm{N}_{K/\mathbb{Q}}(\alpha) = \pm 1.$$

We let $\mathcal{O}_K^\times$ denote the set of units of $\mathcal{O}_K$.

Question 58.1.1. Show that $\mathcal{O}_K^\times$ is a group under multiplication. Hence we name it the **unit group** of $\mathcal{O}_K$.

What are some examples of units?

Example 58.1.2 (Examples of units in a number field)

1. ± 1 are certainly units, present in any number field.
2. If $\mathcal{O}_K$ contains a root of unity ω (i.e. $\omega^n = 1$), then ω is a unit. (In fact, ± 1 are special cases of this.)
3. Of course, not all units of $\mathcal{O}_K$ are roots of unity. For example, if $\mathcal{O}_K = \mathbb{Z}[\sqrt{3}]$ (from $K = \mathbb{Q}(\sqrt{3})$) then the number $2 + \sqrt{3}$ is a unit, as its norm is
$$\mathrm{N}_{K/\mathbb{Q}}(2 + \sqrt{3}) = 2^2 - 3 \cdot 1^2 = 1.$$
Alternatively, just note that the inverse $2 - \sqrt{3} \in \mathcal{O}_K$ as well:
$$\left(2 - \sqrt{3}\right)\left(2 + \sqrt{3}\right) = 1.$$
Either way, $2 - \sqrt{3}$ is a unit.
4. Given any unit $u \in \mathcal{O}_K^\times$, all its powers are also units. So for example, $(3 - 2\sqrt{2})^n$ is always a unit of $\mathbb{Z}[\sqrt{2}]$, for any n. If u is not a root of unity, then this generates infinitely many new units in $\mathcal{O}_K^\times$.

Question 58.1.3. Verify the claims above that

(a) Roots of unity are units, and

(b) Powers of units are units.

One can either proceed from the definition or use the characterization $\mathrm{N}_{K/\mathbb{Q}}(\alpha) = \pm 1$. If one definition seems more natural to you, use the other.

§58.2 Dirichlet's unit theorem

Prototypical example for this section: The units of $\mathbb{Z}[\sqrt{3}]$ *are* $\pm(2+\sqrt{3})^n$.

Definition 58.2.1. Let $\mu(\mathcal{O}_K)$ denote the set of roots of unity contained in a number field K (equivalently, in $\mathcal{O}_K$).

Example 58.2.2 (Examples of $\mu(\mathcal{O}_K)$)

(a) If $K = \mathbb{Q}(i)$, then $\mathcal{O}_K = \mathbb{Z}[i]$. So

$$\mu(\mathcal{O}_K) = \{\pm 1, \pm i\} \quad \text{where } K = \mathbb{Q}(i).$$

(b) If $K = \mathbb{Q}(\sqrt{3})$, then $\mathcal{O}_K = \mathbb{Z}[\sqrt{3}]$. So

$$\mu(\mathcal{O}_K) = \{\pm 1\} \quad \text{where } K = \mathbb{Q}(\sqrt{3}).$$

(c) If $K = \mathbb{Q}(\sqrt{-3})$, then $\mathcal{O}_K = \mathbb{Z}[\frac{1}{2}(1+\sqrt{-3})]$. So

$$\mu(\mathcal{O}_K) = \left\{ \pm 1, \frac{\pm 1 \pm \sqrt{-3}}{2} \right\} \quad \text{where } K = \mathbb{Q}(\sqrt{-3})$$

where the $\pm$'s in the second term need not depend on each other; in other words $\mu(\mathcal{O}_K) = \{z \mid z^6 = 1\}$.

Exercise 58.2.3. Show that we always have that $\mu(\mathcal{O}_K)$ comprises the roots to $x^n - 1$ for some integer n. (First, show it is a finite group under multiplication.)

We now quote, without proof, the so-called Dirichlet's unit theorem, which gives us a much more complete picture of what the units in $\mathcal{O}_K$ are. Legend says that Dirichlet found the proof of this theorem during an Easter concert in the Sistine Chapel.

Theorem 58.2.4 (Dirichlet's unit theorem)

Let K be a number field with signature (r_1, r_2) and set

$$s = r_1 + r_2 - 1.$$

Then there exist units $u_1, \dots, u_s$ such that every unit $\alpha \in \mathcal{O}_K^\times$ can be written *uniquely* in the form

$$\alpha = \omega \cdot u_1^{n_1} \dots u_s^{n_s}$$

for $\omega \in \mu(\mathcal{O}_K)$ is a root of unity, and $n_1, \dots, n_s \in \mathbb{Z}$.

More succinctly:

We have $\mathcal{O}_K^\times \cong \mathbb{Z}^{r_1+r_2-1} \times \mu(\mathcal{O}_K)$.

A choice of $u_1, \dots, u_s$ is called a choice of **fundamental units**.

Here are some example applications.

Example 58.2.5 (Some unit groups)

(a) Let $K = \mathbb{Q}(i)$ with signature $(0,1)$. Then we obtain $s = 0$, so Dirichlet's Unit theorem says that there are no units other than the roots of unity. Thus

$$\mathcal{O}_K^\times = \{\pm 1, \pm i\} \quad \text{where } K = \mathbb{Q}(i).$$

This is not surprising, since $a + bi \in \mathbb{Z}[i]$ is a unit if and only if $a^2 + b^2 = 1$.

(b) Let $K = \mathbb{Q}(\sqrt{3})$, which has signature $(2,0)$. Then $s = 1$, so we expect exactly one fundamental unit. A fundamental unit is $2 + \sqrt{3}$ (or $2 - \sqrt{3}$, its inverse) with norm 1, and so we find

$$\mathcal{O}_K^\times = \left\{ \pm(2+\sqrt{3})^n \mid n \in \mathbb{Z} \right\}.$$

(c) Let $K = \mathbb{Q}(\sqrt[3]{2})$ with signature $(1,1)$. Then $s = 1$, so we expect exactly one fundamental unit. The choice $1 + \sqrt[3]{2} + \sqrt[3]{4}$. So

$$\mathcal{O}_K^\times = \left\{ \pm\left(1+\sqrt[3]{2}+\sqrt[3]{4}\right)^n \mid n \in \mathbb{Z} \right\}.$$

I haven't actually shown you that these are fundamental units, and indeed computing fundamental units is in general hard.

§58.3 Finding fundamental units

Here is a table with some fundamental units.

d	Unit
$d = 2$	$1 + \sqrt{2}$
$d = 3$	$2 + \sqrt{3}$
$d = 5$	$\frac{1}{2}(1 + \sqrt{5})$
$d = 6$	$5 + 2\sqrt{6}$
$d = 7$	$8 + 3\sqrt{7}$
$d = 10$	$3 + \sqrt{10}$
$d = 11$	$10 + 3\sqrt{11}$

In general, determining fundamental units is computationally hard.

However, once I tell you what the fundamental unit is, it's not too bad (at least in the case $s = 1$) to verify it. For example, suppose we want to show that $10 + 3\sqrt{11}$ is a fundamental unit of $K = \mathbb{Q}(\sqrt{11})$, which has ring of integers $\mathbb{Z}[\sqrt{11}]$. If not, then for some $n > 1$, we would have to have

$$10 + 3\sqrt{11} = \pm\left(x + y\sqrt{11}\right)^n.$$

For this to happen, at the very least we would need $|y| < 3$. We would also have $x^2 - 11y^2 = \pm 1$. So one can just verify (using $y = 1, 2$) that this fails.

The point is that: Since $(10, 3)$ is the *smallest* (in the sense of $|y|$) integer solution to $x^2 - 11y^2 = \pm 1$, it must be the fundamental unit. This holds more generally, although in the case that $d \equiv 1 \pmod 4$ a modification must be made as x, y might be half-integers (like $\frac{1}{2}(1 + \sqrt{5})$).

Theorem 58.3.1 (Fundamental units of Pell's equations)

Assume d is a squarefree integer.

(a) If $d \equiv 2, 3 \pmod 4$, and (x, y) is a minimal integer solution to $x^2 - dy^2 = \pm 1$, then $x + y\sqrt{d}$ is a fundamental unit.

(b) If $d \equiv 1 \pmod 4$, and (x, y) is a minimal *half-integer* solution to $x^2 - dy^2 = \pm 1$, then $x + y\sqrt{d}$ is a fundamental unit. (Equivalently, the minimal integer solution to $a^2 - db^2 = \pm 4$ gives $\frac{1}{2}(a + b\sqrt{d})$.)

(Any reasonable definition of "minimal" will work, such as sorting by $|y|$.)

§58.4 Pell's equation

This class of results completely eradicates Pell's Equation. After all, solving

$$a^2 - d \cdot b^2 = \pm 1$$

amounts to finding elements of $\mathbb{Z}[\sqrt{d}]$ with norm ± 1. It's a bit weirder in the $d \equiv 1 \pmod 4$ case, since in that case $K = \mathbb{Q}(\sqrt{d})$ gives $\mathcal{O}_K = \mathbb{Z}[\frac{1}{2}(1 + \sqrt{d})]$, and so the fundamental unit may not actually be a solution. (For example, when $d = 5$, we get the solution $(\frac{1}{2}, \frac{1}{2})$.) Nonetheless, all *integer* solutions are eventually generated.

To make this all concrete, here's a simple example.

Example 58.4.1 ($x^2 - 5y^2 = \pm 1$)

Set $K = \mathbb{Q}(\sqrt{5})$, so $\mathcal{O}_K = \mathbb{Z}[\frac{1}{2}(1 + \sqrt{5})]$. By Dirichlet's unit theorem, $\mathcal{O}_K^\times$ is generated by a single element u. The choice

$$u = \frac{1}{2} + \frac{1}{2}\sqrt{5}$$

serves as a fundamental unit, as there are no smaller integer solutions to $a^2 - 5b^2 = \pm 4$.

The first several powers of u are

n	u^n	Norm
-2	$\frac{1}{2}(3 - \sqrt{5})$	1
-1	$\frac{1}{2}(1 - \sqrt{5})$	-1
0	1	1
1	$\frac{1}{2}(1 + \sqrt{5})$	-1
2	$\frac{1}{2}(3 + \sqrt{5})$	1
3	$2 + \sqrt{5}$	-1
4	$\frac{1}{2}(7 + 3\sqrt{5})$	1
5	$\frac{1}{2}(11 + 5\sqrt{5})$	-1
6	$9 + 4\sqrt{5}$	1

One can see that the first integer solution is $(2, 1)$, which gives -1. The first solution with $+1$ is $(9, 4)$. Continuing the pattern, we find that every third power of u gives an integer solution (see also Problem 58B), with the odd ones giving a solution to $x^2 - 5y^2 = -1$ and the even ones a solution to $x^2 - 5y^2 = +1$. All solutions are generated this way, up to $\pm$ signs (by considering $\pm u^{\pm n}$).

§58.5 A few harder problems to think about

Problem 58A (Fictitious account of the battle of Hastings)**.** Determine the number of soldiers in the following battle:

> The men of Harold stood well together, as their wont was, and formed thirteen squares, with a like number of men in every square thereof, and woe to the hardy Norman who ventured to enter their redoubts; for a single blow of Saxon war-hatched would break his lance and cut through his coat of mail . . . when Harold threw himself into the fray the Saxons were one might square of men, shouting the battle-cries, "Ut!", "Olicrosse!", "Godemite!"

Problem 58B. Let $d > 0$ be a squarefree integer, and let u denote the fundamental unit of $\mathbb{Q}(\sqrt{d})$. Show that either $u \in \mathbb{Z}[\sqrt{d}]$, or $u^n \in \mathbb{Z}[\sqrt{d}] \iff 3 \mid n$.

Problem 58C. Show that there are no integer solutions to

$$x^2 - 34y^2 = -1$$

despite the fact that -1 is a quadratic residue mod 34.

XV

Algebraic NT II: Galois and Ramification Theory

Part XV: Contents

59 Things Galois

§59.1 Motivation

Prototypical example for this section: $\mathbb{Q}(\sqrt{2})$ *and* $\mathbb{Q}(\sqrt[3]{2})$.

The key idea in Galois theory is that of *embeddings*, which give us another way to get at the idea of the "conjugate" we described earlier.

Let K be a number field. An **embedding** $\sigma\colon K \hookrightarrow \mathbb{C}$, is an *injective field homomorphism*: it needs to preserve addition and multiplication, and in particular it should fix 1.

> **Question 59.1.1.** Show that in this context, $\sigma(q) = q$ for any rational number q.

Example 59.1.2 (Examples of embeddings)

(a) If $K = \mathbb{Q}(i)$, the two embeddings of K into $\mathbb{C}$ are $z \mapsto z$ (the identity) and $z \mapsto \overline{z}$ (complex conjugation).

(b) If $K = \mathbb{Q}(\sqrt{2})$, the two embeddings of K into $\mathbb{C}$ are $a + b\sqrt{2} \mapsto a + b\sqrt{2}$ (the identity) and $a + b\sqrt{2} \mapsto a - b\sqrt{2}$ (conjugation).

(c) If $K = \mathbb{Q}(\sqrt[3]{2})$, there are three embeddings:

- The identity embedding, which sends $1 \mapsto 1$ and $\sqrt[3]{2} \mapsto \sqrt[3]{2}$.
- An embedding which sends $1 \mapsto 1$ and $\sqrt[3]{2} \mapsto \omega\sqrt[3]{2}$, where ω is a cube root of unity. Note that this is enough to determine the rest of the embedding.
- An embedding which sends $1 \mapsto 1$ and $\sqrt[3]{2} \mapsto \omega^2\sqrt[3]{2}$.

I want to make several observations about these embeddings, which will form the core ideas of Galois theory. Pay attention here!

- First, you'll notice some duality between roots: in the first example, i gets sent to $\pm i$, $\sqrt{2}$ gets sent to $\pm\sqrt{2}$, and $\sqrt[3]{2}$ gets sent to the other roots of $x^3 - 2$. This is no coincidence, and one can show this occurs in general. Specifically, suppose α has minimal polynomial

$$0 = c_n\alpha^n + c_{n-1}\alpha^{n-1} + \cdots + c_1\alpha + c_0$$

where the c_i are rational. Then applying any embedding σ to both sides gives

$$\begin{aligned} 0 &= \sigma(c_n\alpha^n + c_{n-1}\alpha^{n-1} + \cdots + c_1\alpha + c_0) \\ &= \sigma(c_n)\sigma(\alpha)^n + \sigma(c_{n-1})\sigma(\alpha)^{n-1} + \cdots + \sigma(c_1)\sigma(\alpha) + \sigma(c_0) \\ &= c_n\sigma(\alpha)^n + c_{n-1}\sigma(\alpha)^{n-1} + \cdots + c_1\sigma(\alpha) + c_0 \end{aligned}$$

where in the last step we have used the fact that $c_i \in \mathbb{Q}$, so they are fixed by σ. *So, roots of minimal polynomials go to other roots of that polynomial.*

- Next, I want to draw out a contrast between the second and third examples. Specifically, in example (b) where we consider embeddings $K = \mathbb{Q}(\sqrt{2})$ to $\mathbb{C}$. The image of these embeddings lands entirely in K: that is, we could just as well have looked at $K \to K$ rather than looking at $K \to \mathbb{C}$. However, this is not true in (c): indeed $\mathbb{Q}(\sqrt[3]{2}) \subset \mathbb{R}$, but the non-identity embeddings have complex outputs!

  The key difference is to again think about conjugates. Key observation:

  > **The field $K = \mathbb{Q}(\sqrt[3]{2})$ is "deficient" because the minimal polynomial $x^3 - 2$ has two other roots $\omega\sqrt[3]{2}$ and $\omega^2\sqrt[3]{2}$ not contained in K.**

  On the other hand $K = \mathbb{Q}(\sqrt{2})$ is just fine because both roots of $x^2 - 2$ are contained inside K. Finally, one can actually fix the deficiency in $K = \mathbb{Q}(\sqrt[3]{2})$ by completing it to a field $\mathbb{Q}(\sqrt[3]{2}, \omega)$. Fields like $\mathbb{Q}(i)$ or $\mathbb{Q}(\sqrt{2})$ which are "self-contained" are called *Galois extensions*, as we'll explain shortly.

- Finally, you'll notice that in the examples above, *the number of embeddings from K to $\mathbb{C}$ happens to be the degree of K*. This is an important theorem, Theorem 59.3.1.

In this chapter we'll develop these ideas in full generality, for any field other than $\mathbb{Q}$.

§59.2 Field extensions, algebraic extension, and splitting fields

Prototypical example for this section: *$\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}$ is an extension, $\mathbb{C}$ is an algebraic extension of any number field.*

First, we define a notion of one field sitting inside another, in order to generalize the notion of a number field.

Definition 59.2.1. Let K and F be fields. If $F \subseteq K$, we write K/F and say K is a **field extension** of F.

Thus K is automatically an F-vector space (just like $\mathbb{Q}(\sqrt{2})$ is automatically a $\mathbb{Q}$-vector space). The **degree** is the dimension of this space, denoted $[K : F]$. If $[K : F]$ is finite, we say K/F is a **finite (field) extension**.

That's really all. There's nothing tricky at all.

> **Question 59.2.2.** What do you call a finite extension of $\mathbb{Q}$?

Degrees of finite extensions are multiplicative.

Theorem 59.2.3 (Field extensions have multiplicative degree)
Let $F \subseteq K \subseteq L$ be fields with L/K, K/F finite. Then

$$[L : K][K : F] = [L : F].$$

Proof. Basis bash: you can find a basis of L over K, and then expand that into a basis L over F. (Diligent readers can fill in details.) □

Next, given a field (like $\mathbb{Q}(\sqrt[3]{2})$) we want something to embed it into (in our case $\mathbb{C}$). So we just want a field that contains all the roots of all the polynomials. Let's agree that

a field E is **algebraically closed** if every polynomial with coefficients in E is a product of linear polynomials in E, with the classic example is:

Example 59.2.4 ($\mathbb{C}$)
$\mathbb{C}$ is algebraically closed.

A major theorem is that any field F can be extended to an algebraically closed one $\overline{F}$; since all roots of polynomials in $\overline{F}[x]$ live in $\overline{F}$, in particular so do all roots of polynomials in $F[x]$. Here is the result:

Theorem 59.2.5 (Algebraic closures)
Any field F has algebraically closed field extensions. In fact, there is a unique such extension which is minimal by inclusion, called the **algebraic closure** and denoted $\overline{F}$. (Here "minimal" means any other algebraically closed extension of F contains an isomorphic copy of $\overline{F}$.) It has the property that every element of $\overline{F}$ is indeed the root of some polynomial with coefficients in F.

Example 59.2.6 ($\overline{\mathbb{R}} = \overline{\mathbb{C}} = \mathbb{C} \supsetneq \overline{\mathbb{Q}}$)
$\mathbb{C}$ is the algebraic closure of $\mathbb{R}$ (and itself). But the algebraic closure $\overline{\mathbb{Q}}$ of $\mathbb{Q}$ (i.e. the set of algebraic numbers) is a proper subfield of $\mathbb{C}$ (some complex numbers aren't the root of any rational-coefficient polynomial).

Usually we won't care much about what these extensions look like, and merely be satisfied they exist. Often we won't even use the algebraic closure, just any big enough field; for example, when working with a polynomial f with $\mathbb{Q}$-coefficients, we simply consider roots of f as elements of $\mathbb{C}$ for convenience and concreteness, even though it may be less wasteful to use the smaller $\overline{\mathbb{Q}}$ in place of $\mathbb{C}$.

§59.3 Embeddings into algebraic closures for number fields

Now that I've defined all these ingredients, I can prove:

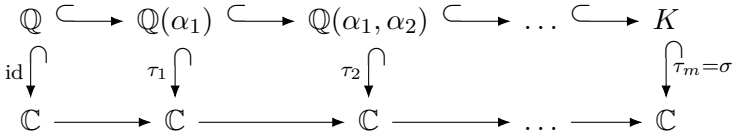
Theorem 59.3.1 (The n embeddings of a number field)
Let K be a number field of degree n. Then there are exactly n field homomorphisms $K \hookrightarrow \mathbb{C}$, say $\sigma_1, \dots, \sigma_n$ which fix $\mathbb{Q}$.

Remark 59.3.2 — Note that a nontrivial homomorphism of fields is necessarily injective (the kernel is an ideal). This justifies the use of "$\hookrightarrow$", and we call each σ_i an **embedding** of K into $\mathbb{C}$.

Proof. This is actually kind of fun! Recall that any irreducible polynomial over $\mathbb{Q}$ has distinct roots (Lemma 54.1.2). We'll adjoin elements $\alpha_1, \alpha_2, \dots, \alpha_m$ one at a time to $\mathbb{Q}$, until we eventually get all of K, that is,

$$K = \mathbb{Q}(\alpha_1, \dots, \alpha_n).$$

Diagrammatically, this is



First, we claim there are exactly

$$[\mathbb{Q}(\alpha_1) : \mathbb{Q}]$$

ways to pick τ_1. Observe that τ_1 is determined by where it sends α_1 (since it has to fix $\mathbb{Q}$). Letting p_1 be the minimal polynomial of α_1, we see that there are $\deg p_1$ choices for τ_1, one for each (distinct) root of p_1. That proves the claim.

Similarly, given a choice of τ_1, there are

$$[\mathbb{Q}(\alpha_1, \alpha_2) : \mathbb{Q}(\alpha_1)]$$

ways to pick τ_2. (It's a little different: τ_1 need not be the identity. But it's still true that τ_2 is determined by where it sends α_2, and as before there are $[\mathbb{Q}(\alpha_1, \alpha_2) : \mathbb{Q}(\alpha_1)]$ possible ways.)

Multiplying these all together gives the desired $[K : \mathbb{Q}]$. □

Remark 59.3.3 — The primitive element theorem actually implies that $m = 1$ is sufficient; we don't need to build a whole tower. This simplifies the proof somewhat.

It's common to see expressions like "let K be a number field of degree n, and $\sigma_1, \dots, \sigma_n$ its n embeddings" without further explanation. The relation between these embeddings and the Galois conjugates is given as follows.

Theorem 59.3.4 (Embeddings are evenly distributed over conjugates)
Let K be a number field of degree n with n embeddings $\sigma_1, \dots, \sigma_n$, and let $\alpha \in K$ have m Galois conjugates over $\mathbb{Q}$.
Then $\sigma_j(\alpha)$ is "evenly distributed" over each of these m conjugates: for any Galois conjugate β, exactly $\frac{n}{m}$ of the embeddings send α to β.

Proof. In the previous proof, adjoin $\alpha_1 = \alpha$ first. □

So, now we can define the trace and norm over $\mathbb{Q}$ in a nice way: given a number field K, we set

$$\operatorname{Tr}_{K/\mathbb{Q}}(\alpha) = \sum_{i=1}^{n} \sigma_i(\alpha) \quad \text{and} \quad \operatorname{N}_{K/\mathbb{Q}}(\alpha) = \prod_{i=1}^{n} \sigma_i(\alpha)$$

where σ_i are the n embeddings of K into $\mathbb{C}$.

§59.4 Everyone hates characteristic 2: separable vs irreducible

Prototypical example for this section: $\mathbb{Q}$ has characteristic zero, hence irreducible polynomials are separable.

Now, we want a version of the above theorem for any field F. If you read the proof, you'll see that the only thing that ever uses anything about the field $\mathbb{Q}$ is Lemma 54.1.2, where we use the fact that

Irreducible polynomials over F have no double roots.

Let's call a polynomial with no double roots **separable**; thus we want irreducible polynomials to be separable. We did this for $\mathbb{Q}$ in the last chapter by taking derivatives. Should work for any field, right?

Nope. Suppose we took the derivative of some polynomial like $2x^3 + 24x + 9$, namely $6x^2 + 24$. In $\mathbb{C}$ it's obvious that the derivative of a nonconstant polynomial f' isn't zero. But suppose we considered the above as a polynomial in $\mathbb{F}_3$, i.e. modulo 3. Then the derivative is zero. Oh, no!

We have to impose a condition that prevents something like this from happening.

Definition 59.4.1. For a field F, the **characteristic** of F is the smallest positive integer p such that,

$$\underbrace{1_F + \cdots + 1_F}_{p \text{ times}} = 0$$

or zero if no such integer p exists.

Example 59.4.2 (Field characteristics)

Old friends $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$ all have characteristic zero. But $\mathbb{F}_p$, the integers modulo p, is a field of characteristic p.

Exercise 59.4.3. Let F be a field of characteristic p. Show that if $p > 0$ then p is a prime number. (A proof is given next chapter.)

With the assumption of characteristic zero, our earlier proof works.

Lemma 59.4.4 (Separability in characteristic zero)

Any irreducible polynomial in a characteristic zero field is separable.

Unfortunately, this lemma is false if the "characteristic zero" condition is dropped.

Remark 59.4.5 — The reason it's called *separable* is (I think) this picture: I have a polynomial and I want to break it into irreducible parts. Normally, if I have a double root in a polynomial, that means it's not irreducible. But in characteristic $p > 0$ this fails. So inseparable polynomials are strange when you think about them: somehow you have double roots that can't be separated from each other.

We can get this to work for any field extension in which separability is not an issue.

Definition 59.4.6. A **separable extension** K/F is one where for each $\alpha \in K$, the minimal polynomial of α over F is separable (for example, if F has characteristic zero). A field F is **perfect** if any finite field extension K/F is separable.

In fact, as we see in the next chapter:

Theorem 59.4.7 (Finite fields are perfect)

Suppose F is a field with finitely many elements. Then it is perfect.

Thus, we will almost never have to worry about separability since every field we see in the Napkin is either finite or characteristic 0. So the inclusion of the word "separable" is mostly a formality.

Proceeding onwards, we obtain

Theorem 59.4.8 (The n embeddings of any separable extension)
Let K/F be a separable extension of degree n and let $\overline{F}$ be an algebraic closure of F. Then there are exactly n field homomorphisms $K \hookrightarrow \overline{F}$, say $\sigma_1, \dots, \sigma_n$, which fix F.

In any case, this lets us define the trace for *any* separable normal extension.

Definition 59.4.9. Let K/F be a separable extension of degree n, and let $\sigma_1, \dots, \sigma_n$ be the n embeddings into an algebraic closure of F. Then we define

$$\mathrm{Tr}_{K/F}(\alpha) = \sum_{i=1}^{n} \sigma_i(\alpha) \quad \text{and} \quad \mathrm{N}_{K/F}(\alpha) = \prod_{i=1}^{n} \sigma_i(\alpha).$$

When $F = \mathbb{Q}$ and the algebraic closure is $\mathbb{C}$, this coincides with our earlier definition!

§59.5 Automorphism groups and Galois extensions

Prototypical example for this section: $\mathbb{Q}(\sqrt{2})$ *is Galois but* $\mathbb{Q}(\sqrt[3]{2})$ *is not.*

We now want to get back at the idea we stated at the beginning of this section that $\mathbb{Q}(\sqrt[3]{2})$ is deficient in a way that $\mathbb{Q}(\sqrt{2})$ is not.

First, we define the "internal" automorphisms.

Definition 59.5.1. Suppose K/F is a finite extension. Then $\mathrm{Aut}(K/F)$ is the set of *field isomorphisms* $\sigma \colon K \to K$ which fix F. In symbols

$$\mathrm{Aut}(K/F) = \{\sigma \colon K \to K \mid \sigma \text{ is identity on } F\}.$$

This is a group under function composition!

Note that this time, we have a condition that F is fixed by σ. (This was not there before when we considered $F = \mathbb{Q}$, because we got it for free.)

Example 59.5.2 (Old examples of automorphism groups)
Reprising the example at the beginning of the chapter in the new notation, we have:

(a) $\mathrm{Aut}(\mathbb{Q}(i)/\mathbb{Q}) \cong \mathbb{Z}/2\mathbb{Z}$, with elements $z \mapsto z$ and $z \mapsto \overline{z}$.

(b) $\mathrm{Aut}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}) \cong \mathbb{Z}/2\mathbb{Z}$ in the same way.

(c) $\mathrm{Aut}(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q})$ is the trivial group, with only the identity embedding!

Example 59.5.3 (Automorphism group of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$)
Here's a new example: let $K = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. It turns out that $\mathrm{Aut}(K/\mathbb{Q}) =$

$\{1, \sigma, \tau, \sigma\tau\}$, where

$$\sigma : \begin{cases} \sqrt{2} & \mapsto -\sqrt{2} \\ \sqrt{3} & \mapsto \sqrt{3} \end{cases} \quad \text{and} \quad \tau : \begin{cases} \sqrt{2} & \mapsto \sqrt{2} \\ \sqrt{3} & \mapsto -\sqrt{3}. \end{cases}$$

In other words, $\mathrm{Aut}(K/\mathbb{Q})$ is the Klein Four Group.

First, let's repeat the proof of the observation that these embeddings shuffle around roots (akin to the first observation in the introduction):

Lemma 59.5.4 (Root shuffling in $\mathrm{Aut}(K/F)$)

Let $f \in F[x]$, suppose K/F is a finite extension, and assume $\alpha \in K$ is a root of f. Then for any $\sigma \in \mathrm{Aut}(K/F)$, $\sigma(\alpha)$ is also a root of f.

Proof. Let $f(x) = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_0$, where $c_i \in F$. Thus,

$$0 = \sigma(f(\alpha)) = \sigma\left(c_n\alpha^n + \cdots + c_0\right) = c_n\sigma(\alpha)^n + \cdots + c_0 = f(\sigma(\alpha)). \qquad \square$$

In particular, taking f to be the minimal polynomial of α we deduce

An embedding $\sigma \in \mathrm{Aut}(K/F)$ sends an $\alpha \in K$ to one of its various Galois conjugates (over F).

Next, let's look again at the "deficiency" of certain fields. Look at $K = \mathbb{Q}(\sqrt[3]{2})$. So, again $K/\mathbb{Q}$ is deficient for two reasons. First, while there are three maps $\mathbb{Q}(\sqrt[3]{2}) \hookrightarrow \mathbb{C}$, only one of them lives in $\mathrm{Aut}(K/\mathbb{Q})$, namely the identity. In other words, $|\mathrm{Aut}(K/\mathbb{Q})|$ is *too small*. Secondly, K is missing some Galois conjugates ($\omega\sqrt[3]{2}$ and $\omega^2\sqrt[3]{2}$).

The way to capture the fact that there are missing Galois conjugates is the notion of a splitting field.

Definition 59.5.5. Let F be a field and $p(x) \in F[x]$ a polynomial of degree n. Then $p(x)$ has roots $\alpha_1, \ldots, \alpha_n$ in the algebraic closure of F. The **splitting field** of $p(x)$ over F is defined as $F(\alpha_1, \ldots, \alpha_n)$.

In other words, the splitting field is the smallest field in which $p(x)$ splits.

Example 59.5.6 (Examples of splitting fields)

(a) The splitting field of $x^2 - 5$ over $\mathbb{Q}$ is $\mathbb{Q}(\sqrt{5})$. This is a degree 2 extension.

(b) The splitting field of $x^2 + x + 1$ over $\mathbb{Q}$ is $\mathbb{Q}(\omega)$, where ω is a cube root of unity. This is a degree 2 extension.

(c) The splitting field of $x^2 + 3x + 2 = (x+1)(x+2)$ is just $\mathbb{Q}$! There's nothing to do.

Example 59.5.7 (Splitting fields: a cautionary tale)

The splitting field of $x^3 - 2$ over $\mathbb{Q}$ is in fact

$$\mathbb{Q}(\sqrt[3]{2}, \omega)$$

and not just $\mathbb{Q}(\sqrt[3]{2})$! One must really adjoin *all* the roots, and it's not necessarily the case that these roots will generate each other.
To be clear:

- For $x^2 - 5$, we adjoin $\sqrt{5}$ and this will automatically include $-\sqrt{5}$.
- For $x^2 + x + 1$, we adjoin ω and get the other root ω^2 for free.
- But for $x^3 - 2$, if we adjoin $\sqrt[3]{2}$, we do NOT get $\omega\sqrt[3]{2}$ and $\omega^2\sqrt[3]{2}$ for free. Indeed, $\mathbb{Q}(\sqrt[3]{2}) \subset \mathbb{R}$!

Note that in particular, the splitting field of $x^3 - 2$ over $\mathbb{Q}$ is *degree six*, not just degree three.

In general, **the splitting field of a polynomial can be an extension of degree up to $n!$**. The reason is that if $p(x)$ has n roots and no "coincidental" relations between them then any permutation of the roots will work.

Now, we obtain:

Theorem 59.5.8 (Galois extensions are splitting)
For finite extensions K/F, $|\mathrm{Aut}(K/F)|$ divides $[K : F]$, with equality if and only if K is the *splitting field* of some separable polynomial with coefficients in F.

The proof of this is deferred to an optional section at the end of the chapter. If K/F is a finite extension and $|\mathrm{Aut}(K/F)| = [K : F]$, we say the extension K/F is **Galois**. In that case, we denote $\mathrm{Aut}(K/F)$ by $\mathrm{Gal}(K/F)$ instead and call this the **Galois group** of K/F.

Example 59.5.9 (Examples and non-examples of Galois extensions)
(a) The extension $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ is Galois, since it's the splitting field of $x^2 - 2$ over $\mathbb{Q}$. The Galois group has order two, $\sqrt{2} \mapsto \pm\sqrt{2}$.

(b) The extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}$ is Galois, since it's the splitting field of $(x^2-5)^2-6$ over $\mathbb{Q}$. As discussed before, the Galois group is $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.

(c) The extension $\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}$ is *not* Galois.

To explore $\mathbb{Q}(\sqrt[3]{2})$ one last time:

Example 59.5.10 (Galois closures, and the automorphism group of $\mathbb{Q}(\sqrt[3]{2}, \omega)$)
Let's return to the field $K = \mathbb{Q}(\sqrt[3]{2}, \omega)$, which is a field with $[K : \mathbb{Q}] = 6$. Consider

the two automorphisms:

$$\sigma : \begin{cases} \sqrt[3]{2} & \mapsto \omega\sqrt[3]{2} \\ \omega & \mapsto \omega \end{cases} \quad \text{and} \quad \tau : \begin{cases} \sqrt[3]{2} & \mapsto \sqrt[3]{2} \\ \omega & \mapsto \omega^2. \end{cases}$$

Notice that $\sigma^3 = \tau^2 = \mathrm{id}$. From this one can see that the automorphism group of K must have order 6 (it certainly has order ≤ 6; now use Lagrange's theorem). So, $K/\mathbb{Q}$ is Galois! Actually one can check explicitly that

$$\mathrm{Gal}(K/\mathbb{Q}) \cong S_3$$

is the symmetric group on 3 elements, with order $3! = 6$.

This example illustrates the fact that given a non-Galois field extension, one can "add in" missing conjugates to make it Galois. This is called taking a **Galois closure**.

§59.6 Fundamental theorem of Galois theory

After all this stuff about Galois Theory, I might as well tell you the fundamental theorem, though I won't prove it. Basically, it says that if K/F is Galois with Galois group G, then:

Subgroups of G correspond exactly to fields E with $F \subseteq E \subseteq K$.

To tell you how the bijection goes, I have to define a fixed field.

Definition 59.6.1. Let K be a field and H a subgroup of $\mathrm{Aut}(K/F)$. We define the **fixed field** of H, denoted K^H, as

$$K^H \coloneqq \{x \in K : \sigma(x) = x \ \forall \sigma \in H\}.$$

Question 59.6.2. Verify quickly that K^H is actually a field.

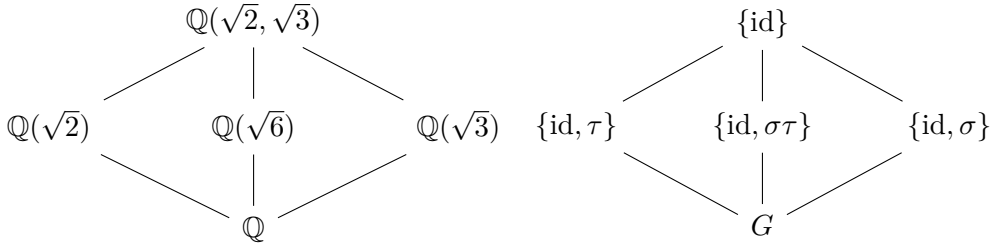
Now let's look at examples again. Consider $K = \mathbb{Q}(\sqrt{2}, \sqrt{3})$, where

$$G = \mathrm{Gal}(K/\mathbb{Q}) = \{\mathrm{id}, \sigma, \tau, \sigma\tau\}$$

is the Klein four group (where $\sigma(\sqrt{2}) = -\sqrt{2}$ but $\sigma(\sqrt{3}) = \sqrt{3}$; τ goes the other way).

Question 59.6.3. Let $H = \{\mathrm{id}, \sigma\}$. What is K^H?

In that case, the diagram of fields between $\mathbb{Q}$ and K matches exactly with the subgroups of G, as follows:



We see that subgroups correspond to fixed fields. That, and much more, holds in general.

Theorem 59.6.4 (Fundamental theorem of Galois theory)
Let K/F be a Galois extension with Galois group $G = \text{Gal}(K/F)$.

(a) There is a bijection between field towers $F \subseteq E \subseteq K$ and subgroups $H \subseteq G$:

$$\left\{ \begin{array}{c} K \\ | \\ E \\ | \\ F \end{array} \right\} \iff \left\{ \begin{array}{c} 1 \\ | \\ H \\ | \\ G \end{array} \right\}$$

The bijection sends H to its fixed field K^H, and hence is inclusion reversing.

(b) Under this bijection, we have $[K : E] = |H|$ and $[E : F] = |G/H|$.

(c) K/E is always Galois, and its Galois group is $\text{Gal}(K/E) = H$.

(d) E/F is Galois if and only if H is normal in G. If so, $\text{Gal}(E/F) = G/H$.

Exercise 59.6.5. Suppose we apply this theorem for

$$K = \mathbb{Q}(\sqrt[3]{2}, \omega).$$

Verify that the fact $E = \mathbb{Q}(\sqrt[3]{2})$ is not Galois corresponds to the fact that S_3 does not have normal subgroups of order 2.

§59.7 A few harder problems to think about

Problem 59A* (Galois group of the cyclotomic field)**.** Let p be an odd rational prime and ζ_p a primitive pth root of unity. Let $K = \mathbb{Q}(\zeta_p)$. Show that

$$\text{Gal}(K/\mathbb{Q}) \cong (\mathbb{Z}/p\mathbb{Z})^\times.$$

Problem 59B. Give an example of a degree-three Galois extension of $\mathbb{Q}$.

Problem 59C (Greek constructions)**.** Prove that the three Greek constructions

(a) doubling the cube,

(b) squaring the circle, and

(c) trisecting an angle

are all impossible. (Assume π is transcendental.)

Problem 59D (China Hong Kong Math Olympiad)**.** Prove that there are no rational numbers p, q, r satisfying

$$\cos\left(\frac{2\pi}{7}\right) = p + \sqrt{q} + \sqrt[3]{r}.$$

Problem 59E. Show that the only automorphism of $\mathbb{R}$ is the identity. Hence $\text{Aut}(\mathbb{R}/\mathbb{Q})$ is the trivial group.

Problem 59F (Artin's primitive element theorem)**.** Let K be a number field. Show that $K \cong \mathbb{Q}(\gamma)$ for some γ.

§59.8 (Optional) Proof that Galois extensions are splitting

We prove Theorem 59.5.8. First, we extract a useful fragment from the fundamental theorem.

Theorem 59.8.1 (Fixed field theorem)
Let K be a field and G a subgroup of $\operatorname{Aut}(K)$. Then $[K : K^G] = |G|$.

The inequality itself is not difficult:

Exercise 59.8.2. Show that $[K : F] \geq |\operatorname{Aut}(K/F)|$, and that equality holds if and only if the set of elements fixed by all $\sigma \in \operatorname{Aut}(K/F)$ is exactly F. (Use Theorem 59.8.1.)

The equality case is trickier.

The easier direction is when K is a splitting field. Assume $K = F(\alpha_1, \dots, \alpha_n)$ is the splitting field of some separable polynomial $p \in F[x]$ with n distinct roots $\alpha_1, \dots, \alpha_n$. Adjoin them one by one:

$$\begin{array}{ccccccccc}
F & \hookrightarrow & F(\alpha_1) & \hookrightarrow & F(\alpha_1,\alpha_2) & \hookrightarrow & \dots & \hookrightarrow & K \\
\downarrow{\scriptstyle \mathrm{id}} & & \downarrow{\scriptstyle \tau_1} & & \downarrow{\scriptstyle \tau_2} & & & & \downarrow{\scriptstyle \tau_n=\sigma} \\
F & \hookrightarrow & F(\alpha_1) & \hookrightarrow & F(\alpha_1,\alpha_2) & \hookrightarrow & \dots & \hookrightarrow & K
\end{array}$$

(Does this diagram look familiar?) Every map $K \to K$ which fixes F corresponds to an above commutative diagram. As before, there are exactly $[F(\alpha_1) : F]$ ways to pick τ_1. (You need the fact that the minimal polynomial p_1 of α_1 is separable for this: there need to be exactly $\deg p_1 = [F(\alpha_1) : F]$ distinct roots to nail p_1 into.) Similarly, given a choice of τ_1, there are $[F(\alpha_1, \alpha_2) : F(\alpha_1)]$ ways to pick τ_2. Multiplying these all together gives the desired $[K : F]$.

Now assume K/F is Galois. First, we state:

Lemma 59.8.3
Let K/F be Galois, and $p \in F[x]$ irreducible. If any root of p (in $\overline{F}$) lies in K, then all of them do, and in fact p is separable.

Proof. Let $\alpha \in K$ be the prescribed root. Consider the set

$$S = \{\sigma(\alpha) \mid \sigma \in \operatorname{Gal}(K/F)\}.$$

(Note that $\alpha \in S$ since $\operatorname{Gal}(K/F) \ni \mathrm{id}$.) By construction, any $\tau \in \operatorname{Gal}(K/F)$ fixes S. So if we construct

$$\tilde{p}(x) = \prod_{\beta \in S} (x - \beta),$$

then by Vieta's Formulas, we find that all the coefficients of $\tilde{p}$ are fixed by elements of σ. By the *equality case* we specified in the exercise, it follows that $\tilde{p}$ has coefficients in F! (This is where we use the condition.) Also, by Lemma 59.5.4, $\tilde{p}$ divides p.

Yet p was irreducible, so it is the minimal polynomial of α in $F[x]$, and therefore we must have that p divides $\tilde{p}$. Hence $p = \tilde{p}$. Since $\tilde{p}$ was built to be separable, so is p. □

Now we're basically done – pick a basis $\omega_1, \dots, \omega_n$ of K/F, and let p_i be their minimal polynomials; by the above, we don't get any roots outside K. Consider $P = p_1 \dots p_n$, removing any repeated factors. The roots of P are $\omega_1, \dots, \omega_n$ and some other guys in K. So K is the splitting field of P.

60 Finite fields

In this short chapter, we classify all fields with finitely many elements and compute the Galois groups. Nothing in here is very hard, and so most of the proofs are just sketches; if you like, you should check the details yourself.

The whole point of this chapter is to prove:

- A finite field F must have order p^n, with p prime and n an integer.
- In this case, F has characteristic p.
- All such fields are isomorphic, so it's customary to use the notation $\mathbb{F}_{p^n}$ for "the" finite field of order p^n if we only care up to isomorphism.
- The extension $F/\mathbb{F}_p$ is Galois, and $\mathrm{Gal}(F/\mathbb{F}_p)$ is a cyclic group of order n. The generator is the automorphism

$$\sigma \colon F \to F \quad \text{by} \quad x \mapsto x^p.$$

If you're in a hurry you can just remember these results and skip to the next chapter.

§60.1 Example of a finite field

Before diving in, we give some examples.

Recall that the *characteristic* of a field F is the smallest positive integer p such that

$$\underbrace{1_F + \cdots + 1_F}_{p \text{ times}} = 0$$

or 0 if no such integer p exists.

Example 60.1.1 (Base field)

Let $\mathbb{F}_p$ denote the field of integers modulo p. This is a field with p elements, with characteristic p.

Example 60.1.2 (The finite field of nine elements)

Let

$$F \cong \mathbb{F}_3[X]/(X^2+1) \cong \mathbb{Z}[i]/(3).$$

We can think of its elements as

$$\{a + bi \mid 0 \le a, b \le 2\}.$$

Since (3) is prime in $\mathbb{Z}[i]$, the ring of integers of $\mathbb{Q}(i)$, we see F is a field with $3^2 = 9$ elements inside it. Note that, although this field has 9 elements, every element x has the property that

$$3x = \underbrace{x + \cdots + x}_{3 \text{ times}} = 0.$$

In particular, F has characteristic 3.

§60.2 Finite fields have prime power order

Lemma 60.2.1
If the characteristic of a field F isn't zero, it must be a prime number.

Proof. Assume not, so $n = ab$ for $a, b < n$. Then let

$$A = \underbrace{1_F + \cdots + 1_F}_{a \text{ times}} \neq 0$$

and

$$B = \underbrace{1_F + \cdots + 1_F}_{b \text{ times}} \neq 0.$$

Then $AB = 0$, contradicting the fact that F is a field. □

We like fields of characteristic zero, but unfortunately for finite fields we are doomed to have nonzero characteristic.

Lemma 60.2.2 (Finite fields have prime power orders)
Let F be a finite field. Then

(a) Its characteristic is nonzero, and hence some prime p.

(b) The field F is a finite extension of $\mathbb{F}_p$, and in particular it is an $\mathbb{F}_p$-vector space.

(c) We have $|F| = p^n$ for some prime p, integer n.

Proof. Very briefly, since this is easy:

(a) Apply Lagrange's theorem (or pigeonhole principle!) to $(F, +)$ to get the characteristic isn't zero.

(b) The additive subgroup of $(F, +)$ generated by 1_F is an isomorphic copy of $\mathbb{F}_p$.

(c) Since it's a field extension, F is a finite-dimensional vector space over $\mathbb{F}_p$, with some basis $e_1, \ldots, e_n$. It follows that there are p^n elements of F. □

Remark 60.2.3 — An amusing alternate proof of (c) by contradiction: if a prime $q \neq p$ divides $|F|$, then by Cauchy's theorem (Problem 17A⋆) on $(F, +)$ there's a (nonzero) element x of order q. Evidently

$$x \cdot (\underbrace{1_F + \cdots + 1_F}_{q \text{ times}}) = 0$$

then, but $x \neq 0$, and hence the characteristic of F also divides q, which is impossible.

An important point in the above proof is that

Lemma 60.2.4 (Finite fields are field extensions of $\mathbb{F}_p$)
If $|F| = p^n$ is a finite field, then there is an isomorphic copy of $\mathbb{F}_p$ sitting inside F. Thus F is a field extension of $\mathbb{F}_p$.

We want to refer a lot to this copy of $\mathbb{F}_p$, so in what follows:

Abuse of Notation 60.2.5. Every integer n can be identified as an element of F, namely

$$n \coloneqq \underbrace{1_F + \cdots + 1_F}_{n \text{ times}}.$$

Note that (as expected) this depends only on $n \pmod p$.

This notation makes it easier to think about statements like the following.

Theorem 60.2.6 (Freshman's dream)
For any $a, b \in F$ we have

$$(a+b)^p = a^p + b^p.$$

Proof. Use the Binomial theorem, and the fact that $\binom{p}{i}$ is divisible by p for $0 < i < p$. □

Exercise 60.2.7. Convince yourself that this proof works.

§60.3 All finite fields are isomorphic

We next proceed to prove "Fermat's little theorem":

Theorem 60.3.1 (Fermat's little theorem in finite fields)
Let F be a finite field of order p^n. Then every element $x \in F$ satisfies

$$x^{p^n} - x = 0.$$

Proof. If $x = 0$ it's true; otherwise, use Lagrange's theorem on the abelian group $(F, \times)$ to get $x^{p^n - 1} = 1_F$. □

We can now prove the following result, which is the "main surprise" about finite fields: that there is a unique one up to isomorphism for each size.

Theorem 60.3.2 (Complete classification of finite fields)
A field F is a finite field with p^n elements if and only if it is a splitting field of $x^{p^n} - x$ over $\mathbb{F}_p$.

Proof. By "Fermat's little theorem", all the elements of F satisfy this polynomial. So we just have to show that the roots of this polynomial are distinct (i.e. that it is separable).

To do this, we use the derivative trick again: the derivative of this polynomial is

$$p^n \cdot x^{p^n-1} - 1 = -1$$

which has no roots at all, so the polynomial cannot have any double roots. □

Definition 60.3.3. For this reason, it's customary to denote *the* field with p^n elements by $\mathbb{F}_{p^n}$.

Note that the polynomial $x^{p^n} - x \pmod p$ is far from irreducible, but the computation above shows that it's separable.

Example 60.3.4 (The finite field of order nine again)
The polynomial $x^9 - x$ is separable modulo 3 and has factorization

$$x(x+1)(x+2)(x^2+1)(x^2+x+2)(x^2+2x+2) \pmod 3.$$

So if F has order 9, then we intuitively expect it to be the field generated by adjoining all the roots: 0, 1, 2, as well as $\pm i$, $1 \pm i$, $2 \pm i$. Indeed, that's the example we had at the beginning of this chapter.
(Here i denotes *an* element of $\mathbb{F}_9$ satisfying $i^2 = -1$. The notation is deliberately similar to the usual imaginary unit.)

§60.4 The Galois theory of finite fields

Retain the notation $\mathbb{F}_{p^n}$ now (instead of F like before). By the above theorem, it's the splitting field of a separable polynomial, hence we know that $\mathbb{F}_{p^n}/\mathbb{F}_p$ is a Galois extension. We would like to find the Galois group.

In fact, we are very lucky: it is cyclic. First, we exhibit one such element $\sigma_p \in \mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$:

Theorem 60.4.1 (The pth power automorphism)
The map $\sigma_p \colon \mathbb{F}_{p^n} \to \mathbb{F}_{p^n}$ defined by

$$\sigma_p(x) = x^p$$

is an automorphism, and moreover fixes $\mathbb{F}_p$.

This is called the Frobenius automorphism, and will re-appear later on in Chapter 62.

Proof. It's a homomorphism since it fixes 1, respects multiplication, and respects addition.

> **Question 60.4.2.** Why does it respect addition?

Next, we claim that it is injective. To see this, note that

$$x^p = y^p \iff x^p - y^p = 0 \iff (x-y)^p = 0 \iff x = y.$$

Here we have again used the Freshman's Dream. Since $\mathbb{F}_{p^n}$ is finite, this injective map is automatically bijective. The fact that it fixes $\mathbb{F}_p$ is Fermat's little theorem. □

Now we're done:

Theorem 60.4.3 (Galois group of the extension $\mathbb{F}_{p^n}/\mathbb{F}_p$)
We have $\mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p) \cong \mathbb{Z}/n\mathbb{Z}$ with generator σ_p.

Proof. Since $[\mathbb{F}_{p^n} : \mathbb{F}_p] = n$, the Galois group G has order n. So we just need to show $\sigma_p \in G$ has order n.

Note that σ_p applied k times gives $x \mapsto x^{p^k}$. Hence, σ_p applied n times is the identity, as all elements of $\mathbb{F}_{p^n}$ satisfy $x^{p^n} = x$. But if $k < n$, then σ_p applied k times cannot be the identity or $x^{p^k} - x$ would have too many roots. □

We can see an example of this again with the finite field of order 9.

Example 60.4.4 (Galois group of finite field of order 9)
Let $\mathbb{F}_9$ be the finite field of order 9, and represent it concretely by $\mathbb{F}_9 = \mathbb{Z}[i]/(3)$. Let $\sigma_3 \colon \mathbb{F}_9 \to \mathbb{F}_9$ be $x \mapsto x^3$. We can witness the fate of all nine elements:

$$\begin{array}{ccccccc} 0 & & 1 & & 2 & & \begin{array}{c} i \\ \updownarrow \sigma \\ -i \end{array} \qquad \begin{array}{c} 1+i \\ \updownarrow \sigma \\ 1-i \end{array} \qquad \begin{array}{c} 2+i \\ \updownarrow \sigma \\ 2-i \end{array} \end{array}$$

(As claimed, 0, 1, 2 are the fixed points, so I haven't drawn arrows for them.) As predicted, the Galois group has order two:

$$\mathrm{Gal}(\mathbb{F}_9/\mathbb{F}_3) = \{\mathrm{id}, \sigma_3\} \cong \mathbb{Z}/2\mathbb{Z}.$$

This concludes the proof of all results stated at the beginning of this chapter.

§60.5 Extra: The multiplicative group of a finite field

In this section we prove a result which is interesting by its own right, even though it is not used in the following chapters.

Consider the field F of order $p = 17$. We may want to ask the following questions about F:

- How many nonzero elements in F is a quadratic residue (can be written as a square of another element in F)?
- Is there any element x in p such that $x^2 = -1$?

With the following proposition, the questions above become easy.

Proposition 60.5.1
Let F be a finite field, then the multiplicative group $F^\times$ is a cyclic group.

Essentially, it says that the multiplicative group $F^\times$ is as nice as possible — the cyclic group is the simplest Abelian group!

If we look back at the examples above, we can see how this knowledge can help us.

Example 60.5.2 (The multiplicative group of $\mathbb{F}_{17}$)

The group $F^\times$ is cyclic of order 16, so there is some element $g \in F^\times$ such that all of the elements in F are

$$0, g^0 = 1, g^1, g^2, \dots, g^{15}.$$

Using this knowledge, if we square the nonzero elements, we can easily see that the result are the following (note that $g^{16} = g^0 = 1$):

$$g^0, g^2, g^4, \dots, g^{14}, g^0, g^2, \dots, g^{14}.$$

As such, exactly half of the elements — 8 elements — in $F^\times$ are quadratic residues! Checking whether there is an element x such that $x^2 = -1$ isn't much harder. First, where may -1 appear in the sequence $\{g^0, g^1, \dots, g^{15}\}$? If you find an explicit value of g (you can pick $g = 3$ for instance), you will see that $g^8 = -1$. But, even without an explicit calculation, you can still see that, because:

> **Question 60.5.3.** Note that the equation $x^2 = 1$ has only 2 roots, 1 and -1. Using group operations in the group $F^\times$, what does the equation say?

We have $g^8 = -1$, so $(g^4)^2 = (g^{12})^2 = -1$, so we're done.

So, why is the proposition true? Consider a finite field F of order a prime power q. First, using an idea similar to the above, we have:

> **Question 60.5.4.** Show that, for any positive $d < |F^\times| = q - 1$, then there are at most d elements x such that $x^d = 1$.

The rest is easily handled using group theory. Recall from Theorem 18.1.5, because $F^\times$ is a finite Abelian group, it is in particular finitely generated, and can be written in the form

$$F^\times \cong \mathbb{Z}^{\oplus r} \oplus \mathbb{Z}/q_1\mathbb{Z} \oplus \mathbb{Z}/q_2\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/q_m\mathbb{Z}.$$

Of course, $r = 0$ here.

Now, assume $F^\times$ is not cyclic, so the lowest common denominator of all the q_i values are less than $|F^\times|$. Then, for all $x \in F^\times$, $x^{\operatorname{lcm}(q_1,\dots,q_m)} = 1$, which gives a contradiction.

Remark 60.5.5 — If you look up an elementary proof of why there are exactly $p - 1$ quadratic residues modulo p, most of the time, you will get some argument using "$x^d - 1$ has at most d roots" similar to the proof above, but by not going all the way and show the structure of the multiplicative group, it hides the spirit of what is really going on.

The proposition above takes a step further — now, without any calculation, you know for instance the finite field $\mathbb{F}_{17^2}$ has exactly $\frac{17^2-1}{2} = 144$ nonzero quadratic residues.

§60.6 A few harder problems to think about

Problem 60A† (HMMT 2017)**.** What is the period of the Fibonacci sequence modulo 127?

61 Ramification theory

We're very interested in how rational primes p factor in a bigger number field K. Some examples of this behavior: in $\mathbb{Z}[i]$ (which is a UFD!), we have factorizations

$$\begin{aligned}(2) &= (1+i)^2 \\ (3) &= (3) \\ (5) &= (2+i)(2-i).\end{aligned}$$

In this chapter we'll learn more about how primes break down when they're thrown into bigger number fields. Using weapons from Galois Theory, this will culminate in a proof of Quadratic Reciprocity.

§61.1 Ramified / inert / split primes

Prototypical example for this section: In $\mathbb{Z}[i]$, 2 is ramified, 3 is inert, and 5 splits.

Let p be a rational prime, and toss it into $\mathcal{O}_K$. Thus we get a factorization into prime ideals

$$p \cdot \mathcal{O}_K = \mathfrak{p}_1^{e_1} \dots \mathfrak{p}_g^{e_g}.$$

We say that each $\mathfrak{p}_i$ is **above** (p).[1] Pictorially, you might draw this as follows:

$$\begin{array}{ccccc} K & \supset & \mathcal{O}_K & & \mathfrak{p}_i \\ | & & | & & | \\ \mathbb{Q} & \supset & \mathbb{Z} & & (p) \end{array}$$

Some names for various behavior that can happen:

- We say p is **ramified** if $e_i > 1$ for some i. For example 2 is ramified in $\mathbb{Z}[i]$.
- We say p is **inert** if $g = 1$ and $e_1 = 1$; i.e. (p) remains prime. For example 3 is inert in $\mathbb{Z}[i]$.
- We say p is **split** if $g > 1$. For example 5 is split in $\mathbb{Z}[i]$.

Question 61.1.1. More generally, for a prime p in $\mathbb{Z}[i]$:

- p is ramified exactly when $p = 2$.
- p is inert exactly when $p \equiv 3 \pmod 4$.
- p is split exactly when $p \equiv 1 \pmod 4$.

Prove this.

[1] Reminder that $p \cdot \mathcal{O}_K$ and (p) mean the same thing, and I'll use both interchangeably.

§61.2 Primes ramify if and only if they divide Δ_K

The most unusual case is ramification: Just like we don't expect a randomly selected polynomial to have a double root, we don't expect a randomly selected prime to be ramified. In fact, the key to understanding ramification is the discriminant.

For the sake of discussion, let's suppose that K is monogenic, $\mathcal{O}_K = \mathbb{Z}[\theta]$, where θ has minimal polynomial f. Let p be a rational prime we'd like to factor. If f factors as $f_1^{e_1} \dots f_g^{e_g}$, then we know that the prime factorization of (p) is given by

$$p \cdot \mathcal{O}_K = \prod_i (p, f_i(\theta))^{e_i}.$$

In particular, p ramifies exactly when f *has a double root mod* p! To detect whether this happens, we look at the polynomial discriminant of f, namely

$$\Delta(f) = \prod_{i<j} (z_i - z_j)^2$$

and see whether it is zero mod p – thus p ramifies if and only if this is true.

It turns out that the naïve generalization to any number field works if we replace $\Delta(f)$ by just the discriminant Δ_K of K; (these are the same for monogenic $\mathcal{O}_K$ by Problem 57C⋆). That is,

Theorem 61.2.1 (Discriminant detects ramification)
Let p be a rational prime and K a number field. Then p is ramified if and only if p divides Δ_K.

Example 61.2.2 (Ramification in the Gaussian integers)
Let $K = \mathbb{Q}(i)$ so $\mathcal{O}_K = \mathbb{Z}[i]$ and $\Delta_K = -4$. As predicted, the only prime ramifying in $\mathbb{Z}[i]$ is 2, the only prime factor of Δ_K.

In particular, only finitely many primes ramify.

§61.3 Inertial degrees

Prototypical example for this section: (7) *has inertial degree* 2 *in* $\mathbb{Z}[i]$ *and* $(2+i)$ *has inertial degree* 1 *in* $\mathbb{Z}[i]$.

Recall that we were able to define an ideal norm $\mathrm{N}(\mathfrak{a}) = |\mathcal{O}_K/\mathfrak{a}|$ measuring how "roomy" the ideal $\mathfrak{a}$ is. For example, (5) has ideal norm $5^2 = 25$ in $\mathbb{Z}[i]$, since

$$\mathbb{Z}[i]/(5) \cong \{a + bi \mid a, b \in \mathbb{Z}/5\mathbb{Z}\}$$

has $5^2 = 25$ elements.

Now, let's look at

$$p \cdot \mathcal{O}_K = \mathfrak{p}_1^{e_1} \dots \mathfrak{p}_g^{e_g}$$

in $\mathcal{O}_K$, where K has degree n. Taking the ideal norms of both sides, we have that

$$p^n = \mathrm{N}(\mathfrak{p}_1)^{e_1} \dots \mathrm{N}(\mathfrak{p}_g)^{e_g}.$$

We conclude that $\mathrm{N}(\mathfrak{p}_i) = p^{f_i}$ for some integer $f_i \geq 1$, and moreover that

$$n = \sum_{i=1}^{g} e_i f_i.$$

Definition 61.3.1. We say f_i is the **inertial degree** of $\mathfrak{p}_i$, and e_i is the **ramification index**.

Example 61.3.2 (Examples of inertial degrees)

Work in $\mathbb{Z}[i]$, which is degree 2. The inertial degree detects how "spacy" the given $\mathfrak{p}$ is when interpreted in $\mathcal{O}_K$.

(a) The prime $7 \cdot \mathbb{Z}[i]$ has inertial degree 2. Indeed, $\mathbb{Z}[i]/(7)$ has $7^2 = 49$ elements, those of the form $a + bi$ for a, b modulo 7. It gives "two degrees" of space.

(b) Let $(5) = (2+i)(2-i)$. The inertial degrees of $(2+i)$ and $(2-i)$ are both 1. Indeed, $\mathbb{Z}[i]/(2+i)$ only gives "one degree" of space, since each of its elements can be viewed as integers modulo 5, and there are only $5^1 = 5$ elements.

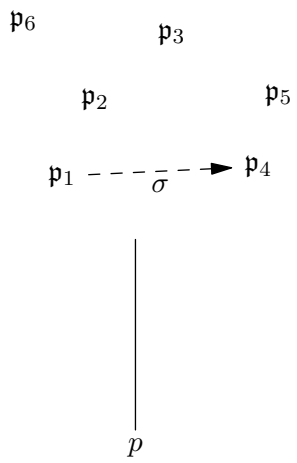
If you understand this, it should be intuitively clear why the sum of $e_i f_i$ should equal n.

§61.4 The magic of Galois extensions

OK, that's all fine and well. But something *really magical* happens when we add the additional hypothesis that $K/\mathbb{Q}$ is *Galois*: all the inertial degrees and ramification indices are equal. We set about proving this.

Let $K/\mathbb{Q}$ be Galois with $G = \mathrm{Gal}(K/\mathbb{Q})$. Note that if $\mathfrak{p} \subseteq \mathcal{O}_K$ is a prime above p, then the image $\sigma^{\mathrm{img}}(\mathfrak{p})$ is also prime for any $\sigma \in G$ (since σ is an automorphism!). Moreover, since $p \in \mathfrak{p}$ and σ fixes $\mathbb{Q}$, we know that $p \in \sigma^{\mathrm{img}}(\mathfrak{p})$ as well.

Thus, by the pointwise mapping, **the Galois group acts on the prime ideals above a rational prime** p. Picture:



The notation $\sigma^{\mathrm{img}}(\mathfrak{p})$ is hideous in this context, since we're really thinking of σ as just doing a group action, and so we give the shorthand:

Abuse of Notation 61.4.1. Let $\sigma\mathfrak{p}$ be shorthand for $\sigma^{\mathrm{img}}(\mathfrak{p})$.

Since the σ's are all bijections (they are automorphisms!), it should come as no surprise that the prime ideals which are in the same orbit are closely related. But miraculously, it turns out there is only one orbit!

Theorem 61.4.2 (Galois group acts transitively)
Let $K/\mathbb{Q}$ be Galois with $G = \mathrm{Gal}(K/\mathbb{Q})$. Let $\{\mathfrak{p}_i\}$ be the set of distinct prime ideals in the factorization of $p \cdot \mathcal{O}_K$ (in $\mathcal{O}_K$).
Then G acts transitively on the $\mathfrak{p}_i$: for every i and j, we can find σ such that $\sigma\mathfrak{p}_i = \mathfrak{p}_j$.

In other words,

All of the $\{\mathfrak{p}_i\}$ are Galois conjugates of each other.

Before proving this, let us consider the easier problem of factorization into elements.

> Suppose $\mathcal{O}_K$ is an UFD, and p factors as $up_1p_2\cdots p_n$ in $\mathcal{O}_K$, where p_i are irreducibles and u is an unit. Show that the p_i are all conjugates of each other, up to multiplication by an unit.

Question 61.4.3. Try to prove it before reading it below. (Hint: Galois theory. Alternatively, take the norm of p_1.)

Proof. Let $q = \mathrm{N}_{K/\mathbb{Q}}(p_1)$ be the product of all conjugates of p_1, then $q \in \mathbb{Q}$. Thus $p \mid q$, so each p_i is a factor of q, and we're done by unique factorization. □

Unfortunately, the product of all conjugates of an ideal $\mathfrak{p}_1$ is not necessarily of the form $p \cdot \mathcal{O}_K$ (for example, $K = \mathbb{Q}[i]$ and $(1+i)$ has no other conjugates). So in the proof, we pick x which is an "representative" of $\mathfrak{p}_1$.

Proof of Theorem 61.4.2. Because $\mathfrak{p}_i$ are distinct primes, by the Chinese remainder theorem, we can find an $x \in \mathcal{O}_K$ such that

$$\begin{aligned} x &\equiv 0 \pmod{\mathfrak{p}_1} \\ x &\equiv 1 \pmod{\mathfrak{p}_i} \text{ for } i \geq 2 \end{aligned}$$

Then, compute the norm

$$\mathrm{N}_{K/\mathbb{Q}}(x) = \prod_{\sigma \in \mathrm{Gal}(K/\mathbb{Q})} \sigma(x).$$

Each $\sigma(x)$ is in K because $K/\mathbb{Q}$ is Galois!

Since $\mathrm{N}_{K/\mathbb{Q}}(x)$ is an integer and divisible by $\mathfrak{p}_1$, we should have that $\mathrm{N}_{K/\mathbb{Q}}(x)$ is divisible by p. Thus it should be divisible by $\mathfrak{p}_2$ as well. Thus, for some $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$, $\sigma(x)$ is divisible by $\mathfrak{p}_2$, equivalently, x is divisible by $\sigma^{-1}\mathfrak{p}_2$. But by the way we selected x, we have within the factors of p, x is divisible by only $\mathfrak{p}_1$! So $\sigma^{-1}\mathfrak{p}_2 = \mathfrak{p}_1$, and we're done. □

Theorem 61.4.4 (Inertial degree and ramification indices are all equal)
Assume $K/\mathbb{Q}$ is Galois. Then for any rational prime p we have

$$p \cdot \mathcal{O}_K = (\mathfrak{p}_1\mathfrak{p}_2 \dots \mathfrak{p}_g)^e$$

for some e, where the $\mathfrak{p}_i$ are distinct prime ideals with the same inertial degree f. Hence

$$[K : \mathbb{Q}] = efg.$$

Proof. To see that the inertial degrees are equal, note that each σ induces an isomorphism

$$\mathcal{O}_K/\mathfrak{p} \cong \mathcal{O}_K/\sigma(\mathfrak{p}).$$

Because the action is transitive, all f_i are equal.

Exercise 61.4.5. Using the fact that $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$, show that

$$\sigma^{\mathrm{img}}(p \cdot \mathcal{O}_K) = p \cdot \sigma^{\mathrm{img}}(\mathcal{O}_K) = p \cdot \mathcal{O}_K.$$

So for every σ, we have that $p \cdot \mathcal{O}_K = \prod \mathfrak{p}_i^{e_i} = \prod(\sigma\mathfrak{p}_i)^{e_i}$. Since the action is transitive, all e_i are equal. □

Let's see an illustration of this.

Example 61.4.6 (Factoring 5 in a Galois/non-Galois extension)
Let $p = 5$ be a prime.

(a) Let $E = \mathbb{Q}(\sqrt[3]{2})$. One can show that $\mathcal{O}_E = \mathbb{Z}[\sqrt[3]{2}]$, so we use the Factoring Algorithm on the minimal polynomial $x^3 - 2$. Since $x^3 - 2 \equiv (x-3)(x^2+3x+9) \pmod 5$ is the irreducible factorization, we have that

$$(5) = (5, \sqrt[3]{2} - 3)(5, \sqrt[3]{4} + 3\sqrt[3]{2} + 9)$$

which have inertial degrees 1 and 2, respectively. The fact that this is not uniform reflects that E is not Galois.

(b) Now let $K = \mathbb{Q}(\sqrt[3]{2}, \omega)$, which is the splitting field of $x^3 - 2$ over $\mathbb{Q}$; now K is Galois. It turns out that

$$\mathcal{O}_K = \mathbb{Z}[\varepsilon] \quad \text{where} \quad \varepsilon \text{ is a root of } t^6 + 3t^5 - 5t^3 + 3t + 1.$$

(this takes a lot of work to obtain, so we won't do it here). Modulo 5 this has an irreducible factorization $(x^2 + x + 2)(x^2 + 3x + 3)(x^2 + 4x + 1) \pmod 5$, so by the Factorization Algorithm,

$$(5) = (5, \varepsilon^2 + \varepsilon + 2)(5, \varepsilon^2 + 3\varepsilon + 3)(5, \varepsilon^2 + 4\varepsilon + 1).$$

This time all inertial degrees are 2, as the theorem predicts for K Galois.

§61.5 (Optional) Decomposition and inertia groups

Let p be a rational prime. Thus

$$p \cdot \mathcal{O}_K = (\mathfrak{p}_1 \dots \mathfrak{p}_g)^e$$

and all the $\mathfrak{p}_i$ have inertial degree f. Let $\mathfrak{p}$ denote a choice of the $\mathfrak{p}_i$.

We can look at both the fields $\mathcal{O}_K/\mathfrak{p}$ and $\mathbb{Z}/p = \mathbb{F}_p$. Naturally, since $\mathcal{O}_K/\mathfrak{p}$ is a finite field we can view it as a field extension of $\mathbb{F}_p$. So we can get the diagram

$$\begin{array}{ccccc} K & \supset & \mathcal{O}_K & \mathfrak{p} & \mathcal{O}_K/\mathfrak{p} \cong \mathbb{F}_{p^f} \\ | & & | & | & | \\ \mathbb{Q} & \supset & \mathbb{Z} & (p) & \mathbb{F}_p. \end{array}$$

At the far right we have finite field extensions, which we know are *really* well behaved. So we ask:

> *How are* $\mathrm{Gal}\left((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p\right)$ *and* $\mathrm{Gal}(K/\mathbb{Q})$ *related?*

First, every $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$ induces an automorphism of $\mathcal{O}_K$, which induces a map $\mathcal{O}_K \to \mathcal{O}_K/\mathfrak{p}$ by

$$\alpha \mapsto \sigma(\alpha) \pmod{\mathfrak{p}}.$$

For this to induce a map in $\mathrm{Gal}\left((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p\right)$, it's necessary that $\sigma(\mathfrak{p}) \subseteq \mathfrak{p}$. So, we consider the subset of automorphisms that fixes $\mathfrak{p}$:

Definition 61.5.1. Let $D_\mathfrak{p} \subseteq \mathrm{Gal}(K/\mathbb{Q})$ be the stabilizer of $\mathfrak{p}$, that is

$$D_\mathfrak{p} \coloneqq \{\sigma \in \mathrm{Gal}(K/\mathbb{Q}) \mid \sigma\mathfrak{p} = \mathfrak{p}\}.$$

We say $D_\mathfrak{p}$ is the **decomposition group** of $\mathfrak{p}$.

Note that this definition is in fact equivalent to the set of σ such that $\sigma(\mathfrak{p}) \subseteq \mathfrak{p}$, because a field isomorphism fixes the ideal norm $\mathrm{N}(\mathfrak{p})$.

So there's a natural map

$$D_\mathfrak{p} \xrightarrow{\theta} \mathrm{Gal}\left((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p\right)$$

by declaring $\theta(\sigma)$ to just be "$\sigma \pmod{\mathfrak{p}}$". The fact that $\sigma \in D_\mathfrak{p}$ (i.e. σ fixes $\mathfrak{p}$) ensures this map is well-defined.

Surprisingly, every element of $\mathrm{Gal}\left((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p\right)$ arises this way from some field automorphism of K.

Theorem 61.5.2 (Decomposition group and Galois group)

Define θ as above. Then

- θ is surjective, and
- its kernel is a group of order e, the ramification index.

In particular, if p is unramified then $D_\mathfrak{p} \cong \mathrm{Gal}\left((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p\right)$.

(The proof is not hard, but a bit lengthy and in my opinion not very enlightening.)

If p is unramified, then taking modulo $\mathfrak{p}$ gives $D_{\mathfrak{p}} \cong \operatorname{Gal}((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p)$.

But we know exactly what $\operatorname{Gal}((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p)$ is! We already have $\mathcal{O}_K/\mathfrak{p} \cong \mathbb{F}_{p^f}$, and the Galois group is

$$\operatorname{Gal}((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p) \cong \operatorname{Gal}\left(\mathbb{F}_{p^f}/\mathbb{F}_p\right) \cong \langle x \mapsto x^p \rangle \cong \mathbb{Z}/f\mathbb{Z}.$$

So

$$D_{\mathfrak{p}} \cong \mathbb{Z}/f\mathbb{Z}$$

as well.

Let's now go back to

$$D_{\mathfrak{p}} \xrightarrow{\theta} \operatorname{Gal}((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p).$$

The kernel of θ is called the **inertia group** and denoted $I_{\mathfrak{p}} \subseteq D_{\mathfrak{p}}$; it has order e.

This gives us a pretty cool sequence of subgroups $\{1\} \subseteq I \subseteq D \subseteq G$ where G is the Galois group (I'm dropping the $\mathfrak{p}$-subscripts now). Let's look at the corresponding *fixed fields* via the Fundamental theorem of Galois theory. Picture:

$$\begin{array}{lcc}
\mathfrak{p} \subset \mathcal{O}_K \subset & K & \{1\} \\
& \uparrow \text{Ramify} & \downarrow e \\
& K^I & I \\
& \uparrow \text{Inert} & \downarrow f \\
& K^D & D \\
& \uparrow \text{Split} & \downarrow g \\
(p) \subset \mathbb{Z} \subset & \mathbb{Q} & G
\end{array}$$

Something curious happens:

- If $D \trianglelefteq G$, when (p) is lifted into K^D it splits completely into g unramified primes. Each of these has inertial degree 1.
- If $I \trianglelefteq G$ as well, when the primes in K^D are lifted to K^I, they remain inert, and now have inertial degree f.
- When they're then lifted to K, they ramify with exponent e (but don't split at all).

In other words, the process of going from 1 to efg can be very nicely broken into the three steps above. To draw this in the picture, we get

$$(p) \longrightarrow \mathfrak{p}_1' \dots \mathfrak{p}_g' \longrightarrow \mathfrak{p}_1'' \dots \mathfrak{p}_g'' \longrightarrow (\mathfrak{p}_1 \dots \mathfrak{p}_g)^e$$

$$\{f_i\}: \qquad 1, \dots, 1 \qquad f, \dots, f \qquad f, \dots, f$$

$$\mathbb{Q} \overset{\text{Split}}{\text{———}} K^D \overset{\text{Inert}}{\text{———}} K^I \overset{\text{Ramify}}{\text{———}} K$$

In any case, in the "typical" case that there is no ramification, we just have $K^I = K$.

Example 61.5.3 (Primes split before remaining inert)

Let $K = \mathbb{Q}[\zeta_5]$ where ζ_5 is a primitive 5th root of unity. From Problem 59A*, we know that the Galois group $\mathrm{Gal}(K/\mathbb{Q})$ is isomorphic to $(\mathbb{Z}/5\mathbb{Z})^* \cong \mathbb{Z}/4\mathbb{Z}$.
Let $p = 19$. In K, p factors as $19 = (2\sqrt{5}+1)(2\sqrt{5}-1)$, and luckily for us, $\mathcal{O}_K$ is a principal ideal domain, which means the ideal (19) factors as $(19) = \mathfrak{p}_1\mathfrak{p}_2 = (2\sqrt{5}+1)(2\sqrt{5}-1)$.
In this case, we have $K^{D_{\mathfrak{p}_1}} = K^D = \mathbb{Q}[\sqrt{5}]$ and $K^I = K$, and indeed:

- When (19) is lifted to K^D, it already splits into $(2\sqrt{5}+1)(2\sqrt{5}-1)$ — because $2\sqrt{5}+1 \in K^D$. As $[K^D : \mathbb{Q}] = 2$ and (19) already split into 2 primes, each of the prime necessarily have inertial degree 1.
- When each of $(2\sqrt{5}+1)$ and $(2\sqrt{5}-1)$ is lifted from K^D to K, they remains inert. Again, as $[K : K^D] = 2$, the inertial degree must be 2.

Part of the theorem can be seen very easily: by the fundamental theorem of Galois theory, because all of the field automorphisms in D fixes $2\sqrt{5}+1$, then tautologically, $2\sqrt{5}+1$ must belong to the fixed field of D! In other words, $2\sqrt{5}+1 \in K^D$, which means p already splits when lifted to K^D.
The argument only need to be modified a little to show $\mathfrak{p}_1' = \mathfrak{p}_1 \cap K^D$ does not split when lifted from K^D to K: because the extension K/K^D is Galois, the Galois group $\mathrm{Gal}(K/K^D)$ acts transitively on the primes $\mathfrak{p}_i$ above $\mathfrak{p}_1' = (2\sqrt{5}+1) \subseteq K^D$, but once again, $\mathfrak{p}_1$ is the only prime in the orbit by the definition of D.

Example 61.5.4 (Different primes have different K^D)

When $D \not\trianglelefteq G$, there need not be a single subfield K^D that p splits cleanly into $\mathfrak{p}_1 \dots \mathfrak{p}_g$ when lifted to that field.
The reason is simple — each prime $\mathfrak{p}_i$ gets split from the product in its *own* $K^{D_{\mathfrak{p}_i}}$, but if $D_{\mathfrak{p}_1}$ is not normal in G, then the different $D_{\mathfrak{p}_i}$ are not the same — instead, they're conjugate subgroups of G.
Let us take a concrete example: let $K = \mathbb{Q}(\sqrt[3]{2}, \omega)$ be the splitting field of $x^3 - 2$ over $\mathbb{Q}$. The rational prime $p = (5)$ splits as $p = \mathfrak{p}_1\mathfrak{p}_2\mathfrak{p}_3$ in K, and each has inertial degree 2. Thus $|D_{\mathfrak{p}_i}| = 2$ for each i.
We know that $\mathrm{Gal}(K/\mathbb{Q}) \cong S_3$, and S_3 has no subgroups of order 2, so obviously $D_{\mathfrak{p}_i}$ is not normal in G!
As mentioned above, what happens here is: when p is lifted to $K^{D_{\mathfrak{p}_1}}$, it splits into $\mathfrak{p}_1'\mathfrak{p}_{23}'$, with $\mathfrak{p}_1$ above $\mathfrak{p}_1'$ and both $\mathfrak{p}_2$ and $\mathfrak{p}_3$ above $\mathfrak{p}_{23}'$. In the extension $K^{D_{\mathfrak{p}_1}}/\mathbb{Q}$, $\mathfrak{p}_1'$ has inertial degree 1 as before, but $\mathfrak{p}_{23}'$ has inertial degree 2.

§61.6 Tangential remark: more general Galois extensions

All the discussion about Galois extensions carries over if we replace $K/\mathbb{Q}$ by some different Galois extension K/F. Instead of a rational prime p breaking down in $\mathcal{O}_K$, we would have a prime ideal $\mathfrak{p}$ of F breaking down as

$$\mathfrak{p} \cdot \mathcal{O}_L = (\mathfrak{P}_1 \dots \mathfrak{P}_g)^e$$

in $\mathcal{O}_L$ and then all results hold verbatim. (The $\mathfrak{P}_i$ are primes in L above $\mathfrak{p}$.) Instead of $\mathbb{F}_p$ we would have $\mathcal{O}_F/\mathfrak{p}$.

The reason I choose to work with $F = \mathbb{Q}$ is that capital Gothic P's ($\mathfrak{P}$) look *really* terrifying.

§61.7 A few harder problems to think about

more problems

Problem 61A†. Prove that no rational prime p can remain inert in $K = \mathbb{Q}(\sqrt[3]{2}, \omega)$, the splitting field of $x^3 - 2$. How does this generalize?

62 The Frobenius element

Throughout this chapter $K/\mathbb{Q}$ is a Galois extension with Galois group G, p is an *unramified* rational prime in K, and $\mathfrak{p}$ is a prime above it. Picture:

$$\begin{array}{ccccc} K & \supset & \mathcal{O}_K & \mathfrak{p} & \mathcal{O}_K/\mathfrak{p} \cong \mathbb{F}_{p^f} \\ | & & | & | & | \\ \mathbb{Q} & \supset & \mathbb{Z} & (p) & \mathbb{F}_p \end{array}$$

We recall that the p-th power map $\sigma\colon \mathbb{F}_{p^f} \to \mathbb{F}_{p^f}$ is an automorphism, and it's called the Frobenius map on $\mathbb{F}_{p^f}$. We can try to extend this map to a $K \to K$ map by $\sigma(x) = x^p$, unfortunately this doesn't make it a field automorphism.

Surprisingly, it is nevertheless possible to extend this to some field automorphism $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$.

If p is unramified, then one can show there is a unique $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$ such that $\sigma(\alpha) \equiv \alpha^p \pmod{\mathfrak{p}}$ for every prime p.

§62.1 Frobenius elements

Prototypical example for this section: $\mathrm{Frob}_{\mathfrak{p}}$ *in* $\mathbb{Z}[i]$ *depends on* $p \pmod 4$.

Here is the theorem statement again:

Theorem 62.1.1 (The Frobenius element)

Assume $K/\mathbb{Q}$ is Galois with Galois group G. Let p be a rational prime unramified in K, and $\mathfrak{p}$ a prime above it. There is a *unique* element $\mathrm{Frob}_{\mathfrak{p}} \in G$ with the property that, for all $\alpha \in \mathcal{O}_K$,

$$\mathrm{Frob}_{\mathfrak{p}}(\alpha) \equiv \alpha^p \pmod{\mathfrak{p}}.$$

It is called the **Frobenius element** at $\mathfrak{p}$, and has order f.

The *uniqueness* part is pretty important: it allows us to show that a given $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$ is the Frobenius element by just observing that it satisfies the above functional equation.

Let's see an example of this:

Example 62.1.2 (Frobenius elements of the Gaussian integers)

Let's actually compute some Frobenius elements for $K = \mathbb{Q}(i)$, which has $\mathcal{O}_K = \mathbb{Z}[i]$. This is a Galois extension, with $G = (\mathbb{Z}/2\mathbb{Z})^\times$, corresponding to the identity and complex conjugation.

If p is an odd prime with $\mathfrak{p}$ above it, then $\mathrm{Frob}_{\mathfrak{p}}$ is the unique element such that

$$(a+bi)^p \equiv \mathrm{Frob}_{\mathfrak{p}}(a+bi) \pmod{\mathfrak{p}}$$

in $\mathbb{Z}[i]$. In particular,

$$\text{Frob}_{\mathfrak{p}}(i) = i^p = \begin{cases} i & p \equiv 1 \pmod 4 \\ -i & p \equiv 3 \pmod 4. \end{cases}$$

From this we see that $\text{Frob}_{\mathfrak{p}}$ is the identity when $p \equiv 1 \pmod 4$ and $\text{Frob}_{\mathfrak{p}}$ is complex conjugation when $p \equiv 3 \pmod 4$.

Note that we really only needed to compute $\text{Frob}_{\mathfrak{p}}$ on i. If this seems too good to be true, a philosophical reason is "freshman's dream" where $(x+y)^p \equiv x^p + y^p \pmod p$ (and hence mod $\mathfrak{p}$). So if σ satisfies the functional equation on generators, it satisfies the functional equation everywhere.

We also have an important lemma:

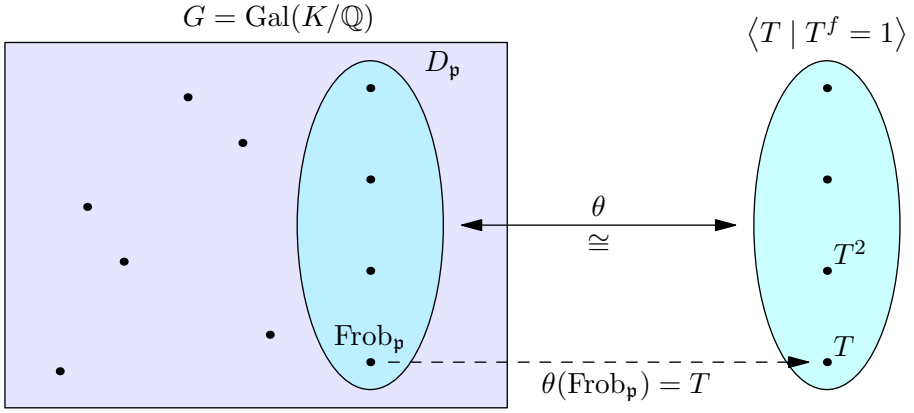
Lemma 62.1.3 (Order of the Frobenius element)
Let $\text{Frob}_{\mathfrak{p}}$ be a Frobenius element from an extension $K/\mathbb{Q}$. Then the order of $\text{Frob}_{\mathfrak{p}}$ is equal to the inertial degree $f_{\mathfrak{p}}$. In particular, (p) splits completely in $\mathcal{O}_K$ if and only if $\text{Frob}_{\mathfrak{p}} = \text{id}$.

This lemma allows us to tell the splitting behavior of $\mathfrak{p}$ just by computing $\text{Frob}_{\mathfrak{p}}$, which will later be seen in Lemma 62.4.1 and Section 62.6.iii.

Exercise 62.1.4. Prove this lemma as by using the fact that $\mathcal{O}_K/\mathfrak{p}$ is the finite field of order $f_{\mathfrak{p}}$, and the Frobenius element is just $x \mapsto x^p$ on this field.

Let us now prove the main theorem. This will only make sense in the context of decomposition groups, so readers which skipped that part should omit this proof.

Proof of existence of Frobenius element. The entire theorem is just a rephrasing of the fact that the map θ defined in the last section is an isomorphism when p is unramified. Picture:



In here we can restrict our attention to $D_{\mathfrak{p}}$ since we need to have $\sigma(\alpha) \equiv 0 \pmod{\mathfrak{p}}$ when $\alpha \equiv 0 \pmod{\mathfrak{p}}$. Thus we have the isomorphism

$$D_{\mathfrak{p}} \xrightarrow{\theta} \text{Gal}\left((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p\right).$$

But we already know $\operatorname{Gal}((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p)$, according to the string of isomorphisms

$$\operatorname{Gal}((\mathcal{O}_K/\mathfrak{p})/\mathbb{F}_p) \cong \operatorname{Gal}\left(\mathbb{F}_{p^f}/\mathbb{F}_p\right) \cong \langle T = x \mapsto x^p \rangle \cong \mathbb{Z}/f\mathbb{Z}.$$

So the unique such element is the pre-image of T under θ. □

§62.2 Conjugacy classes

Now suppose $\mathfrak{p}_1$ and $\mathfrak{p}_2$ are *two* primes above an unramified rational prime p. Then we can define $\operatorname{Frob}_{\mathfrak{p}_1}$ and $\operatorname{Frob}_{\mathfrak{p}_2}$. Since the Galois group acts transitively, we can select $\sigma \in \operatorname{Gal}(K/\mathbb{Q})$ be such that

$$\sigma(\mathfrak{p}_1) = \mathfrak{p}_2.$$

We claim that

$$\operatorname{Frob}_{\mathfrak{p}_2} = \sigma \circ \operatorname{Frob}_{\mathfrak{p}_1} \circ \sigma^{-1}.$$

Note that this is an equation in G.

> **Question 62.2.1.** Prove this.

More generally, for a given unramified rational prime p, we obtain:

Theorem 62.2.2 (Conjugacy classes in Galois groups)
The set

$$\{\operatorname{Frob}_{\mathfrak{p}} \mid \mathfrak{p} \text{ above } p\}$$

is one of the conjugacy classes of G.

Proof. We've used the fact that $G = \operatorname{Gal}(K/\mathbb{Q})$ is transitive to show that $\operatorname{Frob}_{\mathfrak{p}_1}$ and $\operatorname{Frob}_{\mathfrak{p}_2}$ are conjugate if they both lie above p; hence it's *contained* in some conjugacy class. So it remains to check that for any $\mathfrak{p}$, σ, we have $\sigma \circ \operatorname{Frob}_{\mathfrak{p}} \circ \sigma^{-1} = \operatorname{Frob}_{\mathfrak{p}'}$ for some $\mathfrak{p}'$. For this, just take $\mathfrak{p}' = \sigma\mathfrak{p}$. Hence the set is indeed a conjugacy class. □

In summary,

$\operatorname{Frob}_{\mathfrak{p}}$ is determined up to conjugation by the prime p from which $\mathfrak{p}$ arises.

So even though the Gothic letters look scary, the content of $\operatorname{Frob}_{\mathfrak{p}}$ really just comes from the more friendly-looking rational prime p.

Example 62.2.3 (Frobenius elements in $\mathbb{Q}(\sqrt[3]{2}, \omega)$)
With those remarks, here is a more involved example of a Frobenius map. Let $K = \mathbb{Q}(\sqrt[3]{2}, \omega)$ be the splitting field of

$$t^3 - 2 = (t - \sqrt[3]{2})(t - \omega\sqrt[3]{2})(t - \omega^2\sqrt[3]{2}).$$

Thus $K/\mathbb{Q}$ is Galois. We've seen in an earlier example that

$$\mathcal{O}_K = \mathbb{Z}[\varepsilon] \quad \text{where} \quad \varepsilon \text{ is a root of } t^6 + 3t^5 - 5t^3 + 3t + 1.$$

Let's consider the prime 5 which factors (trust me here) as

$$(5) = (5, \varepsilon^2 + \varepsilon + 2)(5, \varepsilon^2 + 3\varepsilon + 3)(5, \varepsilon^2 + 4\varepsilon + 1) = \mathfrak{p}_1\mathfrak{p}_2\mathfrak{p}_3.$$

Note that all the prime ideals have inertial degree 2. Thus $\text{Frob}_{\mathfrak{p}_i}$ will have order 2 for each i.
Note that

$$\text{Gal}(K/\mathbb{Q}) = \text{permutations of } \{\sqrt[3]{2}, \omega\sqrt[3]{2}, \omega^2\sqrt[3]{2}\} \cong S_3.$$

In this S_3 there are 3 elements of order two: fixing one root and swapping the other two. These correspond to each of $\text{Frob}_{\mathfrak{p}_1}$, $\text{Frob}\mathfrak{p}_2$, $\text{Frob}_{\mathfrak{p}_3}$.
In conclusion, the conjugacy class $\{\text{Frob}_{\mathfrak{p}_1}, \text{Frob}_{\mathfrak{p}_2}, \text{Frob}_{\mathfrak{p}_3}\}$ associated to (5) is the cycle type $(\bullet)(\bullet\,\bullet)$ in S_3.

§62.3 Chebotarev density theorem

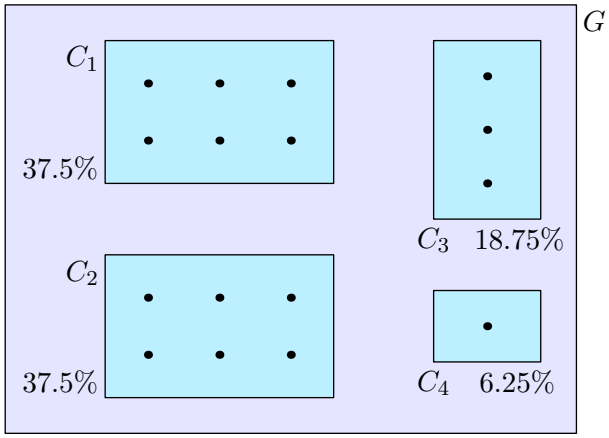
Natural question: can we represent every conjugacy class in this way? In other words, is every element of G equal to $\text{Frob}_{\mathfrak{p}}$ for some $\mathfrak{p}$?

Miraculously, not only is the answer "yes", but in fact it does so in the nicest way possible: the $\text{Frob}_{\mathfrak{p}}$'s are "equally distributed" when we pick a random $\mathfrak{p}$.

Theorem 62.3.1 (Chebotarev density theorem over $\mathbb{Q}$)
Let C be a conjugacy class of $G = \text{Gal}(K/\mathbb{Q})$. The density of (unramified) primes p such that $\{\text{Frob}_{\mathfrak{p}} \mid \mathfrak{p} \text{ above } p\} = C$ is exactly $|C| / |G|$. In particular, for any $\sigma \in G$ there are infinitely many rational primes p with $\mathfrak{p}$ above p so that $\text{Frob}_{\mathfrak{p}} = \sigma$.

By density, I mean that the proportion of primes $p \le x$ that work approaches $\frac{|C|}{|G|}$ as $x \to \infty$. Note that I'm throwing out the primes that ramify in K. This is no issue, since the only primes that ramify are those dividing Δ_K, of which there are only finitely many.

In other words, if I pick a random prime p and look at the resulting conjugacy class, it's a lot like throwing a dart at G: the probability of hitting any conjugacy class depends just on the size of the class.



Remark 62.3.2 — Happily, this theorem (and preceding discussion) also works if we replace $K/\mathbb{Q}$ with any Galois extension K/F; in that case we replace "$\mathfrak{p}$ over p" with "$\mathfrak{P}$ over $\mathfrak{p}$". In that case, we use $\mathrm{N}(\mathfrak{p}) \le x$ rather than $p \le x$ as the way to define density.

§62.4 Example: Frobenius elements of cyclotomic fields

Let q be a prime, and consider $L = \mathbb{Q}(\zeta_q)$, with ζ_q a primitive qth root of unity. You should recall from various starred problems that

- $\Delta_L = \pm q^{q-2}$,
- $\mathcal{O}_L = \mathbb{Z}[\zeta_q]$, and
- The map
$$\sigma_n \colon L \to L \quad \text{by} \quad \zeta_q \mapsto \zeta_q^n$$
is an automorphism of L whenever $\gcd(n, q) = 1$, and depends only on $n \pmod q$. In other words, the automorphisms of $L/\mathbb{Q}$ just shuffle around the qth roots of unity. In fact the Galois group consists exactly of the elements $\{\sigma_n\}$, namely
$$\mathrm{Gal}(L/\mathbb{Q}) = \{\sigma_n \mid n \not\equiv 0 \pmod q\}.$$
As a group,
$$\mathrm{Gal}(L/\mathbb{Q}) = (\mathbb{Z}/q\mathbb{Z})^\times \cong \mathbb{Z}/(q-1)\mathbb{Z}.$$

This is surprisingly nice, because **elements of** $\mathrm{Gal}(L/\mathbb{Q})$ **look a lot like Frobenius elements already**. Specifically:

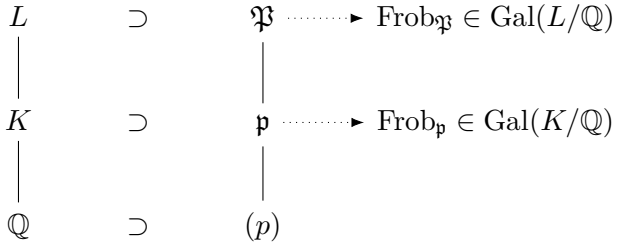
Lemma 62.4.1 (Cyclotomic Frobenius elements)
In the cyclotomic setting $L = \mathbb{Q}(\zeta_q)$, let p be a rational unramified prime and $\mathfrak{p}$ above it. Then
$$\mathrm{Frob}_{\mathfrak{p}} = \sigma_p.$$

Proof. Observe that σ_p satisfies the functional equation (check on generators). Done by uniqueness. □

Question 62.4.2. Conclude that a rational prime p splits completely in $\mathcal{O}_L$ if and only if $p \equiv 1 \pmod q$.

§62.5 Frobenius elements behave well with restriction

Let $L/\mathbb{Q}$ and $K/\mathbb{Q}$ be Galois extensions, and consider the setup



Here $\mathfrak{p}$ is above (p) and $\mathfrak{P}$ is above $\mathfrak{p}$. We may define

$$\text{Frob}_{\mathfrak{p}} \colon K \to K \quad \text{and} \quad \text{Frob}_{\mathfrak{P}} \colon L \to L$$

and want to know how these are related.

Both maps $\text{Frob}_{\mathfrak{P}}$ and $\text{Frob}_{\mathfrak{p}}$ induce the power-of-p map in the corresponding quotient field, hence we would expect them to be naturally the same.

Theorem 62.5.1 (Restrictions of Frobenius elements)
Assume $L/\mathbb{Q}$ and $K/\mathbb{Q}$ are both Galois. Let $\mathfrak{P}$ and $\mathfrak{p}$ be unramified as above. Then $\text{Frob}_{\mathfrak{P}}\restriction_K = \text{Frob}_{\mathfrak{p}}$, i.e. for every $\alpha \in K$,

$$\text{Frob}_{\mathfrak{p}}(\alpha) = \text{Frob}_{\mathfrak{P}}(\alpha).$$

Proof. First, $K/\mathbb{Q}$ is normal, so $\text{Frob}_{\mathfrak{P}}$ fixes the image of K, that is, $\text{Frob}_{\mathfrak{P}}\restriction_K \in \text{Gal}(K/\mathbb{Q})$ is well-defined.

We have the natural map $\phi \colon \mathcal{O}_K \to \mathcal{O}_L \to \mathcal{O}_L/\mathfrak{P}$, and the quotient map $q \colon \mathcal{O}_K \to \mathcal{O}_K/\mathfrak{p}$. Since $\mathfrak{p} \subseteq \mathfrak{P} \cap \mathcal{O}_K \subseteq \ker \phi$, it follows ϕ factors through q to give a natural field homomorphism $\mathcal{O}_K/\mathfrak{p} \to \mathcal{O}_L/\mathfrak{P}$.

Since a field homomorphism is injective, $\text{Frob}_{\mathfrak{P}}$ induces the power-of-p map on $\mathcal{O}_L/\mathfrak{P}$, and everything is commutative, the theorem follows. □

In short, the point of this section is that

Frobenius elements upstairs restrict to Frobenius elements downstairs.

§62.6 Application: Quadratic reciprocity

We now aim to prove:

Theorem 62.6.1 (Quadratic reciprocity)
Let p and q be distinct odd primes. Then

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}}.$$

(See, e.g. [Le] for an exposition on quadratic reciprocity, if you're not familiar with it.)

§62.6.i Step 1: Setup

For this proof, we first define

$$L = \mathbb{Q}(\zeta_q)$$

where ζ_q is a primitive qth root of unity. Then $L/\mathbb{Q}$ is Galois, with Galois group G.

Question 62.6.2. Show that G has a unique subgroup H of index two.

In fact, we can describe it exactly: viewing $G \cong (\mathbb{Z}/q\mathbb{Z})^\times$, we have

$$H = \{\sigma_n \mid n \text{ quadratic residue mod } q\}.$$

By the fundamental theorem of Galois Theory, there ought to be a degree 2 extension of $\mathbb{Q}$ inside $\mathbb{Q}(\zeta_q)$ (that is, a quadratic field). Call it $\mathbb{Q}(\sqrt{q^*})$, for q^* squarefree:

$$\begin{array}{ccc} L = \mathbb{Q}(\zeta_q) & \longleftrightarrow & \{1\} \\ {\scriptstyle \frac{q-1}{2}}\big| & & \big| \\ K = \mathbb{Q}(\sqrt{q^*}) & \longleftrightarrow & H \\ {\scriptstyle 2}\big| & & \big| \\ \mathbb{Q} & \longleftrightarrow & G \end{array}$$

Exercise 62.6.3. Note that if a rational prime ℓ ramifies in K, then it ramifies in L. Use this to show that

$$q^* = \pm q \text{ and } q^* \equiv 1 \pmod 4.$$

Together these determine the value of q^*.

(Actually, it is true in general Δ_K divides Δ_L in a tower $L/K/\mathbb{Q}$.)

§62.6.ii Step 2: Reformulation

Now we are going to prove:

Theorem 62.6.4 (Quadratic reciprocity, equivalent formulation)
For distinct odd primes p, q we have

$$\left(\frac{p}{q}\right) = \left(\frac{q^*}{p}\right).$$

Exercise 62.6.5. Using the fact that $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}$, show that this is equivalent to quadratic reciprocity as we know it.

We look at the rational prime p in $\mathbb{Z}$. Either it splits into two in K or is inert; either way let $\mathfrak{p}$ be a prime factor in the resulting decomposition (so $\mathfrak{p}$ is either $p \cdot \mathcal{O}_K$ in the inert case, or one of the primes in the split case). Then let $\mathfrak{P}$ be above $\mathfrak{p}$. It could possibly also split in K: the picture looks like

$$\begin{array}{ccccc} \mathcal{O}_L = \mathbb{Z}[\zeta_q] & \supset & \mathfrak{P} & \dashrightarrow & \mathbb{Z}[\zeta_q]/\mathfrak{P} \cong \mathbb{F}_{p^f} \\ \mathcal{O}_K = \mathbb{Z}[\frac{1+\sqrt{q^*}}{2}] & \supset & \mathfrak{p} & \dashrightarrow & \mathbb{F}_p \text{ or } \mathbb{F}_{p^2} \\ \mathbb{Z} & \supset & (p) & \dashrightarrow & \mathbb{F}_p \end{array}$$

Question 62.6.6. Why is p not ramified in either K or L?

§62.6.iii Step 3: Introducing the Frobenius

Now, we take the Frobenius

$$\sigma_p = \text{Frob}_{\mathfrak{P}} \in \text{Gal}(L/\mathbb{Q}).$$

We claim that

$$\text{Frob}_{\mathfrak{P}} \in H \iff p \text{ splits in } K.$$

To see this, note that $\text{Frob}_{\mathfrak{P}}$ is in H if and only if it acts as the identity on K. But $\text{Frob}_{\mathfrak{P}}\restriction_K$ is $\text{Frob}_{\mathfrak{p}}$! So

$$\text{Frob}_{\mathfrak{P}} \in H \iff \text{Frob}_{\mathfrak{p}} = \text{id}_K.$$

Finally, by Lemma 62.1.3, $\text{Frob}_{\mathfrak{p}}$ has order 1 if p splits ($\mathfrak{p}$ has inertial degree 1) and order 2 if p is inert. This completes the proof of the claim.

§62.6.iv Finishing up

We already know by Lemma 62.4.1 that $\text{Frob}_{\mathfrak{P}} = \sigma_p \in H$ if and only if p is a quadratic residue. On the other hand,

> **Exercise 62.6.7.** Show that p splits in $\mathcal{O}_K = \mathbb{Z}[\frac{1}{2}(1+\sqrt{q^*})]$ if and only if $\left(\frac{q^*}{p}\right) = 1$. (Use the factoring algorithm. You need the fact that $p \neq 2$ here.)

In other words,

$$\begin{aligned}
\left(\frac{p}{q}\right) = 1 &\iff \sigma_p \in H \\
&\iff \text{Frob}_{\mathfrak{P}} \in H \\
&\iff \text{Frob}_{\mathfrak{p}} = \text{id}_K \\
&\iff \text{ord}\,\text{Frob}_{\mathfrak{p}} = 1 \\
&\iff f_{\mathfrak{p}} = 1 \\
&\iff p \text{ splits in } \mathbb{Z}\left[\tfrac{1}{2}(1+\sqrt{q^*})\right] \\
&\iff \left(\frac{q^*}{p}\right) = 1.
\end{aligned}$$

This completes the proof.

§62.7 Frobenius elements control factorization

Prototypical example for this section: $\text{Frob}_{\mathfrak{p}}$ *controlled the splitting of* p *in the proof of quadratic reciprocity; the same holds in general.*

In the proof of quadratic reciprocity, we used the fact that Frobenius elements behaved well with restriction in order to relate the splitting of p with properties of $\text{Frob}_{\mathfrak{p}}$.

In fact, there is a much stronger statement for any intermediate field $\mathbb{Q} \subseteq E \subseteq K$ which works even if $E/\mathbb{Q}$ is not Galois. It relies on the notion of a *factorization pattern*. Here is how it goes.

Set $n = [E : \mathbb{Q}]$, and let p be a rational prime unramified in K. Then p can be broken in E as

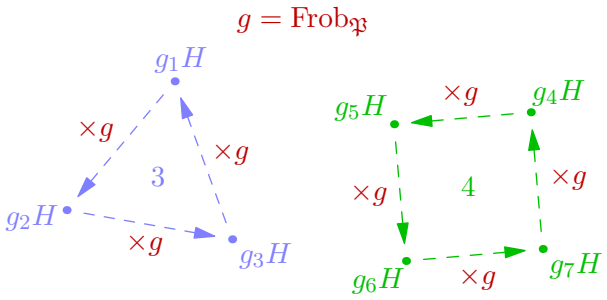
$$p \cdot \mathcal{O}_E = \mathfrak{p}_1\mathfrak{p}_2\ldots\mathfrak{p}_g$$

with inertial degrees $f_1, \dots, f_g$: (these inertial degrees might be different since $E/\mathbb{Q}$ isn't Galois). The numbers $f_1 + \cdots + f_g = n$ form a partition of the number n. For example, in the quadratic reciprocity proof we had $n = 2$, with possible partitions $1+1$ (if p split) and 2 (if p was inert). We call this the **factorization pattern** of p in E.

Next, we introduce a Frobenius $\text{Frob}_{\mathfrak{P}}$ above (p), all the way in K; this is an element of $G = \text{Gal}(K/\mathbb{Q})$. Then let H be the group corresponding to the field E. Diagram:

$$\begin{array}{ccc} K & \longleftrightarrow & \{1\} \\ | & & | \\ E & \longleftrightarrow & H \\ {\scriptstyle n}| & & |{\scriptstyle n} \\ \mathbb{Q} & \longleftrightarrow & G \end{array} \qquad \begin{array}{c} \text{Frob}_{\mathfrak{P}} \\ \\ \mathfrak{p}_1 \dots \mathfrak{p}_g \\ | \\ (p) \end{array} \qquad f_1 + \cdots + f_g = n$$

Then $\text{Frob}_{\mathfrak{P}}$ induces a *permutation* of the n left cosets gH by left multiplication (after all, $\text{Frob}_{\mathfrak{P}}$ is an element of G too!). Just as with any permutation, we may look at the resulting cycle decomposition, which has a natural "cycle structure": a partition of n.



$$\boxed{n = 7 = 3 + 4}$$

The theorem is that these coincide:

Theorem 62.7.1 (Frobenius elements control decomposition)
Let $\mathbb{Q} \subseteq E \subseteq K$ an extension of number fields and assume $K/\mathbb{Q}$ is Galois (though $E/\mathbb{Q}$ need not be). Pick an unramified rational prime p; let $G = \text{Gal}(K/\mathbb{Q})$ and H the corresponding intermediate subgroup. Finally, let $\mathfrak{P}$ be a prime above p in K. Then the *factorization pattern* of p in E is given by the *cycle structure* of $\text{Frob}_{\mathfrak{P}}$ acting on the left cosets of H.

Often, we take $E = K$, in which case this is just asserting that the decomposition of the prime p is controlled by a Frobenius element over it.

Sketch of Proof. Let α be an algebraic integer and f its minimal polynomial (of degree n). Set $E = \mathbb{Q}(\alpha)$ (which has degree n over $\mathbb{Q}$). Suppose we're lucky enough that $\mathcal{O}_E = \mathbb{Z}[\alpha]$, i.e. that E is monogenic. Then we know by the Factoring Algorithm, to factor any p in E, all we have to do is factor f modulo p, since if $f = f_1^{e_1} \dots f_g^{e_g} \pmod p$ then we have

$$(p) = \prod_i \mathfrak{p}_i = \prod_i (f_i(\alpha), p)^{e_i}.$$

This gives us complete information about the ramification indices and inertial degrees; the e_i are the ramification indices, and $\deg f_i$ are the inertial degrees (since $\mathcal{O}_E/\mathfrak{p}_i \cong \mathbb{F}_p[X]/(f_i(X))$).

In particular, if p is unramified then all the e_i are equal to 1, and we get

$$n = \deg f = \deg f_1 + \deg f_2 + \cdots + \deg f_g.$$

Once again we have a partition of n; we call this the **factorization pattern** of f modulo p. So, to see the factorization pattern of an unramified p in $\mathcal{O}_E$, we just have to know the factorization pattern of $f \pmod p$.

To prove our theorem, we will show that the factorization pattern of $f \pmod p$ corresponds exactly to the cycle decomposition of the action of $\text{Frob}_{\mathfrak{P}}$ on the roots of f and that the roots of f correspond exactly to the cosets of H in G.

To do this, suppose $S = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ are the roots of f (distinct roots since f is irreducible over $\mathbb{Q}$). We let $\text{Frob}_{\mathfrak{P}}$ act on S. This splits S into orbits S_1, S_2, …, S_k. Construct polynomials f_i with coefficients in E having roots exactly the elements of S_i. This forms a factorization of f over E, say

$$f = f_1 f_2 \dots f_k.$$

We claim that this in fact induces a factorization of $f \pmod p$. To see this, consider the images of these polynomials f_i under the quotient $\mathcal{O}_K \to \mathcal{O}_K/\mathfrak{P}$, denote them by $\overline{f_i}$. Then since p is unramified, we know that the decomposition group $D(\mathfrak{P}|p)$ is isomorphic to the Galois group $\mathcal{G} = \text{Gal}((\mathcal{O}_E/\mathfrak{P})/(\mathbb{Z}/p\mathbb{Z}))$. Thus $\text{Frob}_{\mathfrak{P}}$ corresponds to the generator σ of $\mathcal{G}$. It is not hard to believe that the action of $\text{Frob}_{\mathfrak{P}}$ on the roots of f is the same as that of σ on the roots of $\overline{f}$. Since the roots of f_i form an orbit under the action of $\text{Frob}_{\mathfrak{P}}$, we see that the roots of $\overline{f_i}$ form an orbit under the action of σ and hence under the action of $\mathcal{G}$. It is now a standard fact of Galois theory that $\overline{f_i}$ is an irreducible polynomial over $\mathbb{F}_p$ (since it is fixed by $\mathcal{G}$), thus the claim is proved.

Now we just need to observe that the roots of f correspond to the cosets of H, this will be established later. □

We saw above that given the factorization pattern of $f \pmod p$, we can determine the factorization pattern of an unramified prime p in $\mathcal{O}_E$.

Turning this on its head, if we want to know the factorization pattern of $f \pmod p$, we just need to know how p decomposes. And it turns out these coincide even without the assumption that E is monogenic.

Theorem 62.7.2 (Frobenius controls polynomial factorization)

Let α be an algebraic integer with minimal polynomial f, and let $E = \mathbb{Q}(\alpha)$. Then for any prime p unramified in the splitting field K of f, the following coincide:

(i) The factorization pattern of p in E.

(ii) The factorization pattern of $f \pmod p$.

(iii) The cycle structure associated to the action of $\text{Frob}_{\mathfrak{P}} \in \text{Gal}(K/\mathbb{Q})$ on the roots of f, where $\mathfrak{P}$ is above p in K.

Example 62.7.3 (Factoring $x^3 - 2 \pmod 5$)

Let $\alpha = \sqrt[3]{2}$ and $f = x^3 - 2$, so $E = \mathbb{Q}(\sqrt[3]{2})$. Set $p = 5$ and finally, let $K = \mathbb{Q}(\sqrt[3]{2}, \omega)$ be the splitting field. Setup:

$$\begin{array}{ccccc}
K = \mathbb{Q}(\sqrt[3]{2}, \omega) & & \mathfrak{P} & & x^3 - 2 = (x - \sqrt[3]{2})(x - \sqrt[3]{2}\omega)(x - \sqrt[3]{2}\omega^2) \\
\Big| 2 & & \Big| & & \\
E = \mathbb{Q}(\sqrt[3]{2}) & & \mathfrak{p} & & x^3 - 2 = (x - \sqrt[3]{2})(x^2 + \sqrt[3]{2}x + \sqrt[3]{4}) \\
\Big| 3 & & \Big| & & \\
\mathbb{Q} & & (5) & & x^3 - 2 \text{ irreducible over } \mathbb{Q}
\end{array}$$

The three claimed objects now all have shape $2+1$:

(i) By the Factoring Algorithm, we have $(5) = (5, \sqrt[3]{2} - 3)(5, 9 + 3\sqrt[3]{2} + \sqrt[3]{4})$.

(ii) We have $x^3 - 2 \equiv (x-3)(x^2+3x+9) \pmod 5$.

(iii) We saw before that $\text{Frob}_{\mathfrak{P}} = (\bullet)(\bullet\ \bullet)$.

Sketch of Proof. Letting $n = \deg f$. Let H be the subgroup of $G = \text{Gal}(K/\mathbb{Q})$ corresponding to E, so $|G/H| = n$. Pictorially, we have

$$\begin{array}{ccccc}
K & & \{1\} & & \mathfrak{P} \\
\Big| & & \Big| & & \Big| \\
E = \mathbb{Q}(\alpha) & & H & & \mathfrak{p} \\
\Big| & & \Big| & & \Big| \\
\mathbb{Q} & & G & & (p)
\end{array}$$

We claim that (i), (ii), (iii) are all equivalent to

(iv) The pattern of the action of $\text{Frob}_{\mathfrak{P}}$ on the G/H.

In other words we claim the cosets correspond to the n roots of f in K. Indeed H is just the set of $\tau \in G$ such that $\tau(\alpha) = \alpha$, so there's a bijection between the roots and the cosets G/H by $\tau H \mapsto \tau(\alpha)$. Think of it this way: if $G = S_n$, and $H = \{\tau : \tau(1) = 1\}$, then G/H has order $n!/(n-1)! = n$ and corresponds to the elements $\{1, \dots, n\}$. So there is a natural bijection from (iii) to (iv).

The fact that (i) is in bijection to (iv) was the previous theorem, Theorem 62.7.1. The correspondence (i) $\iff$ (ii) is a fact of Galois theory, so we omit the proof here. □

All this can be done in general with $\mathbb{Q}$ replaced by F; for example, in [Le02].

§62.8 Example application: IMO 2003 problem 6

As an example of the power we now have at our disposal, let's prove:

Problem 6. Let p be a prime number. Prove that there exists a prime number q such that for every integer n, the number $n^p - p$ is not divisible by q.

We will show, much more strongly, that there exist infinitely many primes q such that $X^p - p$ is irreducible modulo q.

Solution. Okay! First, we draw the tower of fields

$$\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[p]{p}) \subseteq K$$

where K is the splitting field of $f(x) = x^p - p$. Let $E = \mathbb{Q}(\sqrt[p]{p})$ for brevity and note it has degree $[E : \mathbb{Q}] = p$. Let $G = \mathrm{Gal}(K/\mathbb{Q})$.

> **Question 62.8.1.** Show that p divides the order of G. (Look at E.)

Hence by Cauchy's theorem (Problem 17A*, which is a purely group-theoretic fact) we can find a $\sigma \in G$ of order p. By Chebotarev, there exist infinitely many rational (unramified) primes $q \neq p$ and primes $\mathfrak{Q} \subseteq \mathcal{O}_K$ above q such that $\mathrm{Frob}_{\mathfrak{Q}} = \sigma$. (Yes, that's an uppercase Gothic Q. Sorry.)

We claim that all these q work.

By Theorem 62.7.2, the factorization of $f \pmod q$ is controlled by the action of $\sigma = \mathrm{Frob}_{\mathfrak{Q}}$ on the roots of f. But σ has prime order p in G! So all the lengths in the cycle structure have to divide p. Thus the possible factorization patterns of f are

$$p = \underbrace{1 + 1 + \cdots + 1}_{p \text{ times}} \quad \text{or} \quad p = p.$$

So we just need to rule out the $p = 1 + \cdots + 1$ case now: this only happens if f breaks into linear factors mod q. Intuitively this edge case seems highly unlikely (are we really so unlucky that f factors into *linear* factors when we want it to be irreducible?). And indeed this is easy to see: this means that σ fixes all of the roots of f in K, but that means σ fixes K altogether, and hence is the identity of G, contradiction. □

> **Remark 62.8.2 —** In fact $K = \mathbb{Q}(\sqrt[p]{p}, \zeta_p)$, and $|G| = p(p-1)$. With a little more group theory, we can show that in fact the density of primes q that work is $\frac{1}{p}$.

§62.9 A few harder problems to think about

Problem 62A. Show that for an odd prime p,

$$\left(\frac{2}{p}\right) = (-1)^{\frac{1}{8}(p^2-1)}.$$

Problem 62B. Let f be a nonconstant polynomial with integer coefficients. Suppose $f \pmod p$ splits completely into linear factors for all sufficiently large primes p. Show that f splits completely into linear factors.

Problem 62C† (Dirichlet's theorem on arithmetic progressions)**.** Let a and m be relatively prime positive integers. Show that the density of primes $p \equiv a \pmod m$ is exactly $\frac{1}{\phi(m)}$.

Problem 62D. Let n be an odd integer which is not a prime power. Show that the nth cyclotomic polynomial is not irreducible modulo *any* rational prime.

Problem 62E (Putnam 2012 B6)**.** Let p be an odd prime such that $p \equiv 2 \pmod 3$. Let π be a permutation of $\mathbb{F}_p$ by $\pi(x) = x^3 \pmod p$. Show that π is even if and only if $p \equiv 3 \pmod 4$.

63 Bonus: A Bit on Artin Reciprocity

In this chapter, I'm going to state some big theorems of global class field theory and use them to deduce the Kronecker-Weber plus Hilbert class fields. No proofs, but hopefully still appreciable. For experts: this is global class field theory, without ideles.

Here's the executive summary: let K be a number field. Then all abelian extensions L/K can be understood using solely information intrinsic to K: namely, the ray class groups (generalizing ideal class groups).

§63.1 Overview

At the end of this section, for an Abelian field extension L/K, we will define the Artin symbol

$$\left(\frac{L/K}{\mathfrak{p}}\right),$$

which generalizes the Legendre symbol $\left(\frac{a}{p}\right)$:

- Above the solidus, instead of an integer a, we have a field extension L/K.
- Below the solidus, instead of a rational prime p, we have a prime ideal $\mathfrak{p}$ of K.

We require $\mathfrak{p}$ to not ramify in the extension L/K for the symbol to be defined.

And, at the end, we want to state the Artin reciprocity theorem, which looks something like the following:

For primes $\mathfrak{p}$, $\left(\frac{L/K}{\mathfrak{p}}\right)$ depends only on "$\mathfrak{p} \pmod{\mathfrak{f}}$".

Here, $\mathfrak{f}$ is a "modulus", which only depends on the field extension L/K.

In order to do that, we first need to define what it means for two ideals to be coprime modulo something. We will divide up the ideals of $\mathcal{O}_K$ that is "coprime" to $\mathfrak{f}$ into "residue classes modulo $\mathfrak{f}$" (we will call them "ray classes" from now on) in such a way that:

- It generalizes the class group – two ideals that belong to different ideal classes (i.e. are nonisomorphic as $\mathcal{O}_K$-modules) belong to different ray classes.
- It respects the multiplicative structure – if $\mathfrak{p}$ is in the same ray class as $\mathfrak{p}'$, and $\mathfrak{q}$ is in the same ray class as $\mathfrak{q}'$, then $\mathfrak{pq}$ is in the same ray class as $\mathfrak{p}'\mathfrak{q}'$.

  Note that there is no analogue of element addition for the ideals (for instance, $(1) = (-1)$ but $(1) + (1) \neq (1) + (-1)$), so this is the best we can hope for.

  In other words, the ray classes will form an *abelian group* under multiplication, with the operation induced from ideal multiplication.
- For a fixed modulus $\mathfrak{f}$, there are only finitely many ray classes.

In the section above, you may think of a prime ideal $\mathfrak{p} \in \mathcal{O}_K$ as an irreducible factor, such that all ideals can be written as products of. However, they can also naturally be used as a modulus:

A prime $\mathfrak{p}$ gives a way to divide the elements of $\mathcal{O}_K$ into residue classes that respects the addition and multiplication of elements.

This can further be generalized to divide up the *ideals* of $\mathcal{O}_K$ into ray classes – unfortunately, using only the finite primes is insufficient to divide up the ideals the way we want, as later seen in Example 63.3.3. So, the infinite primes will be introduced in order to divide up the *elements*, as well as the ideals, into classes that satisfies the multiplicative structure.

§63.2 Infinite primes

Prototypical example for this section: $\mathbb{Q}(\sqrt{-5})$ has a complex infinite prime, $\mathbb{Q}(\sqrt{5})$ has two real infinite ones.

Let K be a number field of degree n and signature (r, s). We know what a prime ideal of $\mathcal{O}_K$ is, but we now allow for the so-called infinite primes, which I'll describe using the embeddings.[1] Recall there are n embeddings $\sigma \colon K \to \mathbb{C}$, which consist of

- r real embeddings where $\operatorname{im}\sigma \subseteq \mathbb{R}$, and
- s pairs of conjugate complex embeddings.

Hence $r + 2s = n$. The first class of embeddings form the **real infinite primes**, while the **complex infinite primes** are the second type. We say K is **totally real** (resp **totally complex**) if all its infinite primes are real (resp complex).

Example 63.2.1 (Examples of infinite primes)

- $\mathbb{Q}$ has a single real infinite prime. We often write it as ∞.
- $\mathbb{Q}(\sqrt{-5})$ has a single complex infinite prime, and no real infinite primes. Hence totally complex.
- $\mathbb{Q}(\sqrt{5})$ has two real infinite primes, and no complex infinite primes. Hence totally real.

§63.3 Modular arithmetic with infinite primes

A **modulus** (or **module**) of K is a formal product

$$\mathfrak{m} = \prod_{\mathfrak{p}} \mathfrak{p}^{\nu(\mathfrak{p})}$$

where the product runs over all primes, finite and infinite. (Here $\nu(\mathfrak{p})$ is a nonnegative integer, of which only finitely many are nonzero.) We also require that

- $\nu(\mathfrak{p}) = 0$ for any complex infinite prime $\mathfrak{p}$, and
- $\nu(\mathfrak{p}) \le 1$ for any real infinite prime $\mathfrak{p}$.

[1] This is not really the right definition; the "correct" way to think of primes, finite or infinite, is in terms of valuations. But it'll be sufficient for me to state the theorems I want.

Obviously, every $\mathfrak{m}$ can be written as $\mathfrak{m} = \mathfrak{m}_0\mathfrak{m}_\infty$ by separating the finite from the (real) infinite primes.

We say $a \equiv b \pmod{\mathfrak{p}}$ if

- If $\mathfrak{p}$ is a finite prime, then $a \equiv b \pmod{\mathfrak{p}^{\nu(\mathfrak{p})}}$ means exactly what you think it should mean: $a - b \in \mathfrak{p}^{\nu(\mathfrak{p})}$.
- If $\mathfrak{p}$ is a *real* infinite prime $\sigma \colon K \to \mathbb{R}$, then $a \equiv b \pmod{\mathfrak{p}}$ means that $\sigma(a/b) > 0$.

A real infinite prime $\mathfrak{p} = \sigma$ divides up the elements of $K^\times$ into two classes $\{k \in K^\times \mid \sigma(k) > 0\}$ and $\{k \in K^\times \mid \sigma(k) < 0\}$, this division satisfies the multiplicative operation.

Of course, $a \equiv b \pmod{\mathfrak{m}}$ means $a \equiv b$ modulo each prime power in $\mathfrak{m}$. With this, we can define a generalization of the class group:

Definition 63.3.1. Let $\mathfrak{m}$ be a modulus of a number field K.

- Let $I_K(\mathfrak{m})$ denote the set of all fractional ideals of K which are relatively prime to $\mathfrak{m}$, which is an abelian group.
- Let $P_K(\mathfrak{m})$ be the subgroup of $I_K(\mathfrak{m})$ generated by

$$\left\{\alpha\mathcal{O}_K \mid \alpha \in K^\times \text{ and } \alpha \equiv 1 \pmod{\mathfrak{m}}\right\}.$$

  This is sometimes called a "ray" of principal ideals.[2]

Finally define the **ray class group** of $\mathfrak{m}$ to be $C_K(\mathfrak{m}) = I_K(\mathfrak{m})/P_K(\mathfrak{m})$.

This group is known to always be finite. Note the usual class group is $C_K(1)$.

One last definition that we'll use right after Artin reciprocity:

Definition 63.3.2. A **congruence subgroup** of $\mathfrak{m}$ is a subgroup H with

$$P_K(\mathfrak{m}) \subseteq H \subseteq I_K(\mathfrak{m}).$$

Thus $C_K(\mathfrak{m})$ is a group which contains a lattice of various quotients $I_K(\mathfrak{m})/H$, where H is a congruence subgroup.

This definition takes a while to get used to, so here are examples.

Example 63.3.3 (Ray class groups in $\mathbb{Q}$, finite modulus)
Consider $K = \mathbb{Q}$ with infinite prime ∞. Then

- If we take $\mathfrak{m} = 1$ then $I_{\mathbb{Q}}(1)$ is all fractional ideals, and $P_{\mathbb{Q}}(1)$ is all principal fractional ideals. Their quotient is the usual class group of $\mathbb{Q}$, which is trivial.
- Now take $\mathfrak{m} = 8$. Thus $I_{\mathbb{Q}}(8) \cong \left\{\frac{a}{b}\mathbb{Z} \mid a/b \equiv 1, 3, 5, 7 \pmod{8}\right\}$. Moreover

$$P_{\mathbb{Q}}(8) \cong \left\{\frac{a}{b}\mathbb{Z} \mid a/b \equiv 1 \pmod{8}\right\}.$$

  You might at first glance think that the quotient is thus $(\mathbb{Z}/8\mathbb{Z})^\times$. But the

[2] Probably because, similar to a geometrical ray, it only extends infinitely in one direction – at least when there is an infinite prime in the modulus $\mathfrak{m}$.

issue is that we are dealing with *ideals*: specifically, we have

$$7\mathbb{Z} = -7\mathbb{Z} \in P_{\mathbb{Q}}(8)$$

because $-7 \equiv 1 \pmod 8$. So *actually*, we get

$$C_{\mathbb{Q}}(8) \cong \{1,3,5,7 \bmod 8\} / \{1,7 \bmod 8\} \cong (\mathbb{Z}/4\mathbb{Z})^{\times}.$$

More generally,

$$C_{\mathbb{Q}}(m) = (\mathbb{Z}/m\mathbb{Z})^{\times}/\{\pm 1\}.$$

Example 63.3.4 (Ray class groups in $\mathbb{Q}$, infinite moduli)
Consider $K = \mathbb{Q}$ with infinite prime ∞ again.

- Now take $\mathfrak{m} = \infty$. As before $I_{\mathbb{Q}}(\infty) = \mathbb{Q}^{\times}$. Now, by definition we have

  $$P_{\mathbb{Q}}(\infty) = \left\{ \frac{a}{b}\mathbb{Z} \mid a/b > 0 \right\}.$$

  At first glance you might think this was $\mathbb{Q}_{>0}$, but the same behavior with ideals shows in fact $P_{\mathbb{Q}}(\infty) = \mathbb{Q}^{\times}$. So in this case, $P_{\mathbb{Q}}(\infty)$ still has all principal fractional ideals. Therefore, $C_{\mathbb{Q}}(\infty)$ is still trivial.

- Finally, let $\mathfrak{m} = 8\infty$. As before $I_{\mathbb{Q}}(8\infty) \cong \{\frac{a}{b}\mathbb{Z} \mid a/b \equiv 1,3,5,7 \pmod 8\}$. Now in this case:

  $$P_{\mathbb{Q}}(8\infty) \cong \left\{ \frac{a}{b}\mathbb{Z} \mid a/b \equiv 1 \pmod 8 \text{ and } a/b > 0 \right\}.$$

  This time, we really do have $-7\mathbb{Z} \notin P_{\mathbb{Q}}(8\infty)$: we have $7 \not\equiv 1 \pmod 8$ and also $-7 < 0$. So neither of the generators of $7\mathbb{Z}$ are in $P_{\mathbb{Q}}(8\infty)$. Thus we finally obtain

  $$C_{\mathbb{Q}}(8\infty) \cong \{1,3,5,7 \bmod 8\} / \{1 \bmod 8\} \cong (\mathbb{Z}/8\mathbb{Z})^{\times}$$

  with the bijection $C_{\mathbb{Q}}(8\infty) \to (\mathbb{Z}/8\mathbb{Z})^{\times}$ given by $a\mathbb{Z} \mapsto |a| \pmod 8$.

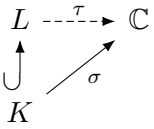
More generally,

$$C_{\mathbb{Q}}(m\infty) = (\mathbb{Z}/m\mathbb{Z})^{\times}.$$

§63.4 Infinite primes in extensions

I want to emphasize that everything above is *intrinsic* to a particular number field K. After this point we are going to consider extensions L/K but it is important to keep in mind the distinction that the concept of modulus and ray class group are objects defined solely from K rather than the above L.

Now take a *Galois* extension L/K of degree m. We already know prime ideals $\mathfrak{p}$ of K break into a product of prime ideals $\mathfrak{P}$ of L in a nice way, so we want to do the same thing with infinite primes. This is straightforward: each of the n infinite primes $\sigma\colon K \to \mathbb{C}$ lifts to m infinite primes $\tau\colon L \to \mathbb{C}$, by which I mean the diagram



commutes. Hence like before, each infinite prime σ of K has m infinite primes τ of L which lie above it.

For a real prime σ of K, if any of the resulting τ above it are complex, we say that the prime σ **ramifies** in the extension L/K. Otherwise it is **unramified** in L/K. An infinite prime of K is always unramified in L/K. In this way, we can talk about an unramified Galois extension L/K: it is one where all primes (finite or infinite) are unramified.

Example 63.4.1 (Ramification of ∞)
Let ∞ be the real infinite prime of $\mathbb{Q}$.

- ∞ is ramified in $\mathbb{Q}(\sqrt{-5})/\mathbb{Q}$.
- ∞ is unramified in $\mathbb{Q}(\sqrt{5})/\mathbb{Q}$.

Note also that if K is totally complex then any extension L/K is unramified.

§63.5 Frobenius element and Artin symbol

Recall the key result:

Theorem 63.5.1 (Frobenius element)
Let L/K be a *Galois* extension. If $\mathfrak{p}$ is a prime unramified in L/K, and $\mathfrak{P}$ a prime above it in L, then there is a unique element of $\mathrm{Gal}(L/K)$, denoted $\mathrm{Frob}_{\mathfrak{P}}$, obeying

$$\mathrm{Frob}_{\mathfrak{P}}(\alpha) \equiv \alpha^{\mathrm{N}(\mathfrak{p})} \pmod{\mathfrak{P}} \qquad \forall \alpha \in \mathcal{O}_L.$$

Recall some examples from Example 62.1.2 and Lemma 62.4.1.

Example 63.5.2 (Example of Frobenius elements)
Let $L = \mathbb{Q}(i)$, $K = \mathbb{Q}$. We have $\mathrm{Gal}(L/K) \cong \mathbb{Z}/2\mathbb{Z}$.
If p is an odd prime with $\mathfrak{P}$ above it, then $\mathrm{Frob}_{\mathfrak{P}}$ is the unique element such that

$$(a+bi)^p \equiv \mathrm{Frob}_{\mathfrak{P}}(a+bi) \pmod{\mathfrak{P}}$$

in $\mathbb{Z}[i]$. In particular,

$$\mathrm{Frob}_{\mathfrak{P}}(i) = i^p = \begin{cases} i & p \equiv 1 \pmod 4 \\ -i & p \equiv 3 \pmod 4. \end{cases}$$

From this we see that $\mathrm{Frob}_{\mathfrak{P}}$ is the identity when $p \equiv 1 \pmod 4$ and $\mathrm{Frob}_{\mathfrak{P}}$ is complex conjugation when $p \equiv 3 \pmod 4$.

Example 63.5.3 (Cyclotomic Frobenius element)

Generalizing previous example, let $L = \mathbb{Q}(\zeta)$ and $K = \mathbb{Q}$, with ζ an mth root of unity. It's well-known that L/K is unramified outside ∞ and prime factors of m. Moreover, the Galois group $\mathrm{Gal}(L/K)$ is $(\mathbb{Z}/m\mathbb{Z})^\times$: the Galois group consists of elements of the form

$$\sigma_n \colon \zeta \mapsto \zeta^n$$

and $\mathrm{Gal}(L/K) = \{\sigma_n \mid n \in (\mathbb{Z}/m\mathbb{Z})^\times\}$.
Then it follows just like before that if $p \nmid m$ is prime and $\mathfrak{P}$ is above p

$$\mathrm{Frob}_{\mathfrak{P}}(x) = \sigma_p.$$

Here, as hinted in Section 61.6, we have to generalize the theory where the base field K is not necessarily $\mathbb{Q}$ (for example, in Example 63.5.8, we need $K = \mathbb{Q}(\omega)$). In this case, $\mathfrak{p}$ is not necessarily an integer, and the induced map on the quotient is the "power-by-$\mathrm{N}(\mathfrak{p})$" map.

Example 63.5.4 (Frobenius element when the base field is $\mathbb{Q}(\omega)$)

Let $L = \mathbb{Q}(\omega, \sqrt[3]{2})$ and $K = \mathbb{Q}(\omega)$.
Consider $\mathfrak{p} = (5)$, which is prime in K, and $\mathrm{N}(\mathfrak{p}) = 25$. The field $\mathcal{O}_K/\mathfrak{p}$ is isomorphic to $\mathbb{F}_{25}$. In L, $\mathfrak{p}$ splits to $\mathfrak{P}_1\mathfrak{P}_2\mathfrak{P}_3$, and each residue field $\mathcal{O}_L/\mathfrak{P}_i$ is isomorphic to $\mathbb{F}_{25}$.
The Frobenius element $\mathrm{Frob}_{\mathfrak{P}} \in \mathrm{Gal}(L/K)$ induces the power-of-25 isomorphism in the quotient field, thus is the identity.

An important property of the Frobenius element is its order is related to the decomposition of $\mathfrak{p}$ in the higher field L in the nicest way possible:

Lemma 63.5.5 (Order of the Frobenius element)

The Frobenius element $\mathrm{Frob}_{\mathfrak{P}} \in \mathrm{Gal}(L/K)$ of an extension L/K has order equal to the inertial degree of $\mathfrak{P}$, that is,

$$\operatorname{ord} \mathrm{Frob}_{\mathfrak{P}} = f(\mathfrak{P} \mid \mathfrak{p}).$$

In particular, $\mathrm{Frob}_{\mathfrak{P}} = \mathrm{id}$ if and only if $\mathfrak{p}$ splits completely in L/K.

This naturally generalizes Lemma 62.1.3.

Proof. We want to understand the order of the map $T \colon x \mapsto x^{\mathrm{N}(\mathfrak{p})}$ on the field $\mathcal{O}_L/\mathfrak{P}$. But the latter is isomorphic to the splitting field of $X^{\mathrm{N}(\mathfrak{P})} - X$ in $\mathbb{F}_p$, by Galois theory of finite fields. Hence the order is $\log_{\mathrm{N}(\mathfrak{p})}(\mathrm{N}(\mathfrak{P})) = f(\mathfrak{P} \mid \mathfrak{p})$. □

The Galois group acts transitively among the set of $\mathfrak{P}$ above a given $\mathfrak{p}$, so that we have

$$\mathrm{Frob}_{\sigma(\mathfrak{P})} = \sigma \circ (\mathrm{Frob}_{\mathfrak{P}}) \circ \sigma^{-1}.$$

Thus $\mathrm{Frob}_{\mathfrak{P}}$ is determined by its underlying $\mathfrak{p}$ up to conjugation.

In class field theory, we are interested in **abelian extensions**, i.e. those for which $\mathrm{Gal}(L/K)$ is abelian. Here the theory becomes extra nice: the conjugacy classes have size one.

Definition 63.5.6. Assume L/K is an **abelian** extension. Then for a given unramified prime $\mathfrak{p}$ in K, the element $\text{Frob}_{\mathfrak{P}}$ doesn't depend on the choice of $\mathfrak{P}$. We denote the resulting $\text{Frob}_{\mathfrak{P}}$ by the **Artin symbol**,

$$\left(\frac{L/K}{\mathfrak{p}}\right).$$

The definition of the Artin symbol is written deliberately to look like the Legendre symbol. To see why:

Example 63.5.7 (Legendre symbol subsumed by Artin symbol)
Suppose we want to understand $\left(\frac{2}{p}\right) \equiv 2^{\frac{p-1}{2}}$ where $p > 2$ is prime. Consider the element

$$\left(\frac{\mathbb{Q}(\sqrt{2})/\mathbb{Q}}{p\mathbb{Z}}\right) \in \text{Gal}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}).$$

It is uniquely determined by where it sends $\sqrt{2}$. But in fact we have

$$\left(\frac{\mathbb{Q}(\sqrt{2})/\mathbb{Q}}{p\mathbb{Z}}\right)\left(\sqrt{2}\right) \equiv \left(\sqrt{2}\right)^p \equiv 2^{\frac{p-1}{2}} \cdot \sqrt{2} \equiv \left(\frac{2}{p}\right)\sqrt{2} \pmod{\mathfrak{P}}$$

where $\left(\frac{2}{p}\right)$ is the usual Legendre symbol, and $\mathfrak{P}$ is above p in $\mathbb{Q}(\sqrt{2})$. Thus the Artin symbol generalizes the quadratic Legendre symbol.

Example 63.5.8 (Cubic Legendre symbol subsumed by Artin symbol)
Similarly, it also generalizes the cubic Legendre symbol. To see this, assume θ is a primary prime in $K = \mathbb{Q}(\sqrt{-3}) = \mathbb{Q}(\omega)$ (thus $\mathcal{O}_K = \mathbb{Z}[\omega]$ is the Eisenstein integers). Then for example

$$\left(\frac{K(\sqrt[3]{2})/K}{\theta\mathcal{O}_K}\right)\left(\sqrt[3]{2}\right) \equiv \left(\sqrt[3]{2}\right)^{\text{N}(\theta)} \equiv 2^{\frac{\text{N}(\theta)-1}{3}} \cdot \sqrt[3]{2} \equiv \left(\frac{2}{\theta}\right)_3 \sqrt[3]{2}. \pmod{\mathfrak{P}}$$

where $\mathfrak{P}$ is above (θ) in $K(\sqrt[3]{2})$.

§63.6 Artin reciprocity

Now, we further capitalize on the fact that $\text{Gal}(L/K)$ is abelian. For brevity, in what follows let $\text{Ram}(L/K)$ denote the primes of K (either finite or infinite) which ramify in L.

Definition 63.6.1. Let L/K be an abelian extension and let $\mathfrak{m}$ be divisible by every prime in $\text{Ram}(L/K)$. Then since L/K is abelian we can extend the Artin symbol multiplicatively to a map

$$\left(\frac{L/K}{\bullet}\right) : I_K(\mathfrak{m}) \twoheadrightarrow \text{Gal}(L/K).$$

This is called the **Artin map**, and it is surjective (for example by Chebotarev Density).

Let $H(L/K, \mathfrak{m}) \subseteq I_K(\mathfrak{m})$ denote the kernel of this map, so

$$\text{Gal}(L/K) \cong I_K(\mathfrak{m})/H(L/K, \mathfrak{m}).$$

We can now present the long-awaited Artin reciprocity theorem.

Theorem 63.6.2 (Artin reciprocity)
Let L/K be an abelian extension. Then there is a modulus $\mathfrak{f} = \mathfrak{f}(L/K)$, divisible by exactly the primes of $\operatorname{Ram}(L/K)$, such that: for any modulus $\mathfrak{m}$ divisible by all primes of $\operatorname{Ram}(L/K)$, we have

$$P_K(\mathfrak{m}) \subseteq H(L/K, \mathfrak{m}) \subseteq I_K(\mathfrak{m}) \quad \text{if and only if} \quad \mathfrak{f} \mid \mathfrak{m}.$$

We call $\mathfrak{f}$ the **conductor** of L/K.

So the conductor $\mathfrak{f}$ plays a similar role to the discriminant (divisible by exactly the primes which ramify), and when $\mathfrak{m}$ is divisible by the conductor, $H(L/K, \mathfrak{m})$ is a *congruence subgroup*.

Here's the reason this is called a "reciprocity" theorem. The above theorem applies on $\mathfrak{m} = \mathfrak{f}$ tells us $P_K(\mathfrak{f}) \subseteq H(L/K, \mathfrak{f})$, so the Artin map factors through the quotient map $I_K(\mathfrak{f}) \twoheadrightarrow I_K(\mathfrak{f})/P_K(\mathfrak{f})$. Recalling that $C_K(\mathfrak{f}) = I_K(\mathfrak{f})/P_K(\mathfrak{f})$, we get a sequence of maps

$$\begin{array}{ccccc} I_K(\mathfrak{f}) & \twoheadrightarrow & C_K(\mathfrak{f}) & \xrightarrow{\left(\frac{L/K}{\bullet}\right)} & \operatorname{Gal}(L/K) \\ & & & \searrow \quad \nearrow_{\cong} & \\ & & & I_K(\mathfrak{f})/H(L/K,\mathfrak{f}) & \end{array}$$

Consequently:

For primes $\mathfrak{p} \in I_K(\mathfrak{f})$, $\left(\frac{L/K}{\mathfrak{p}}\right)$ depends only on "$\mathfrak{p} \pmod{\mathfrak{f}}$".

Let's see how this result relates to quadratic reciprocity.

Example 63.6.3 (Artin reciprocity implies quadratic reciprocity)
The big miracle of quadratic reciprocity states that: for a fixed (squarefree) a, the Legendre symbol $\left(\frac{a}{p}\right)$ should only depend the residue of p modulo something. Let's see why Artin reciprocity tells us this *a priori*.
Let $L = \mathbb{Q}(\sqrt{a})$, $K = \mathbb{Q}$. Then we've already seen that the Artin symbol

$$\left(\frac{\mathbb{Q}(\sqrt{a})/\mathbb{Q}}{\bullet}\right)$$

is the correct generalization of the Legendre symbol. Thus, Artin reciprocity tells us that there is a conductor $\mathfrak{f} = \mathfrak{f}(\mathbb{Q}(\sqrt{a})/\mathbb{Q})$ such that $\left(\frac{\mathbb{Q}(\sqrt{a})/\mathbb{Q}}{p}\right)$ depends only on the residue of p modulo $\mathfrak{f}$, which is what we wanted.

Here is an example along the same lines.

Example 63.6.4 (Cyclotomic field)
Let ζ be a primitive mth root of unity. For primes p, we know that $\operatorname{Frob}_p \in$

$\mathrm{Gal}(\mathbb{Q}(\zeta)/\mathbb{Q})$ is "exactly" $p \pmod m$. Let's translate this idea into the notation of Artin reciprocity.
We are going to prove

$$H(\mathbb{Q}(\zeta)/\mathbb{Q}, m\infty) = P_{\mathbb{Q}}(m\infty) = \left\{ \frac{a}{b}\mathbb{Z} \mid a/b \equiv 1 \pmod m \right\}.$$

This is the generic example of achieving the lower bound in Artin reciprocity. It also implies that $\mathfrak{f}(\mathbb{Q}(\zeta)/\mathbb{Q}) \mid m\infty$.
It's well-known $\mathbb{Q}(\zeta)/\mathbb{Q}$ is unramified outside finite primes dividing m, so that the Artin symbol is defined on $I_K(\mathfrak{m})$. Now the Artin map is given by

$$\begin{array}{ccccc} I_{\mathbb{Q}}(\mathfrak{m}) & \xrightarrow{\left(\frac{\mathbb{Q}(\zeta)/\mathbb{Q}}{\bullet}\right)} & \mathrm{Gal}(\mathbb{Q}(\zeta)/\mathbb{Q}) & \xrightarrow{\cong} & (\mathbb{Z}/m\mathbb{Z})^\times \\ p & \longmapsto & (x \mapsto x^p) & \longmapsto & p \pmod m. \end{array}$$

So we see that the kernel of this map is trivial, i.e. it is given by the identity of the Galois group, corresponding to $1 \pmod m$. On the other hand, we've also computed $P_{\mathbb{Q}}(m\infty)$ already, so we have the desired equality.

In fact, we also have the following "existence theorem": every congruence subgroup appears uniquely once we fix $\mathfrak{m}$.

Theorem 63.6.5 (Takagi existence theorem)
Fix K and let $\mathfrak{m}$ be a modulus. Consider any congruence subgroup H, i.e.

$$P_K(\mathfrak{m}) \subseteq H \subseteq I_K(\mathfrak{m}).$$

Then $H = H(L/K, \mathfrak{m})$ for a *unique* abelian extension L/K.

Finally, such subgroups reverse inclusion in the best way possible:

Lemma 63.6.6 (Inclusion-reversing congruence subgroups)
Fix a modulus $\mathfrak{m}$. Let L/K and M/K be abelian extensions and suppose $\mathfrak{m}$ is divisible by the conductors of L/K and M/K. Then

$$L \subseteq M \quad \text{if and only if} \quad H(M/K, \mathfrak{m}) \subseteq H(L/K, \mathfrak{m}).$$

Here by $L \subseteq M$ we mean that L is isomorphic to some subfield of M.

Sketch of proof. Let us first prove the equivalence with $\mathfrak{m}$ fixed. In one direction, assume $L \subseteq M$; one can check from the definitions that the diagram

$$\begin{array}{ccc} I_K(\mathfrak{m}) & \xrightarrow{\left(\frac{M/K}{\bullet}\right)} & \mathrm{Gal}(M/K) \\ & \searrow^{\left(\frac{L/K}{\bullet}\right)} & \downarrow \\ & & \mathrm{Gal}(L/K) \end{array}$$

commutes, because it suffices to verify this for prime powers, which is just saying that Frobenius elements behave well with respect to restriction. Then the inclusion of kernels

follows directly. The reverse direction is essentially the Takagi existence theorem. □

Note that we can always take $\mathfrak{m}$ to be the product of conductors here.

If you didn't realize it: Apart from generalizing quadratic reciprocity, Artin reciprocity and Takagi existence theorem together enumerates *all abelian field extensions*! Now if you are given a field K and want to list all (finite) abelian field extensions of K, you can list all the modulus $\mathfrak{m}$ of K, list all subgroups of $C_K(\mathfrak{m})$, then each subgroup corresponds to a field extension.

(Of course, the question of how to compute the field L given a modulus and a congruence subgroup is still difficult. At least when $K = \mathbb{Q}$, Problem 63A† gives the answer: all finite abelian field extensions $L/\mathbb{Q}$ are contained in some cyclotomic field.

To finish, here is a quote from Emil Artin on his reciprocity law:

> I will tell you a story about the Reciprocity Law. After my thesis, I had the idea to define L-series for non-abelian extensions. But for them to agree with the L-series for abelian extensions, a certain isomorphism had to be true. I could show it implied all the standard reciprocity laws. So I called it the General Reciprocity Law and tried to prove it but couldn't, even after many tries. Then I showed it to the other number theorists, but they all laughed at it, and I remember Hasse in particular telling me it couldn't possibly be true.
>
> Still, I kept at it, but nothing I tried worked. Not a week went by — *for three years!* — that I did not try to prove the Reciprocity Law. It was discouraging, and meanwhile I turned to other things. Then one afternoon I had nothing special to do, so I said, 'Well, I try to prove the Reciprocity Law again.' So I went out and sat down in the garden. You see, from the very beginning I had the idea to use the cyclotomic fields, but they never worked, and now I suddenly saw that all this time I had been using them in the wrong way — and in half an hour I had it.

§63.7 Application: Generalization of sum of two squares

We start with the follow classical theorem:

Theorem 63.7.1 (Fermat's theorem on sums of two squares)
An odd prime p can be expressed as $p = x^2 + y^2$ for integers x and y if and only if $p \equiv 1 \pmod 4$.

You may see a proof that goes something like the following. Because we have learnt number theory and quadratic reciprocity, this should be intuitive to follow.

Proof. Note that if $p = x^2 + y^2$, then $\left(\frac{x}{y}\right)^2 \equiv -1 \pmod p$, so a necessary condition is that -1 is a quadratic residue modulo p.

We will show that this condition is also sufficient.

Let $a \in \mathbb{Z}$ be such that $a^2 \equiv -1 \pmod p$. Note that $\mathrm{N}_{\mathbb{Q}(i)/\mathbb{Q}}(a+i) = a^2+1$ is divisible by p, and $\mathrm{N}_{\mathbb{Q}(i)/\mathbb{Q}}(p) = p^2$.

Assume it is possible to write $p = x^2 + y^2$. Then p can be factored in $\mathbb{Z}[i]$ as $(x + yi)(x - yi)$, for integers x and y.

We claim that letting $x + yi = \gcd(p, a + i)$ works. Indeed, $p \mid (a + i)(a - i) = a^2 + 1$ but p does not divide either of the factor, which means p is not a prime in $\mathbb{Z}[i]$ and taking the gcd with either $a + i$ or $a - i$ should extract a nontrivial factor.

Note that $\mathrm{N}_{\mathbb{Q}(i)/\mathbb{Q}}(x + yi) = p$, thus $x + yi$ and $x - yi$ are already primes, so the factor extraction above must already give us a prime factor, which is what we want.

Finally, we know that -1 is a quadratic residue modulo p precisely when $p \equiv 1 \pmod 4$, so we're done. □

You may dismiss it as an arcane trick... until you realize that it can be generalized perfectly well to many other cases! Try to prove the following theorem using the same method.

Theorem 63.7.2

An odd prime $p > 7$ can be expressed as $p = x^2 + 7y^2$ for integers x and y if and only if -7 is a quadratic residue modulo p.

Which, by quadratic reciprocity, would boil down to whether $(p \bmod 7) \in \{1, 2, 4\}$.

Nevertheless, it isn't always that nice.

Example 63.7.3

Let $p = 3$. Then $1^2 \equiv -5 \pmod p$, but there is no integers x and y such that $p = x^2 + 5y^2$.

Question 63.7.4. If you haven't, try to figure out what went wrong in the proof before reading the explanation below.

The bug, of course, is to assume that $\gcd(p, 1 + \sqrt{-5})$ is an element — that is, in this case, the ring of integers of $\mathbb{Q}(\sqrt{-5})$ is not an unique factorization domain. But we have all the tools of ideal theory to fix it: the ideal $(p) = (3) \subseteq \mathbb{Q}$ splits into $(p) = \mathfrak{p}_1\mathfrak{p}_2$ when lifted to $\mathbb{Q}(\sqrt{-5})$, where $\mathfrak{p}_1 = (3, 1 + \sqrt{-5})$ and $\mathfrak{p}_2 = (3, 1 - \sqrt{-5})$.

Thus,

Proposition 63.7.5

A prime $p \in \mathbb{Q}$ can be written as $p = x^2 + 5y^2$ if and only if (p) splits into $\mathfrak{p}_1\mathfrak{p}_2$ when lifted to $\mathbb{Q}(\sqrt{-5})$, where both $\mathfrak{p}_1$ and $\mathfrak{p}_2$ are principal ideals.

This is where Artin reciprocity and the Hilbert class field shines — we want to determine the class of $\mathfrak{p}_1$, in other words, $\mathfrak{p}_1 \pmod 1$.

Question 63.7.6. Check that $\mathfrak{p} \equiv (1) \pmod 1$ if and only if $\mathfrak{p} \subseteq \mathbb{Q}(\sqrt{-5})$ is principal. (Definition chasing.)

Question 63.7.7. If $\mathfrak{p}_1$ is principal, then we automatically have $\mathfrak{p}_2$ principal. Why?

From now on, let $K = \mathbb{Q}(\sqrt{-5})$, and let L be some abelian extension of K.

Recall we defined above the group $H(L/K, \mathfrak{m}) = \ker\left(\frac{L/K}{\bullet}\right)$, and the statement of Artin reciprocity claims, among others, that $P_K(\mathfrak{m}) \subseteq H(L/K, \mathfrak{m})$. Naturally, you may wonder, if all we cares is that "$\left(\frac{L/K}{\mathfrak{p}}\right)$ depends only on $\mathfrak{p} \pmod{\mathfrak{f}}$", then why would we need to define yet another piece of notation for H?

Well, the simplified version of Artin reciprocity theorem above states that we can compute $\left(\frac{L/K}{\mathfrak{p}}\right)$ once we know $\mathfrak{p} \pmod{\mathfrak{f}}$. Of course there is more than that:

If $P_K(\mathfrak{f}) = H(L/K, \mathfrak{f})$, then we can compute $\mathfrak{p} \pmod{\mathfrak{f}}$ once we know $\left(\frac{L/K}{\mathfrak{p}}\right)$.

In other words, if L is such that the congruence subgroup reaches the "lower bound", then we also get the converse.

Question 63.7.8. Check that the algebra above works out.

We have seen one example above, Example 63.6.4, where the congruence subgroup $H(\mathbb{Q}(\zeta_m)/\mathbb{Q}, m\infty)$ is equal to the lower bound $P_{\mathbb{Q}}(m\infty)$. We will see one more example below.

Example 63.7.9

In the example above, we can vary both the modulus $\mathfrak{m}$ and the abelian field extension L over K to get different congruence subgroups. This can be confusing, so let us take an example.

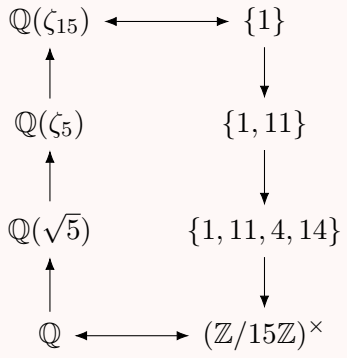
Consider abelian field extensions $L/\mathbb{Q}$. Let the modulus in $\mathbb{Q}$ be $\mathfrak{m} = 15\infty$.

The ray class group $C_K(\mathfrak{m})$ is of course isomorphic to $(\mathbb{Z}/15\mathbb{Z})^\times \cong (\mathbb{Z}/3\mathbb{Z})^\times \times (\mathbb{Z}/5\mathbb{Z})^\times \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$.

As small as this group is (with only 8 elements), it has 8 subgroups.[a] Nevertheless, we will only focus on the relevant parts of the subgroup lattice.

By Artin reciprocity and Takagi existence theorem, each congruence subgroup corresponds to some abelian extension over $L/\mathbb{Q}$.

We draw the correspondence between abelian field extension and the congruence subgroup $H(L/\mathbb{Q}, 15\infty)$ below, depicted using the fact that $H(L/\mathbb{Q}, 15\infty)/P_{\mathbb{Q}}(15\infty)$ is a subgroup of $C_{\mathbb{Q}}(15\infty)$, which is canonically isomorphic to $(\mathbb{Z}/15\mathbb{Z})^\times$.



(Where does the diagram above come from? Well, if the base field is $\mathbb{Q}$, Problem 63A† gives a way.)

Interested readers may want to try to work out the canonical isomorphism between the Galois group $\mathrm{Gal}(L/K)$ and the ray class group $C_K(\mathfrak{f}(L/K))$ in the general

case of an abelian extension.
Next, how does this relate to the abelian extensions that corresponds to different modulus, let's say 5∞? Intuitively speaking, if we know the value of an ideal mod 15∞, we would know its value mod 5∞. Formally, we have this diagram:

$$\begin{array}{ccccc} P_K(15\infty) & \hookrightarrow & I_K(15\infty) & \twoheadrightarrow & C_K(15\infty) \\ \downarrow & & \downarrow & & \downarrow \\ P_K(5\infty) & \hookrightarrow & I_K(5\infty) & \twoheadrightarrow & C_K(5\infty) \end{array}$$

(If you have read the category theory chapter: Morphism of short exact sequence appears everywhere! You just have to look for it.)
That is, we get an induced $C_K(15\infty) \twoheadrightarrow C_K(5\infty)$ map, or equivalently, $(\mathbb{Z}/15\mathbb{Z})^\times \twoheadrightarrow (\mathbb{Z}/5\mathbb{Z})^\times$. This time around, the abelian field extensions that corresponds to the modulus 5∞ are:

$$\begin{array}{ccc} \mathbb{Q}(\zeta_5) & \longleftrightarrow & \{1\} \\ \uparrow & & \downarrow \\ \mathbb{Q}(\sqrt{5}) & & \{1,4\} \\ \uparrow & & \downarrow \\ \mathbb{Q} & \longleftrightarrow & (\mathbb{Z}/5\mathbb{Z})^\times \end{array}$$

[a] https://beta.lmfdb.org/Groups/Abstract/diagram/8.2 has a diagram.

In our case, given $p \in \mathbb{Q}$ be a prime factors as $(p) = \mathfrak{p}_1\mathfrak{p}_2$ when lifted to $K = \mathbb{Q}(\sqrt{-5})$, we want to determine if $\mathfrak{p}_1$ is principal — in other words, we want to compute "$\mathfrak{p}_1 \pmod 1$". With the insight above, we will rephrase the condition in terms of the Artin symbol.

Let $L = K(i)$. (Later on, we will know that L is the **Hilbert class field** of K.) We claim the following is true:

- L/K is an abelian extension,
- the discriminant is $\mathfrak{f} = \mathfrak{f}(L/K) = 1$,
- $H(L/K, \mathfrak{f}) = P_K(\mathfrak{f})$ — that is, this is exactly the situation where we can determine $\mathfrak{p} \pmod 1$ for $\mathfrak{p} \subseteq K$ based on $\left(\frac{L/K}{\mathfrak{p}}\right)$.

(In the general case, the field L exists according to Problem 63B†.)
Then, for a prime $\mathfrak{p} \subseteq K$, the following are equivalent:

1. $\mathfrak{p}$ is principal;
2. $\left(\frac{L/K}{\mathfrak{p}}\right) = \mathrm{id}$;
3. $\mathfrak{p}$ splits completely when lifted to L.

Notice that we used Artin reciprocity (and its "converse") for the abelian extension L/K to prove the equivalence of the first and the second statement.

Exercise 63.7.10. Why is the second and the third statement equivalent? (Problem 63B†.)

Thus, the condition that $(p) = \mathfrak{p}_1\mathfrak{p}_2$ for principal ideals $\mathfrak{p}_1$ and $\mathfrak{p}_2$ is equivalent to that $(p) \subseteq \mathbb{Q}$ splits completely when lifted to L.

Reasoning similar to above for the abelian extension $L/\mathbb{Q}$, the following are equivalent:

1. $(p) \subseteq \mathbb{Q}$ splits completely when lifted to L;
2. $\left(\frac{L/\mathbb{Q}}{(p)}\right) = \mathrm{id}$.

This time, we don't have the first bullet point anymore — L is *not* the Hilbert class field of $\mathbb{Q}$ — but, by Artin reciprocity, we do know:

The value of $\left(\frac{L/\mathbb{Q}}{(p)}\right)$ only depends on $(p) \pmod{\mathfrak{f}(L/\mathbb{Q})}$.

In this case, the discriminant of the extension $L/\mathbb{Q}$ is $\mathfrak{f}(L/\mathbb{Q}) = 20\infty$.

So, in summary:

$$\begin{aligned}
&p \text{ can be written as } x^2 + 5y^2 \\
\iff &(p) = \mathfrak{p}_1\mathfrak{p}_2 \text{ for principal } \mathfrak{p}_1 \text{ when lifted to } \mathbb{Q}(\sqrt{-5}) \\
\iff &(p) = \mathfrak{p}_1\mathfrak{p}_2, \text{ and } \mathfrak{p}_1 \subseteq \mathbb{Q}(\sqrt{-5}) \text{ splits completely when lifted to } \mathbb{Q}(\sqrt{-5}, i) \\
\iff &(p) \subseteq \mathbb{Q} \text{ splits completely when lifted to } \mathbb{Q}(\sqrt{-5}, i) \\
\iff &\left(\frac{\mathbb{Q}(\sqrt{-5}, i)/\mathbb{Q}}{(p)}\right) = \mathrm{id} \\
\iff &(p \bmod 20) \in \{1, 9\}.
\end{aligned}$$

We're done! The final form of the theorem is:

Theorem 63.7.11

Let p be a prime with $p \nmid 20$, then p can be written as $x^2 + 5y^2$ if and only if $(p \bmod 20) \in \{1, 9\}$.

§63.8 A few harder problems to think about

Problem 63A†. [Kronecker-Weber theorem] Let L be an abelian extension of $\mathbb{Q}$. Then L is contained in a cyclic extension $\mathbb{Q}(\zeta)$ where ζ is an mth root of unity (for some m).

Problem 63B† (Hilbert class field)**.** Let K be any number field. Then there exists a unique abelian extension E/K which is unramified at all primes (finite or infinite) and such that

- E/K is the maximal such extension by inclusion.
- $\mathrm{Gal}(E/K)$ is isomorphic to the class group of K.
- A prime $\mathfrak{p}$ of K splits completely in E if and only if it is a principal ideal of $\mathcal{O}_K$.

We call E the **Hilbert class field** of K.

Problem 63C. There is no positive integer m such that whether a prime number $p \nmid m$ can be written as $p = x^2 + 23y^2$ depends only on $p \bmod m$. Guess why.

XVI

Algebraic Topology I: Homotopy

Part XVI: Contents

64 Some topological constructions

In this short chapter we briefly describe some common spaces and constructions in topology that we haven't yet discussed.

§64.1 Spheres

Recall that

$$S^n = \left\{ (x_0, \dots, x_n) \mid x_0^2 + \cdots + x_n^2 = 1 \right\} \subset \mathbb{R}^{n+1}$$

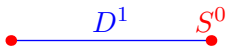
is the surface of an n-sphere while

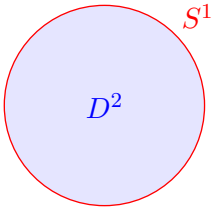
$$D^{n+1} = \left\{ (x_0, \dots, x_n) \mid x_0^2 + \cdots + x_n^2 \le 1 \right\} \subset \mathbb{R}^{n+1}$$

is the corresponding *closed ball* (So for example, D^2 is a disk in a plane while S^1 is the unit circle.)

Exercise 64.1.1. Show that the open ball $D^n \setminus S^{n-1}$ is homeomorphic to $\mathbb{R}^n$.

In particular, S^0 consists of two points, while D^1 can be thought of as the interval $[-1, 1]$.





§64.2 Quotient topology

Prototypical example for this section: $D^n/S^{n-1} = S^n$, *or the torus.*

Given a space X, we can *identify* some of the points together by any equivalence relation $\sim$; for an $x \in X$ we denote its equivalence class by $[x]$. Geometrically, this is the space achieved by welding together points equivalent under $\sim$.

Formally,

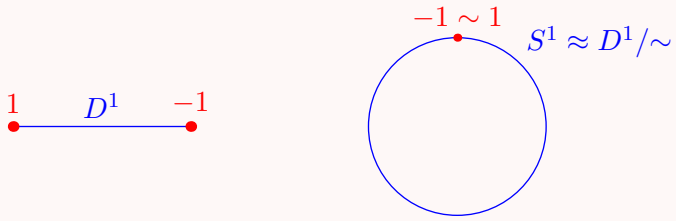
Definition 64.2.1. Let X be a topological space, and $\sim$ an equivalence relation on the points of X. Then $X/\sim$ is the space whose

- Points are equivalence classes of X, and
- $U \subseteq X/\sim$ is open if and only if $\{x \in X \text{ such that } [x] \in U\}$ is open in X.

As far as I can tell, this definition is mostly useless for intuition, so here are some examples.

Example 64.2.2 (Interval modulo endpoints)

Suppose we take $D^1 = [-1, 1]$ and quotient by the equivalence relation which identifies the endpoints -1 and 1. (Formally, $x \sim y \iff (x = y)$ or $\{x, y\} = \{-1, 1\}$.) In that case, we simply recover S^1:



Observe that a small open neighborhood around $-1 \sim 1$ in the quotient space corresponds to two half-intervals at -1 and 1 in the original space D^1. This should convince you the definition we gave is the right one.

Example 64.2.3 (More quotient spaces)

Convince yourself that:

- Generalizing the previous example, D^n modulo its boundary S^{n-1} is S^n.
- Given a square $ABCD$, suppose we identify segments AB and DC together. Then we get a cylinder. (Think elementary school, when you would tape up pieces of paper together to get cylinders.)
- In the previous example, if we also identify BC and DA together, then we get a torus. (Imagine taking our cylinder and putting the two circles at the end together.)
- Let $X = \mathbb{R}$, and let $x \sim y$ if $y - x \in \mathbb{Z}$. Then $X/\sim$ is S^1 as well.

One special case that we did above:

Definition 64.2.4. Let $A \subseteq X$. Consider the equivalence relation which identifies all the points of A with each other while leaving all remaining points inequivalent. (In other words, $x \sim y$ if $x = y$ or $x, y \in A$.) Then the resulting quotient space is denoted X/A.

So in this notation,

$$D^n/S^{n-1} = S^n.$$

Abuse of Notation 64.2.5. Note that I'm deliberately being sloppy, and saying "$D^n/S^{n-1} = S^n$" or "D^n/S^{n-1} *is* S^n", when I really ought to say "D^n/S^{n-1} is homeomorphic to S^n". This is a general theme in mathematics: objects which are homoeomorphic/isomorphic/etc. are generally not carefully distinguished from each other.

Example 64.2.6 (Weirder quotient spaces)

If the subset A is not closed in X, X/A would be quite weird.
For instance, let $X = \mathbb{R}$ and $A = (0, 1)$. Then the space X/A consists of the points $(-\infty, 0] \cup \{A/A\} \cup [1, \infty)$. Here, the points 0 and A/A are different; yet every open

set that contains 0, also contains A/A.
We say this space X/A is not Hausdorff.

§64.3 Product topology

Prototypical example for this section: $\mathbb{R} \times \mathbb{R}$ *is* $\mathbb{R}^2$, $S^1 \times S^1$ *is the torus.*

Definition 64.3.1. Given topological spaces X and Y, the **product topology** on $X \times Y$ is the space whose

- Points are pairs (x, y) with $x \in X$, $y \in Y$, and
- Topology is given as follows: the *basis* of the topology for $X \times Y$ is $U \times V$, for $U \subseteq X$ open and $V \subseteq Y$ open.

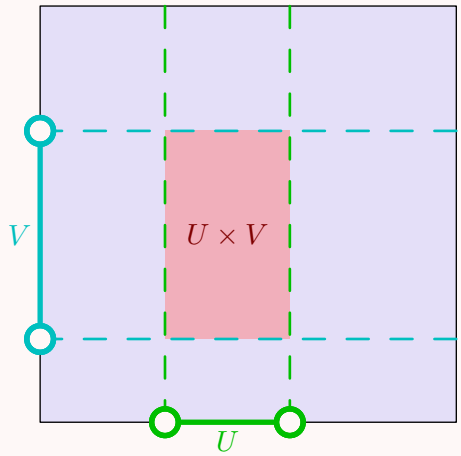
Remark 64.3.2 — It is not hard to show that, in fact, one need only consider basis elements for U and V. That is to say,

$$\{U \times V \mid U, V \text{ basis elements for } X, Y\}$$

is also a basis for $X \times Y$.
We really do need to fiddle with the basis: in $\mathbb{R} \times \mathbb{R}$, an open unit disk better be open, despite not being of the form $U \times V$.

This does exactly what you think it would.

Example 64.3.3 (The unit square)
Let $X = [0, 1]$ and consider $X \times X$. We of course expect this to be the unit square. Pictured below is an open set of $X \times X$ in the basis.



Exercise 64.3.4. Convince yourself this basis gives the same topology as the product metric on $X \times X$. So this is the "right" definition.

Example 64.3.5 (More product spaces)
(a) $\mathbb{R} \times \mathbb{R}$ is the Euclidean plane.
(b) $S^1 \times [0,1]$ is a cylinder.
(c) $S^1 \times S^1$ is a torus! (Why?)

§64.4 Disjoint union and wedge sum

Prototypical example for this section: $S^1 \vee S^1$ *is the figure eight.*

The disjoint union of two spaces is geometrically exactly what it sounds like: you just imagine the two spaces side by side. For completeness, here is the formal definition.

Definition 64.4.1. Let X and Y be two topological spaces. The **disjoint union**, denoted $X \amalg Y$, is defined by

- The points are the disjoint union $X \amalg Y$, and
- A subset $U \subseteq X \amalg Y$ is open if and only if $U \cap X$ and $U \cap Y$ are open.

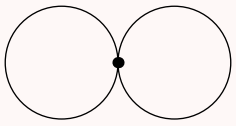
Exercise 64.4.2. Show that the disjoint union of two nonempty spaces is disconnected.

More interesting is the wedge sum, where two topological spaces X and Y are fused together only at a single base point.

Definition 64.4.3. Let X and Y be topological spaces, and $x_0 \in X$ and $y_0 \in Y$ be points. We define the equivalence relation $\sim$ by declaring $x_0 \sim y_0$ only. Then the **wedge sum** of two spaces is defined as

$$X \vee Y = (X \amalg Y)/\sim.$$

Example 64.4.4 ($S^1 \vee S^1$ is a figure eight)
Let $X = S^1$ and $Y = S^1$, and let $x_0 \in X$ and $y_0 \in Y$ be any points. Then $X \vee Y$ is a "figure eight": it is two circles fused together at one point.



Abuse of Notation 64.4.5. We often don't mention x_0 and y_0 when they are understood (or irrelevant). For example, from now on we will just write $S^1 \vee S^1$ for a figure eight.

Remark 64.4.6 — Annoyingly, in LaTeX `\wedge` gives $\wedge$ instead of $\vee$ (which is `\vee`). So this really should be called the "vee product", but too late.

§64.5 CW complexes

Using this construction, we can start building some spaces. One common way to do so is using a so-called **CW complex**. Intuitively, a CW complex is built as follows:

- Start with a set of points X^0.
- Define X^1 by taking some line segments (copies of D^1) and fusing the endpoints (copies of S^0) onto X^0.
- Define X^2 by taking copies of D^2 (a disk) and welding its boundary (a copy of S^1) onto X^1.
- Repeat inductively up until a finite stage n; we say X is **n-dimensional**.

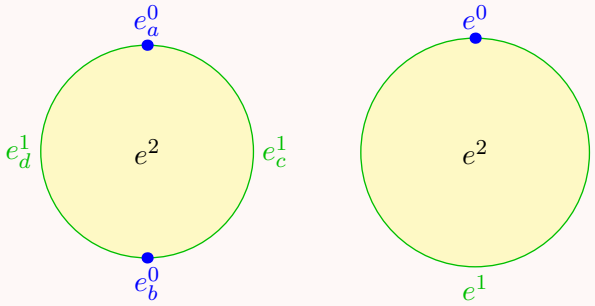
The resulting space X is the CW-complex. The set X^k is called the **k-skeleton** of X. Each D^k is called a **k-cell**; it is customary to denote it by e^k_α where α is some index. We say that X is **finite** if only finitely many cells were used.

Abuse of Notation 64.5.1. Technically, most sources (like [Ha02]) allow one to construct infinite-dimensional CW complexes. We will not encounter any such spaces in the Napkin.

Example 64.5.2 (D^2 with $2+2+1$ and $1+1+1$ cells)

(a) First, we start with X^0 having two points e^0_a and e^0_b. Then, we join them with two 1-cells D^1 (green), call them e^1_c and e^1_d. The endpoints of each 1-cell (the copy of S^0) get identified with distinct points of X^0; hence $X^1 \cong S^1$. Finally, we take a single 2-cell e^2 (yellow) and weld it in, with its boundary fitting into the copy of S^1 that we just drew. This gives the figure on the left.

(b) In fact, one can do this using just $1+1+1=3$ cells. Start with X^0 having a single point e^0. Then, use a single 1-cell e^1, fusing its two endpoints into the single point of X^0. Then, one can fit in a copy of S^1 as before, giving D^2 as on the right.



Example 64.5.3 (S^n as a CW complex)

(a) One can obtain S^n (for $n \geq 1$) with just two cells. Namely, take a single point e^0 for X^0, and to obtain S^n take D^n and weld its entire boundary into e^0.

We already saw this example in the beginning with $n=2$, when we saw that the sphere S^2 was the result when we fuse the boundary of a disk D^2 together.

(b) Alternatively, one can do a "hemisphere" construction, by constructing S^n inductively using two cells in each dimension. So S^0 consists of two points, then S^1 is obtained by joining these two points by two segments (1-cells), and S^2 is obtained by gluing two hemispheres (each a 2-cell) with S^1 as its equator.

Definition 64.5.4. Formally, for each k-cell e_α^k we want to add to X^k, we take its boundary S_α^{k-1} and weld it onto X^{k-1} via an **attaching map** $S_\alpha^{k-1} \to X^{k-1}$. Then

$$X^k = \left(X^{k-1} \amalg \left(\coprod_\alpha e_\alpha^k \right) \right) / \sim$$

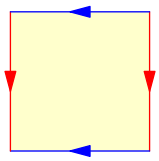
where $\sim$ identifies each boundary point of e_α^k with its image in X^{k-1}.

§64.6 The torus, Klein bottle, $\mathbb{RP}^n$, $\mathbb{CP}^n$

We now present four of the most important examples of CW complexes.

§64.6.i The torus

The **torus** can be formed by taking a square and identifying the opposite edges in the same direction: if you walk off the right edge, you re-appear at the corresponding point in on the left edge. (Think *Asteroids* from Atari!)



Thus the torus is $(\mathbb{R}/\mathbb{Z})^2 \cong S^1 \times S^1$.

Note that all four corners get identified together to a single point. One can realize the torus in 3-space by treating the square as a sheet of paper, taping together the left and right (red) edges to form a cylinder, then bending the cylinder and fusing the top and bottom (blue) edges to form the torus.

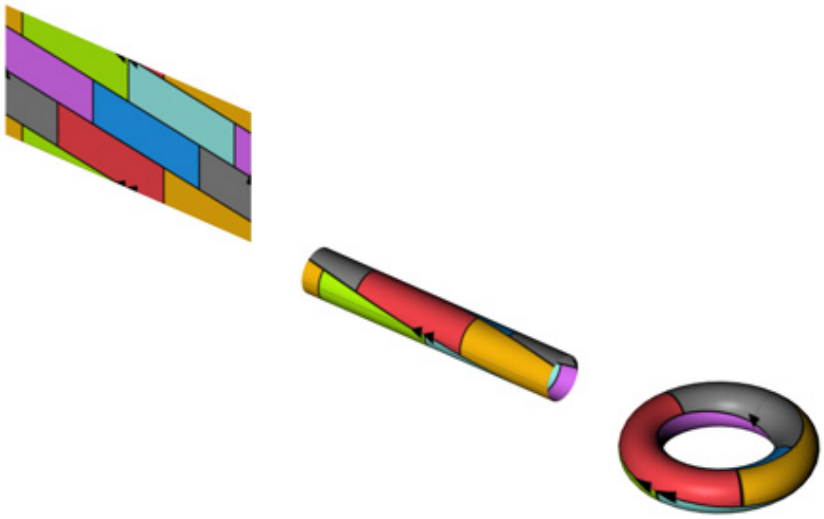
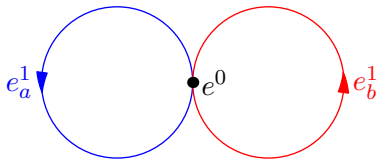


Image from [To]

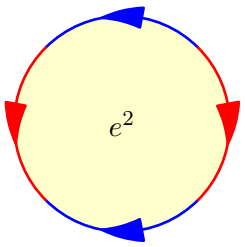
The torus can be realized as a CW complex with

- A 0-skeleton consisting of a single point,

- A 1-skeleton consisting of two 1-cells e_a^1, e_b^1, and



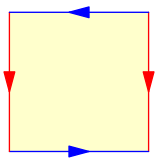
- A 2-skeleton with a single 2-cell e^2, whose circumference is divided into four parts, and welded onto the 1-skeleton "via $aba^{-1}b^{-1}$". This means: wrap a quarter of the circumference around e_a^1, then another quarter around e_b^1, then the third quarter around e_a^1 but in the opposite direction, and the fourth quarter around e_b^1 again in the opposite direction as before.



We say that $aba^{-1}b^{-1}$ is the **attaching word**; this shorthand will be convenient later on.

§64.6.ii The Klein bottle

The **Klein bottle** is defined similarly to the torus, except one pair of edges is identified in the opposite manner, as shown.



Unlike the torus one cannot realize this in 3-space without self-intersecting. One can tape together the red edges as before to get a cylinder, but to then fuse the resulting blue circles in opposite directions is not possible in 3D. Nevertheless, we often draw a picture in 3-dimensional space in which we tacitly allow the cylinder to intersect itself.

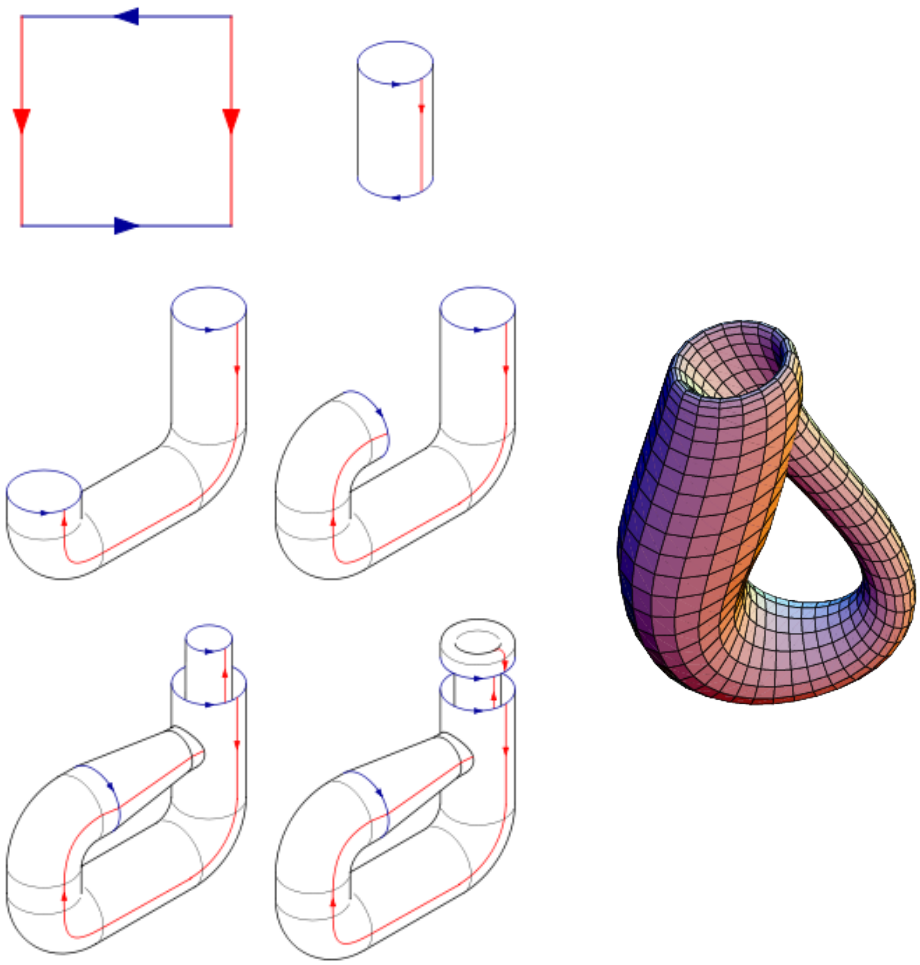


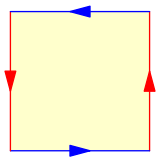
Image from [In; Fr]

Like the torus, the Klein bottle is realized as a CW complex with

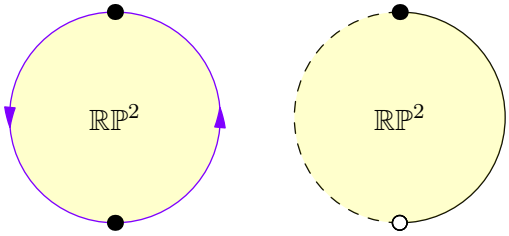
- One 0-cell,
- Two 1-cells e_a^1 and e_b^1, and
- A single 2-cell attached this time via the word $abab^{-1}$.

§64.6.iii Real projective space

Let's start with $n = 2$. The space $\mathbb{RP}^2$ is obtained if we reverse both directions of the square from before, as shown.



However, once we do this the fact that the original polygon is a square is kind of irrelevant; we can combine a red and blue edge to get the single purple edge. Equivalently, one can think of this as a circle with half its circumference identified with the other half:



The resulting space should be familiar to those of you who do projective (Euclidean) geometry. Indeed, there are several possible geometric interpretations:

- One can think of $\mathbb{RP}^2$ as the set of lines through the origin in $\mathbb{R}^3$, with each line being a point in $\mathbb{RP}^2$.

  Of course, we can identify each line with a point on the unit sphere S^2, except for the property that two antipodal points actually correspond to the same line, so that $\mathbb{RP}^2$ can be almost thought of as "half a sphere". Flattening it gives the picture above.

- Imagine $\mathbb{R}^2$, except augmented with "points at infinity". This means that we add some points "infinitely far away", one for each possible direction of a line. Thus in $\mathbb{RP}^2$, any two lines indeed intersect (at a Euclidean point if they are not parallel, and at a point at infinity if they do).

  This gives an interpretation of $\mathbb{RP}^2$, where the boundary represents the *line at infinity* through all of the points at infinity. Here we have used the fact that $\mathbb{R}^2$ and interior of D^2 are homeomorphic.

Exercise 64.6.1. Observe that these formulations are equivalent by considering the plane $z = 1$ in $\mathbb{R}^3$, and intersecting each line in the first formulation with this plane.

We can also express $\mathbb{RP}^2$ using coordinates: it is the set of triples $(x : y : z)$ of real numbers not all zero up to scaling, meaning that

$$(x : y : z) = (\lambda x : \lambda y : \lambda z)$$

for any $\lambda \neq 0$. Using the "lines through the origin in $\mathbb{R}^3$" interpretation makes it clear why this coordinate system gives the right space. The points at infinity are those with $z = 0$, and any point with $z \neq 0$ gives a Cartesian point since

$$(x : y : z) = \left(\frac{x}{z} : \frac{y}{z} : 1\right)$$

hence we can think of it as the Cartesian point $(\frac{x}{z}, \frac{y}{z})$.

In this way we can actually define **real-projective n-space**, $\mathbb{RP}^n$ for any n, as either

(i) The set of lines through the origin in $\mathbb{R}^{n+1}$,

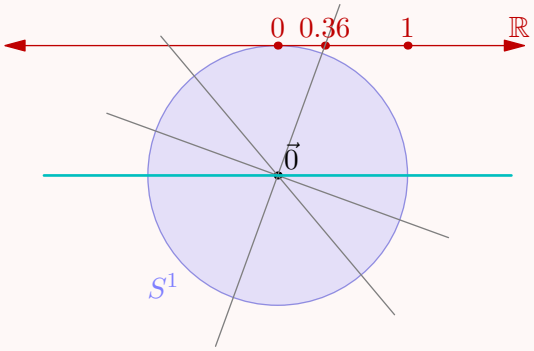
(ii) Using $n + 1$ coordinates as above, or

(iii) As $\mathbb{R}^n$ augmented with points at infinity, which themselves form a copy of $\mathbb{RP}^{n-1}$.

As a possibly helpful example, we give all three pictures of $\mathbb{RP}^1$.

Example 64.6.2 (Real projective 1-Space)

$\mathbb{RP}^1$ can be thought of as S^1 modulo the relation the antipodal points are identified. Projecting onto a tangent line, we see that we get a copy of $\mathbb{R}$ plus a single point at infinity, corresponding to the parallel line (drawn in cyan below).



Thus, the points of $\mathbb{RP}^1$ have two forms:

- $(x : 1)$, which we think of as $x \in \mathbb{R}$ (in dark red above), and
- $(1 : 0)$, which we think of as $1/0 = \infty$, corresponding to the cyan line above.

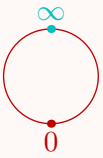
So, we can literally write

$$\mathbb{RP}^1 = \mathbb{R} \cup \{\infty\}.$$

Note that $\mathbb{RP}^1$ is also the boundary of $\mathbb{RP}^2$. In fact, note also that topologically we have

$$\mathbb{RP}^1 \cong S^1$$

since it is the "real line with endpoints fused together".



Since $\mathbb{RP}^n$ is just "$\mathbb{R}^n$ (or D^n) with $\mathbb{RP}^{n-1}$ as its boundary", we can construct $\mathbb{RP}^n$ as a CW complex inductively. Note that $\mathbb{RP}^n$ thus consists of **one cell in each dimension**.

Example 64.6.3 ($\mathbb{RP}^n$ as a cell complex)

(a) $\mathbb{RP}^0$ is a single point.

(b) $\mathbb{RP}^1 \cong S^1$ is a circle, which as a CW complex is a 0-cell plus a 1-cell.

(c) $\mathbb{RP}^2$ can be formed by taking a 2-cell and wrapping its perimeter twice around a copy of $\mathbb{RP}^1$.

§64.6.iv Complex projective space

The **complex projective space** $\mathbb{CP}^n$ is defined like $\mathbb{RP}^n$ with coordinates, i.e.

$$(z_0 : z_1 : \cdots : z_n)$$

under scaling; this time z_i are complex. As before, $\mathbb{CP}^n$ can be thought of as $\mathbb{C}^n$ augmented with some points at infinity (corresponding to $\mathbb{CP}^{n-1}$).

Example 64.6.4 (Complex projective space)

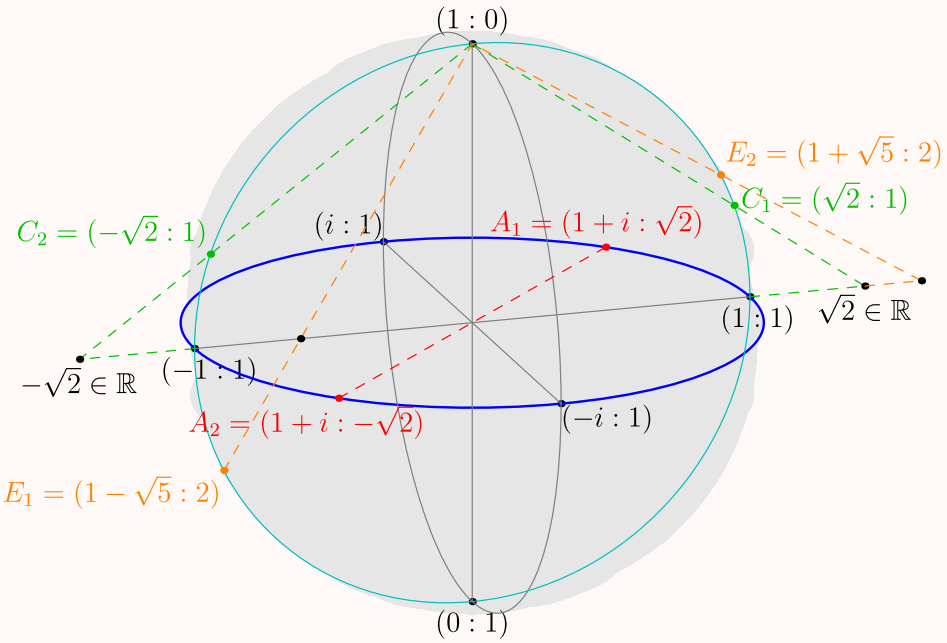
(a) $\mathbb{CP}^0$ is a single point.

(b) $\mathbb{CP}^1$ is $\mathbb{C}$ plus a single point at infinity ("complex infinity" if you will). That means as before we can think of $\mathbb{CP}^1$ as

$$\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}.$$

So, imagine taking the complex plane and then adding a single point to encompass the entire boundary. The result is just sphere S^2.

Here is a picture of $\mathbb{CP}^1$ with its coordinate system, the **Riemann sphere**.



Remark 64.6.5 (For Euclidean geometers) — You may recognize that while $\mathbb{RP}^2$ is the setting for projective geometry, inversion about a circle is done in $\mathbb{CP}^1$ instead. When one does an inversion sending generalized circles to generalized circles, there is only one point at infinity: this is why we work in $\mathbb{CP}^n$.

Like $\mathbb{RP}^n$, $\mathbb{CP}^n$ is a CW complex, built inductively by taking $\mathbb{C}^n$ and welding its boundary onto $\mathbb{CP}^{n-1}$ The difference is that as topological spaces,

$$\mathbb{C}^n \cong \mathbb{R}^{2n} \cong D^{2n}.$$

Thus, we attach the cells D^0, D^2, D^4 and so on inductively to construct $\mathbb{CP}^n$. Thus we see that

$\mathbb{CP}^n$ consists of one cell in each *even* dimension.

§64.7 A few harder problems to think about

Problem 64A. Show that a space X is Hausdorff if and only if the diagonal $\{(x,x) \mid x \in X\}$ is closed in the product space $X \times X$.

Problem 64B. Realize the following spaces as CW complexes:

(a) Möbius strip.

(b) $\mathbb{R}$.

(c) $\mathbb{R}^n$.

Problem 64C†. Show that a finite CW complex is compact.

65 Fundamental groups

Topologists can't tell the difference between a coffee cup and a doughnut. So how do you tell *anything* apart?

This is a very hard question to answer, but one way we can try to answer it is to find some *invariants* of the space. To draw on the group analogy, two groups are clearly not isomorphic if, say, they have different orders, or if one is simple and the other isn't, etc. We'd like to find some similar properties for topological spaces so that we can actually tell them apart.

Two such invariants for a space X are

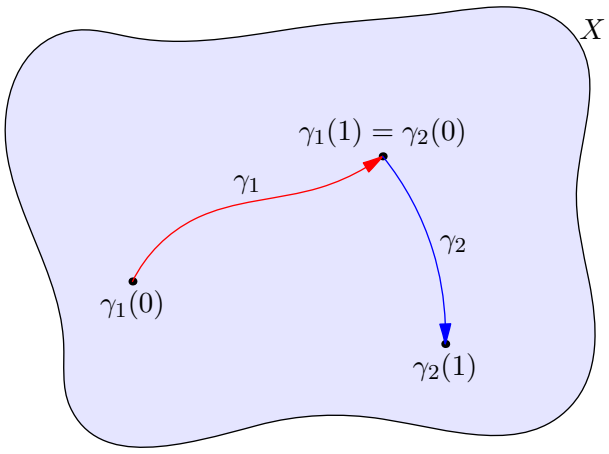
- Defining homology groups $H_1(X)$, $H_2(X)$, ...
- Defining homotopy groups $\pi_1(X)$, $\pi_2(X)$, ...

Homology groups are hard to define, but in general easier to compute. Homotopy groups are easier to define but harder to compute.

This chapter is about the fundamental group π_1.

§65.1 Fusing paths together

Recall that a *path* in a space X is a function $[0,1] \to X$. Suppose we have paths γ_1 and γ_2 such that $\gamma_1(1) = \gamma_2(0)$. We'd like to fuse[1] them together to get a path $\gamma_1 * \gamma_2$. Easy, right?



We unfortunately do have to hack the definition a tiny bit. In an ideal world, we'd have a path $\gamma_1 \colon [0,1] \to X$ and $\gamma_2 \colon [1,2] \to X$ and we could just merge them together to get $\gamma_1 * \gamma_2 \colon [0,2] \to X$. But the "2" is wrong here. The solution is that we allocate $[0, \frac{1}{2}]$ for the first path and $[\frac{1}{2}, 1]$ for the second path; we run "twice as fast".

[1] Almost everyone else in the world uses "gluing" to describe this and other types of constructs. But I was traumatized by Elmer's glue when I was in high school because I hated the stupid "make a poster" projects and hated having to use glue on them. So I refuse to talk about "gluing" paths together, referring instead to "fusing" them together, which sounds cooler anyways.

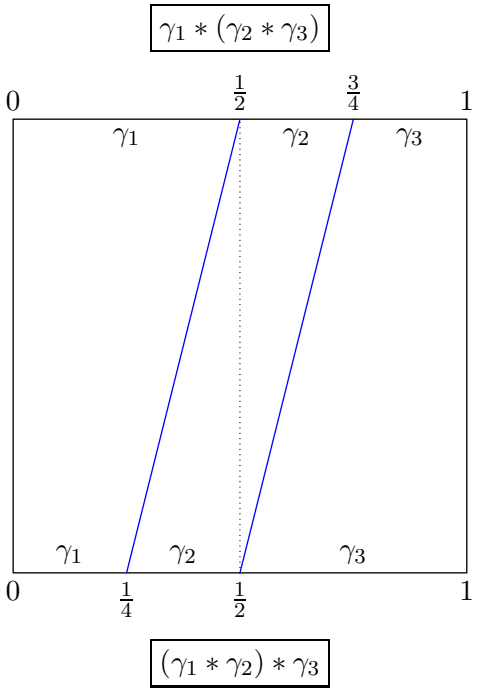
Definition 65.1.1. Given two paths $\gamma_1, \gamma_2 \colon [0,1] \to X$ such that $\gamma_1(1) = \gamma_2(0)$, we define a path $\gamma_1 * \gamma_2 \colon [0,1] \to X$ by

$$(\gamma_1 * \gamma_2)(t) = \begin{cases} \gamma_1(2t) & 0 \le t \le \frac{1}{2} \\ \gamma_2(2t-1) & \frac{1}{2} \le t \le 1. \end{cases}$$

This hack unfortunately reveals a second shortcoming: this "product" is not associative. If we take $(\gamma_1 * \gamma_2) * \gamma_3$ for some suitable paths, then $[0, \frac{1}{4}]$, $[\frac{1}{4}, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$ are the times allocated for γ_1, γ_2, γ_3.

> **Question 65.1.2.** What are the times allocated for $\gamma_1 * (\gamma_2 * \gamma_3)$?

But I hope you'll agree that even though this operation isn't associative, the reason it fails to be associative is kind of stupid. It's just a matter of how fast we run in certain parts.



So as long as we're fusing paths together, we probably don't want to think of $[0,1]$ itself too seriously. And so we only consider everything up to (path) homotopy equivalence. (Recall that two paths α and β are homotopic if there's a path homotopy $F \colon [0,1]^2 \to X$ between them, which is a continuous deformation from α to β.) It is definitely true that

$$(\gamma_1 * \gamma_2) * \gamma_3 \simeq \gamma_1 * (\gamma_2 * \gamma_3).$$

It is also true that if $\alpha_1 \simeq \alpha_2$ and $\beta_1 \simeq \beta_2$ then $\alpha_1 * \beta_1 \simeq \alpha_2 * \beta_2$.

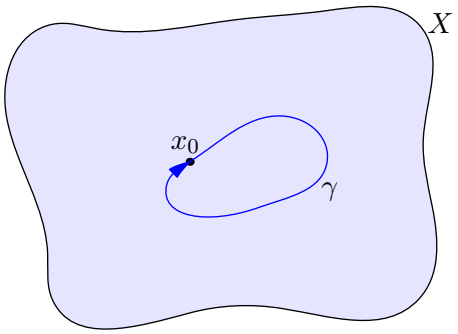
Naturally, homotopy is an equivalence relation, so paths γ lives in some "homotopy type", the equivalence classes under $\simeq$. We'll denote this $[\gamma]$. Then it makes sense to talk about $[\alpha] * [\beta]$. Thus, **we can think of $*$ as an operation on homotopy classes**.

§65.2 Fundamental groups

Prototypical example for this section: *$\pi_1(\mathbb{R}^2)$ is trivial and $\pi_1(S^1) \cong \mathbb{Z}$.*

At this point I'm a little annoyed at keeping track of endpoints, so now I'm going to specialize to a certain type of path.

Definition 65.2.1. A **loop** is a path with $\gamma(0) = \gamma(1)$.

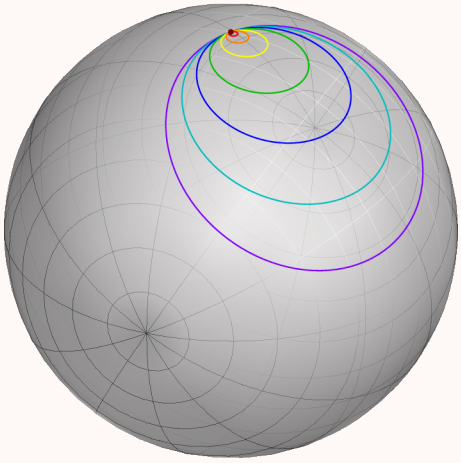


Hence if we restrict our attention to paths starting at a single point x_0, then we can stop caring about endpoints and start-points, since everything starts and stops at x_0. We even have a very canonical loop: the "do-nothing" loop[2] given by standing at x_0 the whole time.

Definition 65.2.2. Denote the trivial "do-nothing loop" by 1. A loop γ is **nulhomotopic** if it is homotopic to 1; i.e. $\gamma \simeq 1$.

For homotopy of loops, you might visualize "reeling in" the loop, contracting it to a single point.

Example 65.2.3 (Loops in S^2 are nulhomotopic)
As the following picture should convince you, every loop in the simply connected space S^2 is nulhomotopic.



(Starting with the purple loop, we contract to the red-brown point.)

Hence to show that spaces are simply connected it suffices to understand the loops of that space. We are now ready to provide:

[2] Fatty.

Definition 65.2.4. The **fundamental group** of X with basepoint x_0, denoted $\pi_1(X, x_0)$, is the set of homotopy classes

$$\{[\gamma] \mid \gamma \text{ a loop at } x_0\}$$

equipped with $*$ as a group operation.

It might come as a surprise that this has a group structure. For example, what is the inverse? Let's define it now.

Definition 65.2.5. Given a path $\alpha\colon [0,1] \to X$ we can define a path $\overline{\alpha}$

$$\overline{\alpha}(t) = \alpha(1-t).$$

In effect, this "runs α backwards". Note that $\overline{\alpha}$ starts at the endpoint of α and ends at the starting point of α.

> **Exercise 65.2.6.** Show that for any path α, $\alpha * \overline{\alpha}$ is homotopic to the "do-nothing" loop at $\alpha(0)$. (Draw a picture.)

Let's check it.

Proof that this is a group structure. Clearly $*$ takes two loops at x_0 and spits out a loop at x_0. We also already took the time to show that $*$ is associative. So we only have to check that (i) there's an identity, and (ii) there's an inverse.

- We claim that the identity is the "do-nothing" loop 1 we described above. The reader can check that for any γ,

$$\gamma \simeq \gamma * 1 = 1 * \gamma.$$

- For a loop γ, recall again we define its "backwards" loop $\overline{\gamma}$ by

$$\overline{\gamma}(t) = \gamma(1-t).$$

  Then we have $\gamma * \overline{\gamma} = \overline{\gamma} * \gamma = 1$.

Hence $\pi_1(X, x_0)$ is actually a group. □

Before going any further I had better give some examples.

Example 65.2.7 (Examples of fundamental groups)

Note that proving the following results is not at all trivial. For now, just try to see intuitively why the claimed answer "should" be correct.

(a) The fundamental group of $\mathbb{C}$ is the trivial group: in the plane, every loop is nulhomotopic. (Proof: imagine it's a piece of rope and reel it in.)

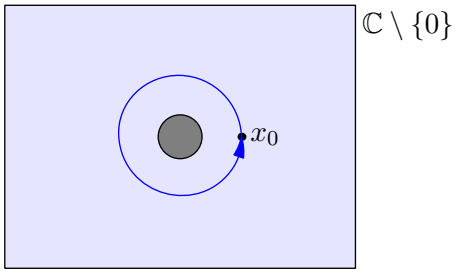
(b) On the other hand, the fundamental group of $\mathbb{C} - \{0\}$ (meteor example from earlier) with any base point is actually $\mathbb{Z}$! We won't be able to prove this for a while, but essentially a loop is determined by the number of times that it winds around the origin – these are so-called *winding numbers*. Think about it!

(c) Similarly, we will soon show that the fundamental group of S^1 (the boundary

of the unit circle) is $\mathbb{Z}$.

Officially, I also have to tell you what the base point is, but by symmetry in these examples, it doesn't matter.

Here is the picture for $\mathbb{C} \setminus \{0\}$, with the hole exaggerated as the meteor from Section 7.7.



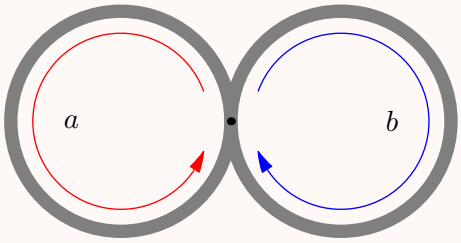
Question 65.2.8. Convince yourself that the fundamental group of S^1 is $\mathbb{Z}$, and understand why we call these "winding numbers". (This will be the most important example of a fundamental group in later chapters, so it's crucial you figure it out now.)

Example 65.2.9 (The figure eight)

Consider a figure eight $S^1 \vee S^1$, and let x_0 be the center. Then

$$\pi_1(S^1 \vee S^1, x_0) \cong \langle a, b \rangle$$

is the *free group* generated on two letters. The idea is that one loop of the eight is a, and the other loop is b, so we expect π_1 to be generated by this loop a and b (and its inverses $\overline{a}$ and $\overline{b}$). These loops don't talk to each other.



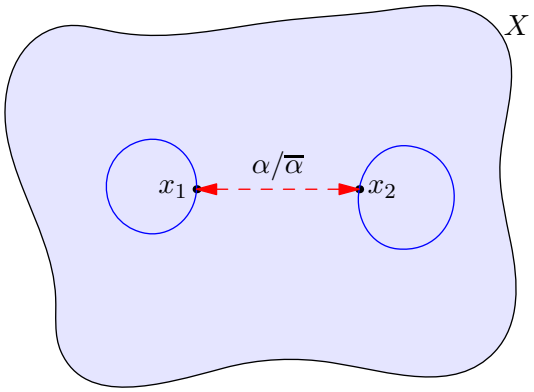
Recall that in graph theory, we usually assume our graphs are connected, since otherwise we can just consider every connected component separately. Likewise, we generally want to restrict our attention to path-connected spaces, since if a space isn't path-connected then it can be broken into a bunch of "path-connected components". (Can you guess how to define this?) Indeed, you could imagine a space X that consists of the objects on my desk (but not the desk itself): π_1 of my phone has nothing to do with π_1 of my mug. They are just totally disconnected, both figuratively and literally.

But on the other hand we claim that in a path-connected space, the groups are very related!

Theorem 65.2.10 (Fundamental groups don't depend on basepoint)
Let X be a path-connected space. Then for any $x_1 \in X$ and $x_2 \in X$, we have

$$\pi_1(X, x_1) \cong \pi_1(X, x_2).$$

Before you read the proof, see if you can guess the isomorphism based just on the picture below.



Proof. Let α be any path from x_1 to x_2 (possible by path-connectedness), and let $\overline{\alpha}$ be its reverse. Then we can construct a map

$$\pi_1(X, x_1) \to \pi_1(X, x_2) \text{ by } [\gamma] \mapsto [\overline{\alpha} * \gamma * \alpha].$$

In other words, given a loop γ at x_1, we can start at x_2, follow $\overline{\alpha}$ to x_1, run γ, then run along α home to x_2. Hence this is a map which builds a loop of $\pi_1(X, x_2)$ from every loop at $\pi_1(X, x_1)$. It is a *homomorphism* of the groups just because

$$(\overline{\alpha} * \gamma_1 * \alpha) * (\overline{\alpha} * \gamma_2 * \alpha) = \overline{\alpha} * \gamma_1 * \gamma_2 * \alpha$$

as $\alpha * \overline{\alpha}$ is nulhomotopic.

Similarly, there is a homomorphism

$$\pi_1(X, x_2) \to \pi_1(X, x_1) \text{ by } [\gamma] \mapsto [\alpha * \gamma * \overline{\alpha}].$$

As these maps are mutual inverses, it follows they must be isomorphisms. End of story. □

This is a bigger reason why we usually only care about path-connected spaces.

Abuse of Notation 65.2.11. For a path-connected space X we will often abbreviate $\pi_1(X, x_0)$ to just $\pi_1(X)$, since it doesn't matter which $x_0 \in X$ we pick.

Finally, recall that we originally defined "simply connected" as saying that any two paths with matching endpoints were homotopic. It's possible to weaken this condition and then rephrase it using fundamental groups.

> **Exercise 65.2.12.** Let X be a path-connected space. Prove that X is **simply connected** if and only if $\pi_1(X)$ is the trivial group. (One direction is easy; the other is a little trickier.)

This is the "usual" definition of simply connected.

§65.3 Fundamental groups are invariant under homeomorphism

One quick shorthand I will introduce to clean up the discussion:

Definition 65.3.1. By $f\colon (X, x_0) \to (Y, y_0)$, we will mean that $f\colon X \to Y$ is a continuous function of spaces which also sends the point x_0 to y_0.

Let X and Y be topological spaces and $f\colon (X, x_0) \to (Y, y_0)$. We now want to relate the fundamental groups of X and Y.

Recall that a loop γ in (X, x_0) is a map $\gamma\colon [0,1] \to X$ with $\gamma(0) = \gamma(1) = x_0$. Then if we consider the composition

$$[0,1] \xrightarrow{\gamma} (X, x_0) \xrightarrow{f} (Y, y_0)$$

then we get straight-away a loop in Y at y_0! Let's call this loop $f_\sharp\gamma$.

Lemma 65.3.2 ($f_\sharp$ is homotopy invariant)
If $\gamma_1 \simeq \gamma_2$ are path-homotopic, then in fact

$$f_\sharp\gamma_1 \simeq f_\sharp\gamma_2.$$

Proof. Just take the homotopy h taking γ_1 to γ_2 and consider $f \circ h$. □

It's worth noting at this point that if X and Y are homeomorphic, then their fundamental groups are all isomorphic. Indeed, let $f\colon X \to Y$ and $g\colon Y \to X$ be mutually inverse continuous maps. Then one can check that $f_\sharp\colon \pi_1(X, x_0) \to \pi_1(Y, y_0)$ and $g_\sharp\colon \pi_1(Y, y_0) \to \pi_1(X, x_0)$ are inverse maps between the groups (assuming $f(x_0) = y_0$ and $g(y_0) = x_0$).

§65.4 Higher homotopy groups

Why the notation π_1 for the fundamental group? And what are π_2, ...? The answer lies in the following rephrasing:

> **Question 65.4.1.** Convince yourself that a loop is the same thing as a continuous function $S^1 \to X$.

It turns out we can define homotopy for things other than paths. Two functions $f, g\colon Y \to X$ are **homotopic** if there exists a continuous function $Y \times [0,1] \to X$ which continuously deforms f to g. So everything we did above was just the special case $Y = S^1$.

For general n, the group $\pi_n(X)$ is defined as the homotopy classes of the maps $S^n \to X$. The group operation is a little harder to specify. You have to show that S^n is homeomorphic to $[0,1]^n$ with some endpoints fused together; for example S^1 is $[0,1]$ with 0 fused to 1. Once you have these cubes, you can merge them together on a face. (Again, I'm being terribly imprecise, deliberately.)

For $n \neq 1$, π_n behaves somewhat differently than π_1. (You might not be surprised, as S^n is simply connected for all $n \geq 2$ but not when $n = 1$.) In particular, it turns out that $\pi_n(X)$ is an abelian group for all $n \geq 2$.

Let's see some examples.

Example 65.4.2 ($\pi_n(S^n) \cong \mathbb{Z}$)

As we saw, $\pi_1(S^1) \cong \mathbb{Z}$; given the base circle S^1, we can wrap a second circle around it as many times as we want. In general, it's true that $\pi_n(S^n) \cong \mathbb{Z}$.

Example 65.4.3 ($\pi_n(S^m) \cong \{1\}$ when $n < m$)

We saw that $\pi_1(S^2) \cong \{1\}$, because a circle in S^2 can just be reeled in to a point. It turns out that similarly, any smaller n-dimensional sphere can be reeled in on the surface of a bigger m-dimensional sphere. So in general, $\pi_n(S^m)$ is trivial for $n < m$.

However, beyond these observations, the groups behave quite weirdly. Here is a table of $\pi_n(S^m)$ for $1 \le m \le 8$ and $2 \le n \le 10$, so you can see what I'm talking about. (Taken from Wikipedia.)

$\pi_n(S^m)$	2	3	4	5	6	7	8	9	10
$m=1$	$\{1\}$	$\{1\}$	$\{1\}$	$\{1\}$	$\{1\}$	$\{1\}$	$\{1\}$	$\{1\}$	$\{1\}$
2	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/12\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/3\mathbb{Z}$	$\mathbb{Z}/15\mathbb{Z}$
3		$\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/12\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/3\mathbb{Z}$	$\mathbb{Z}/15\mathbb{Z}$
4			$\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z} \times \mathbb{Z}/12\mathbb{Z}$	$(\mathbb{Z}/2\mathbb{Z})^2$	$\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/24\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$
5				$\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/24\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$
6					$\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/24\mathbb{Z}$	$\{1\}$
7						$\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/24\mathbb{Z}$
8							$\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$

Actually, it turns out that if you can compute $\pi_n(S^m)$ for every m and n, then you can essentially compute *any* homotopy classes. Thus, computing $\pi_n(S^m)$ is sort of a lost cause in general, and the mixture of chaos and pattern in the above table is a testament to this.

§65.5 Homotopy equivalent spaces

Prototypical example for this section: A disk is homotopy equivalent to a point, an annulus is homotopy equivalent to S^1.

Up to now I've abused notation and referred to "path homotopy" as just "homotopy" for two paths. I will unfortunately continue to do so (and so any time I say two paths are homotopic, you should assume I mean "path-homotopic"). But let me tell you what the general definition of homotopy is first.

Definition 65.5.1. Let $f, g \colon X \to Y$ be continuous functions. A **homotopy** is a continuous function $F \colon X \times [0,1] \to Y$, which we'll write $F_s(x)$ for $s \in [0,1]$, $x \in X$, such that

$$F_0(x) = f(x) \text{ and } F_1(x) = g(x) \text{ for all } x \in X.$$

If such a function exists, then f and g are **homotopic**.

Intuitively this is once again "deforming f to g". You might notice this is almost exactly the same definition as path-homotopy, except that f and g are any functions instead of paths, and hence there's no restriction on keeping some "endpoints" fixed through the deformation.

This homotopy can be quite dramatic:

Example 65.5.2

The zero function $z \mapsto 0$ and the identity function $z \mapsto z$ are homotopic as functions $\mathbb{C} \to \mathbb{C}$. The necessary deformation is

$$[0,1] \times \mathbb{C} \to \mathbb{C} \text{ by } (t,z) \mapsto tz.$$

I bring this up because I want to define:

Definition 65.5.3. Let X and Y be spaces. They are **homotopy equivalent** if there exist continuous functions $f\colon X \to Y$ and $g\colon Y \to X$ such that

(i) $f \circ g\colon X \to X$ is homotopic to the identity map on X, and

(ii) $g \circ f\colon Y \to Y$ is homotopic to the identity map on Y.

If a topological space is homotopy equivalent to a point, then it is said to be **contractible**.

Question 65.5.4. Why are two homeomorphic spaces also homotopy equivalent?

Intuitively, you can think of this as a more generous form of stretching and bending than homeomorphism: we are allowed to compress huge spaces into single points.

Example 65.5.5 ($\mathbb{C}$ is contractible)

Consider the topological spaces $\mathbb{C}$ and the space consisting of the single point $\{0\}$. We claim these spaces are homotopy equivalent (can you guess what f and g are?) Indeed, the two things to check are

(i) $\mathbb{C} \to \{0\} \hookrightarrow \mathbb{C}$ by $z \mapsto 0 \mapsto 0$ is homotopy equivalent to the identity on $\mathbb{C}$, which we just saw, and

(ii) $\{0\} \hookrightarrow \mathbb{C} \to \{0\}$ by $0 \mapsto 0 \mapsto 0$, which *is* the identity on $\{0\}$.

Here by $\hookrightarrow$ I just mean $\to$ in the special case that the function is just an "inclusion".

Remark 65.5.6 — $\mathbb{C}$ cannot be *homeomorphic* to a point because there is no bijection of sets between them.

Example 65.5.7 ($\mathbb{C} \setminus \{0\}$ is homotopy equivalent to S^1)

Consider the topological spaces $\mathbb{C} \setminus \{0\}$, the **punctured plane**, and the circle S^1 viewed as a subset of $\mathbb{C}$. We claim these spaces are actually homotopy equivalent! The necessary functions are the inclusion

$$S^1 \hookrightarrow \mathbb{C} \setminus \{0\}$$

and the function

$$\mathbb{C} \setminus \{0\} \to S^1 \quad \text{by} \quad z \mapsto \frac{z}{|z|}.$$

You can check that these satisfy the required condition.

Remark 65.5.8 — On the other hand, $\mathbb{C} \setminus \{0\}$ cannot be *homeomorphic* to S^1. One can make S^1 disconnected by deleting two points; the same is not true for $\mathbb{C} \setminus \{0\}$.

Example 65.5.9 (Disk = Point, Annulus = Circle)
By the same token, a disk is homotopic to a point; an annulus is homotopic to a circle. (This might be a little easier to visualize, since it's finite.)

I bring these up because it turns out that

Algebraic topology can't distinguish between homotopy equivalent spaces.

More precisely,

Theorem 65.5.10 (Homotopy equivalent spaces have isomorphic fundamental groups)
Let X and Y be path-connected, homotopy-equivalent spaces. Then $\pi_n(X) \cong \pi_n(Y)$ for every positive integer n.

Proof. Let $\gamma \colon [0,1] \to X$ be a loop. Let $f \colon X \to Y$ and $g \colon Y \to X$ be maps witnessing that X and Y are homotopy equivalent (meaning $f \circ g$ and $g \circ f$ are each homotopic to the identity). Then the composition

$$[0,1] \xrightarrow{\gamma} X \xrightarrow{f} Y$$

is a loop in Y and hence f induces a natural homomorphism $\pi_1(X) \to \pi_1(Y)$. Similarly g induces a natural homomorphism $\pi_1(Y) \to \pi_1(X)$. The conditions on f and g now say exactly that these two homomorphisms are inverse to each other, meaning the maps are isomorphisms. □

In particular,

Question 65.5.11. What are the fundamental groups of contractible spaces?

That means, for example, that algebraic topology can't tell the following homotopic subspaces of $\mathbb{R}^2$ apart.

§65.6 The pointed homotopy category

This section is meant to be read by those who know some basic category theory. Those of you that don't should come back after reading Chapters 67 and 68. Those of you that do will enjoy how succinctly we can summarize the content of this chapter using categorical notions.

Definition 65.6.1. The **pointed homotopy category** hTop_* is defined as follows.

- Objects: **pointed spaces**; that is, a pair (X, x_0) of spaces X with a distinguished basepoint x_0, and
- Morphisms: *homotopy classes* of continuous functions $(X, x_0) \to (Y, y_0)$.

In particular, two path-connected spaces are isomorphic in this category exactly when they are homotopy equivalent. Then we can summarize many of the preceding results as follows:

Theorem 65.6.2 (Functorial interpretation of fundamental groups)
There is a functor

$$\pi_1 \colon \mathsf{hTop}_* \to \mathsf{Grp}$$

sending

$$\begin{array}{ccc} (X, x_0) & \dashrightarrow & \pi_1(X, x_0) \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle f_\sharp} \\ (Y, y_0) & \dashrightarrow & \pi_1(Y, y_0) \end{array}$$

The fact that π_1 is a functor instead of merely assigns some group $\pi_1(X, x_0)$ to each pointed topological space (X, x_0) automatically implies several nice things, like:

- The functor bundles the information of $f_\sharp$, including the fact that it respects composition. In the categorical language, $f_\sharp$ is $\pi_1(f)$.
- Homotopic spaces have isomorphic fundamental groups (since the spaces are isomorphic in hTop, and functors preserve isomorphism by Theorem 68.2.8). In fact, you'll notice that the proofs of Theorem 68.2.8 and Theorem 65.5.10 are secretly identical to each other.
- If maps $f, g \colon (X, x_0) \to (Y, y_0)$ are homotopic, then $f_\sharp = g_\sharp$. This is basically Lemma 65.3.2.

Remark 65.6.3 — In fact, $\pi_1(X, x_0)$ is the set of arrows $(S^1, 1) \to (X, x_0)$ in hTop_*, so this is actually a covariant Yoneda functor (Example 68.2.6), except with target Grp instead of Set.

§65.7 A few harder problems to think about

Problem 65A (Harmonic fan)**.** Exhibit a subspace X of the metric space $\mathbb{R}^2$ which is path-connected but for which a point p can be found such that any r-neighborhood of p with $r < 1$ is not path-connected.

Problem 65B† (Special case of Seifert-van Kampen)**.** Let X be a topological space. Suppose U and V are connected open subsets of X, with $X = U \cup V$, so that $U \cap V$ is nonempty and path-connected.

Prove that if $\pi_1(U) = \pi_1(V) = \{1\}$ then $\pi_1(X) = 1$.

Remark 65.7.1 — The **Seifert–van Kampen theorem** generalizes this for $\pi_1(U)$ and $\pi_1(V)$ any groups; it gives a formula for calculating $\pi_1(X)$ in terms of $\pi_1(U)$,

$\pi_1(V)$, $\pi_1(U \cap V)$. The proof is much the same.
Unfortunately, this does not give us a way to calculate $\pi_1(S^1)$, because it is not possible to write $S^1 = U \cup V$ for $U \cap V$ *connected*.

Problem 65C (RMM 2013)**.** Let $n \geq 2$ be a positive integer. A stone is placed at each vertex of a regular $2n$-gon. A move consists of selecting an edge of the $2n$-gon and swapping the two stones at the endpoints of the edge. Prove that if a sequence of moves swaps every pair of stones exactly once, then there is some edge never used in any move.

(This last problem doesn't technically have anything to do with the chapter, but the "gut feeling" which motivates the solution is very similar.)

66 Covering projections

A few chapters ago we talked about what a fundamental group was, but we didn't actually show how to compute any of them except for the most trivial case of a simply connected space. In this chapter we'll introduce the notion of a *covering projection*, which will let us see how some of these groups can be found.

§66.1 Even coverings and covering projections

Prototypical example for this section: $\mathbb{R}$ *covers* S^1.

What we want now is a notion where a big space E, a "covering space", can be projected down onto a base space B in a nice way. Here is the notion of "nice":

Definition 66.1.1. Let $p\colon E \to B$ be a continuous function. Let U be an open set of B. We call U **evenly covered** (by p) if $p^{\text{pre}}(U)$ is a disjoint union of open sets (possibly infinite) such that p restricted to any of these sets is a homeomorphism.

Picture:

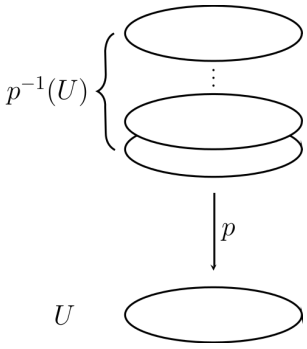


Image from [Wo]

All we're saying is that U is evenly covered if its pre-image is a bunch of copies of it. (Actually, a little more: each of the pancakes is homeomorphic to U, but we also require that p is the homeomorphism.)

Definition 66.1.2. A **covering projection** $p\colon E \to B$ is a surjective continuous map such that every base point $b \in B$ has an open neighborhood $U \ni b$ which is evenly covered by p.

> **Exercise 66.1.3** (On requiring surjectivity of p)**.** Let $p\colon E \to B$ be satisfying this definition, except that p need not be surjective. Show that the image of p is a connected component of B. Thus if B is connected and E is nonempty, then $p\colon E \to B$ is already surjective. For this reason, some authors omit the surjectivity hypothesis as usually B is path-connected.

Here is the most stupid example of a covering projection.

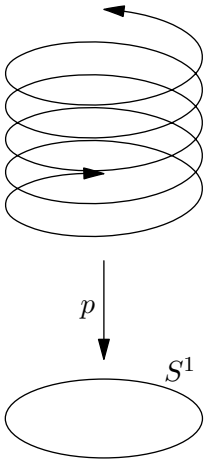
Example 66.1.4 (Tautological covering projection)
Let's take n disconnected copies of any space B: formally, $E = B \times \{1, \dots, n\}$

with the discrete topology on $\{1, \dots, n\}$. Then there exists a tautological covering projection $E \to B$ by $(x, m) \mapsto x$; we just project all n copies.
This is a covering projection because *every* open set in B is evenly covered.

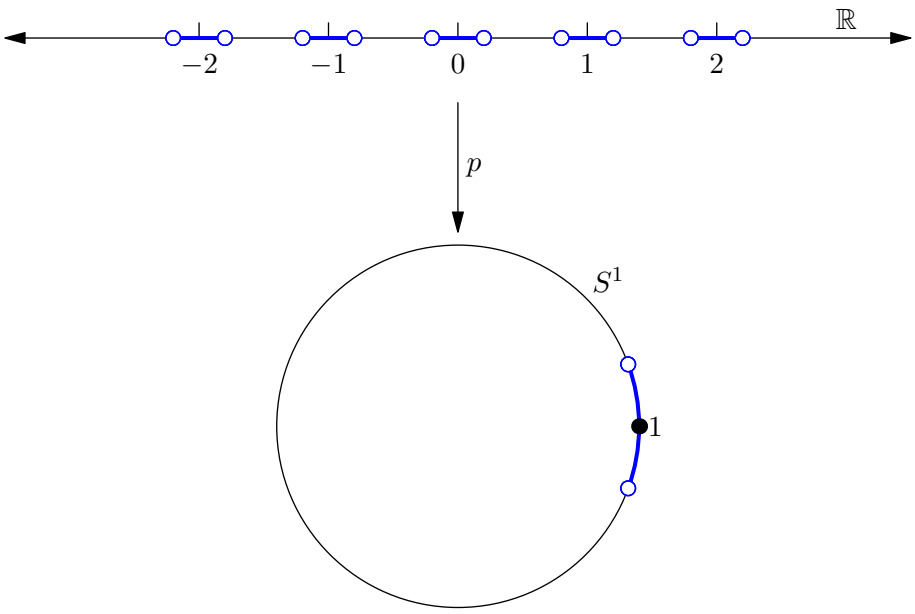
This is not really that interesting because $B \times [n]$ is not path-connected.

A much more interesting example is that of $\mathbb{R}$ and S^1.

Example 66.1.5 (Covering projection of S^1)
Take $p \colon \mathbb{R} \to S^1$ by $\theta \mapsto e^{2\pi i\theta}$. This is essentially wrapping the real line into a single helix and projecting it down.



We claim this is a covering projection. Indeed, consider the point $1 \in S^1$ (where we view S^1 as the unit circle in the complex plane). We can draw a small open neighborhood of it whose pre-image is a bunch of copies in $\mathbb{R}$.



Note that not all open neighborhoods work this time: notably, $U = S^1$ does not work because the pre-image would be the entire $\mathbb{R}$.

Example 66.1.6 (Covering of S^1 by itself)
The map $S^1 \to S^1$ by $z \mapsto z^3$ is also a covering projection. Can you see why?

Example 66.1.7 (Covering projections of $\mathbb{C} \setminus \{0\}$)
For those comfortable with complex arithmetic,

(a) The exponential map $\exp\colon \mathbb{C} \to \mathbb{C} \setminus \{0\}$ is a covering projection.

(b) For each n, the nth power map $-^n : \mathbb{C} \setminus \{0\} \to \mathbb{C} \setminus \{0\}$ is a covering projection.

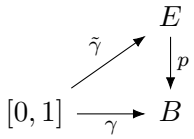
§66.2 Lifting theorem

Prototypical example for this section: $\mathbb{R}$ *covers* S^1.

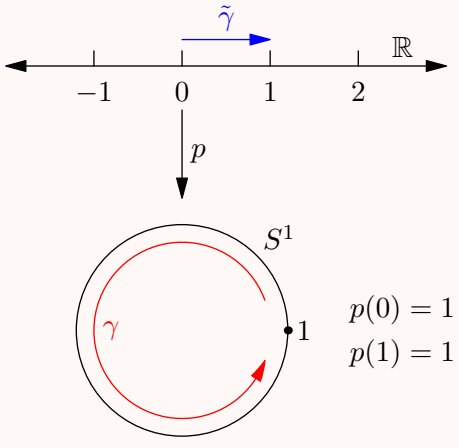
Now here's the key idea: we are going to try to interpret loops in B as paths in $\mathbb{R}$. This is often much simpler. For example, we had no idea how to compute the fundamental group of S^1, but the fundamental group of $\mathbb{R}$ is just the trivial group. So if we can interpret loops in S^1 as paths in $\mathbb{R}$, that might (and indeed it does!) make computing $\pi_1(S^1)$ tractable.

Definition 66.2.1. Let $\gamma\colon [0,1] \to B$ be a path and $p\colon E \to B$ a covering projection. A **lifting** of γ is a path $\tilde{\gamma}\colon [0,1] \to E$ such that $p \circ \tilde{\gamma} = \gamma$.

Picture:



Example 66.2.2 (Typical example of lifting)
Take $p\colon \mathbb{R} \to S^1 \subseteq \mathbb{C}$ by $\theta \mapsto e^{2\pi i\theta}$ (so S^1 is considered again as the unit circle). Consider the path γ in S^1 which starts at $1 \in \mathbb{C}$ and wraps around S^1 once, counterclockwise, ending at 1 again. In symbols, $\gamma\colon [0,1] \to S^1$ by $t \mapsto e^{2\pi it}$.
Then one lifting $\tilde{\gamma}$ is the path which walks from 0 to 1. In fact, *for any integer* n, walking from n to $n+1$ works.



Similarly, the counterclockwise path from $1 \in S^1$ to $-1 \in S^1$ has a lifting: for some integer n, the path from n to $n + \frac{1}{2}$.

The above is the primary example of a lifting. It seems like we have the following structure: given a path γ in B starting at b_0, we start at any point in the fiber $p^{\text{pre}}(b_0)$. (In our prototypical example, $B = S^1$, $b_0 = 1 \in \mathbb{C}$ and that's why we start at any integer n.) After that we just trace along the path in B, and we get a corresponding path in E.

Question 66.2.3. Take a path γ in S^1 with $\gamma(0) = 1 \in \mathbb{C}$. Convince yourself that once we select an integer $n \in \mathbb{R}$, then there is exactly one lifting starting at n.

It turns out this is true more generally.

Theorem 66.2.4 (Lifting paths)
Suppose $\gamma\colon [0,1] \to B$ is a path with $\gamma(0) = b_0$, and $p\colon (E, e_0) \to (B, b_0)$ is a covering projection. Then there exists a *unique* lifting $\tilde{\gamma}\colon [0,1] \to E$ such that $\tilde{\gamma}(0) = e_0$.

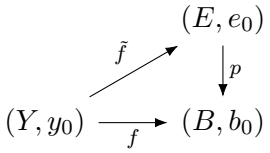
Proof. For every point $b \in B$, consider an evenly covered open neighborhood U_b in B. Then the family of open sets

$$\{\gamma^{\text{pre}}(U_b) \mid b \in B\}$$

is an open cover of $[0,1]$. As $[0,1]$ is compact we can take a finite subcover. Thus we can chop $[0,1]$ into finitely many disjoint closed intervals $[0,1] = I_1 \sqcup I_2 \sqcup \cdots \sqcup I_N$ in that order, such that for every I_k, $\gamma^{\text{img}}(I_k)$ is contained in some U_b.

We'll construct $\tilde{\gamma}$ interval by interval now, starting at I_1. Initially, place a robot at $e_0 \in E$ and a mouse at $b_0 \in B$. For each interval I_k, the mouse moves around according to however γ behaves on I_k. But the whole time it's in some evenly covered U_k; the fact that p is a covering projection tells us that there are several copies of U_k living in E. Exactly one of them, say V_k, contains our robot. So the robot just mimics the mouse until it gets to the end of I_k. Then the mouse is in some new evenly covered U_{k+1}, and we can repeat. □

The theorem can be generalized to a diagram



where Y is some general path-connected space, as follows.

Theorem 66.2.5 (General lifting criterion)

Let $f\colon (Y, y_0) \to (B, b_0)$ be continuous and consider a covering projection $p\colon (E, e_0) \to (B, b_0)$. (As usual, Y, B, E are path-connected.) Then a lifting $\tilde{f}$ with $\tilde{f}(y_0) = e_0$ exists if and only if

$$f_\sharp^{\text{img}}(\pi_1(Y, y_0)) \subseteq p_\sharp^{\text{img}}(\pi_1(E, e_0)),$$

i.e. the image of $\pi_1(Y, y_0)$ under f is contained in the image of $\pi_1(E, e_0)$ under p (both viewed as subgroups of $\pi_1(B, b_0)$). If this lifting exists, it is unique.

As $p_\sharp$ is injective, we actually have $p_\sharp^{\text{img}}(\pi_1(E, e_0)) \cong \pi_1(E, e_0)$. But in this case we are interested in the actual elements, not just the isomorphism classes of the groups.

Question 66.2.6. What happens if we put $Y = [0, 1]$?

Remark 66.2.7 (Lifting homotopies) — Here's another cool special case: Recall that a homotopy can be encoded as a continuous function $[0, 1] \times [0, 1] \to X$. But $[0, 1] \times [0, 1]$ is also simply connected. Hence given a homotopy $\gamma_1 \simeq \gamma_2$ in the base space B, we can lift it to get a homotopy $\tilde{\gamma}_1 \simeq \tilde{\gamma}_2$ in E.

Another nice application of this result is Chapter 33.

§66.3 Lifting correspondence

Prototypical example for this section: $(\mathbb{R}, 0)$ *covers* $(S^1, 1)$.

Let's return to the task of computing fundamental groups. Consider a covering projection $p\colon (E, e_0) \to (B, b_0)$.

A loop γ can be lifted uniquely to $\tilde{\gamma}$ in E which starts at e_0 and ends at some point e in the fiber $p^{\text{pre}}(b_0)$. You can easily check that this $e \in E$ does not change if we pick a different path γ' homotopic to $\tilde{\gamma}$.

Question 66.3.1. Look at the picture in Example 66.2.2.
Put one finger at $1 \in S^1$, and one finger on $0 \in \mathbb{R}$. Trace a loop homotopic to γ in S^1 (meaning, you can go backwards and forwards but you must end with exactly one full counterclockwise rotation) and follow along with the other finger in $\mathbb{R}$.
Convince yourself that you have to end at the point $1 \in \mathbb{R}$.

Thus every homotopy class of a loop at b_0 (i.e. an element of $\pi_1(B, b_0)$) can be associated with some e in the fiber of b_0. The below proposition summarizes this and more.

Proposition 66.3.2

Let $p\colon (E, e_0) \to (B, b_0)$ be a covering projection. Then we have a function of sets

$$\Phi\colon \pi_1(B, b_0) \to p^{\text{pre}}(b_0)$$

by $[\gamma] \mapsto \tilde{\gamma}(1)$, where $\tilde{\gamma}$ is the unique lifting starting at e_0. Furthermore,

- If E is path-connected, then Φ is surjective.
- If E is simply connected, then Φ is injective.

Question 66.3.3. Prove that E path-connected implies Φ is surjective. (This is really offensively easy.)

Proof. To prove the proposition, we've done everything except show that E simply connected implies Φ injective. To do this suppose that γ_1 and γ_2 are loops such that $\Phi([\gamma_1]) = \Phi([\gamma_2])$.

Applying lifting, we get paths $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ both starting at some point $e_0 \in E$ and ending at some point $e_1 \in E$. Since E is simply connected that means they are *homotopic*, and we can write a homotopy $F\colon [0,1] \times [0,1] \to E$ which unites them. But then consider the composition of maps

$$[0,1] \times [0,1] \xrightarrow{F} E \xrightarrow{p} B.$$

You can check this is a homotopy from γ_1 to γ_2. Hence $[\gamma_1] = [\gamma_2]$, done. □

This motivates:

Definition 66.3.4. A **universal cover** of a space B is a covering projection $p\colon E \to B$ where E is simply connected (and in particular path-connected).

Abuse of Notation 66.3.5. When p is understood, we sometimes just say E is the universal cover.

Example 66.3.6 (Fundamental group of S^1)

Let's return to our standard $p\colon \mathbb{R} \to S^1$. Since $\mathbb{R}$ is simply connected, this is a universal cover of S^1. And indeed, the fiber of any point in S^1 is a copy of the integers: naturally in bijection with loops in S^1.
You can show (and it's intuitively obvious) that the bijection

$$\Phi\colon \pi_1(S^1) \leftrightarrow \mathbb{Z}$$

is in fact a group homomorphism if we equip $\mathbb{Z}$ with its additive group structure $\mathbb{Z}$. Since it's a bijection, this leads us to conclude $\pi_1(S^1) \cong \mathbb{Z}$.

§66.4 Regular coverings

Prototypical example for this section: $\mathbb{R} \to S^1$ *comes from* $n \cdot x = n + x$

Here's another way to generate some coverings. Let X be a topological space and G a group acting on its points. Thus for every g, we get a map $X \to X$ by

$$x \mapsto g \cdot x.$$

We require that this map is continuous[1] for every $g \in G$, and that the stabilizer of each point in X is trivial. Then we can consider a quotient space X/G defined by fusing any points in the same orbit of this action. Thus the points of X/G are identified with the orbits of the action. Then we get a natural "projection"

$$X \to X/G$$

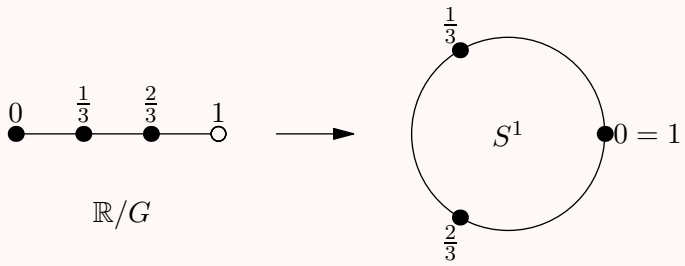
by simply sending every point to the orbit it lives in.

Definition 66.4.1. Such a projection is called **regular**. (Terrible, I know.)

Example 66.4.2 ($\mathbb{R} \to S^1$ is regular)
Let $G = \mathbb{Z}$, $X = \mathbb{R}$ and define the group action of G on X by

$$n \cdot x = n + x$$

You can then think of X/G as "real numbers modulo 1", with $[0, 1)$ a complete set of representatives and $0 \sim 1$.



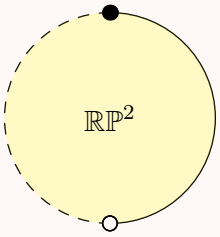
So we can identify X/G with S^1 and the associated regular projection is just our usual exp: $\theta \mapsto e^{2i\pi\theta}$.

Example 66.4.3 (The torus)
Let $G = \mathbb{Z} \times \mathbb{Z}$ and $X = \mathbb{R}^2$, and define the group action of G on X by $(m, n) \cdot (x, y) = (m + x, n + y)$. As $[0, 1)^2$ is a complete set of representatives, you can think of it as a unit square with the edges identified. We obtain the torus $S^1 \times S^1$ and a covering projection $\mathbb{R}^2 \to S^1 \times S^1$.

Example 66.4.4 ($\mathbb{RP}^2$)
Let $G = \mathbb{Z}/2\mathbb{Z} = \langle T \mid T^2 = 1 \rangle$ and let $X = S^2$ be the surface of the sphere, viewed as a subset of $\mathbb{R}^3$. We'll let G act on X by sending $T \cdot \vec{x} = -\vec{x}$; hence the orbits are pairs of opposite points (e.g. North and South pole).
Let's draw a picture of a space. All the orbits have size two: every point below the equator gets fused with a point above the equator. As for the points on the equator, we can take half of them; the other half gets fused with the corresponding antipodes.
Now if we flatten everything, you can think of the result as a disk with half its

[1] Another way of phrasing this: the action, interpreted as a map $G \times X \to X$, should be continuous, where G on the left-hand side is interpreted as a set with the discrete topology.

boundary: this is $\mathbb{RP}^2$ from before. The resulting space has a name: *real projective 2-space*, denoted $\mathbb{RP}^2$.



This gives us a covering projection $S^2 \to \mathbb{RP}^2$ (note that the pre-image of a sufficiently small patch is just two copies of it on S^2.)

Example 66.4.5 (Fundamental group of $\mathbb{RP}^2$)

As above, we saw that there was a covering projection $S^2 \to \mathbb{RP}^2$. Moreover the fiber of any point has size two. Since S^2 is simply connected, we have a natural bijection $\pi_1(\mathbb{RP}^2)$ to a set of size two; that is,

$$\left|\pi_1(\mathbb{RP}^2)\right| = 2.$$

This can only occur if $\pi_1(\mathbb{RP}^2) \cong \mathbb{Z}/2\mathbb{Z}$, as there is only one group of order two!

Question 66.4.6. Show each of the continuous maps $x \mapsto g \cdot x$ is in fact a homeomorphism. (Name its continuous inverse).

§66.5 The algebra of fundamental groups

Prototypical example for this section: S^1, with fundamental group $\mathbb{Z}$.

Next up, we're going to turn functions between spaces into homomorphisms of fundamental groups.

Let X and Y be topological spaces and $f\colon (X, x_0) \to (Y, y_0)$. Recall that we defined a group homomorphism

$$f_\sharp\colon \pi_1(X, x_0) \to \pi_1(Y_0, y_0) \quad \text{by} \quad [\gamma] \mapsto [f \circ \gamma].$$

More importantly, we have:

Proposition 66.5.1

Let $p\colon (E, e_0) \to (B, b_0)$ be a covering projection of path-connected spaces. Then the homomorphism $p_\sharp\colon \pi_1(E, e_0) \to \pi_1(B, b_0)$ is *injective*. Hence $p_\sharp^{\text{img}}(\pi_1(E, e_0))$ is an isomorphic copy of $\pi_1(E, e_0)$ as a subgroup of $\pi_1(B, b_0)$.

Proof. We'll show $\ker p_\sharp$ is trivial. It suffices to show if γ is a nulhomotopic loop in B then its lift is nulhomotopic.

By definition, there's a homotopy $F\colon [0,1] \times [0,1] \to B$ taking γ to the constant loop 1_B. We can lift it to a homotopy $\tilde{F}\colon [0,1] \times [0,1] \to E$ that establishes $\tilde{\gamma} \simeq \tilde{1}_B$. But 1_E

is a lift of 1_B (duh) and lifts are unique. □

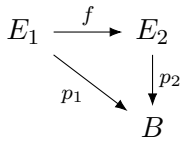
Example 66.5.2 (Subgroups of $\mathbb{Z}$)

Let's look at the space S^1 with fundamental group $\mathbb{Z}$. The group $\mathbb{Z}$ has two types of subgroups:

- The trivial subgroup. This corresponds to the canonical projection $\mathbb{R} \to S^1$, since $\pi_1(\mathbb{R})$ is the trivial group ($\mathbb{R}$ is simply connected) and hence its image in $\mathbb{Z}$ is the trivial group.
- $n\mathbb{Z}$ for $n \geq 1$. This is given by the covering projection $S^1 \to S^1$ by $z \mapsto z^n$. The image of a loop in the covering S^1 is a "multiple of n" in the base S^1.

It turns out that these are the *only* covering projections of S^n by path-connected spaces: there's one for each subgroup of $\mathbb{Z}$. (We don't care about disconnected spaces because, again, a covering projection via disconnected spaces is just a bunch of unrelated "good" coverings.) For this statement to make sense I need to tell you what it means for two covering projections to be equivalent.

Definition 66.5.3. Fix a space B. Given two covering projections $p_1 \colon E_1 \to B$ and $p_2 \colon E_2 \to B$ a **map of covering projections** is a continuous function $f \colon E_1 \to E_2$ such that $p_2 \circ f = p_1$.



Then two covering projections p_1 and p_2 are isomorphic if there are $f \colon E_1 \to E_2$ and $g \colon E_2 \to E_1$ such that $f \circ g = \mathrm{id}_{E_1}$ and $g \circ f = \mathrm{id}_{E_2}$.

Remark 66.5.4 (For category theorists) — The set of covering projections forms a category in this way.

It's an absolute miracle that this is true more generally: the greatest triumph of covering spaces is the following result. Suppose a space X satisfies some nice conditions, like:

Definition 66.5.5. A space X is called **locally connected** if for each point $x \in X$ and open neighborhood V of it, there is a connected open set U with $x \in U \subseteq V$.

Definition 66.5.6. A space X is **semi-locally simply connected** if for every point $x \in X$ there is an open neighborhood U such that all loops in U are nulhomotopic. (But the contraction need not take place in U.)

Example 66.5.7 (These conditions are weak)

Pretty much every space I've shown you has these two properties. In other words, they are rather mild conditions, and you can think of them as just saying "the space is not too pathological".

Then we get:

Theorem 66.5.8 (Group theory via covering spaces)

Suppose B is a locally connected, semi-locally simply connected space. Then:

- Every subgroup $H \subseteq \pi_1(B)$ corresponds to exactly one covering projection $p\colon E \to B$ with E path-connected (up to isomorphism).

  (Specifically, H is the image of $\pi_1(E)$ in $\pi_1(B)$ through $p_\sharp$.)
- Moreover, the *normal* subgroups of $\pi_1(B)$ correspond exactly to the regular covering projections.

Hence it's possible to understand the group theory of $\pi_1(B)$ completely in terms of the covering projections.

Moreover, this is how the "universal cover" gets its name: it is the one corresponding to the trivial subgroup of $\pi_1(B)$. Actually, you can show that it really is universal in the sense that if $p\colon E \to B$ is another covering projection, then E is in turn covered by the universal space. More generally, if $H_1 \subseteq H_2 \subseteq G$ are subgroups, then the space corresponding to H_2 can be covered by the space corresponding to H_1.

§66.6 A few harder problems to think about

problems

XVII

Category Theory

Part XVII: Contents

67 Objects and morphisms

I can't possibly hope to do category theory any justice in these few chapters; thus I'll just give a very high-level overview of how many of the concepts we've encountered so far can be re-cast into categorical terms. So I'll say what a category is, give some examples, then talk about a few things that categories can do. For my examples, I'll be drawing from all the previous chapters; feel free to skip over the examples corresponding to things you haven't seen.

If you're interested in category theory (like I was!), perhaps in what surprising results are true for general categories, I strongly recommend [Le14].

§67.1 Motivation: isomorphisms

From earlier chapters let's recall the definition of an *isomorphism* of two objects:

- Two groups G and H are isomorphic if there was a bijective homomorphism: equivalently, we wanted homomorphisms $\phi\colon G \to H$ and $\psi\colon H \to G$ which were mutual inverses, meaning $\phi \circ \psi = \mathrm{id}_H$ and $\psi \circ \phi = \mathrm{id}_G$.
- Two metric (or topological) spaces X and Y are isomorphic if there is a continuous bijection $f\colon X \to Y$ such that f^{-1} is also continuous.
- Two vector spaces V and W are isomorphic if there is a bijection $T\colon V \to W$ which is a linear map. Again, this can be re-cast as saying that T and T^{-1} are linear maps.
- Two rings R and S are isomorphic if there is a bijective ring homomorphism ϕ; again, we can re-cast this as two mutually inverse ring homomorphisms.

In each case we have some collections of objects and some maps, and the isomorphisms can be viewed as just maps. Let's use this to motivate the definition of a general *category*.

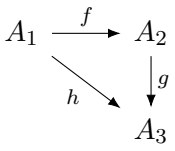
§67.2 Categories, and examples thereof

Prototypical example for this section: Grp *is possibly the most natural example.*

Definition 67.2.1. A **category** $\mathcal{A}$ consists of:

- A class of **objects**, denoted $\mathrm{obj}(\mathcal{A})$.
- For any two objects $A_1, A_2 \in \mathrm{obj}(\mathcal{A})$, a class of **arrows** (also called **morphisms** or **maps**) between them. We'll denote the set of these arrows by $\mathrm{Hom}_{\mathcal{A}}(A_1, A_2)$.
- For any $A_1, A_2, A_3 \in \mathrm{obj}(\mathcal{A})$, if $f\colon A_1 \to A_2$ is an arrow and $g\colon A_2 \to A_3$ is an arrow, we can compose these arrows to get an arrow $g \circ f\colon A_1 \to A_3$.

  We can represent this in a **commutative diagram**



where $h = g \circ f$. The composition operation $\circ$ is part of the data of $\mathcal{A}$; it must be associative. In the diagram above we say that h **factors** through A_2.

- Finally, every object $A \in \text{obj}(\mathcal{A})$ has a special **identity arrow** id_A; you can guess what it does.[1]

Abuse of Notation 67.2.2. From now on, by $A \in \mathcal{A}$ we'll mean $A \in \text{obj}(\mathcal{A})$.

Abuse of Notation 67.2.3. You can think of "class" as just "set". The reason we can't use the word "set" is because of some paradoxical issues with collections which are too large; Cantor's Paradox says there is no set of all sets. So referring to these by "class" is a way of sidestepping these issues.

Now and forever I'll be sloppy and assume all my categories are **locally small**, meaning that $\text{Hom}_{\mathcal{A}}(A_1, A_2)$ is a set for any $A_1, A_2 \in \mathcal{A}$. So elements of $\mathcal{A}$ may not form a set, but the set of morphisms between two *given* objects will always assumed to be a set.

Let's formalize the motivation we began with.

Example 67.2.4 (Basic examples of categories)

(a) There is a category of groups Grp. The data is
- The objects of Grp are the groups.
- The arrows of Grp are the homomorphisms between these groups.
- The composition $\circ$ in Grp is function composition.

(b) In the same way we can conceive a category CRing of (commutative) rings.

(c) Similarly, there is a category Top of topological spaces, whose arrows are the continuous maps.

(d) There is a category Top_* of topological spaces with a *distinguished basepoint*; that is, a pair (X, x_0) where $x_0 \in X$. Arrows are continuous maps $f : X \to Y$ with $f(x_0) = y_0$.

(e) Similarly, there is a category Vect_k of vector spaces (possibly infinite-dimensional) over a field k, whose arrows are the linear maps. There is even a category FDVect_k of *finite-dimensional* vector spaces.

(f) We have a category Set of sets, where the arrows are *any* maps.

And of course, we can now define what an isomorphism is!

Definition 67.2.5. An arrow $A_1 \xrightarrow{f} A_2$ is an **isomorphism** if there exists $A_2 \xrightarrow{g} A_1$ such that $f \circ g = \text{id}_{A_2}$ and $g \circ f = \text{id}_{A_1}$. In that case we say A_1 and A_2 are **isomorphic**, hence $A_1 \cong A_2$.

Remark 67.2.6 — Note that in Set, $X \cong Y \iff |X| = |Y|$.

[1] To be painfully explicit: if $f : A' \to A$ is an arrow then $\text{id}_A \circ f = f$; similarly, if $g : A \to A'$ is an arrow then $g \circ \text{id}_A = g$.

Question 67.2.7. Check that every object in a category is isomorphic to itself. (This is offensively easy.)

More importantly, this definition should strike you as a little impressive. We're able to define whether two groups (rings, spaces, etc.) are isomorphic solely by the functions between the objects. Indeed, one of the key themes in category theory (and even algebra) is that

One can learn about objects by the functions between them. Category theory takes this to the extreme by *only* looking at arrows, and ignoring what the objects themselves are.

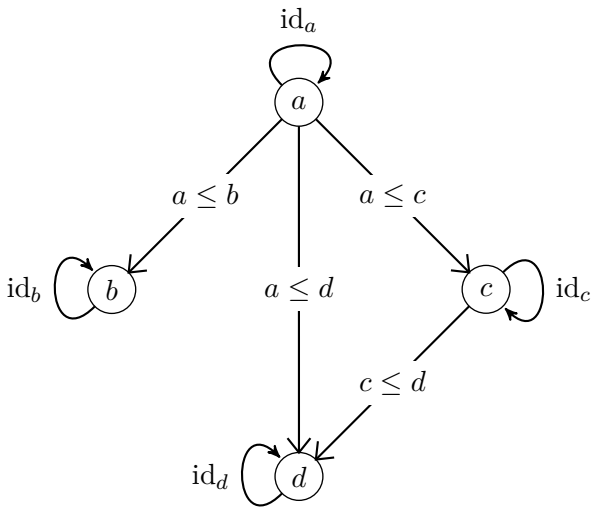
But there are some trickier interesting examples of categories.

Example 67.2.8 (Posets are categories)
Let $\mathcal{P}$ be a partially ordered set. We can construct a category P for it as follows:

- The objects of P are going to be the elements of $\mathcal{P}$.
- The arrows of P are defined as follows:
  - For every object $p \in P$, we add an identity arrow id_p, and
  - For any pair of distinct objects $p \le q$, we add a single arrow $p \to q$.

  There are no other arrows.
- There's only one way to do the composition. What is it?

For example, for the poset $\mathcal{P}$ on four objects $\{a, b, c, d\}$ with $a \le b$ and $a \le c \le d$, we get:



This illustrates the point that

The arrows of a category can be totally different from functions.

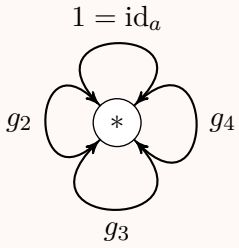
In fact, in a way that can be made precise, the term "concrete category" refers to one where the arrows really are "structure-preserving maps between sets", like Grp, Top, or CRing.

Question 67.2.9. Check that no two distinct objects of a poset are isomorphic.

Here's a second quite important example of a non-concrete category.

Example 67.2.10 (Important: groups are one-Object categories)

A group G can be interpreted as a category $\mathcal{G}$ with one object $*$, all of whose arrows are isomorphisms.



As [Le14] says:

> The first time you meet the idea that a group is a kind of category, it's tempting to dismiss it as a coincidence or a trick. It's not: there's real content. To see this, suppose your education had been shuffled and you took a course on category theory before ever learning what a group was. Someone comes to you and says:
>
> "There are these structures called 'groups', and the idea is this: a group is what you get when you collect together all the symmetries of a given thing."
>
> "What do you mean by a 'symmetry'?" you ask.
>
> "Well, a symmetry of an object X is a way of transforming X or mapping X into itself, in an invertible way."
>
> "Oh," you reply, "that's a special case of an idea I've met before. A category is the structure formed by *lots* of objects and mappings between them – not necessarily invertible. A group's just the very special case where you've only got one object, and all the maps happen to be invertible."

Exercise 67.2.11. Verify the above! That is, show that the data of a one-object category with all isomorphisms is the same as the data of a group.

Finally, here are some examples of categories you can make from other categories.

Example 67.2.12 (Deriving categories)

(a) Given a category $\mathcal{A}$, we can construct the **opposite category** $\mathcal{A}^{\mathrm{op}}$, which is

the same as $\mathcal{A}$ but with all arrows reversed.

(b) Given categories $\mathcal{A}$ and $\mathcal{B}$, we can construct the **product category** $\mathcal{A} \times \mathcal{B}$ as follows: the objects are pairs (A, B) for $A \in \mathcal{A}$ and $B \in \mathcal{B}$, and the arrows from (A_1, B_1) to (A_2, B_2) are pairs

$$\left(A_1 \xrightarrow{f} A_2, B_1 \xrightarrow{g} B_2 \right).$$

What do you think the composition and identities are?

§67.3 Special objects in categories

Prototypical example for this section: Set *has initial object* $\varnothing$ *and final object* $\{*\}$. *An element of* S *corresponds to a map* $\{*\} \to S$.

Certain objects in categories have special properties. Here are a couple examples.

Example 67.3.1 (Initial object)
An **initial object** of $\mathcal{A}$ is an object $A_{\text{init}} \in \mathcal{A}$ such that for any $A \in \mathcal{A}$ (possibly $A = A_{\text{init}}$), there is exactly one arrow from A_{init} to A. For example,

(a) The initial object of Set is the empty set $\varnothing$.

(b) The initial object of Grp is the trivial group $\{1\}$.

(c) The initial object of CRing is the ring $\mathbb{Z}$ (recall that ring homomorphisms $R \to S$ map 1_R to 1_S).

(d) The initial object of Top is the empty space.

(e) The initial object of a partially ordered set is its smallest element, if one exists.

We will usually refer to "the" initial object of a category, since:

Exercise 67.3.2 (Important!)**.** Show that any two initial objects A_1, A_2 of $\mathcal{A}$ are *uniquely isomorphic* meaning there is a unique isomorphism between them.

Remark 67.3.3 — In mathematics, we usually neither know nor care if two objects are actually equal or whether they are isomorphic. For example, there are many competing ways to define $\mathbb{R}$, but we still just refer to it as "the" real numbers. Thus when we define categorical notions, we would like to check they are unique up to isomorphism. This is really clean in the language of categories, and definitions often cause objects to be unique up to isomorphism for elegant reasons like the above.

One can take the "dual" notion, a terminal object.

Example 67.3.4 (Terminal object)

A **terminal object** of $\mathcal{A}$ is an object $A_{\text{final}} \in \mathcal{A}$ such that for any $A \in \mathcal{A}$ (possibly $A = A_{\text{final}}$), there is exactly one arrow from A to A_{final}. For example,

(a) The terminal object of Set is the singleton set $\{*\}$. (There are many singleton sets, of course, but *as sets* they are all isomorphic!)

(b) The terminal object of Grp is the trivial group $\{1\}$.

(c) The terminal object of CRing is the zero ring 0. (Recall that ring homomorphisms $R \to S$ must map 1_R to 1_S).

(d) The terminal object of Top is the single-point space.

(e) The terminal object of a partially ordered set is its maximal element, if one exists.

Again, terminal objects are unique up to isomorphism. The reader is invited to repeat the proof from the preceding exercise here. However, we can illustrate more strongly the notion of duality to give a short proof.

Question 67.3.5. Verify that terminal objects of $\mathcal{A}$ are equivalent to initial objects of $\mathcal{A}^{\text{op}}$. Thus terminal objects of $\mathcal{A}$ are unique up to isomorphism.

In general, one can consider in this way the dual of *any* categorical notion: properties of $\mathcal{A}$ can all be translated to dual properties of $\mathcal{A}^{\text{op}}$ (often by adding the prefix "co" in front).

One last neat construction: suppose we're working in a concrete category, meaning (loosely) that the objects are "sets with additional structure". Now suppose you're sick of maps and just want to think about elements of these sets. Well, I won't let you do that since you're reading a category theory chapter, but I will offer you some advice:

- In Set, arrows from $\{*\}$ to S correspond to elements of S.
- In Top, arrows from $\{*\}$ to X correspond to points of X.
- In Grp, arrows from $\mathbb{Z}$ to G correspond to elements of G.
- In CRing, arrows from $\mathbb{Z}[x]$ to R correspond to elements of R.

and so on. So in most concrete categories, you can think of elements as functions from special sets to the set in question. In each of these cases we call the object in question a **free object**.

§67.4 Binary products

Prototypical example for this section: $X \times Y$ in most concrete categories is the set-theoretic product.

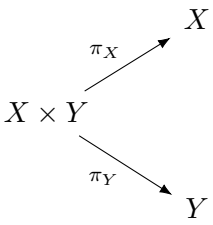
The "universal property" is a way of describing objects in terms of maps in such a way that it defines the object up to unique isomorphism (much the same as the initial and terminal objects).

To show how this works in general, let me give a concrete example. Suppose I'm in a category – let's say Set for now. I have two sets X and Y, and I want to construct the Cartesian product $X \times Y$ as we know it. The philosophy of category theory dictates that I should talk about maps only, and avoid referring to anything about the sets themselves. How might I do this?

Well, let's think about maps into $X \times Y$. The key observation is that

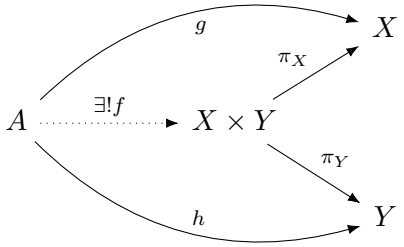
A function $A \xrightarrow{f} X \times Y$ amounts to a pair of functions $(A \xrightarrow{g} X, A \xrightarrow{h} Y)$.

Put another way, there is a natural projection map $X \times Y \to X$ and $X \times Y \to Y$:



(We have to do this in terms of projection maps rather than elements, because category theory forces us to talk about arrows.) Now how do I add A to this diagram? The point is that there is a bijection between functions $A \xrightarrow{f} X \times Y$ and pairs (g, h) of functions. Thus for every pair $A \xrightarrow{g} X$ and $A \xrightarrow{h} Y$ there is a *unique* function $A \xrightarrow{f} X \times Y$.

But $X \times Y$ is special in that it is "universal": for any *other* set A, if you give me functions $A \to X$ and $A \to Y$, I can use it build a *unique* function $A \to X \times Y$. Picture:



We can do this in any general category, defining a so-called product.

Definition 67.4.1. Let X and Y be objects in any category $\mathcal{A}$. The **product** consists of an object $X \times Y$ and arrows π_X, π_Y to X and Y (thought of as projection). We require that for any object A and arrows $A \xrightarrow{g} X$, $A \xrightarrow{h} Y$, there is a *unique* function $A \xrightarrow{f} X \times Y$ such that the above diagram commutes.

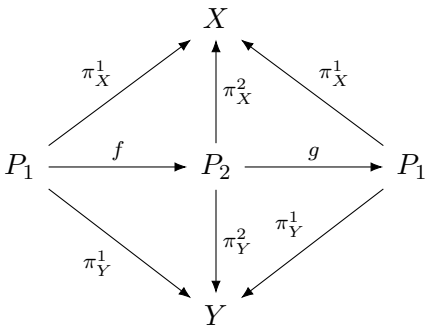
Abuse of Notation 67.4.2. Strictly speaking, the product should consist of *both* the object $X \times Y$ and the projection maps π_X and π_Y. However, if π_X and π_Y are understood, then we often use $X \times Y$ to refer to the object, and refer to it also as the product.

Products do not always exist; for example, take a category with just two objects and no non-identity morphisms. Nonetheless:

Proposition 67.4.3 (Uniqueness of products)

When they exist, products are unique up to isomorphism: given two products P_1 and P_2 of X and Y there is an isomorphism between the two objects.

Proof. This is very similar to the proof that initial objects are unique up to unique isomorphism. Consider two such objects P_1 and P_2, and the associated projection maps. So, we have a diagram



There are unique morphisms f and g between P_1 and P_2 that make the entire diagram commute, according to the universal property.

On the other hand, look at $g \circ f$ and focus on just the outer square. Observe that $g \circ f$ is a map which makes the outer square commute, so by the universal property of P_1 it is the only one. But id_{P_1} works as well. Thus $\mathrm{id}_{P_1} = g \circ f$. Similarly, $f \circ g = \mathrm{id}_{P_2}$ so f and g are isomorphisms. □

Abuse of Notation 67.4.4. Actually, this is not really the morally correct theorem; since we've only showed the objects P_1 and P_2 are isomorphic and have not made any assertion about the projection maps. But I haven't (and won't) define isomorphism of the entire product, and so in what follows if I say "P_1 and P_2 are isomorphic" I really just mean the objects are isomorphic.

> **Exercise 67.4.5.** In fact, show the products are unique up to *unique* isomorphism: the f and g above are the only isomorphisms between the objects P_1 and P_2 respecting the projections.

The nice fact about this "universal property" mindset is that we don't have to give explicit constructions; assuming existence, the "universal property" allows us to bypass all this work by saying "the object with these properties is unique up to unique isomorphism", thus we don't need to understand the internal workings of the object to use its properties.

Of course, that's not to say we can't give concrete examples.

Example 67.4.6 (Examples of products)

(a) In Set, the product of two sets X and Y is their Cartesian product $X \times Y$.

(b) In Grp, the product of G, H is the group product $G \times H$.

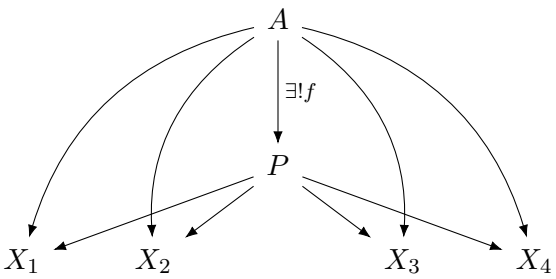
(c) In Vect_k, the product of V and W is $V \oplus W$.

(d) In CRing, the product of R and S is appropriately the ring product $R \times S$.

(e) Let $\mathcal{P}$ be a poset interpreted as a category. Then the product of two objects x and y is the **greatest lower bound**; for example,
- If the poset is $(\mathbb{R}, \leq)$ then it's $\min\{x, y\}$.
- If the poset is the subsets of a finite set by inclusion, then it's $x \cap y$.
- If the poset is the positive integers ordered by division, then it's $\gcd(x, y)$.

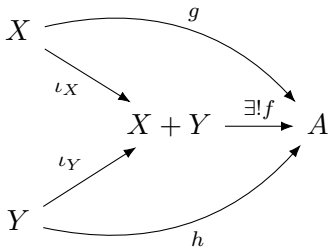
Of course, we can define products of more than just one object. Consider a set of objects $(X_i)_{i \in I}$ in a category $\mathcal{A}$. We define a **cone** on the X_i to be an object A with some "projection" maps to each X_i. Then the **product** is a cone P which is "universal" in the same sense as before: given any other cone A there is a unique map $A \to P$ making the diagram commute. In short, a product is a "universal cone".

The picture of this is



See also Problem 67C.

One can also do the dual construction to get a **coproduct**: given X and Y, it's the object $X + Y$ together with maps $X \xrightarrow{\iota_X} X + Y$ and $Y \xrightarrow{\iota_Y} X + Y$ (that's Greek iota, think inclusion) such that for any object A and maps $X \xrightarrow{g} A$, $Y \xrightarrow{h} A$ there is a unique f for which



commutes. We'll leave some of the concrete examples as an exercise this time, for example:

> **Exercise 67.4.7.** Describe the coproduct in Set.

Predictable terminology: a coproduct is a universal **cocone**.

Spoiler alert later on: this construction can be generalized vastly to so-called "limits", and we'll do so later on.

§67.5 Monic and epic maps

The notion of "injective" doesn't make sense in an arbitrary category since arrows need not be functions. The correct categorical notion is:

Definition 67.5.1. A map $X \xrightarrow{f} Y$ is **monic** (or a monomorphism) if for any commutative diagram

$$A \underset{h}{\overset{g}{\rightrightarrows}} X \xrightarrow{f} Y$$

we must have $g = h$. In other words, $f \circ g = f \circ h \implies g = h$.

> **Question 67.5.2.** Verify that in a *concrete* category, injective $\implies$ monic.

Question 67.5.3. Show that the composition of two monic maps is monic.

In most but not all situations, the converse is also true.

Exercise 67.5.4. Show that in Set, Grp, CRing, monic implies injective. (Take $A = \{*\}$, $A = \mathbb{Z}$, $A = \mathbb{Z}[x]$.)

More generally, as we said before there are many categories with a "free" object that you can use to think of as elements. An element of a set is a function $1 \to S$, and element of a ring is a function $\mathbb{Z}[x] \to R$, et cetera. In all these categories, the definition of monic literally reads "f is injective on $\hom_{\mathcal{A}}(A, X)$". So in these categories, "monic" and "injective" coincide.

That said, here is the standard counterexample. An additive abelian group $G = (G, +)$ is called *divisible* if for every $x \in G$ and integer $n > 0$ there exists $y \in G$ with $ny = x$. Let $\mathsf{DivAbGrp}$ be the category of such groups.

Exercise 67.5.5. Show that the projection $\mathbb{Q} \to \mathbb{Q}/\mathbb{Z}$ is monic but not injective.

Of course, we can also take the dual notion.

Definition 67.5.6. A map $X \xrightarrow{f} Y$ is **epic** (or an epimorphism) if for any commutative diagram

$$X \xrightarrow{f} Y \underset{h}{\overset{g}{\rightrightarrows}} A$$

we must have $g = h$. In other words, $g \circ f = h \circ f \implies g = h$.

This is kind of like surjectivity, although it's a little farther than last time. Note that in concrete categories, surjective $\implies$ epic.

Exercise 67.5.7. Show that in Set, Grp, Ab, Vect_k, Top, the notions of epic and surjective coincide. (For Set, take $A = \{0, 1\}$.)

However, there are more cases where it fails. Most notably:

Example 67.5.8 (Epic but not surjective)

(a) In CRing, for instance, the inclusion $\mathbb{Z} \hookrightarrow \mathbb{Q}$ is epic (and not surjective). Indeed, if two homomorphisms $\mathbb{Q} \to A$ agree on every integer then they agree everywhere (why?),

(b) In the category of *Hausdorff* topological spaces (every two points have disjoint open neighborhoods), in fact epic $\iff$ dense image (like $\mathbb{Q} \hookrightarrow \mathbb{R}$).

Thus failures arise when a function $f : X \to Y$ can be determined by just some of the points of X.

§67.6 A few harder problems to think about

Problem 67A. In the category Vect_k of k-vector spaces (for a field k), what are the initial and terminal objects?

Problem 67B†. What is the coproduct $X + Y$ in the categories Set, Vect_k, and a poset?

Problem 67C. In any category $\mathcal{A}$ where all products exist, show that

$$(X \times Y) \times Z \cong X \times (Y \times Z)$$

where X, Y, Z are arbitrary objects. (Here both sides refer to the objects, as in Abuse of Notation 67.4.2.)

Problem 67D. Consider a category $\mathcal{A}$ with a **zero object**, meaning an object which is both initial and terminal. Given objects X and Y in A, prove that the projection $X \times Y \to X$ is epic.

68 Functors and natural transformations

Functors are maps between categories; natural transformations are maps between functors.

§68.1 Many examples of functors

Prototypical example for this section: Forgetful functors; fundamental groups; $\bullet^\vee$.

Here's the point of a functor:

Pretty much any time you make an object out of another object, you get a functor.

Before I give you a formal definition, let me list (informally) some examples. (You'll notice some of them have opposite categories $\mathcal{A}^{\text{op}}$ appearing in places. Don't worry about those for now; you'll see why in a moment.)

- Given a group G (or vector space, field, ...), we can take its underlying set S; this is a functor from $\mathsf{Grp} \to \mathsf{Set}$.
- Given a set S we can consider a vector space with basis S; this is a functor from $\mathsf{Set} \to \mathsf{Vect}$.
- Given a vector space V we can consider its dual space $V^\vee$. This is a functor $\mathsf{Vect}_k^{\text{op}} \to \mathsf{Vect}_k$.
- Tensor products give a functor from $\mathsf{Vect}_k \times \mathsf{Vect}_k \to \mathsf{Vect}_k$.
- Given a set S, we can build its power set, giving a functor $\mathsf{Set} \to \mathsf{Set}$.
- In algebraic topology, we take a topological space X and build several groups $H_1(X)$, $\pi_1(X)$, etc. associated to it. All these group constructions are functors $\mathsf{Top} \to \mathsf{Grp}$.
- Sets of homomorphisms: let $\mathcal{A}$ be a category.
  - Given two vector spaces V_1 and V_2 over k, we construct the abelian group of linear maps $V_1 \to V_2$. This is a functor from $\mathsf{Vect}_k^{\text{op}} \times \mathsf{Vect}_k \to \mathsf{AbGrp}$.
  - More generally for any category $\mathcal{A}$ we can take pairs (A_1, A_2) of objects and obtain a set $\text{Hom}_{\mathcal{A}}(A_1, A_2)$. This turns out to be a functor $\mathcal{A}^{\text{op}} \times \mathcal{A} \to \mathsf{Set}$.
  - The above operation has two "slots". If we "pre-fill" the first slots, then we get a functor $\mathcal{A} \to \mathsf{Set}$. That is, by fixing $A \in \mathcal{A}$, we obtain a functor (called H^A) from $\mathcal{A} \to \mathsf{Set}$ by sending $A' \in \mathcal{A}$ to $\text{Hom}_{\mathcal{A}}(A, A')$. This is called the covariant Yoneda functor (explained later).
  - As we saw above, for every $A \in \mathcal{A}$ we obtain a functor $H^A \colon \mathcal{A} \to \mathsf{Set}$. It turns out we can construct a category $[\mathcal{A}, \mathsf{Set}]$ whose elements are functors $\mathcal{A} \to \mathsf{Set}$; in that case, we now have a functor $\mathcal{A}^{\text{op}} \to [\mathcal{A}, \mathsf{Set}]$.

That having said, here are some non-functors. Just so that when you see a theorem that says "F is a functor" (in other words, "F is functorial"), you should read it as "F has a deep hidden symmetry behind it! This is very nice!" instead of "this theorem is trivial".

What is that deep symmetry? Keep reading.

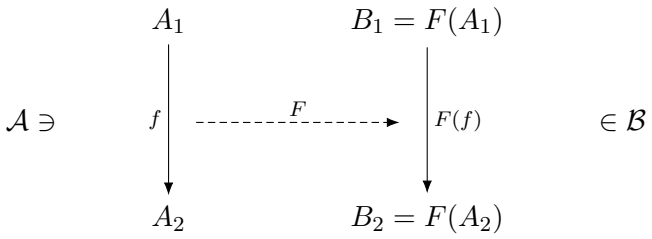
- Given a group G, we can build its automorphism group $\mathrm{Aut}(G)$. But this is not a functor in any natural way.
- Given a group G, we can build its center $Z(G)$, which is the set of elements in G that commutes with everything in G. Again, this is not a functor in any natural way.[1]
- The operation of taking the dual space above is a contravariant functor $\mathsf{Vect}_k^{\text{op}} \to \mathsf{Vect}_k$, but it isn't a covariant functor $\mathsf{Vect}_k \to \mathsf{Vect}_k$. (Don't worry what a contravariant functor is for now.)

§68.2 Covariant functors

Prototypical example for this section: Forgetful/free functors, ...

Category theorists are always asking "what are the maps?", and so we can now think about maps between categories.

Definition 68.2.1. Let $\mathcal{A}$ and $\mathcal{B}$ be categories. Of course, a **functor** F takes every object of $\mathcal{A}$ to an object of $\mathcal{B}$. In addition, though, it must take every arrow $A_1 \xrightarrow{f} A_2$ to an arrow $F(A_1) \xrightarrow{F(f)} F(A_2)$. You can picture this as follows.



(I'll try to use dotted arrows for functors, which cross different categories, for emphasis.) It needs to satisfy the "naturality" requirements:

- Identity arrows get sent to identity arrows: for each identity arrow id_A, we have $F(\mathrm{id}_A) = \mathrm{id}_{F(A)}$.
- The functor respects composition: if $A_1 \xrightarrow{f} A_2 \xrightarrow{g} A_3$ are arrows in $\mathcal{A}$, then $F(g \circ f) = F(g) \circ F(f)$.

So the idea is:

> **Whenever we naturally make an object $A \in \mathcal{A}$ into an object of $B \in \mathcal{B}$, there should usually be a natural way to transform a map $A_1 \to A_2$ into a map $B_1 \to B_2$.**

Let's see some examples of this.

[1] It is easy to find a counterexample based on properties of functor — in particular, identity maps get sent to identity maps. See https://math.stackexchange.com/q/158438 for a proof.

Example 68.2.2 (Free and forgetful functors)

Note that these are both informal terms, and don't have a rigid definition.

(a) We talked about a **forgetful functor** earlier, which takes the underlying set of a category like Vect_k. Let's call it $U \colon \mathsf{Vect}_k \to \mathsf{Set}$.

Now, given a map $T \colon V_1 \to V_2$ in Vect_k, there is an obvious $U(T) \colon U(V_1) \to U(V_2)$ which is just the set-theoretic map corresponding to T.

Similarly there are forgetful functors from Grp, CRing, etc., to Set. There is even a forgetful functor $\mathsf{CRing} \to \mathsf{Grp}$: send a ring R to the abelian group $(R, +)$. The common theme is that we are "forgetting" structure from the original category.

(b) We also talked about a **free functor** in the example. A free functor $F \colon \mathsf{Set} \to \mathsf{Vect}_k$ can be taken by considering $F(S)$ to be the vector space with basis S. Now, given a map $f \colon S \to T$, what is the obvious map $F(S) \to F(T)$? Simple: take each basis element $s \in S$ to the basis element $f(s) \in T$.

Similarly, we can define $F \colon \mathsf{Set} \to \mathsf{Grp}$ by taking the free group generated by a set S.

Remark 68.2.3 — There is also a notion of "injective" and "surjective" for functors (on arrows) as follows. A functor $F \colon \mathcal{A} \to \mathcal{B}$ is **faithful** (resp. **full**) if for any A_1, A_2, $F \colon \operatorname{Hom}_{\mathcal{A}}(A_1, A_2) \to \operatorname{Hom}_{\mathcal{B}}(FA_1, FA_2)$ is injective (resp. surjective).[a]
We can use this to give an exact definition of concrete category: it's a category with a faithful (forgetful) functor $U \colon \mathcal{A} \to \mathsf{Set}$.

[a] Again, experts might object that $\operatorname{Hom}_{\mathcal{A}}(A_1, A_2)$ or $\operatorname{Hom}_{\mathcal{B}}(FA_1, FA_2)$ may be proper classes instead of sets, but I am assuming everything is locally small.

Example 68.2.4 (Functors from $\mathcal{G}$)

Let G be a group and $\mathcal{G} = \{*\}$ be the associated one-object category.

(a) Consider a functor $F \colon \mathcal{G} \to \mathsf{Set}$, and let $S = F(*)$. Then the data of F corresponds to putting a *group action* of G on S.

(b) Consider a functor $F \colon \mathcal{G} \to \mathsf{FDVect}_k$, and let $V = F(*)$ have dimension n. Then the data of F corresponds to embedding G as a subgroup of the $n \times n$ matrices (i.e. the linear maps $V \to V$). This is one way groups historically arose; the theory of viewing groups as matrices forms the field of representation theory.

(c) Let H be a group and construct $\mathcal{H}$ the same way. Then functors $\mathcal{G} \to \mathcal{H}$ correspond to homomorphisms $G \to H$.

Exercise 68.2.5. Check the above group-based functors work as advertised.

Here's a more involved example. If you find it confusing, skip it and come back after reading about its contravariant version.

Example 68.2.6 (Covariant Yoneda functor)

Fix an $A \in \mathcal{A}$. For a category $\mathcal{A}$, define the **covariant Yoneda functor** $H^A \colon \mathcal{A} \to \mathsf{Set}$ by defining

$$H^A(A_1) \coloneqq \mathrm{Hom}_{\mathcal{A}}(A, A_1) \in \mathsf{Set}.$$

Hence each A_1 is sent to the *arrows from A to A_1*; so H^A **describes how A sees the world**.

Now we want to specify how H^A behaves on arrows. For each arrow $A_1 \xrightarrow{f} A_2$, we need to specify Set-map $\mathrm{Hom}_{\mathcal{A}}(A, A_1) \to \mathrm{Hom}(A, A_2)$; in other words, we need to send an arrow $A \xrightarrow{p} A_1$ to an arrow $A \to A_2$. There's only one reasonable way to do this: take the composition

$$A \xrightarrow{p} A_1 \xrightarrow{f} A_2.$$

In other words, $H_A(f)$ is $p \mapsto f \circ p$. In still other words, $H_A(f) = f \circ -$; the $-$ is a slot for the input to go into.

As another example:

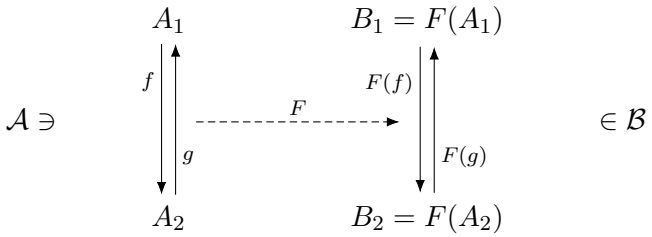
Question 68.2.7. If $\mathcal{P}$ and $\mathcal{Q}$ are posets interpreted as categories, what does a functor from $\mathcal{P}$ to $\mathcal{Q}$ represent?

Now, let me explain why we might care. Consider the following "obvious" fact: if G and H are isomorphic groups, then they have the same size. We can formalize it by saying: if $G \cong H$ in Grp and $U \colon \mathsf{Grp} \to \mathsf{Set}$ is the forgetful functor (mapping each group to its underlying set), then $U(G) \cong U(H)$. The beauty of category theory shows itself: this in fact works *for any functors and categories*, and the proof is done solely through arrows:

Theorem 68.2.8 (Functors preserve isomorphism)

If $A_1 \cong A_2$ are isomorphic objects in $\mathcal{A}$ and $F \colon \mathcal{A} \to \mathcal{B}$ is a functor then $F(A_1) \cong F(A_2)$.

Proof. Try it yourself! The picture is:



You'll need to use both key properties of functors: they preserve composition and the identity map. □

This will give us a great intuition in the future, because

(i) Almost every operation we do in our lifetime will be a functor, and

(ii) We now know that functors take isomorphic objects to isomorphic objects.

Thus, we now automatically know that basically any "reasonable" operation we do will preserve isomorphism (where "reasonable" means that it's a functor). This is super convenient in algebraic topology, for example; see Theorem 65.6.2, where we get for free that homotopic spaces have isomorphic fundamental groups.

Remark 68.2.9 — This lets us construct a category Cat whose objects are categories and arrows are functors.

§68.3 Covariant functors as indexed family of objects

Instead of viewing functor as a *function*, sometimes it is more convenient to view a functor as an *object* (or a family of objects).

For sets A and B, sometimes the notation A^B is used to denote the set $\operatorname{Hom}(B, A)$ being the set of all functions from B to A. This notation is natural because, for finite sets A and B, then $|\operatorname{Hom}(B, A)| = |A|^{|B|}$.

That said, the product set $A \times A$ is sometimes also denoted A^2. Is there a relation?

Certainly! We define the set $\mathbf{2} = \{0, 1\}$ (or any set of two elements). Then we have $|\mathbf{2}| = 2$. It is not difficult to see there is a correspondence between A^2 and $\operatorname{Hom}(\mathbf{2}, A)$.

Now, let $\mathcal{A}$ be a category. Define the category $\mathcal{A} \times \mathcal{A} = \mathcal{A}^2$ the obvious way:

- The objects of $\mathcal{A}^2$ are pairs of objects (A_1, A_2) with $A_1, A_2 \in \mathcal{A}$,
- The morphisms are pairs of morphisms...

Exercise 68.3.1. For $X, Y \in \mathsf{Top}$, we can define the product space $X \times Y \in \mathsf{Top}$. This gives a functor $\mathsf{Top}^2 \to \mathsf{Top}$. Verify this. (From a pair of maps $(f, g) : (X_1, Y_1) \to (X_2, Y_2)$ in Top^2, how do we get a map $X_1 \times Y_1 \to X_2 \times Y_2$? Check this map is continuous i.e. it is indeed a morphism in Top.)

Similar to above, each object in $\mathcal{A}^2$ should correspond to some sort of function $f : 2 \to \mathcal{A}$. But a function's codomain must be an object... $\mathcal{A}$ is a category, so f should be a functor!

So we can make a category 2, and we have $F : 2 \to \mathcal{A}$. There is only one reasonable way to define 2 that do what we want:[2]

- The objects are $\{0, 1\}$;
- There is no morphism, except id_0 and id_1.

More generally,

A functor $F : \mathcal{A} \to \mathcal{B}$ can be viewed as an indexed collection of objects $\{B_A \in \mathcal{B}\}_{A \in \mathcal{A}}$.

This can be most easily seen for a presheaf: "a contravariant functor $\operatorname{OpenSets}(X)^{\mathrm{op}} \to \mathsf{Rings}$" means "a family of rings indexed by open sets of X, satisfying certain niceness conditions".

In fact, just as $\mathcal{A}^2$ is a category, the functors $2 \to \mathcal{A}$ also forms a category. We will see this in Section 68.7.

[2]This is **different** from the category $\mathbf{2}$ that we will define later for natural transformation! Be careful.

§68.4 Contravariant functors

Prototypical example for this section: Dual spaces, contravariant Yoneda functor, etc.

Now I have to explain what the opposite categories were doing earlier. In all the previous examples, we took an arrow $A_1 \to A_2$, and it became an arrow $F(A_1) \to F(A_2)$. Sometimes, however, the arrow in fact goes the other way: we get an arrow $F(A_2) \to F(A_1)$ instead. In other words, instead of just getting a functor $\mathcal{A} \to \mathcal{B}$ we ended up with a functor $\mathcal{A}^{\text{op}} \to \mathcal{B}$.

These functors have a name:

Definition 68.4.1. A **contravariant functor** from $\mathcal{A}$ to $\mathcal{B}$ is a functor $F \colon \mathcal{A}^{\text{op}} \to \mathcal{B}$. (Note that we do *not* write "contravariant functor $F \colon \mathcal{A} \to \mathcal{B}$", since that would be confusing; the function notation will always use the correct domain and codomain.)

Pictorially:

$$\begin{array}{ccccc} & A_1 & & B_1 = F(A_1) & \\ \mathcal{A} \ni & \downarrow f & \dashrightarrow^{F} & \uparrow F(f) & \in \mathcal{B} \\ & A_2 & & B_2 = F(A_2) & \end{array}$$

For emphasis, a usual functor is often called a **covariant functor**. (The word "functor" with no adjective always refers to covariant.)

Let's see why this might happen.

Example 68.4.2 ($V \mapsto V^\vee$ is contravariant)

Consider the functor $\mathsf{Vect}_k \to \mathsf{Vect}_k$ by $V \mapsto V^\vee$.

If we were trying to specify a covariant functor, we would need, for every linear map $T \colon V_1 \to V_2$, a linear map $T^\vee \colon V_1^\vee \to V_2^\vee$. But recall that $V_1^\vee = \text{Hom}(V_1, k)$ and $V_2^\vee = \text{Hom}(V_2, k)$: there's no easy way to get an obvious map from left to right.

However, there *is* an obvious map from right to left: given $\xi_2 \colon V_2 \to k$, we can easily give a map from $V_1 \to k$: just compose with T! In other words, there is a very natural map $V_2^\vee \to V_1^\vee$ according to the composition

$$V_1 \xrightarrow{T} V_2 \xrightarrow{\xi_2} k$$

In summary, a map $T \colon V_1 \to V_2$ induces naturally a map $T^\vee \colon V_2^\vee \to V_1^\vee$ in the opposite direction. So the contravariant functor looks like:

$$\begin{array}{ccc} V_1 & & V_1^\vee \\ \downarrow T & \dashrightarrow^{\bullet^\vee} & \uparrow T^\vee \\ V_2 & & V_2^\vee \end{array}$$

We can generalize the example above in any category by replacing the field k with any chosen object $A \in \mathcal{A}$.

Example 68.4.3 (Contravariant Yoneda functor)

The **contravariant Yoneda functor** on $\mathcal{A}$, denoted $H_A \colon \mathcal{A}^{\text{op}} \to \mathsf{Set}$, is used to describe how objects of $\mathcal{A}$ see A. For each $X \in \mathcal{A}$ it puts

$$H_A(X) \coloneqq \text{Hom}_{\mathcal{A}}(X, A) \in \mathsf{Set}.$$

For $X \xrightarrow{f} Y$ in $\mathcal{A}$, the map $H_A(f)$ sends each arrow $Y \xrightarrow{p} A \in \text{Hom}_{\mathcal{A}}(Y, A)$ to

$$X \xrightarrow{f} Y \xrightarrow{p} A \quad \in \text{Hom}_{\mathcal{A}}(X, A)$$

as we did above. Thus $H_A(f)$ is an arrow from $\text{Hom}_{\mathcal{A}}(Y, A) \to \text{Hom}_{\mathcal{A}}(X, A)$. (Note the flipping!)

Exercise 68.4.4. Check now the claim that $\mathcal{A}^{\text{op}} \times \mathcal{A} \to \mathsf{Set}$ by $(A_1, A_2) \mapsto \text{Hom}(A_1, A_2)$ is in fact a functor.

§68.5 Equivalence of categories

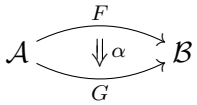
fully faithful and essentially surjective

§68.6 (Optional) Natural transformations

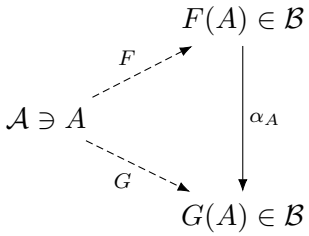
We made categories to keep track of objects and maps, then went a little crazy and asked "what are the maps between categories?" to get functors. Now we'll ask "what are the maps between functors?" to get natural transformations.

It might sound terrifying that we're drawing arrows between functors, but this is actually an old idea. Recall that given two paths $\alpha, \beta \colon [0,1] \to X$, we built a path-homotopy by "continuously deforming" the path α to β; this could be viewed as a function $[0,1] \times [0,1] \to X$. The definition of a natural transformation is similar: we want to pull F to G along a series of arrows in the target space $\mathcal{B}$.

Definition 68.6.1. Let $F, G \colon \mathcal{A} \to \mathcal{B}$ be two functors. A **natural transformation** α from F to G, denoted



consists of, for each $A \in \mathcal{A}$ an arrow $\alpha_A \in \text{Hom}_{\mathcal{B}}(F(A), G(A))$, which is called the **component** of α at A. Pictorially, it looks like this:

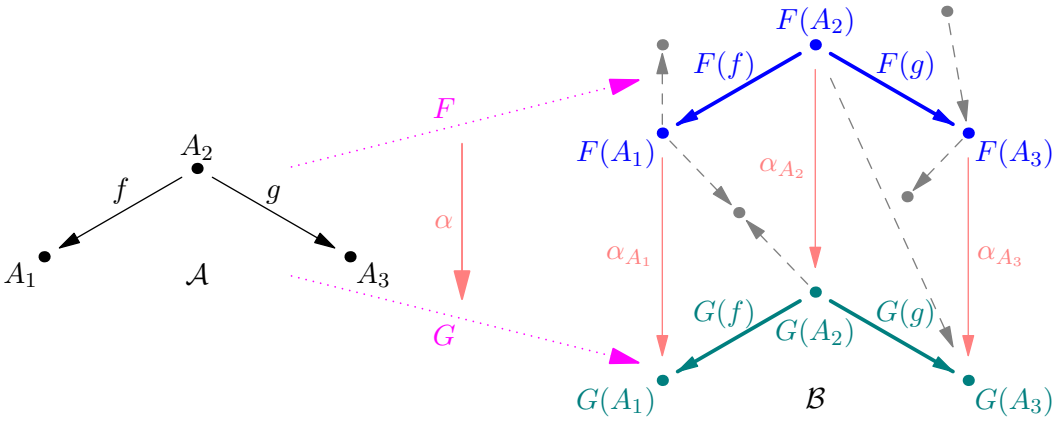


These α_A are subject to the "naturality" requirement that for any $A_1 \xrightarrow{f} A_2$, the diagram

$$\begin{array}{ccc} F(A_1) & \xrightarrow{F(f)} & F(A_2) \\ {\scriptstyle \alpha_{A_1}}\downarrow & & \downarrow{\scriptstyle \alpha_{A_2}} \\ G(A_1) & \xrightarrow[G(f)]{} & G(A_2) \end{array}$$

commutes.

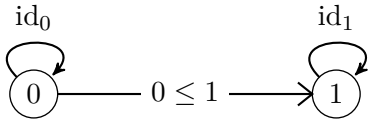
The arrow α_A represents the path that $F(A)$ takes to get to $G(A)$ (just as in a path-homotopy from α to β each *point* $\alpha(t)$ gets deformed to the *point* $\beta(t)$ continuously). A picture might help: consider



Here $\mathcal{A}$ is the small category with three elements and two non-identity arrows f, g (I've omitted the identity arrows for simplicity). The images of $\mathcal{A}$ under F and G are the blue and green "subcategories" of $\mathcal{B}$. Note that $\mathcal{B}$ could potentially have many more objects and arrows in it (grey). The natural transformation α (red) selects an arrow of $\mathcal{B}$ from each $F(A)$ to the corresponding $G(A)$, dragging the entire image of F to the image of G. Finally, we require that any diagram formed by the blue, red, and green arrows is commutative (naturality), so the natural transformation is really "natural".

There is a second equivalent definition that looks much more like the homotopy.

Definition 68.6.2. Let $\mathbf{2}$ denote the category generated by a poset with two elements $0 \le 1$, that is,



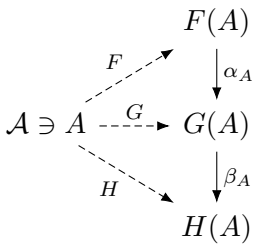
Then a *natural transformation* $\mathcal{A} \overset{F}{\underset{G}{\Downarrow \alpha}} \mathcal{B}$ is just a functor $\alpha \colon \mathcal{A} \times \mathbf{2} \to \mathcal{B}$ satisfying

$$\alpha(A,0) = F(A), \quad \alpha(f,0) = F(f) \quad \text{and} \quad \alpha(A,1) = G(A), \quad \alpha(f,1) = G(f).$$

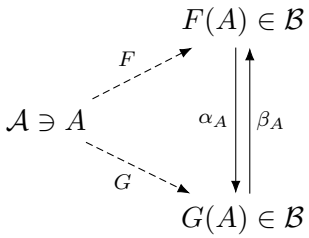
More succinctly, $\alpha(-,0) = F$, $\alpha(-,1) = G$.

The proof that these are equivalent is left as a practice problem.

Naturally, two natural transformations $\alpha \colon F \to G$ and $\beta \colon G \to H$ can get composed.



Now suppose α is a natural transformation such that α_A is an isomorphism for each A. In this way, we can construct an inverse arrow β_A to it.



In this case, we say α is a **natural isomorphism**. We can then say that $F(A) \cong G(A)$ **naturally** in A. (And β is an isomorphism too!) This means that the functors F and G are "really the same": not only are they isomorphic on the level of objects, but these isomorphisms are "natural". As a result of this, we also write $F \cong G$ to mean that the functors are naturally isomorphic.

This is what it really means when we say that "there is a natural / canonical isomorphism". For example, I claimed earlier (in Problem 15A*) that there was a canonical isomorphism $(V^\vee)^\vee \cong V$, and mumbled something about "not having to pick a basis" and "God-given". Category theory, amazingly, lets us formalize this: it just says that $(V^\vee)^\vee \cong \mathrm{id}(V)$ naturally in $V \in \mathsf{FDVect}_k$. Really, we have a natural transformation

$$\mathsf{FDVect}_k \underset{(\bullet^\vee)^\vee}{\overset{\mathrm{id}}{\Downarrow \varepsilon}} \mathsf{FDVect}_k \ .$$

where the component ε_V is given by $v \mapsto \mathrm{ev}_v$ (as discussed earlier, the fact that it is an isomorphism follows from the fact that V and $(V^\vee)^\vee$ have equal dimensions and ε_V is injective).

Another example can be found in Remark 71.2.8.

§68.7 (Optional) The Yoneda lemma

Now that I have natural transformations, I can define:

Definition 68.7.1. The **functor category** of two categories $\mathcal{A}$ and $\mathcal{B}$, denoted $[\mathcal{A}, \mathcal{B}]$, is defined as follows:

- The objects of $[\mathcal{A}, \mathcal{B}]$ are (covariant) functors $F \colon \mathcal{A} \to \mathcal{B}$, and
- The morphisms are natural transformations $\alpha \colon F \to G$.

Question 68.7.2. When are two objects in the functor category isomorphic?

With this, I can make good on the last example I mentioned at the beginning:

Exercise 68.7.3. Construct the following functors:

- $\mathcal{A} \to [\mathcal{A}^{\text{op}}, \mathsf{Set}]$ by $A \mapsto H_A$, which we call $H_\bullet$.
- $\mathcal{A}^{\text{op}} \to [\mathcal{A}, \mathsf{Set}]$ by $A \mapsto H^A$, which we call $H^\bullet$.

Notice that we have opposite categories either way; even if you like H^A because it is covariant, the map $H^\bullet$ is contravariant. So for what follows, we'll prefer to use $H_\bullet$.

The main observation now is that given a category $\mathcal{A}$, $H_\bullet$ provides some *special* functors $\mathcal{A}^{\text{op}} \to \mathsf{Set}$ which are already "built" in to the category A. In light of this, we define:

Definition 68.7.4. A **presheaf** X is just a contravariant functor $\mathcal{A}^{\text{op}} \to \mathsf{Set}$. It is called **representable** if $X \cong H_A$ for some A.

In other words, when we think about representable, the question we're asking is:

> *What kind of presheaves are already "built in" to the category $\mathcal{A}$?*

One way to get at this question is: given a presheaf X and a particular H_A, we can look at the *set* of natural transformations $\alpha \colon X \implies H_A$, and see if we can learn anything about it. In fact, this set can be written explicitly:

Theorem 68.7.5 (Yoneda lemma)

Let $\mathcal{A}$ be a category, pick $A \in \mathcal{A}$, and let H_A be the contravariant Yoneda functor. Let $X \colon \mathcal{A}^{\text{op}} \to \mathsf{Set}$ be a contravariant functor. Then the map

$$\left\{ \text{Natural transformations } \mathcal{A}^{\text{op}} \underset{X}{\overset{H_A}{\Downarrow \alpha}} \mathsf{Set} \right\} \to X(A)$$

defined by $\alpha \mapsto \alpha_A(\text{id}_A) \in X(A)$ is an isomorphism of Set (i.e. a bijection). Moreover, if we view both sides of the equality as functors

$$\mathcal{A}^{\text{op}} \times [\mathcal{A}^{\text{op}}, \mathsf{Set}] \to \mathsf{Set}$$

then this isomorphism is natural.

This might be startling at first sight. Here's an unsatisfying explanation why this might not be too crazy: in category theory, a rule of thumb is that "two objects of the same type that are built naturally are probably the same". You can see this theme when we defined functors and natural transformations, and even just compositions. Now to look at the set of natural transformations, we took a pair of elements $A \in \mathcal{A}$ and $X \in [\mathcal{A}^{\text{op}}, \mathsf{Set}]$ and constructed a *set* of natural transformations. Is there another way we can get a set from these two pieces of information? Yes: just look at $X(A)$. The Yoneda lemma is telling us that our heuristic still holds true here.

Some consequences of the Yoneda lemma are recorded in [Le14]. Since this chapter is already a bit too long, I'll just write down the statements, and refer you to [Le14] for the proofs.

1. As we mentioned before, $H^\bullet$ provides a functor
$$\mathcal{A} \to [\mathcal{A}^{\text{op}}, \mathsf{Set}].$$

It turns out this functor is in fact *fully faithful*; it quite literally embeds the category $\mathcal{A}$ into the functor category on the right (much like Cayley's theorem embeds every group into a permutation group).

2. If $X, Y \in \mathcal{A}$ then
$$H_X \cong H_Y \iff X \cong Y \iff H^X \cong H^Y.$$
To see why this is expected, consider $\mathcal{A} = \mathsf{Grp}$ for concreteness. Suppose A, X, Y are groups such that $H_X(A) \cong H_Y(A)$ for all A. For example,
   - If $A = \mathbb{Z}$, then $|X| = |Y|$.
   - If $A = \mathbb{Z}/2\mathbb{Z}$, then X and Y have the same number of elements of order 2.
   - ...

   Each A gives us some information on how X and Y are similar, but the whole natural isomorphism is strong enough to imply $X \cong Y$.

3. Consider the covariant forgetful functor $U : \mathsf{Grp} \to \mathsf{Set}$.[3] It can be represented by $H^{\mathbb{Z}}$, in the sense that
$$\operatorname{Hom}_{\mathsf{Grp}}(\mathbb{Z}, G) \cong U(G) \qquad \text{by} \qquad \phi \mapsto \phi(1).$$
That is, elements of G are in bijection with maps $\mathbb{Z} \to G$, determined by the image of $+1$ (or -1 if you prefer). So a representation of U was determined by looking at $\mathbb{Z}$ and picking $+1 \in U(\mathbb{Z})$.

   The generalization of this is a follows: let $\mathcal{A}$ be a category and $X : \mathcal{A} \to \mathsf{Set}$ a covariant functor. Then a representation $H^A \cong X$ consists of an object $A \in \mathcal{A}$ and an element $u \in X(A)$ satisfying a certain condition. You can read this off the condition[4] if you know what the inverse map is in Theorem 68.7.5. In the above situation, $X = U$, $A = \mathbb{Z}$ and $u = \pm 1$.

§68.8 A few harder problems to think about

Problem 68A. Show that the two definitions of natural transformation (one in terms of $\mathcal{A} \times \mathbf{2} \to \mathcal{B}$ and one in terms of arrows $F(A) \xrightarrow{\alpha_A} G(A)$) are equivalent.

Problem 68B. Let $\mathcal{A}$ be the category of finite sets whose arrows are bijections between sets. For $A \in \mathcal{A}$, let $F(A)$ be the set of *permutations* of A and let $G(A)$ be the set of *orderings* on A.[5]

(a) Extend F and G to functors $\mathcal{A} \to \mathsf{Set}$.

(b) Show that $F(A) \cong G(A)$ for every A, but this isomorphism is *not* natural.

Problem 68C (Proving the Yoneda lemma)**.** In the context of Theorem 68.7.5:

(a) Prove that the map described is in fact a bijection. (To do this, you will probably have to explicitly write down the inverse map.)

(b) Prove that the bijection is indeed natural. (This is long-winded, but not difficult; from start to finish, there is only one thing you can possibly do.)

[3] Actually, you need to apply a dual version. Theorem 68.7.5 uses contravariant functor.

[4] Just for completeness, the condition is: For all $A' \in \mathcal{A}$ and $x \in X(A')$, there's a unique $f : A \to A'$ with $(Xf)(u) = x$.

[5] A permutation is a bijection $A \to A$, and an ordering is a bijection $\{1, \dots, n\} \to A$, where n is the size of A.

69 Limits in categories (TO DO)

We saw near the start of our category theory chapter the nice construction of products by drawing a bunch of arrows. It turns out that this concept can be generalized immensely, and I want to give a you taste of that here.

write introduction

To run this chapter, we follow the approach of [Le14].

§69.1 Equalizers

Prototypical example for this section: The equalizer of $f, g \colon X \to Y$ is the set of points with $f(x) = g(x)$.

Given two sets X and Y, and maps $X \xrightarrow{f,g} Y$, we define their **equalizer** to be

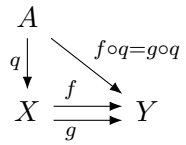
$$\{x \in X \mid f(x) = g(x)\}.$$

We would like a categorical way of defining this, too.

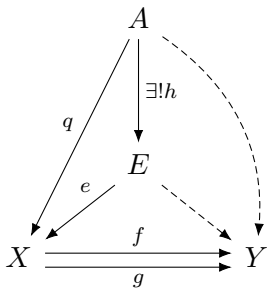
Consider two objects X and Y with two maps f and g between them. Stealing a page from [Le14], we call this a **fork**:

$$X \underset{g}{\overset{f}{\rightrightarrows}} Y$$

A cone over this fork is an object A and arrows over X and Y which make the diagram commute, like so.



Effectively, the arrow over Y is just forcing $f \circ q = g \circ q$. In any case, the **equalizer** of f and g is a "universal cone" over this fork: it is an object E and a map $E \xrightarrow{e} X$ such that for each $A \xrightarrow{q} X$ the diagram



commutes for a unique $A \xrightarrow{h} E$. In other words, any map $A \xrightarrow{q} X$ as above must factor uniquely through E. Again, the dotted arrows can be omitted, and as before equalizers may not exist. But when they do exist:

Exercise 69.1.1. If $E \xrightarrow{e} X$ and $E' \xrightarrow{e'} X$ are equalizers, show that $E \cong E'$.

Example 69.1.2 (Examples of equalizers)

(a) In Set, given $X \xrightarrow{f,g} Y$ the equalizer E can be realized as $E = \{x \mid f(x) = g(x)\}$, with the inclusion $e \colon E \hookrightarrow X$ as the morphism. As usual, by abuse we'll often just refer to E as the equalizer.

(b) Ditto in Top, Grp. One has to check that the appropriate structures are preserved (e.g. one should check that $\{\phi(g) = \psi(g) \mid g \in G\}$ is a group).

(c) In particular, given a homomorphism $\phi \colon G \to H$, the inclusion $\ker \phi \hookrightarrow G$ is an equalizer for the fork $G \to H$ by ϕ and the trivial homomorphism.

According to (c) equalizers let us get at the concept of a kernel if there is a distinguished "trivial map", like the trivial homomorphism in Grp. We'll flesh this idea out in the chapter on abelian categories.

§69.2 Pullback squares (TO DO)

write me

Great example: differentiable functions on $(-3, 1)$ and $(-1, 3)$

Example 69.2.1

§69.3 Limits

We've defined cones over discrete sets of X_i and over forks. It turns out you can also define a cone over any general **diagram** of objects and arrows; we specify a projection from A to each object and require that the projections from A commute with the arrows in the diagram. (For example, a cone over a fork is a diagram with two edges and two arrows.) If you then demand the cone be universal, you have the extremely general definition of a **limit**. As always, these are unique up to unique isomorphism. We can also define the dual notion of a **colimit** in the same way.

§69.4 A few harder problems to think about

Problem 69A* (Equalizers are monic)**.** Show that the equalizer of any fork is monic.

pushout square gives tenor product
p-adic
relative Chinese remainder theorem!!

70 Abelian categories

In this chapter I'll translate some more familiar concepts into categorical language; this will require some additional assumptions about our category, culminating in the definition of a so-called "abelian category". Once that's done, I'll be able to tell you what this "diagram chasing" thing is all about.

Throughout this chapter, "$\hookrightarrow$" will be used for monic maps and "$\twoheadrightarrow$" for epic maps.

§70.1 Zero objects, kernels, cokernels, and images

Prototypical example for this section: In Grp*, the trivial group and homomorphism are the zero objects and morphisms. If G, H are abelian then the cokernel of $\phi\colon G \to H$ is $H/\operatorname{im}\phi$.*

A **zero object** of a category is an object 0 which is both initial and terminal; of course, it's unique up to unique isomorphism. For example, in Grp the zero object is the trivial group, in Vect_k it's the zero-dimensional vector space consisting of one point, and so on.

> **Question 70.1.1.** Show that Set and Top don't have zero objects.

For the rest of this chapter, all categories will have zero objects.

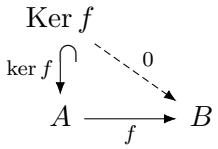
In a category $\mathcal{A}$ with zero objects, any two objects A and B thus have a distinguished morphism

$$A \to 0 \to B$$

which is called the **zero morphism** and also denoted 0. For example, in Grp this is the trivial homomorphism.

We can now define:

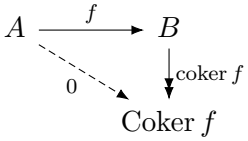
Definition 70.1.2. Consider a map $A \xrightarrow{f} B$. The **kernel** is defined as the equalizer of this map and the map $A \xrightarrow{0} B$. Thus, it's a map $\ker f\colon \operatorname{Ker} f \hookrightarrow A$ such that



commutes, and moreover any other map with the same property factors uniquely through $\operatorname{Ker} A$ (so it is universal with this property). By Problem 69A*, $\ker f$ is a monic morphism, which justifies the use of "$\hookrightarrow$".

Notice that we're using $\ker f$ to represent the map and $\operatorname{Ker} f$ to represent the object Similarly, we define the cokernel, the dual notion:

Definition 70.1.3. Consider a map $A \xrightarrow{f} B$. The **cokernel** of f is a map $\operatorname{coker} f\colon B \twoheadrightarrow \operatorname{Coker} f$ such that



commutes, and moreover any other map with the same property factors uniquely through $\operatorname{Coker} f$ (so it is universal with this property). Thus it is the "coequalizer" of this map and the map $A \xrightarrow{0} B$. By the dual of Problem 69A*, $\operatorname{coker} f$ is an epic morphism, which justifies the use of "$\twoheadrightarrow$".

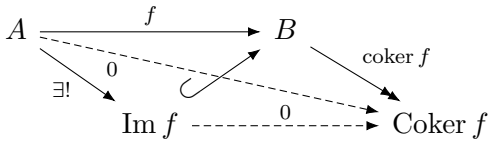
Think of the cokernel of a map $A \xrightarrow{f} B$ as "B modulo the image of f".

Example 70.1.4 (Cokernels)
Consider the map $\mathbb{Z}/6\mathbb{Z} \to D_{12} = \langle r, s \mid r^6 = s^2 = 1, rs = sr^{-1} \rangle$. Then the cokernel of this map in Grp is $D_{12}/\langle r \rangle \cong \mathbb{Z}/2\mathbb{Z}$.

This doesn't always work out quite the way we want since in general the image of a homomorphism need not be normal in the codomain. Nonetheless, we can use this to define:

Definition 70.1.5. The **image** of $A \xrightarrow{f} B$ is the kernel of $\operatorname{coker} f$. We denote $\operatorname{Im} f = \operatorname{Ker}(\operatorname{coker} f)$. This gives a unique map $\operatorname{im} f \colon A \to \operatorname{Im} f$.

When it exists, this coincides with our concrete notion of "image". Picture:



Note that by universality of $\operatorname{Im} f$, we find that there is a unique map $\operatorname{im} f \colon A \to \operatorname{Im} f$ that makes the entire diagram commute.

§70.2 Additive and abelian categories

Prototypical example for this section: Ab, Vect_k, *or more generally* Mod_R.

We can now define the notion of an additive and abelian category, which are the types of categories where this notion is most useful.

Definition 70.2.1. An **additive category** $\mathcal{A}$ is one such that:

- $\mathcal{A}$ has a zero object, and any two objects have a product.
- More importantly: every $\operatorname{Hom}_{\mathcal{A}}(A, B)$ forms an *abelian group* (written additively) such that composition distributes over addition:

$$(g + h) \circ f = g \circ f + h \circ f \quad \text{and} \quad f \circ (g + h) = f \circ g + f \circ h.$$

  The zero map serves as the identity element for each group.

In short:

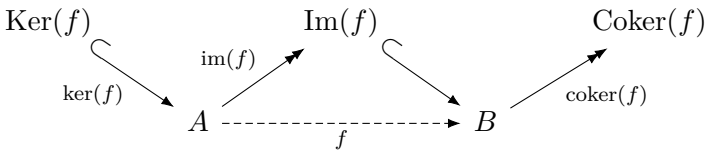
In an additive category, you can add two morphisms.

Which is the only definition that makes sense anyway, we cannot talk about elements.

Definition 70.2.2. An **abelian category** $\mathcal{A}$ is one with the additional properties that for any morphism $A \xrightarrow{f} B$,

- The kernel and cokernel exist, and
- The morphism factors through the image so that $\text{im}(f)$ is epic.

So, this yields a diagram



Example 70.2.3 (Examples of abelian categories)

(a) Vect_k, Ab are abelian categories, where $f+g$ takes its usual meaning.

(b) Generalizing this, the category Mod_R of R-modules is abelian.

(c) Grp is not even additive, because there is no way to assign a commutative addition to pairs of morphisms.

From now on, you can basically forget about additive category, we will be working in abelian category.

In general, once you assume a category is abelian, all the properties you would want of these kernels, cokernels, ... that you would guess hold true. For example,

Proposition 70.2.4 (Monic $\iff$ trivial kernel)

A map $A \xrightarrow{f} B$ is monic if and only if its kernel is $0 \to A$. Dually, $A \xrightarrow{f} B$ is epic if and only if its cokernel is $B \to 0$.

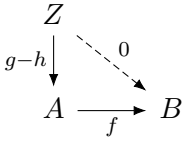
Proof. The easy direction is:

Exercise 70.2.5. Show that if $A \xrightarrow{f} B$ is monic, then $0 \to A$ is a kernel. (This holds even in non-abelian categories.)

Of course, since kernels are unique up to isomorphism, monic $\implies$ 0 kernel. On the other hand, assume that $0 \to A$ is a kernel of $A \xrightarrow{f} B$. For this we can exploit the group structure of the underlying homomorphisms now. Assume the diagram

$$Z \underset{h}{\overset{g}{\rightrightarrows}} A \xrightarrow{f} B$$

commutes. Then $(g-h) \circ f = g \circ f - h \circ f = 0$, and we've arrived at a commutative diagram.



But since $0 \to A$ is a kernel it follows that $g-h$ factors through 0, so $g-h=0 \implies g=h$, which is to say that f is monic. □

Proposition 70.2.6 (Isomorphism $\iff$ monic and epic)
In an abelian category, a map is an isomorphism if and only if it is monic and epic.

Proof. Omitted. (The Mitchell embedding theorem presented later implies this anyways for most situations we care about, by looking at a small sub-category.) □

§70.3 Exact sequences

Prototypical example for this section: $0 \to G \to G \times H \to H \to 0$ *is exact.*

Exact sequences will seem exceedingly unmotivated until you learn about homology groups, which is one of the most natural places that exact sequences appear. In light of this, it might be worth trying to read the chapter on homology groups simultaneously with this one.

First, let me state the definition for groups, to motivate the general categorical definition. A sequence of groups

$$G_0 \xrightarrow{f_1} G_1 \xrightarrow{f_2} G_2 \xrightarrow{f_3} \dots \xrightarrow{f_n} G_n$$

is *exact* at G_k if the image of f_k is the kernel of f_{k+1}. We say the entire sequence is exact if it's exact at $k = 1, \dots, n-1$.

Example 70.3.1 (Exact sequences)

(a) The sequence
$$0 \to \mathbb{Z}/3\mathbb{Z} \xhookrightarrow{\times 5} \mathbb{Z}/15\mathbb{Z} \twoheadrightarrow \mathbb{Z}/5\mathbb{Z} \to 0$$
is exact. Actually, $0 \to G \hookrightarrow G \times H \twoheadrightarrow H \to 0$ is exact in general. (Here 0 denotes the trivial group.)

(b) For groups, the map $0 \to A \to B$ is exact if and only if $A \to B$ is injective.

(c) For groups, the map $A \to B \to 0$ is exact if and only if $A \to B$ is surjective.

If you look at the prototypical example, actually, a **short exact sequence** (an exact sequence of the form $0 \to A \to B \to C \to 0$) is the most natural things ever:

It's basically just an equation $C = B/A$.

Whenever you see "there is a short exact sequence $0 \to A \to B \to C \to 0$", you can mentally translate it to "$C \cong B/A$"; but there's a slight difference: A group has more structures than a number, so the sequence also contains the information of the maps — the map that identifies A with a subgroup of B, and the map that identifies C with the quotient group B/A.

Example 70.3.2 (More exact sequences)

(a) The sequence

$$0 \to \mathbb{Z} \xrightarrow{\times 3} \mathbb{Z} \to \mathbb{Z}/3\mathbb{Z} \to 0$$

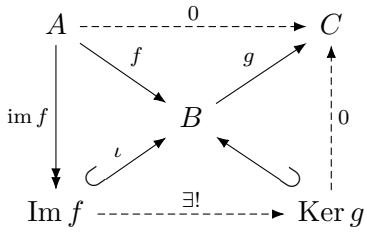
is short exact.

(b) So is

$$0 \to \mathbb{Z} \xrightarrow{\times 5} \mathbb{Z} \to \mathbb{Z}/5\mathbb{Z} \to 0.$$

As you can see, the written equation "$C \cong B/A$" is not completely accurate, the map $A \to B$ also matters in determining what C is. This also explains the common notation: the image of the map $\mathbb{Z} \xrightarrow{\times 3} \mathbb{Z}$ is usually written $3\mathbb{Z}$, thus $\mathbb{Z}/3\mathbb{Z} = \frac{\mathbb{Z}}{3\mathbb{Z}}$.

Now, we want to mimic this definition in a general *abelian* category $\mathcal{A}$. So, let's write down a criterion for when $A \xrightarrow{f} B \xrightarrow{g} C$ is exact. First, we had better have that $g \circ f = 0$, which encodes the fact that $\operatorname{im}(f) \subseteq \ker(g)$. Adding in all the relevant objects, we get the commutative diagram below.



Here the map $A \twoheadrightarrow \operatorname{Im} f$ is epic since we are assuming $\mathcal{A}$ is an abelian category. So, we have that

$$0 = (g \circ \iota) \circ \operatorname{im} f = g \circ (\iota \circ \operatorname{im} f) = g \circ f = 0$$

but since $\operatorname{im} f$ is epic, this means that $g \circ \iota = 0$. So there is a *unique* map $\operatorname{Im} f \to \operatorname{Ker} g$, and we require that this diagram commutes. In short,

Definition 70.3.3. Let $\mathcal{A}$ be an abelian category. The sequence

$$\dots \to A_{n-1} \xrightarrow{f_n} A_n \xrightarrow{f_{n+1}} A_{n+1} \to \dots$$

is **exact** at A_n if $f_n \circ f_{n+1} = 0$ and the canonical map $\operatorname{Im} f_n \to \operatorname{Ker} f_{n+1}$ is an isomorphism. The entire sequence is exact if it is exact at each A_i. (For finite sequences we don't impose condition on the very first and very last object.)

Exercise 70.3.4. Show that, as before, $0 \to A \to B$ is exact $\iff$ $A \to B$ is monic.

§70.4 The Freyd-Mitchell embedding theorem

We now introduce the Freyd-Mitchell embedding theorem, which essentially says that any abelian category can be realized as a concrete one.

Definition 70.4.1. A category is **small** if $\operatorname{obj}(\mathcal{A})$ is a set (as opposed to a class), i.e. there is a "set of all objects in $\mathcal{A}$". For example, Set is not small because there is no set of all sets.

Theorem 70.4.2 (Freyd-Mitchell embedding theorem)

Let $\mathcal{A}$ be a small abelian category. Then there exists a ring R (with 1 but possibly non-commutative) and a full, faithful, exact functor onto the category of left R-modules.

Here a functor is **exact** if it preserves exact sequences. This theorem is good because it means

You can basically forget about all the weird definitions that work in any abelian category.

Any time you're faced with a statement about an abelian category, it suffices to just prove it for a "concrete" category where injective/surjective/kernel/image/exact/etc. agree with your previous notions. A proof by this means is sometimes called *diagram chasing*.

Remark 70.4.3 — The "small" condition is a technical obstruction that requires the objects $\mathcal{A}$ to actually form a set. I'll ignore this distinction, because one can almost always work around it by doing enough set-theoretic technicalities.

For example, let's prove:

Lemma 70.4.4 (Short five lemma)

In an abelian category, consider the commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \overset{p}{\hookrightarrow} & B & \overset{q}{\twoheadrightarrow} & C & \longrightarrow & 0 \\ & & \cong\downarrow{\scriptstyle\alpha} & & \downarrow{\scriptstyle\beta} & & \cong\downarrow{\scriptstyle\gamma} & & \\ 0 & \longrightarrow & A' & \underset{p'}{\hookrightarrow} & B' & \underset{q'}{\twoheadrightarrow} & C' & \longrightarrow & 0 \end{array}$$

and assume the top and bottom rows are exact. If α and γ are isomorphisms, then so is β.

Proof. We prove that β is epic (with a similar proof to get monic). By the embedding theorem we can treat the category as R-modules over some R. This lets us do a so-called "diagram chase" where we move elements around the picture, using the concrete interpretation of our category as R-modules.

Let b' be an element of B'. Then $q'(b') \in C'$, and since γ is surjective, we have a c such that $\gamma(c) = b'$, and finally a $b \in B$ such that $q(b) = c$. Picture:

$$\begin{array}{ccc} b \in B & \overset{q}{\longmapsto} & c \in C \\ \beta\downarrow & & \cong\downarrow\gamma \\ b' \in B' & \overset{q'}{\longmapsto} & c' \in C' \end{array}$$

Now, it is not necessarily the case that $\beta(b) = b'$. However, since the diagram commutes we at least have that

$$q'(b') = q'(\beta(b))$$

so $b' - \beta(b) \in \operatorname{Ker} q' = \operatorname{Im} p'$, and there is an $a' \in A'$ such that $p'(a') = b' - \beta(b)$; use α now to lift it to $a \in A$. Picture:

$$\begin{array}{ccccc} a \in A & & b \in B & & \\ \downarrow & & & & \\ a' \in A' & \longmapsto & b' - \beta(b) \in B' & \longmapsto & 0 \in C \end{array}$$

Then, we have

$$\beta(b + q(a)) = \beta b + \beta p a = \beta b + p' \alpha a = \beta b + (b' - \beta b) = b'$$

so $b' \in \operatorname{Im} \beta$ which completes the proof that β' is surjective. □

§70.5 Breaking long exact sequences

Prototypical example for this section: First isomorphism theorem.

In fact, it turns out that any exact sequence breaks into short exact sequences. This relies on:

Proposition 70.5.1 ("First isomorphism theorem" in abelian categories)

Let $A \xrightarrow{f} B$ be an arrow of an abelian category. Then there is an exact sequence

$$0 \to \operatorname{Ker} f \xrightarrow{\ker f} A \xrightarrow{\operatorname{im} f} \operatorname{Im} f \to 0.$$

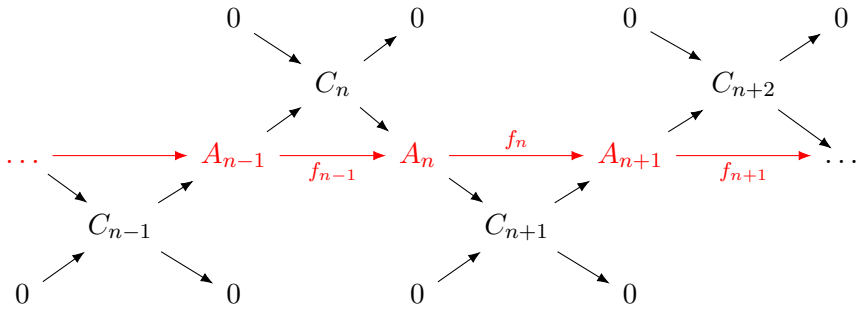
Example 70.5.2

Let's analyze this theorem in our two examples of abelian categories:

(a) In the category of abelian groups, this is basically the first isomorphism theorem.

(b) In the category Vect_k, this amounts to the rank-nullity theorem, Theorem 9.7.7.

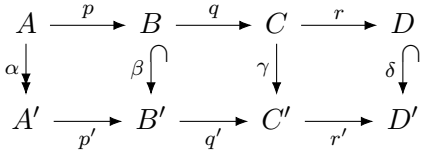
Thus, any exact sequence can be broken into short exact sequences, as



where $C_k = \operatorname{im} f_{k-1} = \ker f_k$ for every k.

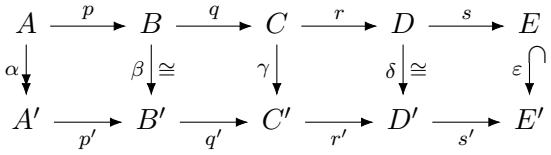
§70.6 A few harder problems to think about

Problem 70A (Four lemma)**.** In an abelian category, consider the commutative diagram



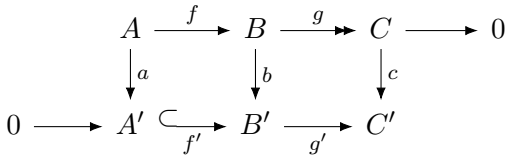
where the first and second rows are exact. Prove that if α is epic, and β and δ are monic, then γ is monic.

Problem 70B (Five lemma)**.** In an abelian category, consider the commutative diagram



where the two rows are exact, β and δ are isomorphisms, α is epic, and ε is monic. Prove that γ is an isomorphism.

Problem 70C* (Snake lemma)**.** In an abelian category, consider the diagram



where the first and second rows are exact sequences. Prove that there is an exact sequence

$$\operatorname{Ker} a \to \operatorname{Ker} b \to \operatorname{Ker} c \to \operatorname{Coker} a \to \operatorname{Coker} b \to \operatorname{Coker} c.$$

Problem 70D (An additive category that is not abelian)**.** Consider a category, where:

- the objects are pairs of abelian groups (B, A) where A is a subgroup of B.
- the morphisms $(B, A) \to (B', A')$ are maps $f : B \to B'$ where $f^{\text{img}}(A) \subseteq A'$.

(You can think of this similar to the PairTop category, seen in Chapter 73. We use abelian groups here to make the category additive.)

This category can be equivalently viewed as the category of short exact sequences $0 \to A \to B \to B/A \to 0$ of abelian groups.

Show that the arrow $(X, 0) \to (X, X)$ is monic and epic, but not an isomorphism. Conclude that the category is not abelian.

XVIII

Algebraic Topology II: Homology

Part XVIII: Contents

71 Singular homology

Now that we've defined $\pi_1(X)$, we turn our attention to a second way of capturing the same idea, $H_1(X)$. We'll then define $H_n(X)$ for $n \geq 2$. The good thing about the H_n groups is that, unlike the π_n groups, they are much easier to compute in practice. The downside is that their definition will require quite a bit of setup, and the "algebraic" part of "algebraic topology" will become a lot more technical.

§71.1 Simplices and boundaries

Prototypical example for this section: $\partial[v_0, v_1, v_2] = [v_0, v_1] - [v_0, v_2] + [v_1, v_2]$.

First things first:

Definition 71.1.1. The **standard n-simplex**, denoted Δ^n, is defined as

$$\{(x_0, x_1, \ldots x_n) \mid x_i \geq 0, x_0 + \cdots + x_n = 1\}.$$

Hence it's the convex hull of some vertices $[v_0, \ldots, v_n]$. Note that we keep track of the order $v_0, \ldots, v_n$ of the vertices, for reasons that will soon become clear.

Given a topological space X, a **singular n-simplex** is a map $\sigma \colon \Delta^n \to X$.

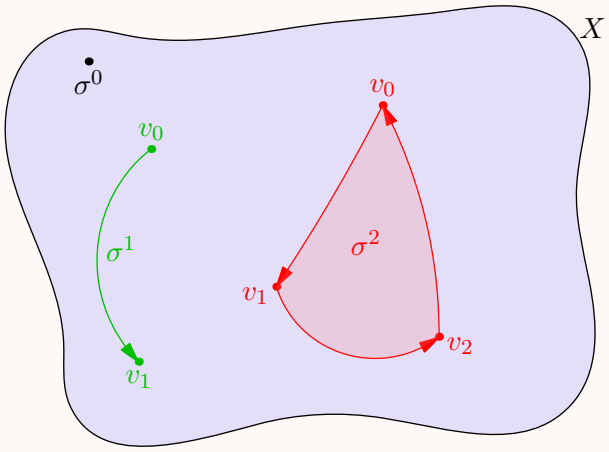
Example 71.1.2 (Singular simplices)

(a) Since $\Delta^0 = [v_0]$ is just a point, a singular 0-simplex X is just a point of X.

(b) Since $\Delta^1 = [v_0, v_1]$ is an interval, a singular 1-simplex X is just a path in X.

(c) Since $\Delta^2 = [v_0, v_1, v_2]$ is an equilateral triangle, a singular 2-simplex X looks a "disk" in X.

Here is a picture of all three in a space X:



The arrows aren't strictly necessary, but I've included them to help keep track of the "order" of the vertices; this will be useful in just a moment.

Now we're going to do something much like when we were talking about Stokes' theorem: we'll put a boundary ∂ operator on the singular n-simplices. This will give us a formal linear sums of n-simplices $\sum_k a_k\sigma_k$, which we call an **n-chain**.

In that case,

Definition 71.1.3. Given a singular n-simplex σ with vertices $[v_0, \dots, v_n]$, note that for every i we have an $(n-1)$ simplex $[v_0, \dots, v_{i-1}, v_{i+1}, \dots, v_n]$. The **boundary operator** ∂ is then defined by

$$\partial(\sigma) \coloneqq \sum_i (-1)^i [v_0, \dots, v_{i-1}, v_{i+1}, \dots, v_n].$$

The boundary operator then extends linearly to n-chains:

$$\partial\left(\sum_k a_k\sigma_k\right) \coloneqq \sum a_k \partial(\sigma_k).$$

By convention, a 0-chain has empty boundary.

Example 71.1.4 (Boundary operator)
Consider the chains depicted in Example 71.1.2. Then

(a) $\partial\sigma^0 = 0$.

(b) $\partial(\sigma^1) = [v_1] - [v_0]$: it's the "difference" of the 0-chain corresponding to point v_1 and the 0-chain corresponding to point v_0.

(c) $\partial(\sigma^2) = [v_0, v_1] - [v_0, v_2] + [v_1, v_2]$; i.e. one can think of it as the sum of the three oriented arrows which make up the "sides" of σ^2.

(d) Notice that if we take the boundary again, we get

$$\begin{aligned}\partial(\partial(\sigma^2)) &= \partial([v_0, v_1]) - \partial([v_0, v_2]) + \partial([v_1, v_2]) \\ &= ([v_1] - [v_0]) - ([v_2] - [v_0]) + ([v_2] - [v_1]) \\ &= 0.\end{aligned}$$

The fact that $\partial^2 = 0$ is of course not a coincidence.

Theorem 71.1.5 ($\partial^2 = 0$)
For any chain c, $\partial(\partial(c)) = 0$.

Proof. Essentially identical to Problem 45B: this is just a matter of writing down a bunch of $\sum$ signs. Diligent readers are welcome to try the computation. □

Remark 71.1.6 — The eerie similarity between the chains used to integrate differential forms and the chains in homology is not a coincidence. The de Rham cohomology, discussed much later, will make the relation explicit.

§71.2 The singular homology groups

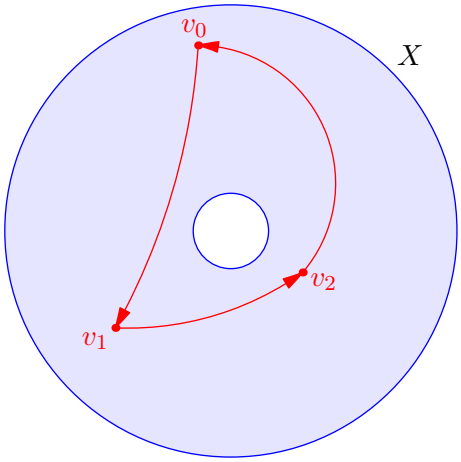
Prototypical example for this section: Probably $H_n(S^m)$, especially the case $m = n = 1$.

Let X be a topological space, and let $C_n(X)$ be the free abelian group of n-chains of X that we defined earlier. Our work above gives us a boundary operator ∂, so we have a sequence of maps

$$\ldots \xrightarrow{\partial} C_3(X) \xrightarrow{\partial} C_2(X) \xrightarrow{\partial} C_1(X) \xrightarrow{\partial} C_0(X) \xrightarrow{\partial} 0$$

(here I'm using 0 to for the trivial group, which is standard notation for abelian groups.) We'll call this the **singular chain complex**.

Now, how does this let us detect holes in the space? To see why, let's consider an annulus, with a 1-chain c drawn in red:



Notice that

$$\partial c = ([v_1] - [v_0]) - ([v_2] - [v_0]) + ([v_2] - [v_1]) = 0$$

and so we can say this 1-chain c is a "cycle", because it has trivial boundary. However, c is not itself the boundary of any 2-chain, because of the hole in the center of the space — it's impossible to "fill in" the interior of c! So, we have detected the hole by the algebraic fact that

$$c \in \ker\left(C_1(X) \xrightarrow{\partial} C_0(X)\right) \qquad \text{but} \qquad c \notin \operatorname{im}\left(C_2(X) \xrightarrow{\partial} C_1(X)\right).$$

Indeed, if the hole was not present then this statement would be false.

Remark 71.2.1 — Note that homotopy and homology captures slightly different notion of "holes". For example, let T be a torus. Then, every map $S^2 \to T$ is nulhomotopic so $\pi_2(T)$ is trivial, but, as we will see in Proposition 72.3.6, $H_2(T) \cong \mathbb{Z}$. At least in the case of $n = 1$, then Theorem 71.2.7 states that for any path-connected space X and $x_0 \in X$, then $H_1(X)$ is the abelianization of $\pi_1(X, x_0)$, which is pretty much the best result you can expect — $H_1(X)$ must be abelian, while $\pi_1(X, x_0)$ need not be abelian. Nevertheless, it is still possible that $\pi_1(X, x_0)$ is nontrivial and $H_1(X)$ is trivial — see https://math.stackexchange.com/q/1052414 for an example.

We can capture this idea in any dimension, as follows.

Definition 71.2.2. Let

$$\ldots \xrightarrow{\partial} C_2(X) \xrightarrow{\partial} C_1(X) \xrightarrow{\partial} C_0(X) \xrightarrow{\partial} 0$$

as above. We say that $c \in C_n(X)$ is:

- a **cycle** if $c \in \ker\left(C_n(X) \xrightarrow{\partial} C_{n-1}(X)\right)$, and
- a **boundary** if $c \in \operatorname{im}\left(C_{n+1}(X) \xrightarrow{\partial} C_n(X)\right)$.

Denote the cycles and boundaries by $Z_n(X), B_n(X) \subseteq C_n(X)$, respectively.[1]

> **Question 71.2.3.** Just to get you used to the notation: check that B_n and Z_n are themselves abelian groups, and that $B_n(X) \subseteq Z_n(X) \subseteq C_n(X)$.

The key point is that we can now define:

Definition 71.2.4. The **nth homology group** $H_n(X)$ is defined as

$$H_n(X) \coloneqq Z_n(X)/B_n(X).$$

Example 71.2.5 (The zeroth homology group)
Let's compute $H_0(X)$ for a topological space X. We take $C_0(X)$, which is just formal linear sums of points of X.
First, we consider the kernel of $\partial\colon C_0(X) \to 0$, so the kernel of ∂ is the entire space $C_0(X)$: that is, every point is a "cycle".
Now, what is the boundary? The main idea is that $[b] - [a] = 0$ if and only if there's a 1-chain which connects a to b, i.e. there is a path from a to b. In particular,

$$X \text{ path connected} \implies H_0(X) \cong \mathbb{Z}.$$

More generally, we have

Proposition 71.2.6 (Homology groups split into path-connected components)
If $X = \bigcup_\alpha X_\alpha$ is a decomposition into path-connected components, then we have

$$H_n(X) \cong \bigoplus_\alpha H_n(X_\alpha).$$

In particular, if X has r path-connected components, then $H_0(X) \cong \mathbb{Z}^{\oplus r}$.

(If it's surprising to see $\mathbb{Z}^{\oplus r}$, remember that an abelian group is the same thing as a $\mathbb{Z}$-module, so the notation $G \oplus H$ is customary in place of $G \times H$ when G, H are abelian.)
Now let's investigate the first homology group.

Theorem 71.2.7 (Hurewicz theorem)
Let X be path-connected. Then $H_1(X)$ is the *abelianization* of $\pi_1(X, x_0)$.

We won't prove this but you can see it roughly from the example. The group $H_1(X)$ captures the same information as $\pi_1(X, x_0)$: a cycle (in $Z_1(X)$) corresponds to the same thing as the loops we studied in $\pi_1(X, x_0)$, and the boundaries (in $B_1(X)$, i.e. the things we mod out by) are exactly the nulhomotopic loops in $\pi_1(X, x_0)$. The difference is that $H_1(X)$ allows loops to commute, whereas $\pi_1(X, x_0)$ does not.

[1] We don't use $C_n(X)$ to denote cycles — apart from the obvious reason that the notation is already used, the letter Z comes from the German word "Zyklus".

Remark 71.2.8 (Digression: category theory interpretation) — From this, you can say that there is a Hurewicz map $\pi_1(X, x_0) \xrightarrow{\phi} H_1(X)$ for each (X, x_0).
But there is more than that: this map is *natural*, in the sense that for $h\colon (X, x_0) \to (Y, y_0)$ map of pointed spaces, then

$$\begin{array}{ccc} \pi_1(X, x_0) & \xrightarrow{h_\sharp} & \pi_1(Y, y_0) \\ \downarrow{\scriptstyle \phi} & & \downarrow{\scriptstyle \phi} \\ H_1(X) & \xrightarrow{h_*} & H_1(Y) \end{array}$$

commutes.
In category theory terms, we say that ϕ is a *natural transformation* from π_1 to H_1.
Another way to say this is: we have families of groups

$$\{\pi_1(X, x_0) \mid (X, x_0) \text{ pointed space}\}$$

and

$$\{H_1(X) \mid (X, x_0) \text{ pointed space}\}$$

then the natural transformation ϕ can be seen as a family of homomorphisms

$$\{\phi\colon \pi_1(X, x_0) \to H_1(X) \mid (X, x_0) \text{ pointed space}\}$$

satisfying the naturality conditions.
Of course, the fact that π_1 is a functor means $\{\pi_1(X, x_0) \mid (X, x_0) \text{ pointed space}\}$ is a lot more than a family of groups indexed by pointed spaces, as explained in Theorem 65.6.2.

Example 71.2.9 (The first homology group of the annulus)

To give a concrete example, consider the annulus X above. We found a chain c that wrapped once around the hole of X. The point is that in fact,

$$H_1(X) = \langle c \rangle \cong \mathbb{Z}$$

which is to say the chains c, $2c$, ... are all not the same in $H_1(X)$, but that any other 1-chain is equivalent to one of these. This captures the fact that X is really just S^1.

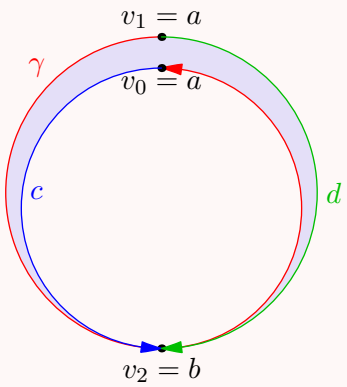
Example 71.2.10 (An explicit boundary in S^1)

In $X = S^1$, let a be the uppermost point and b the lowermost point. Let c be the simplex from a to b along the left half of the circle, and d the simplex from a to b along the right half. Finally, let γ be the simplex which represents a loop γ from a to itself, wrapping once counterclockwise around S^1. We claim that in $H^1(S^1)$ we have

$$\gamma = c - d$$

which geometrically means that $c - d$ represents wrapping once around the circle

(which is of course what we expect).



Indeed this can be seen from the picture above, where we have drawn a 2-simplex whose boundary is exactly $\gamma - c + d$. The picture is somewhat metaphorical: in reality $v_0 = v_1 = a$, and the entire 2-simplex is embedded in S^1. This is why singular homology is so-called: the images of the simplex can sometimes look quite "singular".

Example 71.2.11 (The first homology group of the figure eight)

Consider X_8 (see Example 65.2.9). Both homology and homotopy see the two loops in X_8, call them a and b. The difference is that in $\pi_1(X_8, x_0)$, these two loops are not allowed to commute: we don't have $ab \neq ba$, because the group operation in π_1 is "concatenate paths". But in the homology group $H_1(X)$ the way we add a and b is to add them formally, to get the 1-chain $a + b$. So

$$H_1(X) \cong \mathbb{Z}^{\oplus 2} \quad \text{while} \quad \pi_1(X, x_0) = \langle a, b \rangle .$$

Example 71.2.12 (The homology groups of S^2)

Consider S^2, the two-dimensional sphere. Since it's path connected, we have $H_0(S^2) = \mathbb{Z}$. We also have $H_1(S^2) = 0$, for the same reason that $\pi_1(S^2)$ is trivial as well. On the other hand we claim that

$$H_2(S^2) \cong \mathbb{Z}.$$

The elements of $H_2(S^2)$ correspond to wrapping S^2 in a tetrahedral bag (or two bags, or three bags, etc.). Thus, the second homology group lets us detect the spherical cavity of S^2.[a]

[a] As remarked in Remark 71.2.1, unlike π_2, H_2 also detects other kinds of cavities, not just spherical.

Actually, more generally it turns out that we will have

$$H_n(S^m) \cong \begin{cases} \mathbb{Z} & n = m \text{ or } n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Example 71.2.13 (Contractible spaces)
Given any contractible space X, it turns out that

$$H_n(X) \cong \begin{cases} \mathbb{Z} & n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

The reason is that, like homotopy groups, it turns out that homology groups are homotopy invariant. (We'll prove this next section.) So the homology groups of contractible X are the same as those of a one-point space, which are those above.

Example 71.2.14 (Homology groups of the torus)
While we won't be able to prove it for a while, it turns out that

$$H_n(S^1 \times S^1) \cong \begin{cases} \mathbb{Z} & n = 0, 2 \\ \mathbb{Z}^{\oplus 2} & n = 1 \\ 0 & \text{otherwise.} \end{cases}$$

The homology group at 1 corresponds to our knowledge that $\pi_1(S^1 \times S^1) \cong \mathbb{Z}^2$ and the homology group at 2 detects the "cavity" of the torus.

This is fantastic and all, but how does one go about actually computing any homology groups? This will be a rather long story, and we'll have to do a significant amount of both algebra and geometry before we're really able to compute any homology groups. In what follows, it will often be helpful to keep track of which things are purely algebraic (work for any chain complex), and which parts are actually stating something which is geometrically true.

§71.3 The homology functor and chain complexes

As I mentioned before, the homology groups are homotopy invariant. This will be a similar song and dance as the work we did to create a functor $\pi_1 \colon \mathsf{hTop}_* \to \mathsf{Grp}$. Rather than working slowly and pulling away the curtain to reveal the category theory at the end, we'll instead start with the category theory right from the start just to save some time.

Definition 71.3.1. The category hTop is defined as follows:

- Objects: topological spaces.
- Morphisms: *homotopy classes* of morphisms $X \to Y$.

In particular, X and Y are isomorphic in hTop if and only if they are homotopic.

You'll notice this is the same as hTop_*, except without the basepoints.

Theorem 71.3.2 (Homology is a functor hTop $\to$ Grp)

For any particular n, H_n is a functor $\mathsf{hTop} \to \mathsf{Grp}$. In particular,

- Given any map $f\colon X \to Y$, we get an induced map $f_*\colon H_n(X) \to H_n(Y)$.
- For two homotopic maps $f, g\colon X \to Y$, $f_* = g_*$.
- Two homotopic spaces X and Y have isomorphic homology groups: if $f\colon X \to Y$ is a homotopy then $f_*\colon H_n(X) \to H_n(Y)$ is an isomorphism.
- (Insert your favorite result about functors here.)

In order to do this, we have to describe how to take a map $f\colon X \to Y$ and obtain a map $H_n(f)\colon H_n(X) \to H_n(Y)$. Then we have to show that this map doesn't depend on the choice of homotopy. (This is the analog of the work we did with $f_\sharp$ before.) It turns out that this time around, proving this is much more tricky, and we will have to go back to the chain complex $C_\bullet(X)$ that we built at the beginning.

§71.3.i Algebra of chain complexes

Let's start with the algebra. First, I'll define the following abstraction of the complex to any sequence of abelian groups. Actually, though, it works in any category (not just AbGrp). The strategy is as follows: we'll define everything that we need completely abstractly, then show that the geometry concepts we want correspond to this setting.

Definition 71.3.3. A **chain complex** is a sequence of groups A_n and maps

$$\dots \xrightarrow{\partial} A_{n+1} \xrightarrow{\partial} A_n \xrightarrow{\partial} A_{n-1} \xrightarrow{\partial} \dots$$

such that the composition of any two adjacent maps is the zero morphism. We usually denote this by $A_\bullet$.

The nth homology group $H_n(A_\bullet)$ is defined as $\ker(A_n \to A_{n-1}) / \operatorname{im}(A_{n+1} \to A_n)$. Cycles and boundaries are defined in the same way as before.

Obviously, this is just an algebraic generalization of the structure we previously looked at, rid of all its original geometric context.

Definition 71.3.4. A **morphism of chain complexes** (or chain map) $f\colon A_\bullet \to B_\bullet$ is a sequence of maps f_n for every n such that the diagram

$$\begin{array}{ccccccccc}
\dots & \xrightarrow{\partial_A} & A_{n+1} & \xrightarrow{\partial_A} & A_n & \xrightarrow{\partial_A} & A_{n-1} & \xrightarrow{\partial_A} & \dots \\
 & & \downarrow{\scriptstyle f_{n+1}} & & \downarrow{\scriptstyle f_n} & & \downarrow{\scriptstyle f_{n-1}} & & \\
\dots & \xrightarrow{\partial_B} & B_{n+1} & \xrightarrow{\partial_B} & B_n & \xrightarrow{\partial_B} & B_{n-1} & \xrightarrow{\partial_B} & \dots
\end{array}$$

commutes. Under this definition, the set of chain complexes becomes a category, which we denote Cmplx.

Note that given a morphism of chain complexes $f\colon A_\bullet \to B_\bullet$, every cycle in A_n gets sent to a cycle in B_n, since the square

$$\begin{array}{ccc}
A_n & \xrightarrow{\partial_A} & A_{n-1} \\
{\scriptstyle f_n}\downarrow & & \downarrow{\scriptstyle f_{n-1}} \\
B_n & \xrightarrow[\partial_B]{} & B_{n-1}
\end{array}$$

commutes. Similarly, every boundary in A_n gets sent to a boundary in B_n. Thus,

Every map of $f\colon A_\bullet \to B_\bullet$ gives a map $f_*\colon H_n(A) \to H_n(B)$ for every n.

Exercise 71.3.5. Interpret H_n as a functor $\mathsf{Cmplx} \to \mathsf{Grp}$.

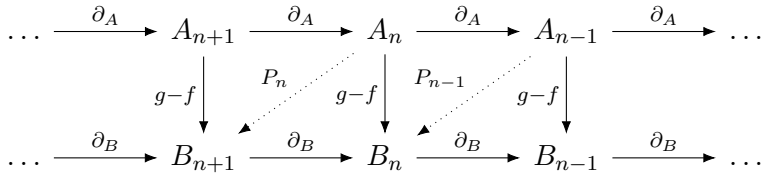
Next, we want to define what it means for two maps f and g to be homotopic. Here's the answer:

Definition 71.3.6. Let $f, g\colon A_\bullet \to B_\bullet$. Suppose that one can find a map $P_n\colon A_n \to B_{n+1}$ for every n such that

$$g_n - f_n = \partial_B \circ P_n + P_{n-1} \circ \partial_A$$

Then P is a **chain homotopy** from f to g and f and g are **chain homotopic**.

We can draw a picture to illustrate this (warning: the diagonal dotted arrows do NOT commute with all the other arrows):



The definition is that in each slanted "parallelogram", the $g - f$ arrow is the sum of the two compositions along the sides.

Remark 71.3.7 — This equation should look terribly unmotivated right now, aside from the fact that we are about to show it does the right algebraic thing. Its derivation comes from the geometric context that we have deferred until the next section, where "homotopy" will naturally give "chain homotopy".

Now, the point of this definition is that

Proposition 71.3.8 (Chain homotopic maps induce the same map on homology groups)
Let $f, g : A_\bullet \to B_\bullet$ be chain homotopic maps $A_\bullet \to B_\bullet$. Then the induced maps $f_*, g_*\colon H_n(A_\bullet) \to H_n(B_\bullet)$ coincide for each n.

Proof. It's equivalent to show $g - f$ gives the zero map on homology groups, In other words, we need to check that every cycle of A_n becomes a boundary of B_n under $g - f$.

Question 71.3.9. Verify that this is true. □

§71.3.ii Geometry of chain complexes

Now let's fill in the geometric details of the picture above. First:

Lemma 71.3.10 (Map of space $\implies$ map of singular chain complexes)
Each $f\colon X \to Y$ induces a map $C_n(X) \to C_n(Y)$.

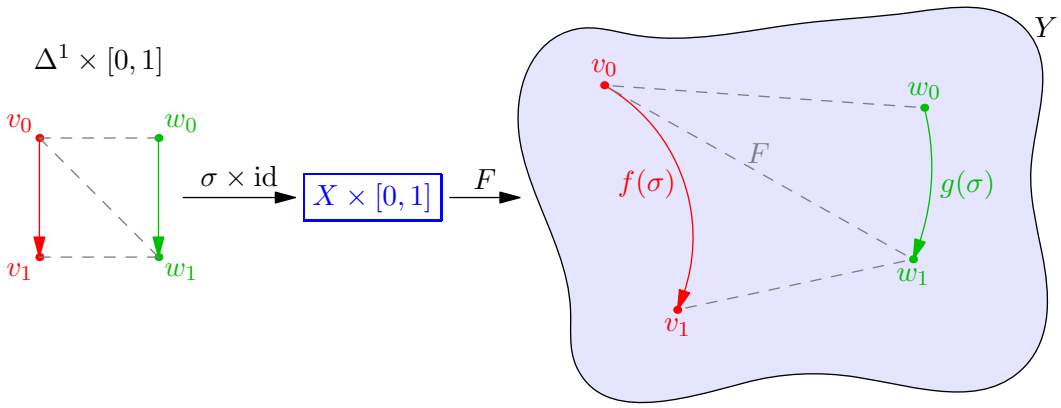
Proof. Take the composition

$$\Delta^n \xrightarrow{\sigma} X \xrightarrow{f} Y.$$

In other words, a path in X becomes a path in Y, et cetera. (It's not hard to see that the squares involving ∂ commute; check it if you like.) □

Now, what we need is to show that if $f, g\colon X \to Y$ are homotopic, then they are chain homotopic. To produce a chain homotopy, we need to take every n-simplex X to an $(n+1)$-chain in Y, thus defining the map P_n.

Let's think about how we might do this. Let's take the n-simplex $\sigma\colon \Delta^n \to X$ and feed it through f and g; pictured below is a 1-simplex σ (i.e. a path in X) which has been mapped into the space Y. Homotopy means the existence of a map $F\colon X \times [0,1] \to Y$ such that $F(-,0) = f$ and $F(-,1) = g$, parts of which I've illustrated below with grey arrows in the image for Y.



This picture suggests how we might proceed: we want to create a 2-chain on Y given the 1-chains we've drawn. The homotopy F provides us with a "square" structure on Y, i.e. the square bounded by v_0, v_1, w_1, w_0. We split this up into two triangles; and that's our 2-chain.

We can make this formal by taking $\Delta^1 \times [0,1]$ (which *is* a square) and splitting it into two triangles. Then, if we apply $\sigma \times \mathrm{id}$, we'll get an 2-chain in $X \times [0,1]$, and then finally applying F will map everything into our space Y. In our example, the final image is the 2-chain, consisting of two triangles, which in our picture can be written as $[v_0, w_0, w_1] - [v_0, v_1, w_1]$; the boundaries are given by the red, green, grey.

More generally, for an n-simplex $\phi = [x_0, \dots, x_n]$ we define the so-called *prism operator* P_n as follows. Set $v_i = f(x_i)$ and $w_i = g(x_i)$ for each i. Then, we let

$$P_n(\phi) \coloneqq \sum_{i=0}^{n} (-1)^i (F \circ (\phi \times \mathrm{id}))\,[v_0, \dots, v_i, w_i, \dots, w_n]\,.$$

This is just the generalization of the construction above to dimensions $n > 1$; we split $\Delta^n \times [0,1]$ into $n+1$ simplices, map it into X by $\phi \times \mathrm{id}$ and then push the whole thing into Y. The $(-1)^i$ makes sure that the "diagonal" faces all cancel off with each other.

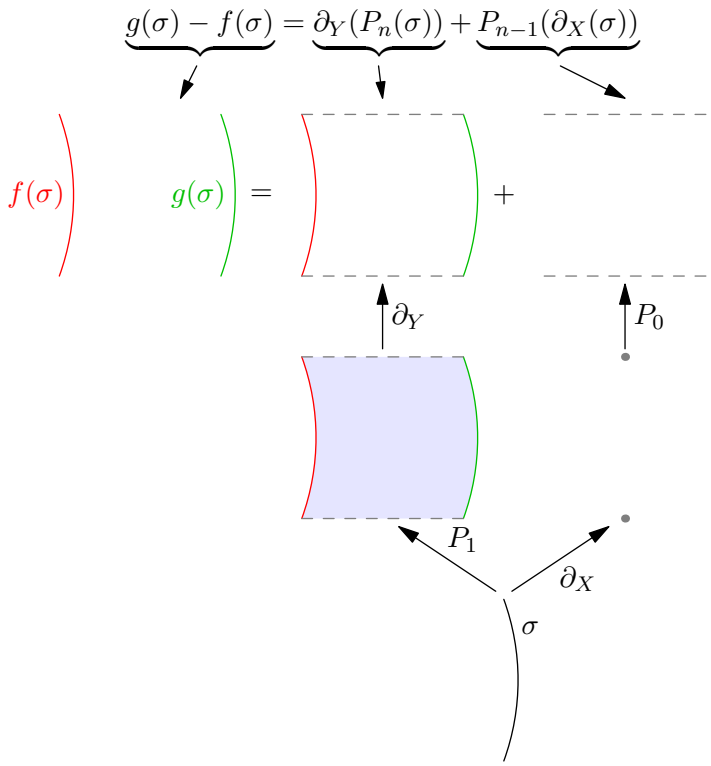
We now claim that for every σ,

$$\partial_Y(P_n(\sigma)) = g(\sigma) - f(\sigma) - P_{n-1}(\partial_X \sigma).$$

In the picture, $\partial_Y \circ P_n$ is the boundary of the entire prism (in the figure, this becomes the red, green, and grey lines, not including diagonal grey, which is cancelled out). The $g - f$ is the green minus the red, and the $P_{n-1} \circ \partial_X$ represents the grey edges of the

prism (not including the diagonal line from v_1 to w_0). Indeed, one can check (just by writing down several $\sum$ signs) that the above identity holds.

As a picture:



So that gives the chain homotopy from f to g, completing the proof of Theorem 71.3.2.

§71.4 More examples of chain complexes

We now end this chapter by providing some more examples of chain complexes, which we'll use in the next chapter to finally compute topological homology groups.

Example 71.4.1 (Reduced homology groups)

Suppose X is a (nonempty) topological space. One can augment the standard singular complex as follows: do the same thing as before, but augment the end by adding a $\mathbb{Z}$, as shown:

$$\cdots \to C_1(X) \to C_0(X) \xrightarrow{\varepsilon} \mathbb{Z} \to 0$$

Here ε is defined by $\varepsilon(\sum n_i p_i) = \sum n_i$ for points $p_i \in X$. (Recall that a 0-chain is just a formal sum of points!) We denote this **augmented singular chain complex** by $\widetilde{C}_\bullet(X)$.

This may seem like a random thing to do, but it can be justified by taking the definitions we started with and "generalizing backwards". Recall that an n-simplex is given by $n+1$ vertices: $[v_0, \ldots, v_n]$. That suggests that a (-1)-simplex is given by 0 vertices: []!

We reach the same conclusion if we apply the definition of the standard n-simplex using $n = -1$. Δ^{-1} must be the subset of $\mathbb{R}^0$ consisting of all points whose coordinates are nonnegative and sum to 1. There are no such points, so $\Delta^{-1} = \{\}$. Consequently, given a topological space X, a singular (-1)-simplex in X must be a

function $\{\} \to X$. There is one such function: the *empty function*, whose image is the empty set.
That is, every topological space X has exactly one (-1)-simplex, which we identify with $\{\}$. Thus, the (-1)st chain group $C_{-1}(X)$ is the free abelian group generated by one element; ie, $\widetilde{C}_{-1}(X) \cong \mathbb{Z}$ (where the isomorphism identifies $\{\}$ with 1).
What about boundaries? To take the boundary of a simplex $[v_0, \dots, v_n]$, we remove each vertex one-by-one, and take the alternating sum. Therefore, $\partial([v]) = []$. Extending it linearly to complexes yields $\partial(\sum n_i p_i) = \sum n_i \cdot 1$ — so ε really is just the boundary operator, generalized to the case $\widetilde{C}_0(X) \xrightarrow{\partial} \widetilde{C}_{-1}(X)$.[a]

[a]What about $n \le -2$? An n-simplex comes from a list of vertices of length $(n+1)$, so a (-2)-simplex would require a list of vertices length (-1) — but there aren't any such lists. So while there is one (-1)-simplex, there are zero (-2)-simplices (ditto for $n < -2$). The free abelian group on zero elements is the trivial group, so $\widetilde{C}_{-2} \cong \mathbf{0}$. In particular, $\partial([]) = 0$.

Question 71.4.2. What's the homology of the above chain at $\mathbb{Z}$? (Hint: you need X nonempty.)

Definition 71.4.3. The homology groups of the augmented chain complex are called the **reduced homology groups** $\widetilde{H}_n(X)$ of the space X.

Obviously $\widetilde{H}_n(X) \cong H_n(X)$ for $n > 0$. But when $n = 0$, the map $H_0(X) \to \mathbb{Z}$ by ε has kernel $\widetilde{H}_0(X)$, thus $H_0(X) \cong \widetilde{H}_0(X) \oplus \mathbb{Z}$.

This is usually just an added convenience. For example, it means that if X is contractible, then all its reduced homology groups vanish, and thus we won't have to keep fussing with the special $n = 0$ case.

Question 71.4.4. Given the claim earlier about $H_n(S^m)$, what should $\widetilde{H}_n(S^m)$ be?

Example 71.4.5 (Relative chain groups)

Suppose X is a topological space, and $A \subseteq X$ a subspace. We can "mod out" by A by defining

$$C_n(X, A) \coloneqq C_n(X)/C_n(A)$$

for every n. Thus chains contained entirely in A are trivial.
Then, the usual ∂ on $C_n(X)$ generates a new chain complex

$$\dots \xrightarrow{\partial} C_{n+1}(X, A) \xrightarrow{\partial} C_n(X, A) \xrightarrow{\partial} C_{n-1}(X, A) \xrightarrow{\partial} \dots.$$

This is well-defined since ∂ takes $C_n(A)$ into $C_{n-1}(A)$.

Definition 71.4.6. The homology groups of the relative chain complex are the **relative homology groups** and denoted $H_n(X, A)$.

One naïve guess is that this might equal $H_n(X)/H_n(A)$. This is not true and in general doesn't even make sense; if we take X to be $\mathbb{R}^2$ and $A = S^1$ a circle inside it, we have $H_1(X) = H_1(\mathbb{R}^2) = 0$ and $H_1(S^1) = \mathbb{Z}$.

Another guess is that $H_n(X, A)$ might just be $\widetilde{H}_n(X/A)$. This will turn out to be true for most reasonable spaces X and A, and we will discuss this when we reach the excision

theorem in Chapter 73.

Example 71.4.7 (Mayer-Vietoris sequence)

Suppose a space X is covered by two open sets U and V. We can define $C_n(U+V)$ as follows: it consists of chains such that each simplex is either entirely contained in U, or entirely contained in V.
Of course, ∂ then defines another chain complex

$$\dots \xrightarrow{\partial} C_{n+1}(U+V) \xrightarrow{\partial} C_n(U+V) \xrightarrow{\partial} C_{n-1}(U+V) \xrightarrow{\partial} \dots.$$

So once again, we can define homology groups for this complex; we denote them by $H_n(U+V)$. Miraculously, it will turn out that $H_n(U+V) \cong H_n(X)$.

§71.5 A few harder problems to think about

Problem 71A. For $n \geq 1$ show that the composition

$$S^{n-1} \hookrightarrow D^n \xrightarrow{F} S^{n-1}$$

cannot be the identity map on S^{n-1} for any continuous F.

Problem 71B (Brouwer fixed point theorem)**.** Use the previous problem to prove that any continuous function $f\colon D^n \to D^n$ has a fixed point.

72 The long exact sequence

In this chapter we introduce the key fact about chain complexes that will allow us to compute the homology groups of any space: the so-called "long exact sequence".

For those that haven't read about abelian categories: a sequence of morphisms of abelian groups

$$\dots \to G_{n+1} \to G_n \to G_{n-1} \to \dots$$

is **exact** if the image of any arrow is equal to the kernel of the next arrow. In particular,

- The map $0 \to A \to B$ is exact if and only if $A \to B$ is injective.
- the map $A \to B \to 0$ is exact if and only if $A \to B$ is surjective.

(On that note: what do you call a chain complex whose homology groups are all trivial?) A short exact sequence is one of the form $0 \to A \hookrightarrow B \twoheadrightarrow C \to 0$.

§72.1 Short exact sequences and four examples

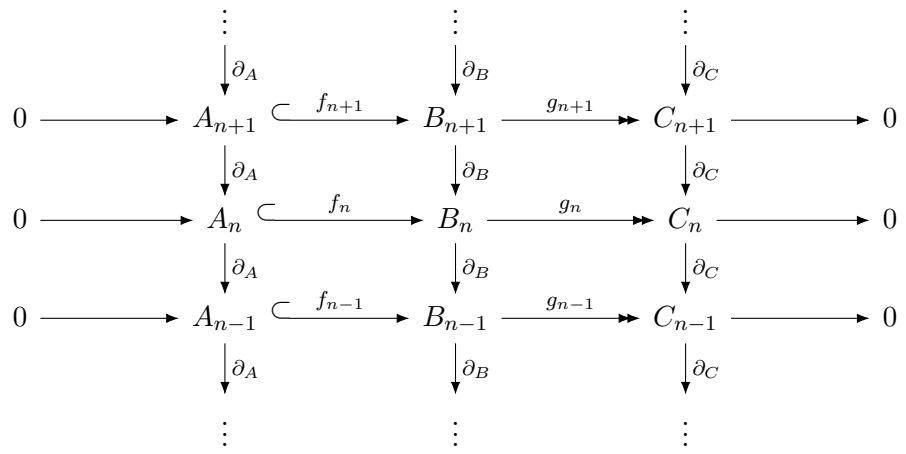
Prototypical example for this section: Relative sequence and Mayer-Vietoris sequence.

Let $\mathcal{A} = \mathsf{AbGrp}$. Recall that we defined a morphism of chain complexes in $\mathcal{A}$ already.

Definition 72.1.1. Suppose we have a map of chain complexes

$$0 \to A_\bullet \xrightarrow{f} B_\bullet \xrightarrow{g} C_\bullet \to 0$$

It is said to be **short exact** if *each row* of the diagram below is short exact.



This basically means $C_\bullet = B_\bullet / A_\bullet$, for suitable definition of / on chain complexes.

This agrees with the definition in Section 70.3.

Example 72.1.2 (Mayer-Vietoris short exact sequence and its augmentation)
Let $X = U \cup V$ be an open cover. For each n consider

$$\begin{aligned} C_n(U \cap V) &\hookrightarrow C_n(U) \oplus C_n(V) \twoheadrightarrow C_n(U+V) \\ c &\mapsto (c, -c) \\ (c,d) &\mapsto c+d \end{aligned}$$

One can easily see (by taking a suitable basis) that the kernel of the latter map is exactly the image of the first map. This generates a short exact sequence

$$0 \to C_\bullet(U \cap V) \hookrightarrow C_\bullet(U) \oplus C_\bullet(V) \twoheadrightarrow C_\bullet(U+V) \to 0.$$

Example 72.1.3 (Augmented Mayer-Vietoris sequence)
We can *augment* each of the chain complexes in the Mayer-Vietoris sequence as well, by appending

$$\begin{array}{ccccccccc} 0 & \to & C_0(U \cap V) & \hookrightarrow & C_0(U) \oplus C_0(V) & \twoheadrightarrow & C_0(U+V) & \to & 0 \\ & & \downarrow{\scriptstyle \varepsilon} & & \downarrow{\scriptstyle \varepsilon \oplus \varepsilon} & & \downarrow{\scriptstyle \varepsilon} & & \\ 0 & \to & \mathbb{Z} & \to & \mathbb{Z} \oplus \mathbb{Z} & \to & \mathbb{Z} & \to & 0 \end{array}$$

to the bottom of the diagram. In other words we modify the above into

$$0 \to \widetilde{C}_\bullet(U \cap V) \hookrightarrow \widetilde{C}_\bullet(U) \oplus \widetilde{C}_\bullet(V) \twoheadrightarrow \widetilde{C}_\bullet(U+V) \to 0$$

where $\widetilde{C}_\bullet$ is the chain complex defined in Definition 71.4.3.

Example 72.1.4 (Relative chain short exact sequence)
Since $C_n(X,A) \coloneqq C_n(X)/C_n(A)$, we have a short exact sequence

$$0 \to C_\bullet(A) \hookrightarrow C_\bullet(X) \twoheadrightarrow C_\bullet(X,A) \to 0$$

for every space X and subspace A. This can be augmented: we get

$$0 \to \widetilde{C}_\bullet(A) \hookrightarrow \widetilde{C}_\bullet(X) \twoheadrightarrow C_\bullet(X,A) \to 0$$

by adding the final row

$$\begin{array}{ccccccccc} 0 & \to & C_0(A) & \hookrightarrow & C_0(X) & \twoheadrightarrow & C_0(X,A) & \to & 0 \\ & & \downarrow{\scriptstyle \varepsilon} & & \downarrow{\scriptstyle \varepsilon} & & & & \\ 0 & \to & \mathbb{Z} & \xrightarrow[\text{id}]{} & \mathbb{Z} & \to & 0 & \to & 0. \end{array}$$

§72.2 The long exact sequence of homology groups

Consider a short exact sequence $0 \to A_\bullet \xrightarrow{f} B_\bullet \xrightarrow{g} C_\bullet \to 0$. Now, we know that we get induced maps of homology groups, i.e. we have

$$\begin{array}{ccccc}
\vdots & & \vdots & & \vdots \\
H_{n+1}(A_\bullet) & \xrightarrow{f_*} & H_{n+1}(B_\bullet) & \xrightarrow{g_*} & H_{n+1}(C_\bullet) \\
H_n(A_\bullet) & \xrightarrow{f_*} & H_n(B_\bullet) & \xrightarrow{g_*} & H_n(C_\bullet) \\
H_{n-1}(A_\bullet) & \xrightarrow{f_*} & H_{n-1}(B_\bullet) & \xrightarrow{g_*} & H_{n-1}(C_\bullet) \\
\vdots & & \vdots & & \vdots
\end{array}$$

But the theorem is that we can string these all together, taking each $H_{n+1}(C_\bullet)$ to $H_n(A_\bullet)$.

Theorem 72.2.1 (Short exact $\implies$ long exact)

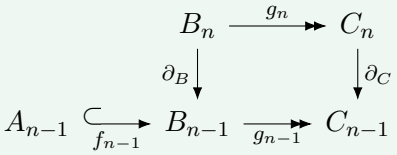
Let $0 \to A_\bullet \xrightarrow{f} B_\bullet \xrightarrow{g} C_\bullet \to 0$ be *any* short exact sequence of chain complexes we like. Then there is an *exact* sequence

$$\begin{array}{ccccc}
 & & \dots & \longrightarrow & H_{n+2}(C_\bullet) \\
 & & & \overset{\partial}{\swarrow} & \\
H_{n+1}(A_\bullet) & \xrightarrow[f_*]{} & H_{n+1}(B_\bullet) & \xrightarrow{g_*} & H_{n+1}(C_\bullet) \\
 & & & \overset{\partial}{\swarrow} & \\
H_n(A_\bullet) & \xrightarrow[f_*]{} & H_n(B_\bullet) & \xrightarrow{g_*} & H_n(C_\bullet) \\
 & & & \overset{\partial}{\swarrow} & \\
H_{n-1}(A_\bullet) & \xrightarrow[f_*]{} & H_{n-1}(B_\bullet) & \xrightarrow{g_*} & H_{n-1}(C_\bullet) \\
 & & & \overset{\partial}{\swarrow} & \\
H_{n-2}(A_\bullet) & \longrightarrow & \dots & &
\end{array}$$

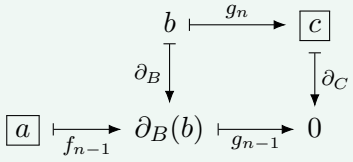
This is called a **long exact sequence** of homology groups.

Proof. A very long diagram chase, valid over any abelian category. (Alternatively, it's actually possible to use the snake lemma twice.) □

Remark 72.2.2 — The map $\partial\colon H_n(C_\bullet) \to H_{n-1}(A_\bullet)$ can be written explicitly as follows. Recall that H_n is "cycles modulo boundaries", and consider the sub-diagram



We need to take every cycle in C_n to a cycle in A_{n-1}. (Then we need to check a ton of "well-defined" issues, but let's put that aside for now.)
Suppose $c \in C_n$ is a cycle (so $\partial_C(c) = 0$). By surjectivity, there is a $b \in B_n$ with $g_n(b) = c$, which maps down to $\partial_B(b)$. Now, the image of $\partial_B(b)$ under g_{n-1} is zero by commutativity of the square, and so we can pull back under f_{n-1} to get a unique element of A_{n-1} (by exactness at B_{n-1}).
In summary: we go "*left, down, left*" to go from c to a:



Exercise 72.2.3. Check quickly that the recovered a is actually a cycle, meaning $\partial_A(a) = 0$. (You'll need another row, and the fact that $\partial_B^2 = 0$.)

The final word is that:

Short exact sequences of chain complexes give long exact sequences of homology groups.

In particular, let us take the four examples given earlier.

Example 72.2.4 (Mayer-Vietoris long exact sequence, provisional version)
The Mayer-Vietoris ones give, for $X = U \cup V$ an open cover,

$$\dots \to H_n(U \cap V) \to H_n(U) \oplus H_n(V) \to H_n(U + V) \to H_{n-1}(U \cap V) \to \dots.$$

and its reduced version

$$\dots \to \widetilde{H}_n(U \cap V) \to \widetilde{H}_n(U) \oplus \widetilde{H}_n(V) \to \widetilde{H}_n(U + V) \to \widetilde{H}_{n-1}(U \cap V) \to \dots.$$

This version is "provisional" because in the next section we will replace $H_n(U + V)$ and $\widetilde{H}_n(U + V)$ with something better. As for the relative homology sequences, we have:

Theorem 72.2.5 (Long exact sequence for relative homology)
Let X be a space, and let $A \subseteq X$ be a subspace. There are long exact sequences

$$\cdots \to H_n(A) \to H_n(X) \to H_n(X,A) \to H_{n-1}(A) \to \dots.$$

and

$$\cdots \to \widetilde{H}_n(A) \to \widetilde{H}_n(X) \to H_n(X,A) \to \widetilde{H}_{n-1}(A) \to \dots.$$

The exactness of these sequences will give **tons of information** about $H_n(X)$ if only we knew something about what $H_n(U+V)$ or $H_n(X,A)$ looked like. This is the purpose of the next chapter.

§72.3 The Mayer-Vietoris sequence

Prototypical example for this section: The computation of $H_n(S^m)$ by splitting S^m into two hemispheres.

Now that we have done so much algebra, we need to invoke some geometry. There are two major geometric results in the Napkin. One is the excision theorem, which we discuss next chapter. The other we present here, which will let us take advantage of the Mayer-Vietoris sequence. The proofs are somewhat involved and are thus omitted; see [Ha02] for details.

The first theorem is that the notation $H_n(U+V)$ that we have kept until now is redundant, and can be replaced with just $H_n(X)$:

Theorem 72.3.1 (Open cover homology theorem)
Consider the inclusion $\iota\colon C_\bullet(U+V) \hookrightarrow C_\bullet(X)$. Then ι induces an isomorphism

$$H_n(U+V) \cong H_n(X).$$

Remark 72.3.2 — In fact, this is true for any open cover (even uncountable), not just those with two covers $U \cup V$. But we only state the special case with two open sets, because this is what is needed for Example 72.1.2.

So, Example 72.1.2 together with the above theorem implies, after replacing all the $H_n(U+V)$'s with $H_n(X)$'s:

Theorem 72.3.3 (Mayer-Vietoris long exact sequence)
If $X = U \cup V$ is an open cover, then we have long exact sequences

$$\cdots \to H_n(U \cap V) \to H_n(U) \oplus H_n(V) \to H_n(X) \to H_{n-1}(U \cap V) \to \dots.$$

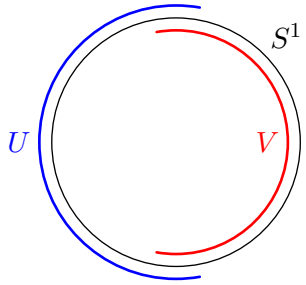
and

$$\cdots \to \widetilde{H}_n(U \cap V) \to \widetilde{H}_n(U) \oplus \widetilde{H}_n(V) \to \widetilde{H}_n(X) \to \widetilde{H}_{n-1}(U \cap V) \to \dots.$$

At long last, we can compute the homology groups of the spheres.

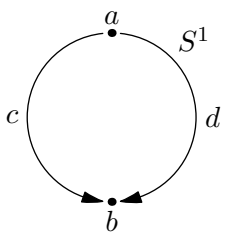
Theorem 72.3.4 (The homology groups of S^m)
For integers m and n,
$$\widetilde{H}_n(S^m) \cong \begin{cases} \mathbb{Z} & n = m \\ 0 & \text{otherwise.} \end{cases}$$
The generator $\widetilde{H}_n(S^n)$ is an n-cell which covers S^n exactly once (for example, the generator for $\widetilde{H}_1(S^1)$ is a loop which wraps around S^1 once).

Proof. This one's fun, so I'll only spoil the case $m = 1$, and leave the rest to you. Decompose the circle S^1 into two arcs U and V, as shown:



Each of U and V is contractible, so all their reduced homology groups vanish. Moreover, $U \cap V$ is homotopy equivalent to two points, hence
$$\widetilde{H}_n(U \cap V) \cong \begin{cases} \mathbb{Z} & n = 0 \\ 0 & \text{otherwise.} \end{cases}$$
Now consider again the segment of the short exact sequence
$$\cdots \to \underbrace{\widetilde{H}_n(U) \oplus \widetilde{H}_n(V)}_{=0} \to \widetilde{H}_n(S^1) \xrightarrow{\partial} \widetilde{H}_{n-1}(U \cap V) \to \underbrace{\widetilde{H}_{n-1}(U) \oplus \widetilde{H}_{n-1}(V)}_{=0} \to \dots.$$
From this we derive that $\widetilde{H}_n(S^1)$ is $\mathbb{Z}$ for $n = 1$ and 0 elsewhere.

It remains to analyze the generators of $\widetilde{H}_1(S^1)$. Note that the isomorphism was given by the connecting homomorphism ∂, which is given by a "left, down, left" procedure (Remark 72.2.2) in the diagram
$$\begin{array}{ccc} & C_1(U) \oplus C_1(V) & \longrightarrow C_1(U+V) \\ & \downarrow{\scriptstyle \partial\oplus\partial} & \\ C_0(U \cap V) & \longrightarrow C_0(U) \oplus C_0(V) & \end{array}$$
Mark the points a and b as shown in the two disjoint paths of $U \cap V$.



Then $a - b$ is a cycle which represents a generator of $H_0(U \cap V)$. We can find the pre-image of ∂ as follows: letting c and d be the chains joining a and b, with c contained in U, and d contained in V, the diagram completes as

$$\begin{array}{ccc} (c,d) & \longmapsto & c-d \\ \downarrow & & \\ a-b & \longmapsto & (a-b,a-b) \end{array}$$

In other words $\partial(c-d) = a-b$, so $c-d$ is a generator for $\widetilde{H}^1(S^1)$.

Thus we wish to show that $c-d$ is (in $H^1(S^1)$) equivalent to the loop γ wrapping around S^1 once, counterclockwise. This was illustrated in Example 71.2.10. □

Thus, the key idea in Mayer-Vietoris is that

Mayer-Vietoris lets us compute $H_n(X)$ by splitting X into two open sets.

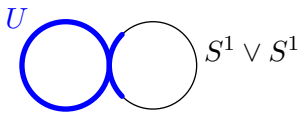
Here are some more examples.

Proposition 72.3.5 (The homology groups of the figure eight)
Let $X = S^1 \vee S^1$ be the figure eight. Then

$$\widetilde{H}_n(X) \cong \begin{cases} \mathbb{Z}^{\oplus 2} & n = 1 \\ 0 & \text{otherwise.} \end{cases}$$

The generators for $\widetilde{H}_1(X)$ are the two loops of the figure eight.

Proof. Again, for simplicity we work with reduced homology groups. Let U be the "left" half of the figure eight plus a little bit of the right, as shown below.



The set V is defined symmetrically. In this case $U \cap V$ is contractible, while each of U and V is homotopic to S^1.

Thus, we can read a segment of the long exact sequence as

$$\cdots \to \underbrace{\widetilde{H}_n(U \cap V)}_{=0} \to \widetilde{H}_n(U) \oplus \widetilde{H}_n(V) \to \widetilde{H}_n(X) \to \underbrace{\widetilde{H}_{n-1}(U \cap V)}_{=0} \to \ldots.$$

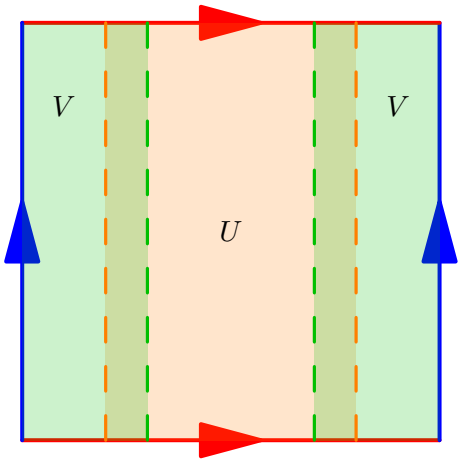
So we get that $\widetilde{H}_n(X) \cong \widetilde{H}_n(S^1) \oplus \widetilde{H}_n(S^1)$, The claim about the generators follows from the fact that, according to the isomorphism above, the generators of $\widetilde{H}_n(X)$ are the generators of $\widetilde{H}_n(U)$ and $\widetilde{H}_n(V)$, which we described geometrically in the last theorem. □

Up until now, we have been very fortunate that we have always been able to make certain parts of the space contractible. This is not always the case, and in the next example we will have to actually understand the maps in question to complete the solution.

Proposition 72.3.6 (Homology groups of the torus)
Let $X = S^1 \times S^1$ be the torus. Then

$$\widetilde{H}_n(X) = \begin{cases} \mathbb{Z}^{\oplus 2} & n = 1 \\ \mathbb{Z} & n = 2 \\ 0 & \text{otherwise.} \end{cases}$$

Proof. To make our diagram look good on 2D paper, we'll represent the torus as a square with its edges identified, though three-dimensionally the picture makes sense as well. Consider U (shaded light orange) and V (shaded green) as shown. (Note that V is connected due to the identification of the left and right (blue) edges, even if it doesn't look connected in the picture).



In the three dimensional picture, U and V are two cylinders which together give the torus. This time, U and V are each homotopic to S^1, and the intersection $U \cap V$ is the disjoint union of two circles: thus $\widetilde{H}_1(U \cap V) \cong \mathbb{Z} \oplus \mathbb{Z}$, and $H_0(U \cap V) \cong \mathbb{Z}^{\oplus 2} \implies \widetilde{H}_0(U \cap V) \cong \mathbb{Z}$.

For $n \geq 3$, we have

$$\cdots \to \underbrace{\widetilde{H}_n(U \cap V)}_{=0} \to \widetilde{H}_n(U) \oplus \widetilde{H}_n(V) \to \widetilde{H}_n(X) \to \underbrace{\widetilde{H}_{n-1}(U \cap V)}_{=0} \to \dots.$$

and so $H_n(X) \cong 0$ for $n \geq 3$. Also, we have $H_0(X) \cong \mathbb{Z}$ since X is path-connected. So it remains to compute $H_2(X)$ and $H_1(X)$.

Let's find $H_2(X)$ first. We first consider the segment

$$\cdots \to \underbrace{\widetilde{H}_2(U) \oplus \widetilde{H}_2(V)}_{=0} \to \widetilde{H}_2(X) \xhookrightarrow{\partial} \underbrace{\widetilde{H}_1(U \cap V)}_{\cong \mathbb{Z} \oplus \mathbb{Z}} \xrightarrow{\phi} \underbrace{\widetilde{H}_1(U) \oplus \widetilde{H}_1(V)}_{\cong \mathbb{Z} \oplus \mathbb{Z}} \to \dots$$

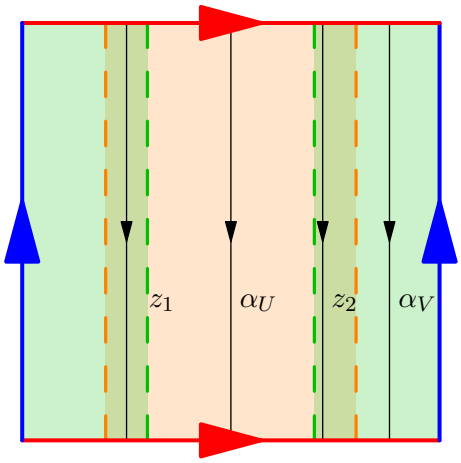
Unfortunately, this time it's not immediately clear what $\widetilde{H}_2(X)$ because we only have one zero at the left. In order to do this, we have to actually figure out what the maps ∂ and ϕ look like. Note that, as we'll see, ϕ isn't an isomorphism even though the groups are isomorphic.

The presence of the zero term has allowed us to make the connecting map ∂ injective. First, $\widetilde{H}_2(X)$ is isomorphic to the image of ∂, which is exactly the kernel of the arrow ϕ inserted. To figure out what $\ker \phi$ is, we have to think back to how the map $C_\bullet(U \cap V) \to C_\bullet(U) \oplus C_\bullet(V)$ was constructed: it was $c \mapsto (c, -c)$. So the induced maps of homology

groups is actually what you would guess: a 1-cycle z in $\widetilde{H}_1(U \cap V)$ gets sent $(z, -z)$ in $\widetilde{H}_1(U) \oplus \widetilde{H}_1(V)$.

In particular, consider the two generators z_1 and z_2 of $\widetilde{H}_1(U \cap V) = \mathbb{Z} \oplus \mathbb{Z}$, i.e. one cycle in each connected component of $U \cap V$. (To clarify: $U \cap V$ consists of two "wristbands"; z_i wraps around the ith one once.) Moreover, let α_U denote a generator of $\widetilde{H}_1(U) \cong \mathbb{Z}$, and α_V a generator of $\widetilde{H}_1(V) \cong \mathbb{Z}$.

The elements are depicted below:



Note that z_1, z_2, α_U, α_V are elements of the homology group, so you can move the paths around a bit — for instance, as elements of $\widetilde{H}_1(U)$, the chain drawn as z_1 and α_U represents the same element.

Then we have that

$$z_1 \mapsto (\alpha_U, -\alpha_V) \qquad \text{and} \qquad z_2 \mapsto (\alpha_U, -\alpha_V).$$

(The signs may differ on which direction you pick for the generators; note that $\mathbb{Z}$ has two possible generators.) We can even format this as a matrix:

$$\phi = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}.$$

And we observe $\phi(z_1 - z_2) = 0$, meaning this map has nontrivial kernel! That is,

$$\ker \phi = \langle z_1 - z_2 \rangle \cong \mathbb{Z}.$$

Thus, $\widetilde{H}_2(X) \cong \operatorname{im} \partial \cong \ker \phi \cong \mathbb{Z}$. We'll also note that $\operatorname{im} \phi$ is the set generated by $(\alpha_U, -\alpha_V)$; (in particular $\operatorname{im} \phi \cong \mathbb{Z}$ and the quotient by $\operatorname{im} \phi$ is $\mathbb{Z}$ too).

The situation is similar with $\widetilde{H}_1(X)$: this time, we have

$$\ldots \xrightarrow{\phi} \underbrace{\widetilde{H}_1(U) \oplus \widetilde{H}_1(V)}_{\cong \mathbb{Z} \oplus \mathbb{Z}} \xrightarrow{\psi} \widetilde{H}_1(X) \xrightarrow{\partial} \underbrace{\widetilde{H}_0(U \cap V)}_{\cong \mathbb{Z}} \to \underbrace{\widetilde{H}_0(U) \oplus \widetilde{H}_0(V)}_{=0} \to \ldots$$

and so we know that the connecting map ∂ is surjective, hence $\operatorname{im} \partial \cong \mathbb{Z}$. Now, we also have

$$\begin{aligned} \ker \partial \cong \operatorname{im} \psi &\cong \left(\widetilde{H}_1(U) \oplus \widetilde{H}_1(V)\right) / \ker \psi \\ &\cong \left(\widetilde{H}_1(U) \oplus \widetilde{H}_1(V)\right) / \operatorname{im} \phi \cong \mathbb{Z} \end{aligned}$$

by what we knew about $\operatorname{im}\phi$ already. To finish off we need some algebraic tricks. The first is Proposition 70.5.1, which gives us a short exact sequence

$$0 \to \underbrace{\ker\partial}_{\cong\operatorname{im}\psi\cong\mathbb{Z}} \hookrightarrow \widetilde{H}_1(X) \twoheadrightarrow \underbrace{\operatorname{im}\partial}_{\cong\mathbb{Z}} \to 0.$$

You should satisfy yourself that $\widetilde{H}_1(X) \cong \mathbb{Z}\oplus\mathbb{Z}$ is the only possibility, but we'll prove this rigorously with Lemma 72.3.8. □

Remark 72.3.7 — Earlier, we remarked (without proof) that $\pi_2(X)$ is trivial — that is, homotopy does not found any "2-dimensional holes" in the torus. Why is it that $H_2(X) \cong \mathbb{Z}$?
You may want to manually compute the nontrivial element in $H_2(X)$ using the long exact sequence using the following method. Look at the long exact sequence:

$$\cdots \longrightarrow \underbrace{H_2(U)\oplus H_2(V)}_{=0} \longrightarrow \underbrace{H_2(X)}_{\cong\mathbb{Z}} \xrightarrow{\partial} \underbrace{H_1(U\cap V)}_{\cong\mathbb{Z}\oplus\mathbb{Z}} \xrightarrow{\phi} \underbrace{H_1(U)\oplus H_1(V)}_{\cong\mathbb{Z}\oplus\mathbb{Z}} \longrightarrow \cdots$$

We wish to find some nontrivial element in $H_2(X)$ — in order to do that, we can take an element in $\ker\phi \subseteq H_1(U\cap V)$ and take its preimage under ∂.
For that, $z_1 - z_2$ would suffice. In order to take its preimage under ∂, we need to recall how ∂ was constructed — it was a "left, down, left" procedure in the diagram:

$$\begin{array}{ccc} C_2(U)\oplus C_2(V) & \twoheadrightarrow & C_2(X) \\ \downarrow & & \\ C_1(U\cap V) & \hookrightarrow & C_1(U)\oplus C_1(V) \end{array}$$

So, we find a (closed) element in $C_1(U\cap V)$ whose image under the quotient map is $z_1 - z_2$, then move it "right, up, right" to an element in $C_2(X)$.
If you did everything correctly, the result should be *the whole torus*!

Which emphasizes the point:

A "hole" detected by homology need not look like the interior of S^n.

Note that the previous example is of a different attitude than the previous ones, because we had to figure out what the maps in the long exact sequence actually were to even compute the groups. In principle, you could also figure out all the isomorphisms in the previous proof and explicitly compute the generators of $\widetilde{H}_1(S^1\times S^1)$, but to avoid getting bogged down in detail I won't do so here.

Finally, to fully justify the last step, we present:

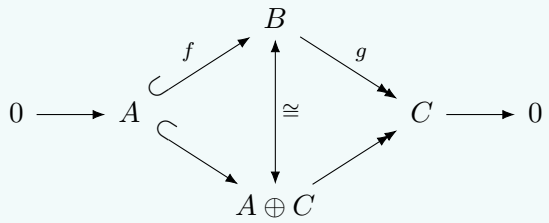
Lemma 72.3.8 (Splitting lemma)

For a short exact sequence $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ of abelian groups, the following are equivalent:

(a) There exists $p\colon B \to A$ such that $A \xrightarrow{f} B \xrightarrow{p} A$ is the identity.

(b) There exists $s\colon C \to B$ such that $C \xrightarrow{s} B \xrightarrow{g} C$ is the identity.

(c) There is an isomorphism from B to $A \oplus C$ such that the diagram



commutes. (The maps attached to $A \oplus C$ are the obvious ones.)

In particular, (b) holds anytime C is free.

In these cases we say the short exact sequence **splits**. The point is that

An exact sequence which splits let us obtain B given A and C.

In particular, for $C = \mathbb{Z}$ or any free abelian group, condition (b) is necessarily true. So, once we obtained the short exact sequence $0 \to \mathbb{Z} \to \widetilde{H}_1(X) \to \mathbb{Z} \to 0$, we were done.

Remark 72.3.9 — Unfortunately, not all exact sequences split: An example of a short exact sequence which doesn't split is

$$0 \to \mathbb{Z}/2\mathbb{Z} \xhookrightarrow{\times 2} \mathbb{Z}/4\mathbb{Z} \twoheadrightarrow \mathbb{Z}/2\mathbb{Z} \to 0$$

since it is not true that $\mathbb{Z}/4\mathbb{Z} \cong \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$.

Remark 72.3.10 — The splitting lemma is true in any abelian category. The "direct sum" is the colimit of the two objects A and C.

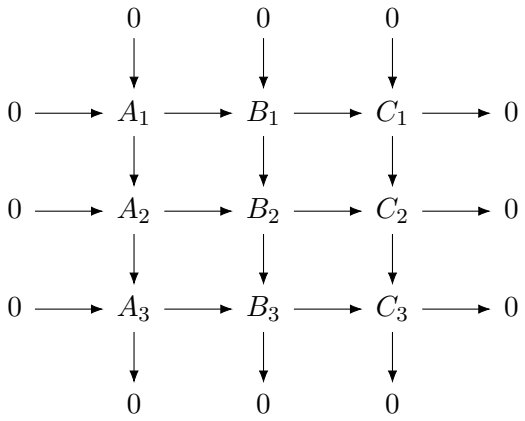
§72.4 A few harder problems to think about

Problem 72A. Complete the proof of Theorem 72.3.4, i.e. compute $H_n(S^m)$ for all m and n. (Try doing $m = 2$ first, and you'll see how to proceed.)

Problem 72B. Compute the reduced homology groups of $\mathbb{R}^n$ with $p \geq 1$ points removed.

Problem 72C*. Let $n \geq 1$ and $k \geq 0$ be integers. Compute $H_k(\mathbb{R}^n, \mathbb{R}^n \setminus \{0\})$.

Problem 72D (Nine lemma)**.** Consider a commutative diagram



and assume that all rows are exact, and two of the columns are exact. Show that the third column is exact as well.

Problem 72E* (Klein bottle)**.** Show that the reduced homology groups of the Klein bottle K are given by

$$\widetilde{H}_n(K) = \begin{cases} \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z} & n = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Problem 72F* (Triple long exact sequence)**.** Let $A \subseteq B \subseteq X$ be subspaces. Show that there is a long exact sequence

$$\dots \to H_n(B, A) \to H_n(X, A) \to H_n(X, B) \to H_{n-1}(B, A) \to \dots.$$

73 Excision and relative homology

We have already seen how to use the Mayer-Vietoris sequence: we started with a sequence

$$\dots \to H_n(U \cap V) \to H_n(U) \oplus H_n(V) \to H_n(U+V) \to H_{n-1}(U \cap V) \to \dots$$

and its reduced version, then appealed to the geometric fact that $H_n(U+V) \cong H_n(X)$. This allowed us to algebraically make computations on $H_n(X)$.

In this chapter, we turn our attention to the long exact sequence associated to the chain complex

$$0 \to C_n(A) \hookrightarrow C_n(X) \twoheadrightarrow C_n(X,A) \to 0.$$

The setup will look a lot like the previous two chapters, except in addition to $H_n \colon \mathsf{hTop} \to \mathsf{Grp}$ we will have a functor $H_n \colon \mathsf{hPairTop} \to \mathsf{Grp}$ which takes a pair (X,A) to $H_n(X,A)$. Then, we state (again without proof) the key geometric result, and use this to make deductions.

§73.1 Motivation

The main motivation is that:

Relative homology is the algebraic analog of quotient space.

So, for instance, when you see a map of pairs $f \colon (X,A) \to (Y,B)$, you should think of $X/A \to Y/B$.

Which explains the "reasonable guess" that for spaces $A \subseteq X$, we have $H_n(X,A) \cong \widetilde{H}_n(X/A)$.

By Theorem 73.4.3, the guess above is indeed true for most spaces. For example:

Question 73.1.1. Let $X = [0,1]$ and $A = \{0,1\}$. Show that $H_1(X/A)$ and $H_1(X,A)$ are isomorphic to $\mathbb{Z}$. (In this example, so is $\pi_1(X/A)$.)

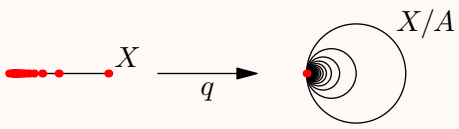
But not all. Similar to Example 64.2.6, if A is not closed, weird things can happen:

Example 73.1.2 ($H_n(X,A)$ where A is open in X)
Let $X = D^2$ be the closed disk.
If A is reasonably nice, for instance $A = S^1$ the boundary of X, we have $H_2(X,A) \cong H_2(X/A) \cong \mathbb{Z}$.
However, if $A = X \setminus \{0\}$ where 0 is the center of X, then $H_2(X,A)$ is still isomorphic to $\mathbb{Z}$; however $H_2(X/A) \cong 0$. (The latter isomorphism is harder to see, mainly because X/A is a weird space — it's not Hausdorff.)

Even when A is closed in X, problems can still happen.

Example 73.1.3 (The shrinking wedge of circles)

Let X be the interval $[0,1]$, and $A \subseteq X$ be $A = \{\frac{1}{n} \mid n \in \mathbb{Z}^+\} \cup \{0\}$. In this case, the quotient X/A would be isomorphic to the shrinking wedge of circles, as depicted below.



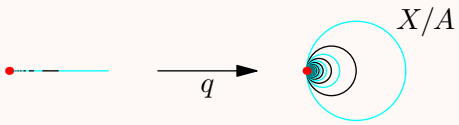
Note that in X/A, any open neighborhood of the red dot A/A must contains all but finitely many circles.

We claim that:

$$H_1(X,A) \ncong \widetilde{H}_1(X/A).$$

What could go wrong? Generally speaking, when you work algebraically then everything are finite, while in topology you have to consider things related to infinity.

Consider the following 1-simplex in $C(X/A)$, depicted in cyan.



Every element of $H(X,A)$ has a representative in $C(X)$ as a 1-cycles, which comprises of finitely many 1-simplices, each 1-simplex is equivalent to a segment $[a,b]$ — modulo a difference of a 1-boundary. Thus, intuitively, every element of $H(X,A)$ can only cover "finitely many circles" (or all but finitely many).

We haven't had enough tools to formalize all these yet. Formally speaking, the quotient maps $q\colon X \to X/A$ and $q\colon A \to A/A$ induces $q_*\colon H_1(X,A) \to H_1(X/A, A/A)$, and q_* is not injective.

Regardless, for nice spaces $A \subseteq X$ such that $H_n(X,A) \cong \widetilde{H}_n(X/A)$, we would be able to compute $H_n(X)$ based on $H_n(A)$ and $\widetilde{H}_n(X/A)$ — note that A and X/A is, in some sense, smaller and simpler than X.

§73.2 The long exact sequences

Recall Theorem 72.2.5, which says that the sequences

$$\cdots \to H_n(A) \to H_n(X) \to H_n(X,A) \to H_{n-1}(A) \to \dots.$$

and

$$\cdots \to \widetilde{H}_n(A) \to \widetilde{H}_n(X) \to H_n(X,A) \to \widetilde{H}_{n-1}(A) \to \dots$$

are long exact. By Problem 72F* we even have a long exact sequence

$$\cdots \to H_n(B,A) \to H_n(X,A) \to H_n(X,B) \to H_{n-1}(B,A) \to \dots.$$

for $A \subseteq B \subseteq X$.

This is the analog of the fact that X/B is homeomorphic to $\frac{X/A}{B/A}$ — we "cancel the common factor in the fraction".

An application of the first long exact sequence above gives:

Lemma 73.2.1 (Homology relative to contractible spaces)
Let X be a topological space, and let $A \subseteq X$ be contractible. For all n,

$$H_n(X, A) \cong \widetilde{H}_n(X).$$

Proof. Since A is contractible, we have $\widetilde{H}_n(A) = 0$ for every n. For each n there's a segment of the long exact sequence given by

$$\cdots \to \underbrace{\widetilde{H}_n(A)}_{=0} \to \widetilde{H}_n(X) \to H_n(X, A) \to \underbrace{\widetilde{H}_{n-1}(A)}_{=0} \to \ldots.$$

So since $0 \to \widetilde{H}_n(X) \to H_n(X, A) \to 0$ is exact, this means $H_n(X, A) \cong \widetilde{H}_n(X)$. □

In particular, the theorem applies if A is a single point. The case $A = \varnothing$ is also worth noting. We compile these results into a lemma:

Lemma 73.2.2 (Relative homology generalizes absolute homology)
Let X be any space, and $* \in X$ a point. Then for all n,

$$H_n(X, \{*\}) \cong \widetilde{H}_n(X) \qquad \text{and} \qquad H_n(X, \varnothing) = H_n(X).$$

§73.3 The category of pairs

Since we now have an $H_n(X, A)$ instead of just $H_n(X)$, a natural next step is to create a suitable category of *pairs* and give ourselves the same functorial setup as before.

Definition 73.3.1. Let $\varnothing \neq A \subseteq X$ and $\varnothing \neq B \subseteq Y$ be subspaces, and consider a map $f \colon X \to Y$. If $f^{\text{img}}(A) \subseteq B$ we write

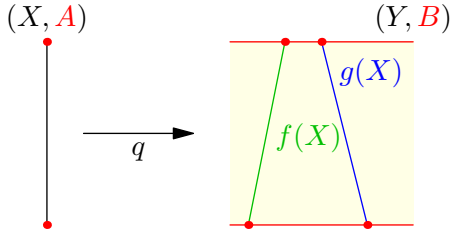
$$f \colon (X, A) \to (Y, B).$$

We say f is a **map of pairs**, between the pairs (X, A) and (Y, B).

Definition 73.3.2. We say that $f, g \colon (X, A) \to (Y, B)$ are **pair-homotopic** if they are "homotopic through maps of pairs".

More formally, a **pair-homotopy** $f, g \colon (X, A) \to (Y, B)$ is a map $F \colon [0, 1] \times X \to Y$, which we'll write as $F_t(X)$, such that F is a homotopy of the maps $f, g \colon X \to Y$ and each F_t is itself a map of pairs.

A typical $f, g \colon (X, A) \to (Y, B)$ that are pair-homotopic might look like this. Note that for all $t \in [0, 1]$, we must have $F_t^{\text{img}}(A) \subseteq B$.



Thus, we naturally arrive at two categories:

- PairTop, the category of *pairs* of topological spaces, and
- hPairTop, the same category except with maps only equivalent up to homotopy.

Definition 73.3.3. As before, we say pairs (X, A) and (Y, B) are **pair-homotopy equivalent** if they are isomorphic in hPairTop. An isomorphism of hPairTop is a **pair-homotopy equivalence**.

Remark 73.3.4 — Pair-homotopy equivalence of pairs is the natural generalization of homotopy equivalence of spaces, as defined in Definition 65.5.3. In fact, if $A = B = \varnothing$ then we have X is homotopy equivalent to Y if and only if $(X, \varnothing)$ is pair-homotopy equivalent to $(Y, \varnothing)$.

We can do the same song and dance as before with the prism operator to obtain:

Lemma 73.3.5 (Induced maps of relative homology)
We have a functor

$$H_n \colon \mathsf{hPairTop} \to \mathsf{Grp}.$$

That is, if $f \colon (X, A) \to (Y, B)$ then we obtain an induced map

$$f_* \colon H_n(X, A) \to H_n(Y, B).$$

and if two such f and g are pair-homotopic then $f_* = g_*$.

Now, we want an analog of contractible spaces for our pairs: i.e. pairs of spaces (X, A) such that $H_n(X, A) = 0$. The correct definition is:

Definition 73.3.6. Let $A \subseteq X$. We say that A is a **deformation retract**[1] of X if there is a map of pairs $r \colon (X, A) \to (A, A)$ which is a pair-homotopy equivalence.

Example 73.3.7 (Examples of deformation retracts)

(a) If a single point p is a deformation retract of a space X, then X is contractible, since the retraction $r \colon X \to \{*\}$ (when viewed as a map $X \to X$) is homotopic to the identity map $\mathrm{id}_X \colon X \to X$.

(b) The punctured disk $D^2 \setminus \{0\}$ deformation retracts onto its boundary S^1.

(c) More generally, $D^n \setminus \{0\}$ deformation retracts onto its boundary S^{n-1}.

(d) Similarly, $\mathbb{R}^n \setminus \{0\}$ deformation retracts onto a sphere S^{n-1}.

Of course in this situation we have that

$$H_n(X, A) \cong H_n(A, A) = 0.$$

Exercise 73.3.8. Show that if $A \subseteq V \subseteq X$, and A is a deformation retract of V, then $H_n(X, A) \cong H_n(X, V)$ for all n. (Use Problem 72F*. Solution in next section.)

[1] This might be called a *deformation retraction in the weak sense* in other resources, such as [Ha02]

§73.4 Excision

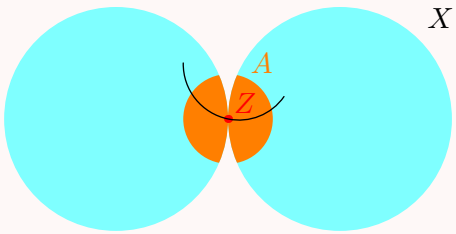
Now for the key geometric result, which is the analog of Theorem 72.3.1 for our relative homology groups.

Theorem 73.4.1 (Excision)
Let $Z \subseteq A \subseteq X$ be subspaces such that the closure of Z is contained in the interior of A. Then the inclusion $\iota(X \setminus Z, A \setminus Z) \hookrightarrow (X, A)$ (viewed as a map of pairs) induces an isomorphism of relative homology groups

$$H_n(X \setminus Z, A \setminus Z) \cong H_n(X, A).$$

This means we can *excise* (delete) a subset Z of A in computing the relative homology groups $H_n(X, A)$. This should intuitively make sense: since we are "modding out by points in A", the internals of the set A should not matter so much.

Example 73.4.2
Excision may seem trivial (for a "relative cycle modulo relative boundary" in $H_n(X, A)$, just tweak the part that lies inside A until it doesn't touch Z), until you realize that it isn't always possible — you may accidentally cut a cycle apart! For example:



The main application of excision is to decide when $H_n(X, A) \cong \widetilde{H}_n(X/A)$. Answer:

Theorem 73.4.3 (Relative homology $\implies$ quotient space)
Let X be a space and A be a closed subspace such that A is a deformation retract of some open set $V \subseteq X$. Then the quotient map $q\colon X \to X/A$ induces an isomorphism

$$H_n(X, A) \cong H_n(X/A, A/A) \cong \widetilde{H}_n(X/A).$$

The key idea of the proof is: While it is not necessarily true that $H(X, A) \cong H(X/A, A/A)$ (indeed, we have seen two counterexamples earlier), if we cut out A, then we trivially have $H(X - A, A - A) \cong H(X/A - A/A, A/A - A/A)$. Unfortunately, this group is not isomorphic to $H(X, A)$, so we fix that using the set V — that is, $H(X - A, V - A) \cong H(X/A - A/A, V/A - A/A)$. The rest of the work is to use excision theorem and deformation retract to show the left hand side is isomorphic to $H(X, A)$, and the right hand side is isomorphic to $H(X/A)$.

Proof. By hypothesis, we can consider the following maps of pairs:

$$\begin{aligned} r &\colon (V, A) \to (A, A) \\ q &\colon (X, A) \to (X/A, A/A) \\ \widehat{q} &\colon (X - A, V - A) \to (X/A - A/A, V/A - A/A). \end{aligned}$$

Moreover, r is a pair-homotopy equivalence. Considering the long exact sequence of a triple (which was Problem 72F*) we have a diagram

$$\begin{array}{ccccccc} H_n(V,A) & \longrightarrow & H_n(X,A) & \xrightarrow{f} & H_n(X,V) & \longrightarrow & H_{n-1}(V,A) \\ \downarrow{\scriptstyle \cong}\,{\scriptstyle r} & & & & & & \downarrow{\scriptstyle \cong}\,{\scriptstyle r} \\ \underbrace{H_n(A,A)}_{=0} & & & & & & \underbrace{H_{n-1}(A,A)}_{=0} \end{array}$$

where the isomorphisms arise since r is a pair-homotopy equivalence. So f is an isomorphism. Similarly the map

$$g \colon H_n(X/A, A/A) \to H_n(X/A, V/A)$$

is an isomorphism.

Now, consider the commutative diagram

$$\begin{array}{ccccc} H_n(X,A) & \xrightarrow{f} & H_n(X,V) & \xleftarrow{\text{Excise}} & H_n(X-A, V-A) \\ \downarrow{\scriptstyle q_*} & & & & \downarrow{\scriptstyle \cong}\,{\scriptstyle \widehat{q}_*} \\ H_n(X/A, A/A) & \xrightarrow[g]{} & H_n(X/A, V/A) & \xleftarrow{\text{Excise}} & H_n(X/A - A/A, V/A - A/A) \end{array}$$

and observe that the rightmost arrow $\widehat{q}_*$ is an isomorphism, because outside of A the map $\widehat{q}$ is the identity. We know f and g are isomorphisms, as are the two arrows marked with "Excise" (by excision). From this we conclude that q_* is an isomorphism. Of course we already know that homology relative to a point is just the relative homology groups (this is the important case of Lemma 73.2.1). □

§73.5 Some applications

One nice application of excision is to compute $\widetilde{H}_n(X \vee Y)$.

Theorem 73.5.1 (Homology of wedge sums)

Let X and Y be spaces with basepoints $x_0 \in X$ and $y_0 \in Y$, and assuming each point is a deformation retract of some open neighborhood. Then for every n we have

$$\widetilde{H}_n(X \vee Y) = \widetilde{H}_n(X) \oplus \widetilde{H}_n(Y).$$

Proof. Apply Theorem 73.4.3 with the subset $\{x_0, y_0\}$ of $X \amalg Y$,

$$\begin{aligned} \widetilde{H}_n(X \vee Y) \cong \widetilde{H}_n((X \amalg Y)/\{x_0, y_0\}) &\cong H_n(X \amalg Y, \{x_0, y_0\}) \\ &\cong H_n(X, \{x_0\}) \oplus H_n(Y, \{y_0\}) \\ &\cong \widetilde{H}_n(X) \oplus \widetilde{H}_n(Y). \end{aligned}$$

□

Another application is to give a second method of computing $H_n(S^m)$. To do this, we will prove that

$$\widetilde{H}_n(S^m) \cong \widetilde{H}_{n-1}(S^{m-1})$$

for any $n, m > 1$. However,

- $\widetilde{H}_0(S^n)$ is $\mathbb{Z}$ for $n = 0$ and 0 otherwise.
- $\widetilde{H}_n(S^0)$ is $\mathbb{Z}$ for $m = 0$ and 0 otherwise.

So by induction on $\min\{m, n\}$ we directly obtain that

$$\widetilde{H}_n(S^m) \cong \begin{cases} \mathbb{Z} & m = n \\ 0 & \text{otherwise} \end{cases}$$

which is what we wanted.

To prove the claim, let's consider the exact sequence formed by the pair $X = D^2$ and $A = S^1$.

Example 73.5.2 (The long exact sequence for $(X, A) = (D^2, S^1)$)

Consider D^2 (which is contractible) with boundary S^1. Clearly S^1 is a deformation retraction of $D^2 \setminus \{0\}$, and if we fuse all points on the boundary together we get $D^2/S^1 \cong S^2$. So we have a long exact sequence

$$\begin{array}{ccccc}
\widetilde{H}_2(S^1) & \longrightarrow & \underbrace{\widetilde{H}_2(D^2)}_{=0} & \longrightarrow & \widetilde{H}_2(S^2) \\
 & \searrow & & & \\
\widetilde{H}_1(S^1) & \longrightarrow & \underbrace{\widetilde{H}_1(D^2)}_{=0} & \longrightarrow & \widetilde{H}_1(S^2) \\
 & \searrow & & & \\
\widetilde{H}_0(S^1) & \longrightarrow & \underbrace{\widetilde{H}_0(D^2)}_{=0} & \longrightarrow & \underbrace{\widetilde{H}_0(S^2)}_{=0}
\end{array}$$

From this diagram we read that

$$\ldots, \quad \widetilde{H}_3(S^2) = \widetilde{H}_2(S^1), \quad \widetilde{H}_2(S^2) = \widetilde{H}_1(S^1), \quad \widetilde{H}_1(S^2) = \widetilde{H}_0(S^1).$$

More generally, the exact sequence for the pair $(X, A) = (D^m, S^{m-1})$ shows that $\widetilde{H}_n(S^m) \cong \widetilde{H}_{n-1}(S^{m-1})$, which is the desired conclusion.

§73.6 Invariance of dimension

Here is one last example of an application of excision.

Definition 73.6.1. Let X be a space and $p \in X$ a point. The kth **local homology group** of p at X is defined as

$$H_k(X, X \setminus \{p\}).$$

Note that for any open neighborhood U of p, we have by excision that

$$H_k(X, X \setminus \{p\}) \cong H_k(U, U \setminus \{p\}).$$

Thus this local homology group only depends on the space near p.

Theorem 73.6.2 (Invariance of dimension, Brouwer 1910)
Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ be nonempty open sets. If U and V are homeomorphic, then $m = n$.

Proof. Consider a point $x \in U$ and its local homology groups. By excision,

$$H_k(\mathbb{R}^n, \mathbb{R}^n \setminus \{x\}) \cong H_k(U, U \setminus \{x\}).$$

But since $\mathbb{R}^n \setminus \{x\}$ is homotopic to S^{n-1}, the long exact sequence of Theorem 72.2.5 tells us that

$$H_k(\mathbb{R}^n, \mathbb{R}^n \setminus \{x\}) \cong \begin{cases} \mathbb{Z} & k = n \\ 0 & \text{otherwise.} \end{cases}$$

Analogously, given $y \in V$ we have

$$H_k(\mathbb{R}^m, \mathbb{R}^m \setminus \{y\}) \cong H_k(V, V \setminus \{y\}).$$

If $U \cong V$, we thus deduce that

$$H_k(\mathbb{R}^n, \mathbb{R}^n \setminus \{x\}) \cong H_k(\mathbb{R}^m, \mathbb{R}^m \setminus \{y\})$$

for all k. This of course can only happen if $m = n$. □

§73.7 A few harder problems to think about

Problem 73A. Let $X = S^1 \times S^1$ and $Y = S^1 \vee S^1 \vee S^2$. Show that

$$H_n(X) \cong H_n(Y)$$

for every integer n.

Problem 73B (Hatcher §2.1 exercise 18)**.** Consider $\mathbb{Q} \subset \mathbb{R}$. Compute $\widetilde{H}_1(\mathbb{R}, \mathbb{Q})$.

Problem 73C*. What are the local homology groups of a topological n-manifold?

Problem 73D. Let

$$X = \{(x, y) \mid x \geq 0\} \subseteq \mathbb{R}^2$$

denote the half-plane. What are the local homology groups of points in X?

Problem 73E (Brouwer-Jordan separation theorem, generalizing Jordan curve theorem)**.** Let $X \subseteq \mathbb{R}^n$ be a subset which is homeomorphic to S^{n-1}. Prove that $\mathbb{R}^n \setminus X$ has exactly two path-connected components.

74 Bonus: Cellular homology

We now introduce cellular homology, which essentially lets us compute the homology groups of any CW complex we like.

§74.1 Degrees

Prototypical example for this section: $z \mapsto z^d$ *has degree* d.

For any $n > 0$ and map $f \colon S^n \to S^n$, consider

$$f_* \colon \underbrace{H_n(S^n)}_{\cong \mathbb{Z}} \to \underbrace{H_n(S^n)}_{\cong \mathbb{Z}}$$

which must be multiplication by some constant d. This d is called the **degree** of f, denoted $\deg f$.

Question 74.1.1. Show that $\deg(f \circ g) = \deg(f)\deg(g)$.

As we mentioned in Example 71.2.12, roughly speaking:

$\deg f$ **counts how many times** $\operatorname{im} f$ **wraps around** S^n**.**

Or, it counts how many "S^n bags" that $\operatorname{im} f$ consists of.

Example 74.1.2 (Degree)

(a) For $n = 1$, the map $z \mapsto z^k$ (viewing $S^1 \subseteq \mathbb{C}$) has degree k.

(b) A reflection map $(x_0, x_1, \dots, x_n) \mapsto (-x_0, x_1, \dots, x_n)$ has degree -1; we won't prove this, but geometrically this should be clear.

(c) The antipodal map $x \mapsto -x$ has degree $(-1)^{n+1}$ since it's the composition of $n+1$ reflections as above. We denote this map by $-\mathrm{id}$.

Obviously, if f and g are homotopic, then $\deg f = \deg g$. In fact, a theorem of Hopf says that this is a classifying invariant: anytime $\deg f = \deg g$, we have that f and g are homotopic.

One nice application of this:

Theorem 74.1.3 (Hairy ball theorem)

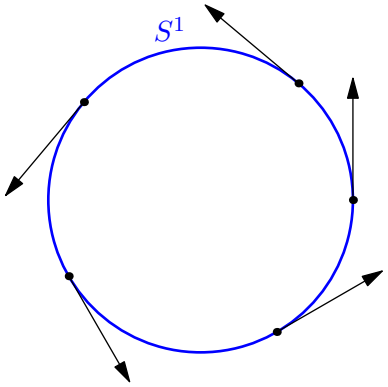
If $n > 0$ is even, then S^n doesn't have a continuous field of nonzero tangent vectors.

Proof. If the vectors are nonzero then WLOG they have norm 1; that is for every x we have an orthogonal unit vector $v(x)$. Then we can construct a homotopy map $F \colon S^n \times [0,1] \to S^n$ by

$$(x, t) \mapsto (\cos \pi t)x + (\sin \pi t)v(x).$$

which gives a homotopy from id to $-\text{id}$. So $\deg(\text{id}) = \deg(-\text{id})$, which means $1 = (-1)^{n+1}$ so n must be odd. □

Of course, the one can construct such a vector field whenever n is odd. For example, when $n = 1$ such a vector field is drawn below.



§74.2 Cellular chain complex

Before starting, we state:

Lemma 74.2.1 (CW homology groups)
Let X be a CW complex. Then

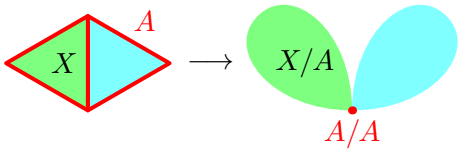
$$H_k(X^n, X^{n-1}) \cong \begin{cases} \mathbb{Z}^{\oplus \#n\text{-cells of } X} & k = n \\ 0 & \text{otherwise.} \end{cases}$$

and

$$H_k(X^n) \cong \begin{cases} H_k(X) & k \le n-1 \\ 0 & k \ge n+1. \end{cases}$$

Proof. The first part is immediate by noting that (X^n, X^{n-1}) satisfies the hypothesis of Theorem 73.4.3, so $H_k(X^n, X^{n-1}) \cong \widetilde{H}_k(X^n/X^{n-1})$, and X^n/X^{n-1} is a wedge sum of several n-spheres.

For an example, for $n = 2$ (the "spheres" are drawn as a balloon-shaped blob here):



For the second part, fix k and note that, as long as $n \le k-1$ or $n \ge k+2$,

$$\underbrace{H_{k+1}(X^n, X^{n-1})}_{=0} \to H_k(X^{n-1}) \to H_k(X^n) \to \underbrace{H_k(X^n, X^{n-1})}_{=0}.$$

So we have isomorphisms

$$H_k(X^{k-1}) \cong H_k(X^{k-2}) \cong \ldots \cong H_k(X^0) = 0$$

and

$$H_k(X^{k+1}) \cong H_k(X^{k+2}) \cong \ldots \cong H_k(X). \qquad \square$$

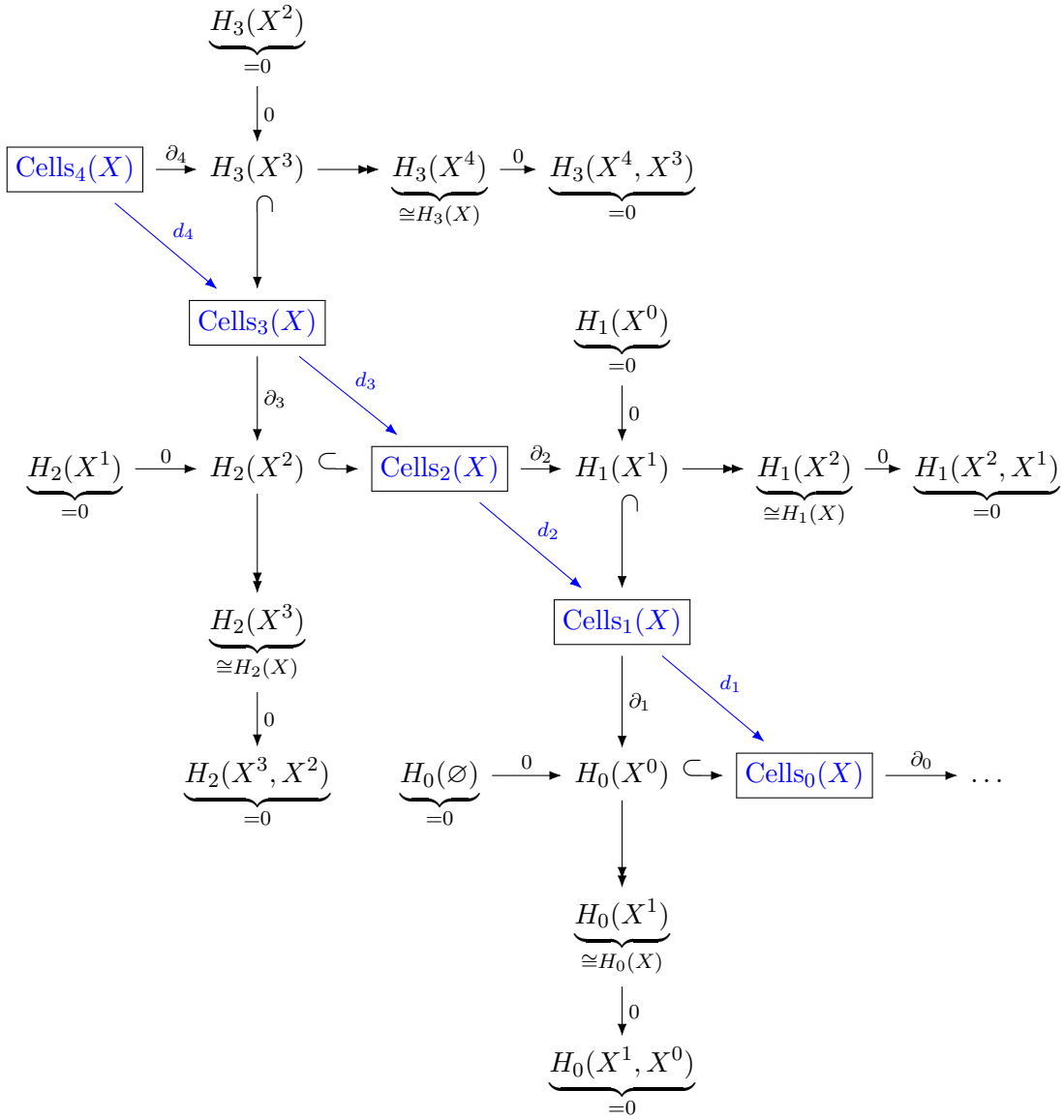
So, we know that the groups $H_k(X^k, X^{k-1})$ are super nice: they are free abelian with basis given by the cells of X. So, we give them a name:

Definition 74.2.2. For a CW complex X, we define

$$\text{Cells}_k(X) = H_k(X^k, X^{k-1})$$

where $\text{Cells}_0(X) = H_0(X^0, \varnothing) = H_0(X^0)$ by convention. So $\text{Cells}_k(X)$ is an abelian group with basis given by the k-cells of X.

Now, using $\text{Cells}_k = H_k(X^k, X^{k-1})$ let's use our long exact sequence and try to string together maps between these. Consider the following diagram.



The idea is that we have taken all the exact sequences generated by adjacent skeletons, and strung them together at the groups $H_k(X^k)$, with half the exact sequences being laid out vertically and the other half horizontally.

In that case, composition generates a sequence of blue maps between the $H_k(X^k, X^{k-1})$ as shown.

> **Question 74.2.3.** Show that the composition of two adjacent blue arrows is zero.

So from the diagram above, we can read off a sequence of arrows

$$\ldots \xrightarrow{d_5} \text{Cells}_4(X) \xrightarrow{d_4} \text{Cells}_3(X) \xrightarrow{d_3} \text{Cells}_2(X) \xrightarrow{d_2} \text{Cells}_1(X) \xrightarrow{d_1} \text{Cells}_0(X) \xrightarrow{d_0} 0.$$

This is a chain complex, called the **cellular chain complex**; as mentioned before all the homology groups are free, but these ones are especially nice because for most reasonable CW complexes, they are also finitely generated (unlike the massive $C_\bullet(X)$ that we had earlier). In other words, the $H_k(X^k, X^{k-1})$ are especially nice "concrete" free groups that one can actually work with.

The other reason we care is that in fact:

Theorem 74.2.4 (Cellular chain complex gives $H_n(X)$)
The kth homology group of the cellular chain complex is isomorphic to $H_k(X)$.

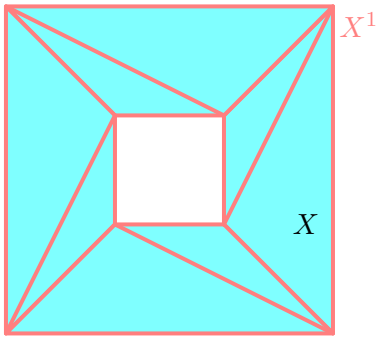
Proof. Follows from the diagram; Problem 74D. □

§74.3 Digression: why are the homology groups equal?

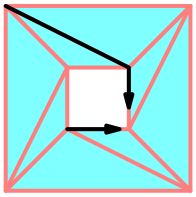
There is another intuition that explains it — roughly speaking,

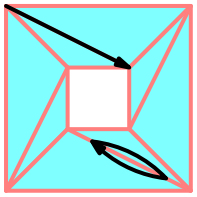
$$H_k(\text{Cells}_\bullet(X)) = \frac{\text{aligned cycle}}{\text{aligned boundary}} = \frac{\text{aligned cycle} \times \text{fuzz}}{\text{aligned boundary} \times \text{fuzz}} = \frac{\text{cycle}}{\text{boundary}} = H_k(X).$$

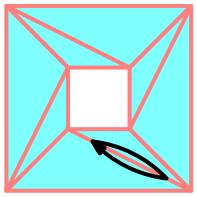
Let me explain. Consider a CW-complex X that looks like the following, where X^1 is drawn in red. Each blue region corresponds to a 2-cell.

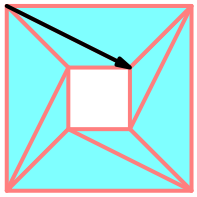


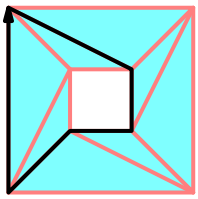
Then, look at the following figure.

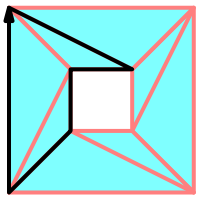

$C_1(X^1)$

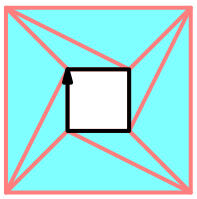

$Z_1(X^1, X^0)$


$B_1(X^1, X^0)$


$\text{Cells}_1(X)$


$\ker(\text{Cells}_1(X) \to \text{Cells}_0(X))$


$\text{im}(\text{Cells}_2(X) \to \text{Cells}_1(X))$


$H_1(\text{Cells}_\bullet(X))$

It looks like a lot, so let me explain.

- The first picture depicts a typical element of $C_1(X^1)$ — that is, a 1-chain that is contained in X^1, being the formal sum of two maps from Δ^1 to X^1, whose image is drawn as black arrows.

  Note that only the image of the maps are depicted, information such as which point of the simplex Δ^1 get mapped to which point inside X^1 is not shown — although different continuous maps give rise to different elements of $C_1(X^1)$.

- The second picture depicts a typical element of $Z_1(X^1, X^0)$ — that is, the *relative cycles*.

  Although we never formally defined what is a relative cycle or the groups $Z_1(X^1, X^0)$, you can guess the definition from the definition of $Z_1(X^1)$ — it is the subgroup of $C_1(X^1, X^0) = C_1(X^1)/C_1(X^0)$ whose boundary vanish.

  The fact that the loop on the bottom is flattened is just to make it look nicer — the whole thing is contained inside the red skeleton i.e. X^1.

  Of course, being an element of the quotient, only a representative element is depicted — the "modded out" parts are the chains that are entirely contained inside X^0 i.e. some vertices.

- The third picture depicts a typical element of $B_1(X^1, X^0)$ i.e. the *relative boundaries*.

  This belongs to im $\left(C_2(X^1, X^0) \xrightarrow{\partial} C_1(X^1, X^0)\right)$ — in words, there is some 2-chain whose boundary equals the depicted element.

- The fourth picture depicts a typical element of $\text{Cells}_1(X)$ — that is, "relative cycles mod relative boundaries".

  Hopefully it is intuitively obvious how this group is isomorphic to the abelian group generated by each 1-cell of X.

  We can in fact think of each of these elements as an "aligned element" of $C_1(X^1)$ where all endpoints lie inside a vertex (that is, the boundary of that element is inside X^0), and for each 1-cell, a canonical 1-simplex is chosen to cover that cell (note that different simplexes with the same image intuitively corresponds to "reparametrization" to change the speed, and the difference between a simplex and its reparametrization is in fact an element of $B_1(X^1, X^0)$ — try to make this rigorous! Hint: use the prism operator.)

- The fifth picture depicts a typical element of
$$\ker\left(\text{Cells}_1(X) \xrightarrow{\partial} \text{Cells}_0(X)\right)$$
which can be thought of as a "cellular cycle", or a 1-cycle (element of $Z_1(X)$) that is "aligned", as explained above.
- The sixth picture depicts a typical element of
$$\text{im}\left(\text{Cells}_2(X) \xrightarrow{\partial} \text{Cells}_1(X)\right)$$
which can be thought of as a "cellular boundary", or a 1-boundary (element of $B_1(X)$[1]) that is "aligned" in the same sense as above.
- Finally, the last picture is $H_1(\text{Cells}_\bullet(X))$, which is
$$H_1(\text{Cells}_\bullet(X)) = \frac{\ker\left(\text{Cells}_1(X) \xrightarrow{\partial} \text{Cells}_0(X)\right)}{\text{im}\left(\text{Cells}_2(X) \xrightarrow{\partial} \text{Cells}_1(X)\right)}.$$
Or, roughly speaking,
$$H_1(\text{Cells}_\bullet(X)) = \frac{\text{aligned cycle}}{\text{aligned boundary}}.$$

That is what we mean by $\frac{\text{aligned cycle}}{\text{aligned boundary}} = \frac{\text{cycle}}{\text{boundary}}$. With a suitable formalization and arbitrary selection of canonical simplices,[2] we can make the argument above rigorous.

What do we mean by "fuzz"? This part is hopefully obvious, but the point is that an aligned cycle can be "moved around" a bit (with reparametrization, or addition of elements in $B_1(X^1, X^0)$) while still keep it a cycle (that is, an element of $Z_1(X)$). Similarly for aligned boundaries.

So, the point is — we can "cancel" the common fuzz factor in the numerator and the denominator, and the result will remain the same.

Refer to [Ha02] for some formal treatment on simplicial approximation.

§74.4 Application: Euler characteristic via Betti numbers

A nice application of this is to define the **Euler characteristic** of a finite CW complex X. Of course we can write
$$\chi(X) = \sum_n (-1)^n \cdot \#(n\text{-cells of } X)$$
which generalizes the familiar $V-E+F$ formula. However, this definition is unsatisfactory because it depends on the choice of CW complex, while we actually want $\chi(X)$ to only depend on the space X itself (and not how it was built). In light of this, we prove that:

Theorem 74.4.1 (Euler characteristic via Betti numbers)
For any finite CW complex X we have
$$\chi(X) = \sum_n (-1)^n \operatorname{rank} H_n(X).$$

[1] This is not an element of $B_1(X^1)$! Think about why.
[2] Technically we need a so-called Δ-complex structure on X, but we don't define Δ-structure in the Napkin. See [Ha02] for details.

Thus $\chi(X)$ does not depend on the choice of CW decomposition. The numbers

$$b_n = \operatorname{rank} H_n(X)$$

are called the **Betti numbers** of X. In fact, we can use this to define $\chi(X)$ for any reasonable space; we are happy because in the (frequent) case that X is a CW complex, the definition coincides with the normal definition of the Euler characteristic.

Proof. We quote the fact that if $0 \to A \to B \to C \to D \to 0$ is exact then $\operatorname{rank} B + \operatorname{rank} D = \operatorname{rank} A + \operatorname{rank} C$. Then for example the row

$$\underbrace{H_2(X^1)}_{=0} \xrightarrow{0} H_2(X^2) \hookrightarrow H_2(X^2, X^1) \xrightarrow{\partial_2} H_1(X^1) \twoheadrightarrow \underbrace{H_1(X^2)}_{\cong H_1(X)} \xrightarrow{0} \underbrace{H_1(X^2, X^1)}_{=0}$$

from the cellular diagram gives

$$\#(\text{2-cells}) + \operatorname{rank} H_1(X) = \operatorname{rank} H_2(X^2) + \operatorname{rank} H_1(X^1).$$

More generally,

$$\#(k\text{-cells}) + \operatorname{rank} H_{k-1}(X) = \operatorname{rank} H_k(X^k) + \operatorname{rank} H_{k-1}(X^{k-1})$$

which holds also for $k = 0$ if we drop the H_{-1} terms (since $\#0\text{-cells} = \operatorname{rank} H_0(X^0)$ is obvious). Multiplying this by $(-1)^k$ and summing across $k \geq 0$ gives the conclusion. □

Example 74.4.2 (Examples of Betti numbers)

(a) The Betti numbers of S^n are $b_0 = b_n = 1$, and zero elsewhere. The Euler characteristic is $1 + (-1)^n$.

(b) The Betti numbers of a torus $S^1 \times S^1$ are $b_0 = 1$, $b_1 = 2$, $b_2 = 1$, and zero elsewhere. Thus the Euler characteristic is 0.

(c) The Betti numbers of $\mathbb{CP}^n$ are $b_0 = b_2 = \cdots = b_{2n} = 1$, and zero elsewhere. Thus the Euler characteristic is $n + 1$.

(d) The Betti numbers of the Klein bottle are $b_0 = 1$, $b_1 = 1$ and zero elsewhere. Thus the Euler characteristic is 0, the same as the sphere (also since their CW structures use the same number of cells).

One notices that in the "nice" spaces S^n, $S^1 \times S^1$ and $\mathbb{CP}^n$ there is a nice symmetry in the Betti numbers, namely $b_k = b_{n-k}$. This is true more generally; see Poincaré duality and Problem 76A†.

§74.5 The cellular boundary formula

In fact, one can describe explicitly what the maps d_n are. Recalling that $H_k(X^k, X^{k-1})$ has a basis the k-cells of X, we obtain:

Theorem 74.5.1 (Cellular boundary formula for $k = 1$)

For $k = 1$,

$$d_1 \colon \operatorname{Cells}_1(X) \to \operatorname{Cells}_0(X)$$

is just the boundary map.

Theorem 74.5.2 (Cellular boundary for $k > 1$)
Let $k > 1$ be a positive integer. Let e^k be an k-cell, and let $\{e_\beta^{k-1}\}_\beta$ denote all $(k-1)$-cells of X. Then

$$d_k \colon \operatorname{Cells}_k(X) \to \operatorname{Cells}_{k-1}(X)$$

is given on basis elements by

$$d_k(e^k) = \sum_\beta d_\beta e_\beta^{k-1}$$

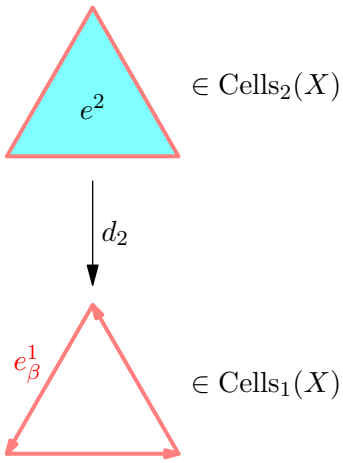
where d_β is be the degree of the composed map

$$S^{k-1} = \partial e^k \xrightarrow{\text{attach}} X^{k-1} \twoheadrightarrow S_\beta^{k-1}.$$

Here the first arrow is the attaching map for e^k and the second arrow is the quotient of collapsing $X^{k-1} \setminus e_\beta^{k-1}$ to a point.

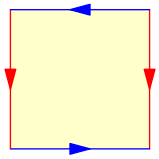
What is the degree doing here? Remember that a basis element $e^k \in \operatorname{Cells}_k(X)$ is just a k-cell, and its boundary should be just the cells that forms its boundary.

With the same visualization as above, we can do something like the following.



But it's not that easy! Note that in a CW complex, the boundary of a k-cell can be fused into *arbitrary points* in X^{k-1}, so an "edge" of a k-cell need not be a $k-1$-cell.

To make matters worse, sometimes there may be a duplicated edge — in the Klein bottle, each pair of two opposing edges depicted actually *the same edge*, possibly in different orientations.



In such a case, we need to count the *multiplicity* of each edge — and this is exactly what the degree of the map counts! We will see an explicit example of computing the homology groups of the Klein bottle in just a moment.

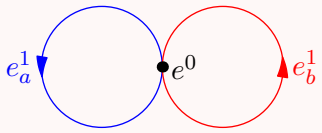
This gives us an algorithm for computing homology groups of a CW complex:

- Construct the cellular chain complex, where $\text{Cells}_k(X)$ is $\mathbb{Z}^{\oplus \#k\text{-cells}}$.
- $d_1\colon \text{Cells}_1(X) \to \text{Cells}_0(X)$ is just the boundary map (so $d_1(e^1)$ is the difference of the two endpoints).
- For any $k > 1$, we compute $d_k\colon \text{Cells}_k(X) \to \text{Cells}_{k-1}(X)$ on basis elements as follows. Repeat the following for each k-cell e^k:
  - For every $k-1$ cell e_β^{k-1}, compute the degree of the boundary of e^k welded onto the boundary of e_β^{k-1}, say d_β.
  - Then $d_k(e^k) = \sum_\beta d_\beta e_\beta^{k-1}$.
- Now we have the maps of the cellular chain complex, so we can compute the homologies directly (by taking the quotient of the kernel by the image).

We can use this for example to compute the homology groups of the torus again, as well as the Klein bottle and other spaces.

Example 74.5.3 (Cellular homology of a torus)

Consider the torus built from e^0, e_a^1, e_b^1 and e^2 as before, where e^2 is attached via the word $aba^{-1}b^{-1}$. For example, X^1 is



The cellular chain complex is

$$0 \longrightarrow \mathbb{Z}e^2 \xrightarrow{d_2} \mathbb{Z}e_a^1 \oplus \mathbb{Z}e_b^1 \xrightarrow{d_1} \mathbb{Z}e^0 \xrightarrow{d_0} 0$$

Now apply the cellular boundary formulas:

- Recall that d_1 was the boundary formula. We have $d_1(e_a^1) = e_0 - e_0 = 0$ and similarly $d_1(e_b^1) = 0$. So $d_1 = 0$.
- For d_2, consider the image of the boundary e^2 on e_a^1. Around X^1, it wraps once around e_a^1, once around e_b^1, again around e_a^1 (in the opposite direction), and again around e_b^1. Once we collapse the entire e_b^1 to a point, we see that the degree of the map is 0. So $d_2(e^2)$ has no e_a^1 coefficient. Similarly, it has no e_b^1 coefficient, hence $d_2 = 0$.

Thus

$$d_1 = d_2 = 0.$$

So at every map in the complex, the kernel of the map is the whole space while the image is $\{0\}$. So the homology groups are $\mathbb{Z}$, $\mathbb{Z}^{\oplus 2}$, $\mathbb{Z}$.

Example 74.5.4 (Cellular homology of the Klein bottle)

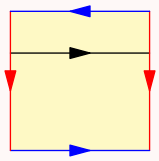
Let X be a Klein bottle. Consider cells e^0, e_a^1, e_b^1 and e^2 as before, but this time e^2 is attached via the word $abab^{-1}$. So d_1 is still zero, but this time we have $d_2(e^2) = 2e_a^1$ instead (why?). So our diagram looks like

$$\begin{aligned} 0 \xrightarrow{0} \mathbb{Z}e^2 &\xrightarrow{d_2} \mathbb{Z}e_a^1 \oplus \mathbb{Z}e_b^1 \xrightarrow{d_1} \mathbb{Z}e^0 \xrightarrow{d_0} 0 \\ e^2 &\longmapsto 2e_a^1 \\ e_1^a &\longmapsto 0 \\ e_1^b &\longmapsto 0 \end{aligned}$$

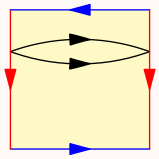
So we get that $H_0(X) \cong \mathbb{Z}$, but

$$H_1(X) \cong \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$$

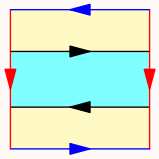
this time (it is $\mathbb{Z}^{\oplus 2}$ modulo a copy of $2\mathbb{Z}$). Also, $\ker d_2 = 0$, and so now $H_2(X) = 0$. Let us sanity check that this makes sense — that is, there is some cycle that is not a boundary, but when doubled it become a boundary.
Actually, most cycles work.



If we double up the path, we get something like the following.



Here is the important part: since the two blue edges are identified in opposite direction, we can pull one of the path across the edge to reverse its direction... but now the region is in fact the boundary of the cyan region! So we're done.



It remains to convince yourself that the difference of two homotopy equivalent path is a boundary.

§74.6 A few harder problems to think about

Problem 74A†. Let n be a positive integer. Show that

$$H_k(\mathbb{CP}^n) \cong \begin{cases} \mathbb{Z} & k = 0, 2, 4, \ldots, 2n \\ 0 & \text{otherwise.} \end{cases}$$

Problem 74B. Show that a non-surjective map $f \colon S^n \to S^n$ has degree zero.

Problem 74C (Moore spaces)**.** Let $G_1, G_2, \ldots, G_N$ be a sequence of finitely generated abelian groups. Construct a space X such that

$$\widetilde{H}_n(X) \cong \begin{cases} G_n & 1 \le n \le N \\ 0 & \text{otherwise.} \end{cases}$$

Problem 74D. Prove Theorem 74.2.4, showing that the homology groups of X coincide with the homology groups of the cellular chain complex.

Problem 74E†. Let n be a positive integer. Show that

$$H_k(\mathbb{RP}^n) \cong \begin{cases} \mathbb{Z} & \text{if } k = 0 \text{ or } k = n \equiv 1 \pmod 2 \\ \mathbb{Z}/2\mathbb{Z} & \text{if } k \text{ is odd and } 0 < k < n \\ 0 & \text{otherwise.} \end{cases}$$

75 Singular cohomology

Here's one way to motivate this chapter. It turns out that:

- $H_n(\mathbb{CP}^2) \cong H_n(S^2 \vee S^4)$ for every n.
- $H_n(\mathbb{CP}^3) \cong H_n(S^2 \times S^4)$ for every n.

This is unfortunate, because if possible we would like to be able to tell these spaces apart (as they are in fact not homotopy equivalent), but the homology groups cannot tell the difference between them.

In this chapter, we'll define a *cohomology group* $H^n(X)$ and $H^n(Y)$. In fact, the H^n's are completely determined by the H_n's by the so-called *universal coefficient theorem*. However, it turns out that one can take all the cohomology groups and put them together to form a *cohomology ring* $H^\bullet$.[1] We will then see that $H^\bullet(X) \ncong H^\bullet(Y)$ as rings.

§75.1 Cochain complexes

Definition 75.1.1. A **cochain complex** $A^\bullet$ is algebraically the same as a chain complex, except that the indices increase. So it is a sequence of abelian groups

$$\ldots \xrightarrow{\delta} A^{n-1} \xrightarrow{\delta} A^n \xrightarrow{\delta} A^{n+1} \xrightarrow{\delta} \ldots.$$

such that $\delta^2 = 0$. Notation-wise, we're now using superscripts, and use δ rather ∂. We define the **cohomology groups** by

$$H^n(A^\bullet) = \ker\left(A^n \xrightarrow{\delta} A^{n+1}\right) / \operatorname{im}\left(A^{n-1} \xrightarrow{\delta} A^n\right).$$

Example 75.1.2 (de Rham cohomology)
We have already met one example of a cochain complex: let M be a smooth manifold and $\Omega^k(M)$ be the additive group of k-forms on M. Then we have a cochain complex

$$0 \xrightarrow{d} \Omega^0(M) \xrightarrow{d} \Omega^1(M) \xrightarrow{d} \Omega^2(M) \xrightarrow{d} \ldots.$$

The resulting cohomology is called **de Rham cohomology**, described later.

Aside from de Rham's cochain complex, **the most common way to get a cochain complex is to *dualize* a chain complex.** Specifically, pick an abelian group G; note that $\operatorname{Hom}(-, G)$ is a contravariant functor, and thus takes every chain complex

$$\ldots \xrightarrow{\partial} A_{n+1} \xrightarrow{\partial} A_n \xrightarrow{\partial} A_{n-1} \xrightarrow{\partial} \ldots$$

into a cochain complex: letting $A^n = \operatorname{Hom}(A_n, G)$ we obtain

$$\ldots \xrightarrow{\delta} A^{n-1} \xrightarrow{\delta} A^n \xrightarrow{\delta} A^{n+1} \xrightarrow{\delta} \ldots.$$

[1] [Ha02] has an explanation why is it that cohomology has more structures than homology — roughly speaking, the natural maps $X \times X \to X$ must be a projection which is not very interesting, but there is a more interesting natural map $X \to X \times X$ given by $p \mapsto (p, p)$.

where $\delta(A_n \xrightarrow{f} G) = A_{n+1} \xrightarrow{\partial} A \xrightarrow{f} G$.

These are the cohomology groups we study most in algebraic topology, so we give a special notation to them.

Definition 75.1.3. Given a chain complex $A_\bullet$ of abelian groups and another group G, we let

$$H^n(A_\bullet; G)$$

denote the cohomology groups of the dual cochain complex $A^\bullet$ obtained by applying $\operatorname{Hom}(-, G)$. In other words, $H^n(A_\bullet; G) = H^n(A^\bullet)$.

§75.2 Cohomology of spaces

Prototypical example for this section: $C^0(X; G)$ *all functions* $X \to G$ *while* $H^0(X)$ *are those functions* $X \to G$ *constant on path components.*

The case of interest is our usual geometric situation, with $C_\bullet(X)$.

Definition 75.2.1. For a space X and abelian group G, we define $C^\bullet(X; G)$ to be the dual to the singular chain complex $C_\bullet(X)$, called the **singular cochain complex** of X; its elements are called **cochains**.

Then we define the **cohomology groups** of the space X as

$$H^n(X; G) \coloneqq H^n(C_\bullet(X); G) = H^n(C^\bullet(X; G)).$$

Remark 75.2.2 — Note that if G is also a ring (like $\mathbb{Z}$ or $\mathbb{R}$), then $H^n(X; G)$ is not only an abelian group but actually a G-module.

Example 75.2.3 ($C^0(X; G)$, $C^1(X; G)$, and $H^0(X; G)$)
Let X be a topological space and consider $C^\bullet(X)$.

- $C_0(X)$ is the free abelian group on X, and $C^0(X) = \operatorname{Hom}(C_0(X), G)$. So a 0-cochain is a function that takes every point of X to an element of G.
- $C_1(X)$ is the free abelian group on 1-simplices in X. So $C^1(X)$ needs to take every 1-simplex to an element of G.

Let's now try to understand $\delta\colon C^0(X) \to C^1(X)$. Given a 0-cochain $\phi \in C^0(X)$, i.e. a homomorphism $\phi\colon C^0(X) \to G$, what is $\delta\phi\colon C^1(X) \to G$? Answer:

$$\delta\phi\colon [v_0, v_1] \mapsto \phi([v_0]) - \phi([v_1]).$$

Hence, elements of $\ker(C^0 \xrightarrow{\delta} C^1) \cong H^0(X; G)$ are those cochains that are *constant on path-connected components.*

In particular, much like $H_0(X)$, we have

$$H^0(X) \cong G^{\oplus r}$$

if X has r path-connected components (where r is finite[2]).

Abuse of Notation 75.2.4. In this chapter the only cochain complexes we will consider are dual complexes as above. So, any time we write a cochain complex $A^\bullet$ it is implicitly given by applying $\Hom(-, G)$ to $A_\bullet$.

The higher cohomology groups $H^n(X; G)$ (or even the cochain groups $C^n(X; G) = \Hom(C_n(X), G)$) are harder to describe concretely.

§75.3 Cohomology of spaces is functorial

We now check that the cohomology groups still exhibit the same nice functorial behavior. First, let's categorize the previous results we had:

Question 75.3.1. Define $\mathsf{CoCmplx}$ the category of cochain complexes.

Exercise 75.3.2. Interpret $\Hom(-, G)$ as a contravariant functor from

$$\Hom(-, G) \colon \mathsf{Cmplx}^{\mathrm{op}} \to \mathsf{CoCmplx}.$$

This means in particular that given a chain map $f \colon A_\bullet \to B_\bullet$, we naturally obtain a dual map $f^\vee \colon B^\bullet \to A^\bullet$.

Question 75.3.3. Interpret $H^n \colon \mathsf{CoCmplx} \to \mathsf{Grp}$ as a functor. Compose these to get a contravariant functor $H^n(-; G) \colon \mathsf{Cmplx}^{\mathrm{op}} \to \mathsf{Grp}$.

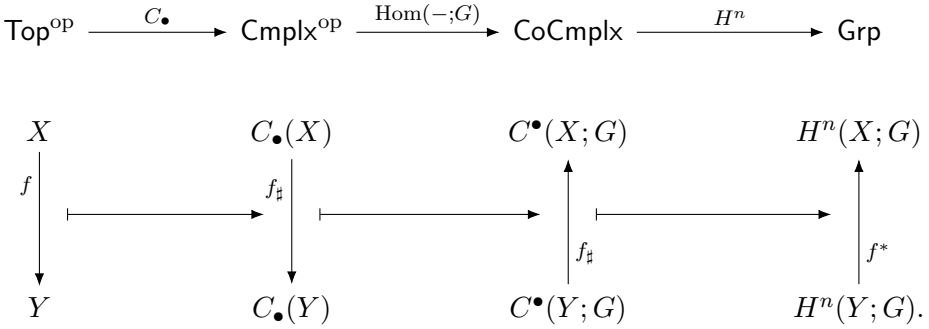
Then in exact analog to our result that $H_n \colon \mathsf{hTop} \to \mathsf{Grp}$ we have:

Theorem 75.3.4 ($H^n(-; G) : \mathsf{hTop}^{\mathrm{op}} \to \mathsf{Grp}$)
For every n, $H^n(-; G)$ is a contravariant functor from $\mathsf{hTop}^{\mathrm{op}}$ to Grp.

Proof. The idea is to leverage the work we already did in constructing the prism operator earlier. First, we construct the entire sequence of functors from $\mathsf{Top}^{\mathrm{op}} \to \mathsf{Grp}$:

[2]Something funny happens if X has *infinitely* many path-connected components: say $X = \coprod_\alpha X_\alpha$ over an infinite indexing set. In this case we have $H_0(X) = \bigoplus_\alpha G$ while $H^0(X) = \prod_\alpha G$. For homology we get a *direct sum* while for cohomology we get a *direct product*.

These are actually different for infinite indexing sets. For general modules $\bigoplus_\alpha M_\alpha$ is *defined* to only allow to have *finitely many* nonzero terms. (This was never mentioned earlier in the Napkin, since I only ever defined $M \oplus N$ and extended it to finite direct sums.) No such restriction holds for $\prod_\alpha G_\alpha$ a product of groups. This corresponds to the fact that $C_0(X)$ is formal linear sums of 0-chains (which, like all formal sums, are finite) from the path-connected components of G. But a cochain of $C^0(X)$ is a *function* from each path-connected component of X to G, where there is no restriction.



Here $f^\sharp = (f_\sharp)^\vee$, and f^* is the resulting induced map on homology groups of the cochain complex.

So as before all we have to show is that $f \simeq g$, then $f^* = g^*$. Recall now that there is a prism operator such that $f_\sharp - g_\sharp = P\partial + \partial P$. If we apply the entire functor $\operatorname{Hom}(-;G)$ we get that $f^\sharp - g^\sharp = \delta P^\vee + P^\vee \delta$ where $P^\vee \colon C^{n+1}(Y;G) \to C^n(X;G)$. So $f^\sharp$ and $g^\sharp$ are chain homotopic thus $f^* = g^*$. □

§75.4 Universal coefficient theorem

We now wish to show that the cohomology groups are determined up to isomorphism by the homology groups: given $H_n(A_\bullet)$, we can extract $H^n(A_\bullet; G)$. This is achieved by the *universal coefficient theorem*.

Theorem 75.4.1 (Universal coefficient theorem)

Let $A_\bullet$ be a chain complex of *free* abelian groups, and let G be another abelian group. Then there is a natural short exact sequence

$$0 \to \operatorname{Ext}(H_{n-1}(A_\bullet), G) \to H^n(A_\bullet; G) \xrightarrow{h} \operatorname{Hom}(H_n(A_\bullet), G) \to 0.$$

In addition, this exact sequence is *split* so in particular

$$H^n(C_\bullet; G) \cong \operatorname{Ext}(H_{n-1}(A_\bullet), G) \oplus \operatorname{Hom}(H_n(A_\bullet), G).$$

Fortunately, in our case of interest, $A_\bullet$ is $C_\bullet(X)$ which is by definition free.

There are two things we need to explain, what the map h is and the map Ext is.

It's not too hard to guess how

$$h \colon H^n(A_\bullet; G) \to \operatorname{Hom}(H_n(A_\bullet), G)$$

is defined. An element of $H^n(A_\bullet; G)$ is represented by a function which sends a cycle in A_n to an element of G. The content of the theorem is to show that h is surjective with kernel $\operatorname{Ext}(H_{n-1}(A_\bullet), G)$.

What about Ext? It turns out that $\operatorname{Ext}(-, G)$ is the so-called **Ext functor**, defined as follows. Let H be an abelian group, and consider a **free resolution** of H, by which we mean an exact sequence

$$\ldots \xrightarrow{f_2} F_1 \xrightarrow{f_1} F_0 \xrightarrow{f_0} H \to 0$$

with each F_i free. Then we can apply $\operatorname{Hom}(-, G)$ to get a cochain complex

$$\ldots \xleftarrow{f_2^\vee} \operatorname{Hom}(F_1, G) \xleftarrow{f_1^\vee} \operatorname{Hom}(F_0, G) \xleftarrow{f_0^\vee} \operatorname{Hom}(H, G) \leftarrow 0.$$

but *this cochain complex need not be exact* (in categorical terms, $\operatorname{Hom}(-, G)$ does not preserve exactness). We define

$$\operatorname{Ext}(H, G) \coloneqq \ker(f_2^\vee)/\operatorname{im}(f_1^\vee)$$

and it's a theorem that this doesn't depend on the choice of the free resolution. There's a lot of homological algebra that goes into this, which I won't take the time to discuss; but the upshot of the little bit that I did include is that the Ext functor is very easy to compute in practice, since you can pick any free resolution you want and compute the above.

Remark 75.4.2 — You have seen a "free resolution" before in a disguised form — in Section 18.3, we proved the structure theorem of finitely-generated modules over PID by writing any module M as $R^{\oplus d}/K$, with both $R^{\oplus d}$ and K free. This gives a free resolution

$$\cdots \to 0 \to K \hookrightarrow R^{\oplus d} \twoheadrightarrow M \to 0.$$

Intuitively, you can think of the Ext functor as measuring the "maps that should be there but aren't" — you will gradually gain some intuitions after seeing some examples.[a]

[a]Taken from https://mathoverflow.net/a/679.

Lemma 75.4.3 (Computing the Ext functor)

For any abelian groups G, H, H' we have

(a) $\operatorname{Ext}(H \oplus H', G) = \operatorname{Ext}(H, G) \oplus \operatorname{Ext}(H', G)$.

(b) $\operatorname{Ext}(H, G) = 0$ for H free, and

(c) $\operatorname{Ext}(\mathbb{Z}/n\mathbb{Z}, G) = G/nG$.

Proof. For (a), note that if $\cdots \to F_1 \to F_0 \to H \to 0$ and $\cdots \to F_1' \to F_0' \to F_0' \to H' \to 0$ are free resolutions, then so is $F_1 \oplus F_1' \to F_0 \oplus F_0' \to H \oplus H' \to 0$.

For (b), note that $0 \to H \to H \to 0$ is a free resolution.

Part (c) follows by taking the free resolution

$$0 \to \mathbb{Z} \xrightarrow{\times n} \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z} \to 0$$

and applying $\operatorname{Hom}(-, G)$ to it.

Question 75.4.4. Finish the proof of (c) from here. □

Question 75.4.5. Some Ext practice: compute $\operatorname{Ext}(\mathbb{Z}^{\oplus 2015}, G)$ and $\operatorname{Ext}(\mathbb{Z}/30\mathbb{Z}, \mathbb{Z}/4\mathbb{Z})$.

§75.5 Explanation for universal coefficient theorem

There is so much unexplained symbols and formulas in the previous chapter that may make you scream:

> I don't care if $\mathbb{CP}^2$ and $S^2 \vee S^4$ are distinct anymore! What are these spaces anyway?

Nevertheless, it is not all that difficult. There are two key points to be read from the theorem:

- Even though $H_n(A_\bullet) = 0$, it is still possible for $H^n(A_\bullet; G) \neq 0$ if $\operatorname{Ext}(H_{n-1}(A_\bullet), G) \neq 0$.

  In low-dimensional cases, we can actually visualize it — Section 75.7 does that for the Klein bottle.

- $H^n(A_\bullet; G)$ is uniquely determined by $H_n(A_\bullet)$ and G, regardless of what $A_\bullet$ is, as long as each A_n is free.

Which means: if you wish, you can forget about the formula in the universal coefficient theorem, and use the cellular chain complex $\text{Cells}_\bullet(X)$ to compute cohomology by:

$$H^n(X; G) = \frac{\ker(\operatorname{Hom}(\text{Cells}_n(X), G) \to \operatorname{Hom}(\text{Cells}_{n+1}(X), G))}{\operatorname{im}(\operatorname{Hom}(\text{Cells}_{n-1}(X), G) \to \operatorname{Hom}(\text{Cells}_n(X), G))}.$$

After all, the cellular chain complex and the singular chain complex are both free and have the same homology groups, so by the universal coefficient theorem they must have the same cohomology groups.

Nevertheless, the formula of the universal coefficient theorem is desirable because, more often than not, the chain complex $A_\bullet$ is more complicated than $H_\bullet(A_\bullet)$.

Example 75.5.1

The Klein bottle's cellular chain complex has the following form:

$$\dots \to \mathbb{Z} \xrightarrow{1 \mapsto (0,2)} \mathbb{Z}^2 \xrightarrow{(a,b) \mapsto 0} \mathbb{Z}.$$

The homology groups is:

$$H_2 = 0, H_1 = \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}, H_0 = \mathbb{Z}.$$

It's indeed simpler, but only marginally (there are 3 generators instead of 4, and we don't need to keep track of the maps) because cellular homology is already so efficient.

Where does the formula come from, again? You can think of it like this. Because the universal coefficient theorem tells us that $H^\bullet(A_\bullet; G)$ only depends on $H_\bullet(A_\bullet)$, if we're given $H_\bullet$, we just construct *any* chain complex of free abelian groups $A_\bullet$ and dualize it.

Assume $H_k = 0$ for every terms, except $H_{n-1} \neq 0$. Then, tautologically, $H^n \cong \operatorname{Ext}(H_{n-1}; G)$ — a free resolution *is* a chain complex!

Exercise 75.5.2. Verify this. (Hint: Starting from the exact sequence $Z_{n-1} \to H_{n-1} \to 0$. Can you extend it to a free resolution of H_{n-1}?)

Assume $H_k = 0$ for every terms, except $H_n \neq 0$. Then we can see $H^n \cong \operatorname{Hom}(H_n, G)$.

The universal coefficient theorem simply states that the choice of free resolution doesn't matter, and that if the other terms can be nonzero, H^n is the direct sum of the two groups in the two cases above.

If you want, you can even prove the fact that the choice of free resolution does not matter yourself — it's a bit tricky, but not all that difficult. It boils down to the

construction of maps between the chain complexes (it's not difficult to ensure the diagram commutes, the groups are free so we can send the basis wherever we want), and show the two free resolutions are chain homotopic.

§75.6 Example computation of cohomology groups

Prototypical example for this section: Possibly $H^n(S^m)$.

The universal coefficient theorem gives us a direct way to compute any cohomology groups, provided we know the homology ones.

Example 75.6.1 (Cohomology groups of S^m)
It is straightforward to compute $H^n(S^m)$ now: all the Ext terms vanish since $H_n(S^m)$ is always free, and hence we obtain that

$$H^n(S^m) \cong \operatorname{Hom}(H_n(S^m), G) \cong \begin{cases} G & n = m, n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Example 75.6.2 (Cohomology groups of torus)
This example has no nonzero Ext terms either, since this time $H^n(S^1 \times S^1)$ is always free. So we obtain

$$H^n(S^1 \times S^1) \cong \operatorname{Hom}(H_n(S^1 \times S^1), G).$$

Since $H_n(S^1 \times S^1)$ is $\mathbb{Z}$, $\mathbb{Z}^{\oplus 2}$, $\mathbb{Z}$ in dimensions $n = 1, 2, 1$ we derive that

$$H^n(S^1 \times S^1) \cong \begin{cases} G & n = 0, 2 \\ G^{\oplus 2} & n = 1. \end{cases}$$

From these examples one might notice that:

Lemma 75.6.3 (0th and 1th cohomology groups are just duals)
For $n = 0$ and $n = 1$, we have

$$H^n(X; G) \cong \operatorname{Hom}(H_n(X), G).$$

Proof. It's already been shown for $n = 0$. For $n = 1$, notice that $H_0(X)$ is free, so the Ext term vanishes. □

Example 75.6.4 (Cohomology groups of Klein bottle)
This example will actually have Ext term. Recall from Example 74.5.4 that if K is a Klein Bottle then its homology groups are $\mathbb{Z}$ in dimension $n = 0$ and $\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ in $n = 1$, and 0 elsewhere.
For $n = 0$, we again just have $H^0(K; G) \cong \operatorname{Hom}(\mathbb{Z}, G) \cong G$. For $n = 1$, the Ext

term is $\operatorname{Ext}(H_0(K), G) \cong \operatorname{Ext}(\mathbb{Z}, G) = 0$ so

$$H^1(K; G) \cong \operatorname{Hom}(\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}, G) \cong G \oplus \operatorname{Hom}(\mathbb{Z}/2\mathbb{Z}, G).$$

We have that $\operatorname{Hom}(\mathbb{Z}/2\mathbb{Z}, G)$ is the subgroup of elements of order 2 in G (and $0 \in G$).
But for $n = 2$, we have our first interesting Ext group: the exact sequence is

$$0 \to \operatorname{Ext}(\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}, G) \to H^2(X; G) \to \underbrace{H_2(X)}_{=0} \to 0.$$

Thus, we have

$$H^2(X; G) \cong (\operatorname{Ext}(\mathbb{Z}, G) \oplus \operatorname{Ext}(\mathbb{Z}/2\mathbb{Z}, G)) \oplus 0 \cong G/2G.$$

All the higher groups vanish. In summary:

$$H^n(X; G) \cong \begin{cases} G & n = 0 \\ G \oplus \operatorname{Hom}(\mathbb{Z}/2\mathbb{Z}, G) & n = 1 \\ G/2G & n = 2 \\ 0 & n \geq 3. \end{cases}$$

§75.7 Visualization of cohomology groups

We try to make sense of $C^n(X; G)$ and $H^n(X; G)$, for higher values of n.

As above, $C_n(X; G)$ is the free abelian group on n-simplices on X, so an element $f \in C^n(X; G)$ is a function that takes each n-simplex to an element of G (and extends linearly to all of $C_n(X; G)$).

This assignment of value need not have any nice properties — recall that a n-simplex is simply a (continuous) map $\sigma \colon \Delta^n \to X$, and different maps σ_1 and σ_2 are considered different even though $\operatorname{im} \sigma_1 = \operatorname{im} \sigma_2$. In particular,

- If $[v_0, v_1, v_2]$ is a singular simplex, it need not be the case that $f([v_0, v_1, v_2]) + f([v_0, v_2, v_1]) = 0$.
- A singular n-simplex ($n \geq 1$) with image contained in a point need not be mapped to 0 by f.

But it *does not matter* that elements of $C_n(X)$ aren't this nice! We will see below why this is the case.

In the homology case (Definition 71.2.2), we defined:

$$\begin{aligned} Z_n(X) &\coloneqq \ker \left(C_n(X) \xrightarrow{\partial} C_{n-1}(X) \right), \\ B_n(X) &\coloneqq \operatorname{im} \left(C_{n+1}(X) \xrightarrow{\partial} C_n(X) \right), \\ H_n(X) &\coloneqq Z_n(X)/B_n(X). \end{aligned}$$

Elements of $Z_n(X)$ and $B_n(X)$ are called cycles and boundaries respectively, with the obvious geometrical interpretation.

So,

$$H_n(X) = \frac{n\text{-cycles}}{n\text{-boundaries}}.$$

For the current section, we will temporarily define:

$$\begin{aligned} Z^n(X;G) &\coloneqq \ker\left(C^n(X;G) \xrightarrow{\delta} C^{n+1}(X;G)\right), \\ B^n(X;G) &\coloneqq \operatorname{im}\left(C^{n-1}(X;G) \xrightarrow{\delta} C^n(X;G)\right), \\ H^n(X;G) &\coloneqq Z^n(X;G)/B^n(X;G). \end{aligned}$$

For this section, we will call elements of $Z^n(X;G)$ the **cocycles** and elements of $B^n(X;G)$ the **coboundaries** respectively. Once again,

$$H^n(X;G) = \frac{n\text{-cocycles}}{n\text{-coboundaries}}.$$

It's less clear geometrically why the elements are named as above, but if we assume the group G is a *field* (where the group operation is the addition operation in the field), then[3] we have:

- a n-cocycle is a map that sends every n-boundary to $0 \in G$;
- a n-coboundary is a map that sends every n-cycle to $0 \in G$.

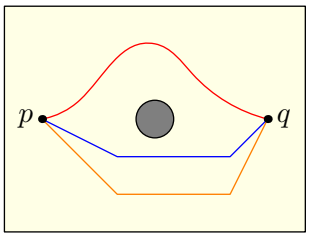
The first statement is clear (definition chasing), the second statement is only generally true in one direction (that a coboundary sends every cycle to 0; but a map that sends every cycle to 0 need not be a coboundary — we will see this later on with the Klein bottle example).

Let us see what a n-cocycle must look like. First,

Homotopic chains with the same boundary are mapped to the same value by cocycles.

We defined what it means for two k-simplices to be homotopic in Section 65.4 — in the current situation, we require in addition that the boundaries are always fixed.

For instance, the blue and the orange 1-simplices below are homotopic, but not the red 1-simplex.

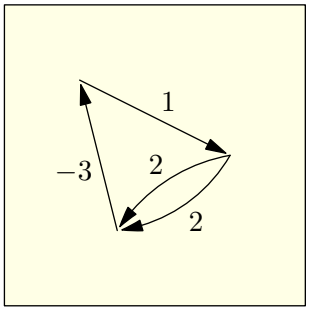


Proof is not difficult — you just need to show that the difference between two homotopic k-simplices is the boundary of something (their interior!), and write the interior as the sum of some $k+1$-simplices. (Hint: The easiest way is actually to write the interior as the difference of two $k+1$-simplices instead, and be careful of vertex ordering issues.)

Exercise 75.7.1. Finish the proof.

A typical 1-cocycle might look something like this, where each arrow is labeled with the value assigned to that 1-simplex. Remember that a cycle must be mapped to 0.

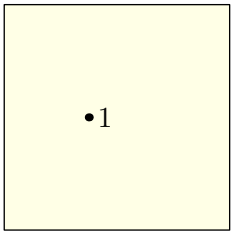
[3] Refer to https://math.stackexchange.com/q/4712676.



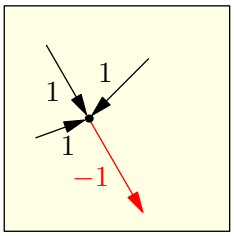
Now, the next observation is that:

If we only consider cocycles modulo coboundaries, we basically only care about values assigned to the cycles.

Why? Remember that a k-coboundary is the δ of some $(k-1)$-cochain. So, given this 0-cochain:



Its δ would look something like this:



So, roughly speaking,

By adding or subtracting a coboundary to a given cochain, we can adjust the value assigned to most chains however we want.

I said "most chains" because, if the chains form a *cycle*, adding a coboundary won't let us change its assigned value.

Fortunately,

- Cycles that are *boundaries* always get assigned the value 0.
- Homotopic cycles get assigned the same value.

  As a generalization, in fact, cycles that are homologous (i.e. they get mapped to the same value under the map $Z_k(X) \twoheadrightarrow H_k(X)$) are assigned the same value.

Therefore,

Knowing the value of a cocycle on each "cycle modulo boundary" almost determines that cocycle, modulo coboundaries.

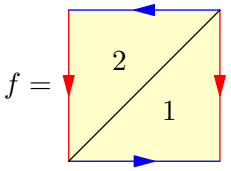
In symbols: $H^n(X;G)$ is "almost isomorphic" to $\mathrm{Hom}(H_n(X),G)$.

In other words, a cocycle modulo coboundary can be "evaluated" on a cycle modulo boundary.

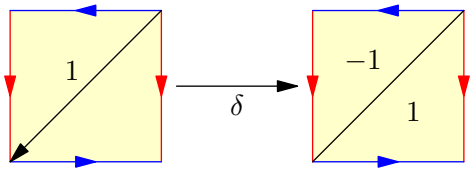
This is precisely what the universal coefficient theorem states, although it says something more: the "error term" is exactly $\mathrm{Ext}(H_{n-1}(X),G)$.

Why would the error term exist? We had an example above, computing $H^2(K;G)$ for K the Klein bottle. Let us work through it geometrically, assume $G=\mathbb{Z}$ for now.

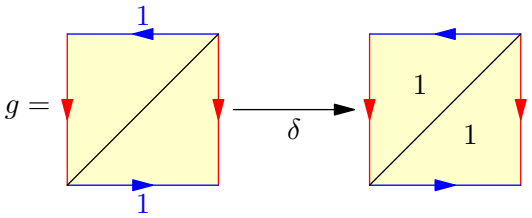
A typical 2-cochain $f\in C^2(K;\mathbb{Z})$ may look something like this. (Only value assigned to a few 2-simplices is depicted, there are too many 2-simplices for us to draw.)



A coboundary may look like this — identical to the situation above, the value assigned to particular simplex doesn't matter, we can "transfer" the assigned value between the two simplices by adding a coboundary.



So, we may just say that the value assigned to the whole surface of the Klein bottle is 3 — formally, let $e^2_K\in C_2(K)$ be the sum of the two 2-simplices above, we can write $f(e^2_K)=3$. However:



The boundary of the 2-chain corresponding to the whole surface of the Klein bottle is 2 times the blue edge, so δ of the 1-cochain whose value on the blue edge is 1 will assign the value 2 to e^2_K.

In symbols: let $e^1_b\in C_1(K)$ be the blue edge, pick $g\in C^1(K;\mathbb{Z})$ such that $g(e^1_b)=1$, then $\delta(g)(e^2_K)=2$. Even though e^2_K is not a cycle, we still need to care about its assigned value modulo 2! Because adding or subtracting the coboundary $\delta(g)$ can only adjust its values in increments of 2.

Therefore,

If the region $e^k\in C_k(X)$ has a boundary $\partial e^k\in C_{k-1}(X)$ divisible by n, then we cares about the value assigned to e^k, modulo n.

This explains where the error term $\text{Ext}(H_{n-1}(X), G)$ comes from.

We have another comparison with de Rham cohomology in Section 76.2 — in that case, the group G is a field, $\mathbb{R}$, so $\text{Ext}(H_{n-1}(X), G)$ is always zero.

§75.8 Relative cohomology groups

One can also define relative cohomology groups in the obvious way: dualize the chain complex

$$\ldots \xrightarrow{\partial} C_1(X, A) \xrightarrow{\partial} C_0(X, A) \to 0$$

to obtain a cochain complex

$$\ldots \xleftarrow{\delta} C^1(X, A; G) \xleftarrow{\delta} C^0(X, A; G) \leftarrow 0.$$

We can take the cohomology groups of this.

Definition 75.8.1. The groups thus obtained are the **relative cohomology groups** are denoted $H^n(X, A; G)$.

In addition, we can define reduced cohomology groups as well. One way to do it is to take the augmented singular chain complex

$$\ldots \xrightarrow{\partial} C_1(X) \xrightarrow{\partial} C_0(X) \xrightarrow{\varepsilon} \mathbb{Z} \to 0$$

and dualize it to obtain

$$\ldots \xleftarrow{\delta} C^1(X; G) \xleftarrow{\delta} C^0(X; G) \xleftarrow{\varepsilon^\vee} \underbrace{\text{Hom}(\mathbb{Z}, G)}_{\cong G} \leftarrow 0.$$

Since the $\mathbb{Z}$ we add is also free, the universal coefficient theorem still applies. So this will give us reduced cohomology groups.

However, since we already defined the relative cohomology groups, it is easiest to simply define:

Definition 75.8.2. The **reduced cohomology groups** of a nonempty space X, denoted $\widetilde{H}^n(X; G)$, are defined to be $H^n(X, \{*\}; G)$ for some point $* \in X$.

§75.9 A few harder problems to think about

Problem 75A* (Wedge product cohomology)**.** For any G and n we have

$$\widetilde{H}^n(X \vee Y; G) \cong \widetilde{H}^n(X; G) \oplus \widetilde{H}^n(Y; G).$$

Problem 75B†. Prove that for a field F of characteristic zero and a space X with finitely generated homology groups:

$$H^k(X, F) \cong (H_k(X))^\vee.$$

Thus over fields cohomology is the dual of homology.

Problem 75C ($\mathbb{Z}/2\mathbb{Z}$-cohomology of $\mathbb{RP}^n$)**.** Prove that

$$H^m(\mathbb{RP}^n, \mathbb{Z}/2\mathbb{Z}) \cong \begin{cases} \mathbb{Z} & m = 0, \text{ or } m \text{ is odd and } m = n \\ \mathbb{Z}/2\mathbb{Z} & 0 < m < n \text{ and } m \text{ is odd} \\ 0 & \text{otherwise.} \end{cases}$$

76 Application of cohomology

In this final chapter on topology, I'll state (mostly without proof) some nice properties of cohomology groups, and in particular introduce the so-called cup product. For an actual treatise on the cup product, see [Ha02] or [Ma13a].

As mentioned in the previous chapter, you can put all the cohomology groups $H^\bullet(X)$ together to form the *cohomology ring*, which gives more structure than the case of homology — enough structure to allow distinguishing between $\mathbb{CP}^2$ and $S^2 \vee S^4$, or between $\mathbb{CP}^3$ and $S^2 \times S^4$.

Even though the description above is completely non-descriptive (it doesn't give you insight into *what* the structure is about), and actually, some people would say:

> It does not matter what homology measures intuitively, as it is a convenient tool that takes something very difficult (topology) and turns it into something simple (abelian group).

Nevertheless, it is interesting that the cup product *is actually visualizable*! At least when the dimension does not exceed 3.

§76.1 Poincaré duality

First cool result: you may have noticed symmetry in the (co)homology groups of "nice" spaces like the torus or S^n. In fact this is predicted by:

Theorem 76.1.1 (Poincaré duality)

If M is a smooth oriented compact n-manifold, then we have a natural isomorphism

$$H^k(M;\mathbb{Z}) \cong H_{n-k}(M)$$

for every k. In particular, $H^k(M) = 0$ for $k > n$.

So for smooth oriented compact manifolds, cohomology and homology groups are not so different.

From this follows the symmetry that we mentioned when we first defined the Betti numbers:

Corollary 76.1.2 (Symmetry of Betti numbers)

Let M be a smooth oriented compact n-manifold, and let b_k denote its Betti number. Then

$$b_k = b_{n-k}.$$

Proof. Problem 76A†. □

§76.2 de Rham cohomology

We now reveal the connection between differential forms and singular cohomology.

Let M be a smooth manifold. We are interested in the homology and cohomology groups of M. We specialize to the case $G = \mathbb{R}$, the additive group of real numbers.

> **Question 76.2.1.** Check that $\text{Ext}(H, \mathbb{R}) = 0$ for any finitely generated abelian group H.

Thus, with real coefficients the universal coefficient theorem says that

$$H^k(M; \mathbb{R}) \cong \text{Hom}(H_k(M), \mathbb{R}) = (H_k(M))^\vee$$

where we view $H_k(X)$ as a real vector space. So, we'd like to get a handle on either $H_k(M)$ or $H^k(M; \mathbb{R})$.

Consider the cochain complex

$$0 \to \Omega^0(M) \xrightarrow{d} \Omega^1(M) \xrightarrow{d} \Omega^2(M) \xrightarrow{d} \Omega^3(M) \xrightarrow{d} \dots$$

and let $H^k_{\text{dR}}(M)$ denote its cohomology groups. Thus the de Rham cohomology is the closed forms modulo the exact forms.

$$\text{Cochain} : \text{Cocycle} : \text{Coboundary} = k\text{-form} : \text{Closed form} : \text{Exact form}.$$

The whole punch line is:

Theorem 76.2.2 (de Rham's theorem)

For any smooth manifold M, we have a natural isomorphism

$$H^k(M; \mathbb{R}) \cong H^k_{\text{dR}}(M).$$

So the theorem is that the real cohomology groups of manifolds M are actually just given by the behavior of differential forms. Thus,

One can metaphorically think of elements of cohomology groups as G-valued differential forms on the space.

Why does this happen? In fact, we observed already behavior of differential forms which reflects holes in the space. For example, let $M = S^1$ be a circle and consider the **angle form** α (see Example 44.7.4). The form α is closed, but not exact, because it is possible to run a full circle around S^1. So the failure of α to be exact is signaling that $H_1(S^1) \cong \mathbb{Z}$.

As another piece of intuition, note that:

- each k-differential form ω can be interpreted as a function that takes each k-smooth submanifold $S \subseteq M$, and returns a real number $\int_S \omega$.
- let us pretend that all k-simplices are smooth for now. Then we have:
  - The k-cochains are the functions that sends each k-simplex to a real number.
  - The k-cocycles are the k-cochains that sends the boundaries to 0.
  - The k-coboundaries are the k-cochains that sends the cycles to 0.

  Meanwhile:
  - The differential forms are the functions that sends each k-simplex to a real number, satisfying certain linearity and smoothness properties — for instance:

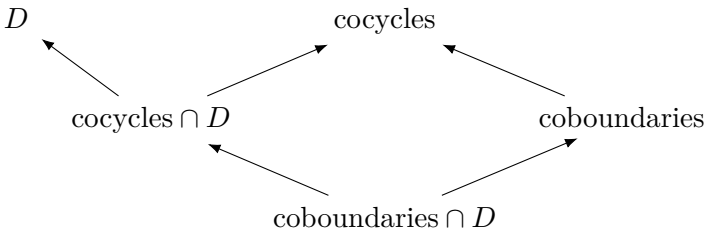
  - if $k \geq 1$ and a k-simplex has the image contained in a point, then it must be sent to 0;
  - if we reparametrize a k-simplex, the assigned value must be the same;
  - if we flip two vertices of a k-simplex, the assigned value must be negated;
  - if a k-simplex can be formed by gluing two k-simplices along a face, then the assigned value must be the sum of the corresponding values assigned to the sub-k-simplices;
  - etc.
- The closed forms are the differential forms that sends the boundaries to 0;
- The exact forms are the differential forms that send the cycles to 0.

We can't help but noticing the parallel — the point is:

$$H^k(M;\mathbb{R}) = \frac{\text{cocycles}}{\text{coboundaries}} \cong \frac{\text{cocycles} \cap \text{differential forms}}{\text{coboundaries} \cap \text{differential forms}} = H^k_{\text{dR}}(M).$$

Roughly speaking, both the numerator and the denominator on the left are bigger, and they *cancels out*. We can compare this with Section 74.3.

Or, as a figure (for space reasons, the group of differential forms is denoted D):



This is precisely the setup of the second isomorphism theorem,[1] and you can try to work out why the two quotients are isomorphic.

§76.3 Graded rings

Prototypical example for this section: Polynomial rings are commutative graded rings, while $\bigwedge^\bullet(V)$ is anticommutative.

In the de Rham cohomology, the differential forms can interact in another way: given a k-form α and an ℓ-form β, we can consider a $(k+\ell)$-form

$$\alpha \wedge \beta.$$

So we can equip the set of forms with a "product", satisfying $\beta \wedge \alpha = (-1)^{k\ell}\alpha \wedge \beta$. This is a special case of a more general structure:

Definition 76.3.1. A **graded pseudo-ring** R is an abelian group

$$R = \bigoplus_{d \geq 0} R^d$$

where R^0, R^1, ..., are abelian groups, with an additional associative binary operation $\times\colon R \to R$. We require that if $r \in R^d$ and $s \in R^e$, we have $rs \in R^{d+e}$. Elements of an R^d are called **homogeneous elements**; if $r \in R^d$ and $r \neq 0$, we write $|r| = d$.

[1] See Section 3.6.

Note that we do *not* assume commutativity. In fact, these "rings" may not even have an identity 1. We use other words if there are additional properties:

Definition 76.3.2. A **graded ring** is a graded pseudo-ring with 1. If it is commutative we say it is a **commutative graded ring**.

Definition 76.3.3. A graded (pseudo-)ring R is **anticommutative** if for any homogeneous r and s we have

$$rs = (-1)^{|r||s|}sr.$$

Remark 76.3.4 — Why not $rs = -sr$? This definition is inspired by the fact that the wedge product is anticommutative. Note that, for $f_1, \dots, f_r, g_1, \dots, g_s$ being 0-forms, let $f = df_1 \wedge df_2 \wedge \cdots \wedge df_r$ be a r-form and $g = dg_1 \wedge dg_2 \wedge \cdots \wedge dg_s$ be a s-form, then starting from the expression

$$f \wedge g = (df_1 \wedge df_2 \wedge \cdots \wedge df_r) \wedge (dg_1 \wedge dg_2 \wedge \cdots \wedge dg_s)$$

if you repeatedly swap two adjacent entries, it will take rs swaps total in order to obtain the expression

$$g \wedge f = (dg_1 \wedge dg_2 \wedge \cdots \wedge dg_s) \wedge (df_1 \wedge df_2 \wedge \cdots \wedge df_r).$$

By linearity, we can prove that in general, for any r-form f and any s-form g, we have $fg = (-1)^{rs}gf$.

To summarize:

Flavors of graded rings	Need not have 1	Must have a 1
No Assumption	graded pseudo-ring	graded ring
Anticommutative	anticommutative pseudo-ring	anticommutative ring
Commutative		commutative graded ring

Example 76.3.5 (Examples of graded rings)

(a) The ring $R = \mathbb{Z}[x]$ is a **commutative graded ring**, with the dth component being the multiples of x^d.

(b) The ring $R = \mathbb{Z}[x, y, z]$ is a **commutative graded ring**, with the dth component being the abelian group of homogeneous degree d polynomials (and 0).

(c) Let V be a vector space, and consider the abelian group

$$\bigwedge^{\bullet}(V) = \bigoplus_{d \geq 0} \bigwedge^{d}(V).$$

For example, $e_1 + (e_2 \wedge e_3) \in \bigwedge^{\bullet}(V)$, say. We endow $\bigwedge^{\bullet}(V)$ with the product $\wedge$, which makes it into an **anticommutative ring**.

(d) Consider the set of differential forms of a manifold M, say

$$\Omega^{\bullet}(M) = \bigoplus_{d \geq 0} \Omega^{d}(M)$$

endowed with the product $\wedge$. This is an **anticommutative ring**.

All four examples have a multiplicative identity.

Let's return to the situation of $\Omega^\bullet(M)$. Consider again the de Rham cohomology groups $H^k_{\text{dR}}(M)$, whose elements are closed forms modulo exact forms. We claim that:

Lemma 76.3.6 (Wedge product respects de Rham cohomology)
The wedge product induces a map

$$\wedge\colon H^k_{\text{dR}}(M) \times H^\ell_{\text{dR}}(M) \to H^{k+\ell}_{\text{dR}}(M).$$

Proof. First, we recall that the operator d satisfies

$$d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + \alpha \wedge (d\beta).$$

Now suppose α and β are closed forms. Then from the above, $\alpha \wedge \beta$ is clearly closed. Also if α is closed and $\beta = d\omega$ is exact, then $\alpha \wedge \beta$ is exact, from the identity

$$d(\alpha \wedge \omega) = d\alpha \wedge \omega + \alpha \wedge d\omega = \alpha \wedge \beta.$$

Similarly if α is exact and β is closed then $\alpha \wedge \beta$ is exact. Thus it makes sense to take the product modulo exact forms, giving the theorem above. □

Therefore, we can obtain a *anticommutative ring*

$$H^\bullet_{\text{dR}}(M) = \bigoplus_{k \geq 0} H^k_{\text{dR}}(M)$$

with $\wedge$ as a product, and $1 \in \bigwedge^0(\mathbb{R}) = \mathbb{R}$ as the identity.

§76.4 Cup products

Inspired by this, we want to see if we can construct a similar product on $\bigoplus_{k \geq 0} H^k(X; R)$ for any topological space X and ring R (where R is commutative with 1 as always). The way to do this is via the *cup product*.

Then this gives us a way to multiply two cochains, as follows.

Definition 76.4.1. Suppose $\phi \in C^k(X; R)$ and $\psi \in C^\ell(X; R)$. Then we can define their **cup product** $\phi \smile \psi \in C^{k+\ell}(X; R)$ to be

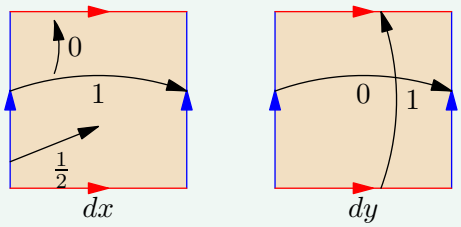
$$(\phi \smile \psi)([v_0, \dots, v_{k+\ell}]) = \phi([v_0, \dots, v_k]) \cdot \psi([v_k, \dots, v_{k+\ell}])$$

where the multiplication is in R.

Question 76.4.2. Assuming R has a 1, which 0-cochain is the identity for $\smile$?

Remark 76.4.3 (Warning) — While you can interpret a n-differential form as a n-cochain the obvious way, the cup product is *not* directly a generalization of the wedge product! For example, let $X = \mathbb{R}^2$, and try to evaluate $dx \smile dy$ on $[v_0, v_1, v_2]$ and $[v_2, v_1, v_0]$ where $v_0 = (1, 0)$, $v_1 = (0, 0)$, $v_2 = (0, 1)$, assume all of the edges are straight lines.
This is because we are not having the alternation operator. Refer to Section 44.5 for details. In this case, the ring G might be $\mathbb{Z}$ where not all nonzero elements have

an inverse, so division would cause trouble.
Nevertheless, the differences will nicely cancel out, and we still have the corresponding element in the cohomology group equal to the element interpreted by the wedge product $dx \wedge dy$ — this is what we mean by $H^\bullet(M;\mathbb{R}) \cong H^\bullet_{\text{dR}}(M)$, stated below.
Let us consider the familiar example of a torus, and the 1-cocycles "dx" and "dy".



From what we know about the wedge product, we want $(dx \wedge dy)(T) = 1$ for T the whole torus (up to a $\pm$ sign). Indeed, with the definition above (work it out! Divide T into two triangles arbitrarily) it will work.
Nevertheless, we don't really care about the cup product itself as much as the induced cup product on the homology ring.

First, we prove an analogous result as before:

Lemma 76.4.4 (δ with cup products)
We have $\delta(\phi \smile \psi) = \delta\phi \smile \psi + (-1)^k \phi \smile \delta\psi$.

Proof. Direct $\sum$ computations. □

Thus, by the same routine we used for de Rham cohomology, we get an induced map

$$\smile : H^k(X;R) \times H^\ell(X;R) \to H^{k+\ell}(X;R).$$

We then define the **singular cohomology ring** whose elements are finite sums in

$$H^\bullet(X;R) = \bigoplus_{k \geq 0} H^k(X;R)$$

and with multiplication given by $\smile$. Thus it is a graded ring (with $1_R \in R$ the identity) and is in fact anticommutative:

Proposition 76.4.5 (Cohomology is anticommutative)
$H^\bullet(X;R)$ is an anticommutative ring, meaning $\phi \smile \psi = (-1)^{k\ell} \psi \smile \phi$.

For a proof, see [Ha02, Theorem 3.11, pages 210-212]. Moreover, we have the de Rham isomorphism

Theorem 76.4.6 (de Rham extends to ring isomorphism)

For any smooth manifold M, the isomorphism of de Rham cohomology groups to singular cohomology groups in facts gives an isomorphism

$$H^\bullet(M;\mathbb{R}) \cong H^\bullet_{\text{dR}}(M)$$

of anticommutative rings.

Therefore, if "differential forms" are the way to visualize the elements of a cohomology group, the wedge product is the correct way to visualize the cup product.

We now present (mostly without proof) the cohomology rings of some common spaces.

Example 76.4.7 (Cohomology of torus)

The cohomology ring $H^\bullet(S^1 \times S^1;\mathbb{Z})$ of the torus is generated by elements $|\alpha| = |\beta| = 1$ which satisfy the relations $\alpha \smile \alpha = \beta \smile \beta = 0$, and $\alpha \smile \beta = -\beta \smile \alpha$. (It also includes an identity 1.) Thus as a $\mathbb{Z}$-module it is

$$H^\bullet(S^1 \times S^1;\mathbb{Z}) \cong \mathbb{Z} \oplus [\alpha\mathbb{Z} \oplus \beta\mathbb{Z}] \oplus (\alpha \smile \beta)\mathbb{Z}.$$

This gives the expected dimensions $1 + 2 + 1 = 4$. It is anti-commutative.

You have already seen the elements α and β as the elements called dx and dy in the remark above.

Example 76.4.8 (Cohomology ring of S^n)

Consider S^n for $n \geq 1$. The nontrivial cohomology groups are given by $H^0(S^n;\mathbb{Z}) \cong H^n(S^n;\mathbb{Z}) \cong \mathbb{Z}$. So as an abelian group

$$H^\bullet(S^n;\mathbb{Z}) \cong \mathbb{Z} \oplus \alpha\mathbb{Z}$$

where α is the generator of $H^n(S^n,\mathbb{Z})$.
Now, observe that $|\alpha \smile \alpha| = 2n$, but since $H^{2n}(S^n;\mathbb{Z}) = 0$ we must have $\alpha \smile \alpha = 0$. So even more succinctly,

$$H^\bullet(S^n;\mathbb{Z}) \cong \mathbb{Z}[\alpha]/(\alpha^2).$$

Confusingly enough, this graded ring is both commutative *and* anti-commutative. The reason is that $\alpha \smile \alpha = 0 = -(\alpha \smile \alpha)$.

Example 76.4.9 (Cohomology ring of real and complex projective space)

It turns out that

$$H^\bullet(\mathbb{RP}^n;\mathbb{Z}/2\mathbb{Z}) \cong \mathbb{Z}/2\mathbb{Z}[\alpha]/(\alpha^{n+1})$$
$$H^\bullet(\mathbb{CP}^n;\mathbb{Z}) \cong \mathbb{Z}[\beta]/(\beta^{n+1})$$

where $|\alpha| = 1$ is a generator of $H^1(\mathbb{RP}^n;\mathbb{Z}/2\mathbb{Z})$ and $|\beta| = 2$ is a generator of $H^2(\mathbb{CP}^n;\mathbb{Z})$.
Confusingly enough, both graded rings are commutative *and* anti-commutative. In

the first case it is because we work in $\mathbb{Z}/2\mathbb{Z}$, for which $1 = -1$, so anticommutative is actually equivalent to commutative. In the second case, all nonzero homogeneous elements have degree 2.

Already we have an interesting example where the cup product $\smile$ is different from the wedge product $\wedge$ — if $n \geq 2$, then the generators α and β above has $\alpha \smile \alpha \neq 0$ and $\beta \smile \beta \neq 0$.

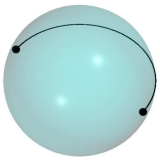
Let us try to see what happens here. The formula above says

$$H^\bullet(\mathbb{RP}^2; \mathbb{Z}/2\mathbb{Z}) \cong \mathbb{Z}/2\mathbb{Z}[\alpha]/(\alpha^3)$$

As an abelian group, there is a single nonzero element in $H^0(\mathbb{RP}^2; \mathbb{Z}/2\mathbb{Z})$, $H^1(\mathbb{RP}^2; \mathbb{Z}/2\mathbb{Z})$, and $H^2(\mathbb{RP}^2; \mathbb{Z}/2\mathbb{Z})$, and the remaining groups are 0.

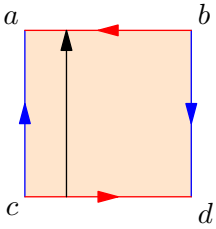
$\mathbb{RP}^2$ isn't too hard to visualize — it's just a 2-sphere, quotient by the relation to identify opposite vertices.

There is a 1-cycle on it that is not homologous to 0:



It's not very easy to show, but every such 1-cycle is homologous to each other, and double of that cycle is homologous to 0.

As such, $H^1(\mathbb{RP}^2; \mathbb{Z}/2\mathbb{Z}) \cong \text{Hom}(H_1(\mathbb{RP}^2), \mathbb{Z}/2\mathbb{Z})$, its only nontrivial element α maps each such 1-cycle to 1.



Consider $\alpha \smile \alpha$. Notice that α acts like both dx and dy at the same time (both the blue edge and the red edge got assigned the value 1), so it assigns the value 1 to the whole surface of the real projective plane! Thus it's nontrivial.

Exercise 76.4.10. Manually compute the cup product $\alpha \smile \alpha$ to verify that. (Divide the surface into some triangles. $[a, c, d] + [a, d, b] - [c, c, d] + [c, d, c]$ is a working choice. Verify that the boundary is nonzero, but is divisible by 2.)

§76.5 Relative cohomology pseudo-rings

For $A \subseteq X$, one can also define a relative cup product

$$H^k(X, A; R) \times H^\ell(X, A; R) \to H^{k+\ell}(X, A; R).$$

After all, if either cochain vanishes on chains in A, then so does their cup product. This lets us define **relative cohomology pseudo-ring** and **reduced cohomology**

pseudo-ring (by $A = \{*\}$), say

$$H^\bullet(X, A; R) = \bigoplus_{k \ge 0} H^k(X, A; R)$$
$$\widetilde{H}^\bullet(X; R) = \bigoplus_{k \ge 0} \widetilde{H}^k(X; R).$$

These are both **anticommutative pseudo-rings**. Indeed, often we have $\widetilde{H}^0(X; R) = 0$ and thus there is no identity at all.

Once again we have functoriality:

Theorem 76.5.1 (Cohomology (pseudo-)rings are functorial)
Fix a ring R (commutative with 1). Then we have functors

$$H^\bullet(-; R) : \mathsf{hTop}^{\mathrm{op}} \to \mathsf{GradedRings}$$
$$H^\bullet(-, -; R) : \mathsf{hPairTop}^{\mathrm{op}} \to \mathsf{GradedPseudoRings}.$$

Unfortunately, unlike with (co)homology groups, it is a nontrivial task to determine[2] the cup product for even nice spaces like CW complexes. So we will not do much in the way of computation. However, there is a little progress we can make.

§76.6 Wedge sums

Our goal is to now compute $\widetilde{H}^\bullet(X \vee Y)$. To do this, we need to define the product of two graded pseudo-rings:

Definition 76.6.1. Let R and S be two graded pseudo-rings. The **product pseudo-ring** $R \times S$ is the graded pseudo-ring defined by taking the underlying abelian group as

$$R \oplus S = \bigoplus_{d \ge 0} (R^d \oplus S^d).$$

Multiplication comes from R and S, followed by declaring $r \cdot s = 0$ for $r \in R$, $s \in S$.

Note that this is just graded version of the product ring defined in Example 4.3.8.

Exercise 76.6.2. Show that if R and S are graded rings (meaning they have 1_R and 1_S), then so is $R \times S$.

Now, the theorem is that:

Theorem 76.6.3 (Cohomology pseudo-rings of wedge sums)
We have

$$\widetilde{H}^\bullet(X \vee Y; R) \cong \widetilde{H}^\bullet(X; R) \times \widetilde{H}^\bullet(Y; R)$$

as graded pseudo-rings.

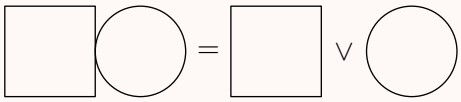
[2] Apart from the method of passing to differential form and back, that is. You have already computed a wedge product above.

Knowing just that the rings are isomorphic doesn't help much, it would be much better if you know what the isomorphism is — so that in simple cases, you can see for yourself the rings are isomorphic.

The isomorphism is the most trivial one: Given $f \in C^\bullet(X \vee Y; R)$ that assigns to each chain c inside $X \vee Y$ a value $f(c) \in R$, we can interpret it as an element of $C^\bullet(X)$, because each chain inside X is trivially a chain inside $X \vee Y$ that can be fed into f — formally, the embedding $X \hookrightarrow X \vee Y$ induces $C_\bullet(X) \hookrightarrow C_\bullet(X \vee Y)$. The map induces a $\widetilde{H}^\bullet(X \vee Y; R) \to \widetilde{H}^\bullet(X; R) \times \widetilde{H}^\bullet(Y; R)$, and it respects the ring multiplication i.e. the cup product.

Example 76.6.4

Let X and Y be depicted as in the following figure.



Let $f \in \widetilde{H}^1(X; \mathbb{Z})$ assigns $f(X) = 2$ to the whole square, and $g \in \widetilde{H}^1(Y; \mathbb{Z})$ assigns $g(Y) = 3$ to the whole circle. Then, of course the element corresponds to (f, g) inside $\widetilde{H}^1(X \vee Y)$ would assigns $2 + 3 = 5$ to the cocycle corresponding to the whole space $X \vee Y$.

This allows us to resolve the first question posed at the beginning. Let $X = \mathbb{CP}^2$ and $Y = S^2 \vee S^4$. We have that

$$H^\bullet(\mathbb{CP}^2; \mathbb{Z}) \cong \mathbb{Z}[\alpha]/(\alpha^3).$$

Hence this is a graded ring generated by there elements:

- 1, in dimension 0.
- α, in dimension 2.
- α^2, in dimension 4.

Next, consider the reduced cohomology pseudo-ring

$$\widetilde{H}^\bullet(S^2 \vee S^4; \mathbb{Z}) \cong \widetilde{H}^\bullet(S^2; \mathbb{Z}) \oplus \widetilde{H}^\bullet(S^4; \mathbb{Z}).$$

Thus the absolute cohomology ring $H^\bullet(S^2 \vee S^4; \mathbb{Z})$ is a graded ring also generated by three elements.

- 1, in dimension 0 (once we add back in the 0th dimension).
- a_2, in dimension 2 (from $H^\bullet(S^2; \mathbb{Z})$).
- a_4, in dimension 4 (from $H^\bullet(S^4; \mathbb{Z})$).

Each graded component is isomorphic, like we expected. However, in the former, the product of two degree 2 generators is

$$\alpha \cdot \alpha = \alpha^2.$$

In the latter, the product of two degree 2 generators is

$$a_2 \cdot a_2 = a_2^2 = 0$$

since $a_2 \smile a_2 = 0 \in H^\bullet(S^2; \mathbb{Z})$.

Thus $S^2 \vee S^4$ and $\mathbb{CP}^2$ are not homotopy equivalent.

Intuitively, what the proof above says is:

The nontrivial 4-cocycle $a_4 \in H^4(S^2 \vee S^4; \mathbb{Z})$ has nothing to do with the 2-cocycle a_2, while the 4-cocycle $\alpha^2 \in H^4(\mathbb{CP}^2)$ is the cup product $\alpha \smile \alpha$ of the 2-cocycle α with itself.

The exercise below would be much easier to visualize, apart from the fact that $\mathbb{RP}^2$ is nonorientable — in fact, we have already seen above why $\alpha \smile \alpha \neq 0$ for the nonzero element $\alpha \in H^1(\mathbb{RP}^2)$.

Exercise 76.6.5. Similarly, show that $S^1 \vee S^2$ and $\mathbb{RP}^2$ are not homotopy equivalent by showing $\widetilde{H}^\bullet(S^1 \vee S^2; \mathbb{Z}/2\mathbb{Z}) \ncong \widetilde{H}^\bullet(\mathbb{RP}^2; \mathbb{Z}/2\mathbb{Z})$, even though each graded component is isomorphic.

§76.7 Cross product

In this section, we will define the cross product.

§76.7.i Motivation

Roughly speaking, the motivation is the following:

If X has a m-dimensional hole and Y has a n-dimensional hole, then $X \times Y$ has a $(m+n)$-dimensional hole.

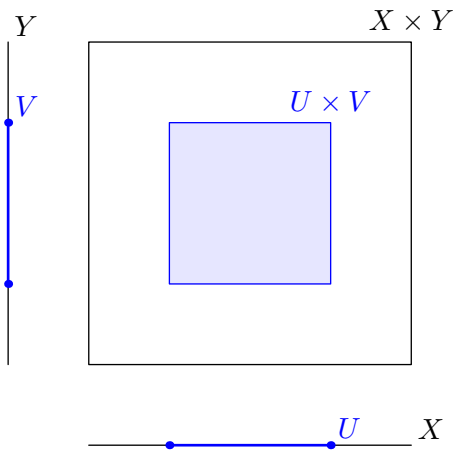
Which is true in most common cases under suitable interpretation of "holes" (either with homology, or with cohomology).

We will formalize and prove the statement above.

§76.7.ii Cross product on singular homology

First, we define the **cross product**, that takes a m-simplex $f\colon \Delta^m \to X$ and a n-simplex $g\colon \Delta^n \to Y$, and returns a $(m+n)$-chain $f \times g \in C_{m+n}(X \times Y)$.[3] This is really the most natural way you might define it: intuitively, the product of a m-dimensional cube in X and a n-dimensional cube in Y is a $(m+n)$-dimensional cube in $X \times Y$.

[3] As far as I know, this is just because the symbol $\times$ is a cross, and it has nothing to do with the cross product of vectors in $\mathbb{R}^3$.



In the case of a simplex, we need to subdivide $\Delta^m \times \Delta^n$ into finitely many copies of Δ^{m+n}.

If $n = 1$, we have already seen a subdivision when we worked with the prism operator. For the general case, refer to [Ha02, page 277] — the number blows up quickly, for example, you need $\binom{30}{15} = 155117520$ simplices to cover $\Delta^{15} \times \Delta^{15}$!

Formally, we can define the cross product of chains: that is, a function

$$C_m(X) \times C_n(Y) \xrightarrow{\times} C_{m\times n}(X \times Y).$$

We can prove that this induces a map on homology groups:

$$H_m(X) \times H_n(Y) \xrightarrow{\times} H_{m\times n}(X \times Y).$$

Exercise 76.7.1. Let $X = Y = S^1$, so that $X \times Y$ is a torus. Let α be a generator of $H_1(X)$, and β be a generator of $H_1(Y)$. Show that $\alpha \times \beta$ is the generator of $H_2(X \times Y)$.

Actually, we have the following:

Theorem 76.7.2

If X and Y are CW complexes and R is a PID, then the cross product of two nonzero elements in $H_m(X)$ and $H_n(Y)$ is nonzero.

Thus formalize our intuition earlier — at least, if we use homology as a measure of "holes".

§76.7.iii Cross product is not a $\mathbb{Z}$-module homomorphism

For this section, if a and b are elements of the $\mathbb{Z}$-module $C_m(X)$ and $C_n(Y)$ respectively, we write $\times(a, b)$ to mean $a \times b \in C_{m+n}(X \times Y)$, and (a, b) to be the element that corresponds in the product $C_m(X) \times C_n(Y)$.

There is a little technical detail that we need to sort out — above, we writes

$$\times : C_m(X) \times C_n(Y) \to C_{m+n}(X \times Y).$$

But written this way, $\times$ is not a $\mathbb{Z}$-module homomorphism!

Example 76.7.3

Let a and b be any nonzero elements in $C_m(X)$ and $C_n(Y)$ respectively.
Then,

$$\begin{aligned} \times(a,b) &= a \times b \\ 2 \cdot (a,b) &= (2a, 2b) \\ \times(2 \cdot (a,b)) &= 4(a \times b). \end{aligned}$$

If we want to talk about isomorphism, or do anything with the $\mathbb{Z}$-module structure of $C_{m+n}(X \times Y)$ or $H_{m+n}(X \times Y)$, we'd better having a $\mathbb{Z}$-module homomorphism.

This is easy enough to fix: $\times$ is bilinear, so it's natural to consider the tensor product:

$$\times : C_m(X) \otimes_{\mathbb{Z}} C_n(Y) \to C_{m+n}(X \times Y).$$

With this notation, $\times(a \otimes b) = a \times b$. (As a side effect, we can also write $\times(a \otimes b + c \otimes d) = a \times b + c \times d$ now.)

And so, let us restate Theorem 76.7.2:

Theorem 76.7.4

If X and Y are CW complexes, then

$$\times : H_m(X) \otimes_{\mathbb{Z}} H_n(Y) \to H_{m+n}(X \times Y)$$

is an injective $\mathbb{Z}$-module homomorphism.

§76.7.iv Cross product on cellular homology

The definition with singular homology is quite clumsy — because we use simplices as the building blocks for the chains, the product of two simplices in X and Y becomes a huge collection of simplices in $X \times Y$.

We will now redefine the cross product using cellular homology — it can be safely skipped, since both definitions of the cross product gives identical result on the homology groups.

If X and Y are CW complexes, we can do better. We see that $X \times Y$ has a natural CW complex structure: for each cell e^m of X and cell e^n of Y, their product makes for a cell e^{m+n} of $X \times Y$.

Example 76.7.5

If X and Y are both line segments built from two 0-cells and one 1-cell, then their product $X \times Y$ has a natural CW complex structure containing:

- 4 0-cells,
- 4 1-cells,
- 1 2-cell.

Recall the cellular groups $\text{Cells}_\bullet(X)$ from Chapter 74, each basis element corresponds

to a cell in X. Then, we can define the cross product on the basis elements:

$$\times\colon \operatorname{Cells}_m(X) \otimes_{\mathbb{Z}} \operatorname{Cells}_n(Y) \to \operatorname{Cells}_{m+n}(X \times Y).$$

To be painfully explicit: let $e^m \in \operatorname{Cells}_m(X)$, $e^n \in \operatorname{Cells}_m(Y)$, then the cross product is defined by $e^m \times e^n = e^m \times e^n \in \operatorname{Cells}_{m+n}(X \times Y)$ — even the notation used is trivial.

Of course, this induces a map on the homology groups:

$$\times\colon H_m(X) \otimes_{\mathbb{Z}} H_n(Y) \to H_{m+n}(X \times Y).$$

This map is the same as the map we defined earlier.

§76.7.v Cross product on cellular cohomology

We do the same thing as above, but this time with cohomology — remember that homology and cohomology are slightly different measures of "holes", for K the Klein bottle then $H_2(X) = 0$ but $H^2(X; \mathbb{Z}) \neq 0$.

Given two cellular cochains $f \in \operatorname{Hom}(\operatorname{Cells}_m(X); R)$ and $g \in \operatorname{Hom}(\operatorname{Cells}_n(Y); R)$, we want to obtain a cochain $f \times g \in \operatorname{Hom}(\operatorname{Cells}_{m+n}(X \times Y); R)$.

Of course, it is defined in the most natural way possible: for a cell e^m of X and a cell e^n of Y, we have $(f \times g)(e^m \times e^n) = f(e^m) \cdot g(e^n)$.

Sounds good? Not yet — since not all $(m+n)$-cells e^{m+n} of $X \times Y$ is formed as a product of a m-cell in X and a n-cell in Y. For those, we simply declare that $(f \times g)(e^{m+n}) = 0$.

As usual, this map induces a R-module homomorphism on the cohomology groups:

$$\times\colon H^m(X; R) \otimes_R H^n(Y; R) \to H^{m+n}(X \times Y; R).$$

§76.7.vi Motivation: cross product of differential forms

The definition of the cross product of two cellular cochains above are clean, but may appear to be dry and unmotivated.

Turns out you can do the same thing on differential form. What's more, it gives a clean way of defining the wedge product $\alpha \wedge \beta$! Let's see it in action.

Instead of the definition, here are a few examples. Motivated readers may try to define the concept formally.

Example 76.7.6 (Examples of cross product of differential form)

Here are a few examples.

- If X and Y are the x-axis and the y-axis of the plane respectively, the cross product $dx \times 2dy$ is equal to $2(dx \wedge dy)$.

  Certainly this is natural — as dx assigns the value 1 to the vector $\mathbf{e}_1$, and $2dy$ assigns the value 2 to the vector $\mathbf{e}_2$, we get that $dx \times 2dy$ should assigns the value $1 \cdot 2 = 2$ to the unit square spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$ — that is, $\mathbf{e}_1 \wedge \mathbf{e}_2$.

- Let X be the xy-plane, and let Y be the z-axis. Consider the cross product $dx \times dz$. What 2-form should the result be?

  Certainly, we should have $(dx \times dz)(\mathbf{e}_1 \wedge \mathbf{e}_3) = 1$ and $(dx \times dz)(\mathbf{e}_2 \wedge \mathbf{e}_3) = 0$. But this isn't enough to uniquely determines $dx \times dz$.

  And so, we declares: $(dx \times dz)(\mathbf{e}_1 \wedge \mathbf{e}_2) = 0$. With this, we get $dx \times dz = dx \wedge dz$.

More generally, we can define the cross product by picking a basis for X and Y, and define the value of $\alpha \times \beta$ on the basis elements.

As promised — you can define the wedge product using the cross product. There's only one thing you can do:

Definition 76.7.7 (Definition of wedge product using the cross product)**.** For X a $\mathbb{R}$-vector space, let $\alpha \in (\bigwedge^m(X))^\vee$ and $\beta \in (\bigwedge^n(X))^\vee$, then $\alpha \wedge \beta \in (\bigwedge^{m+n}(X)$ is defined by

$$\alpha \wedge \beta = \Delta^*(\alpha \times \beta)$$

where $\Delta\colon X \to X \times X$, $\Delta(x) = (x, x)$ is the diagonal map. Recall that Δ^* denotes the pullback operation.

In simpler terms: to evaluate $\alpha \wedge \beta$ on a $(m+n)$-wedge in X, push it to $X \times X$ using the diagonal map, and give it to $\alpha \times \beta$.

§76.7.vii Piecing the cohomology groups together

Recall that we have above the R-module homomorphism

$$\times\colon H^m(X;R) \otimes_R H^n(Y;R) \to H^{m+n}(X \times Y;R).$$

We know that it is in fact possible to piece all the $H^\bullet(X;R)$ together to form an anticommutative graded ring, the cohomology ring. So we wish to extend the map to a R-algebra homomorphism

$$\times\colon H^\bullet(X;R) \otimes_R H^\bullet(Y;R) \to H^\bullet(X \times Y;R).$$

We haven't defined what the tensor product of two graded rings is yet — we will formally do that in the next section, but intuitively, it consists of all the $H^m(X;R) \otimes_R H^n(Y;R)$ pieced together.

§76.8 Künneth formula

We now wish to tell apart the spaces $S^2 \times S^4$ and $\mathbb{CP}^3$. In order to do this, we will need a formula for $H^n(X \times Y;R)$ in terms of $H^n(X;R)$ and $H^n(Y;R)$. These formulas are called **Künneth formulas**. In this section we will only use a very special case, which involves the tensor product of two graded rings.

Definition 76.8.1. Let A and B be two graded rings which are also R-modules (where R is a commutative ring with 1). We define the **tensor product** $A \otimes_R B$ as follows. As an abelian group, it is

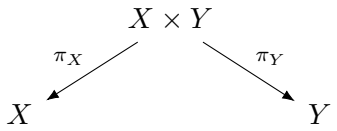
$$A \otimes_R B = \bigoplus_{d \geq 0} \left(\bigoplus_{k=0}^{d} A^k \otimes_R B^{d-k} \right).$$

The multiplication is given on basis elements by

$$(a_1 \otimes b_1)\,(a_2 \otimes b_2) = (a_1 a_2) \otimes (b_1 b_2).$$

Of course the multiplicative identity is $1_A \otimes 1_B$.

Now let X and Y be topological spaces, and take the product: we have a diagram



where π_X and π_Y are projections. As $H^k(-;R)$ is functorial, this gives induced maps

$$\pi_X^* : H^k(X \times Y; R) \to H^k(X; R)$$
$$\pi_Y^* : H^k(X \times Y; R) \to H^k(Y; R)$$

for every k.

By using this, we can define a so-called cross product.

Definition 76.8.2. Let R be a ring, and X and Y spaces. Let π_X and π_Y be the projections of $X \times Y$ onto X and Y. Then the **cross product** is the map

$$H^\bullet(X;R) \otimes_R H^\bullet(Y;R) \xrightarrow{\times} H^\bullet(X \times Y; R)$$

acting on cocycles as follows: $\phi \times \psi = \pi_X^*(\phi) \smile \pi_Y^*(\psi)$.

This is just the most natural way to take a k-cocycle on X and an ℓ-cocycle on Y, and create a $(k+\ell)$-cocycle on the product space $X \times Y$.

Remark 76.8.3 — Of course, this definition coincides with the definition above using cellular cohomology, but the proof is omitted.

Theorem 76.8.4 (Künneth formula)

Let X and Y be CW complexes such that $H^k(Y;R)$ is a finitely generated free R-module for every k. Then the cross product is an isomorphism of anticommutative rings

$$H^\bullet(X;R) \otimes_R H^\bullet(Y;R) \to H^\bullet(X \times Y; R).$$

That is:

There is a one-to-one correspondence between pair of holes in X and Y and holes of $X \times Y$. Furthermore, the correspondence respects the cup product.

Where "holes" is to be understood as "generators of cohomology groups" in this case.

In any case, this finally lets us resolve the question set out at the beginning. We saw that $H_n(\mathbb{CP}^3) \cong H_n(S^2 \times S^4)$ for every n, and thus it follows that $H^n(\mathbb{CP}^3;\mathbb{Z}) \cong H^n(S^2 \times S^4;\mathbb{Z})$ too.

But now let us look at the cohomology rings. First, we have

$$H^\bullet(\mathbb{CP}^3;\mathbb{Z}) \cong \mathbb{Z}[\alpha]/(\alpha^4) \cong \mathbb{Z} \oplus \alpha\mathbb{Z} \oplus \alpha^2\mathbb{Z} \oplus \alpha^3\mathbb{Z}$$

where $|\alpha| = 2$; hence this is a graded ring generated by

- 1, in degree 0.
- α, in degree 2.

- α^2, in degree 4.
- α^3, in degree 6.

Now let's analyze

$$H^\bullet(S^2 \times S^4; \mathbb{Z}) \cong \mathbb{Z}[\beta]/(\beta^2) \otimes \mathbb{Z}[\gamma]/(\gamma^2).$$

It is thus generated thus by the following elements:

- $1 \otimes 1$, in degree 0.
- $\beta \otimes 1$, in degree 2.
- $1 \otimes \gamma$, in degree 4.
- $\beta \otimes \gamma$, in degree 6.

Again in each dimension we have the same abelian group. But notice that if we square $\beta \otimes 1$ we get

$$(\beta \otimes 1)(\beta \otimes 1) = \beta^2 \otimes 1 = 0.$$

Yet the degree 2 generator of $H^\bullet(\mathbb{CP}^3; \mathbb{Z})$ does not have this property. Hence these two graded rings are not isomorphic.

The nontrivial 4-cocycle $1 \otimes \gamma$ of $S^2 \times S^4$ is orthogonal to the 2-cocycle $\beta \otimes 1$, while the 4-cocycle α^2 of $\mathbb{CP}^3$ is the cup product $\alpha \smile \alpha$ of the 2-cocycle α with itself.

So it follows that $\mathbb{CP}^3$ and $S^2 \times S^4$ are not homotopy equivalent.

Exercise 76.8.5. Do the same procedure with $H^\bullet(\mathbb{RP}^3; \mathbb{Z}/2\mathbb{Z})$ and $H^\bullet(S^1 \times S^2; \mathbb{Z}/2\mathbb{Z})$. (Visualize $S^1 \times S^2$ as a thickened sphere with the outer and inner face fused together, and RP^3 as a closed 3-ball with opposing points on the boundary surface fused together. Try to stretch your mind and guess what the homology and cohomology groups are before formally compute it.)

§76.9 A few harder problems to think about

Problem 76A† (Symmetry of Betti numbers by Poincaré duality)**.** Let M be a smooth oriented compact n-manifold, and let b_k denote its Betti number. Prove that $b_k = b_{n-k}$.

Problem 76B. Show that $\mathbb{RP}^n$ is not orientable for even n.

Problem 76C. Show that $\mathbb{RP}^3$ is not homotopy equivalent to $\mathbb{RP}^2 \vee S^3$.

Problem 76D. Show that $S^m \vee S^n$ is not a deformation retract of $S^m \times S^n$ for any $m, n \geq 1$.

XIX

Algebraic Geometry I: Classical Varieties

Part XIX: Contents

77 Affine varieties

In this chapter we introduce affine varieties. We introduce them in the context of coordinates, but over the course of the other chapters we'll gradually move away from this perspective to viewing varieties as "intrinsic objects", rather than embedded in coordinates.

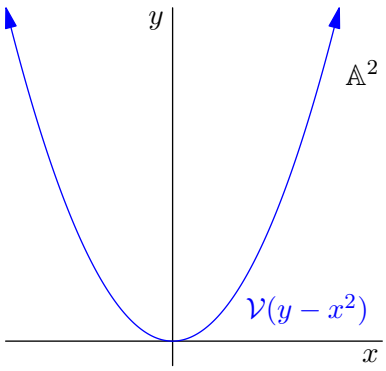
For simplicity, we'll do almost everything over the field of complex numbers, but the discussion generalizes to any algebraically closed field.

§77.1 Affine varieties

Prototypical example for this section: $\mathcal{V}(y - x^2)$ *is a parabola in* $\mathbb{A}^2$.

Definition 77.1.1. Given a set of polynomials $S \subseteq \mathbb{C}[x_1, \dots, x_n]$ (not necessarily finite or even countable), we let $\mathcal{V}(S)$ denote the set of points vanishing on *all* the polynomials in S. Such a set is called an **affine variety**. It lives in **n-dimensional affine space**, denoted $\mathbb{A}^n$ (to distinguish it from projective space later).

For example, a parabola is the zero locus of the polynomial $y - x^2$. Picture:



Example 77.1.2 (Examples of affine varieties)
These examples are in two-dimensional space $\mathbb{A}^2$, whose points are pairs (x, y).

(a) A straight line can be thought of as $\mathcal{V}(Ax + By + C)$.

(b) A parabola as above can be pictured as $\mathcal{V}(y - x^2)$.

(c) A hyperbola might be $\mathcal{V}(xy - 1)$.

(d) The two axes can be thought of as $\mathcal{V}(xy)$; this is the set of points such that $x = 0$ *or* $y = 0$.

(e) A point (x_0, y_0) can be thought of as $\mathcal{V}(x - x_0, y - y_0)$.

(f) The entire space $\mathbb{A}^2$ can be thought of as $\mathcal{V}(0)$.

(g) The empty set is the zero locus of the constant polynomial 1, that is $\mathcal{V}(1)$.

§77.2 Naming affine varieties via ideals

Prototypical example for this section: $\mathcal{V}(I)$ *is a parabola, where* $I = (y - x^2)$.

As you might have already noticed, a variety can be named by $\mathcal{V}(-)$ in multiple ways. For example, the set of solutions to

$$x = 3 \text{ and } y = 4$$

is just the point $(3, 4)$. But this is also the set of solutions to

$$x = 3 \text{ and } y = x + 1.$$

So, for example

$$\{(3,4)\} = \mathcal{V}(x-3, y-4) = \mathcal{V}(x-3, y-x-1).$$

That's a little annoying, because in an ideal[1] world we would have *one* name for every variety. Let's see if we can achieve this.

A partial solution is to use *ideals* rather than small sets. That is, consider the ideal

$$I = (x-3, y-4) = \{p(x,y)\cdot(x-3) + q(x,y)\cdot(y-4) \mid p, q \in \mathbb{C}[x,y]\}$$

and look at $\mathcal{V}(I)$.

> **Question 77.2.1.** Convince yourself that $\mathcal{V}(I) = \{(3,4)\}$.

So rather than writing $\mathcal{V}(x-3, y-4)$ it makes sense to think about this as $\mathcal{V}(I)$, where $I = (x-3, y-4)$ is the *ideal* generated by the two polynomials $x-3$ and $y-4$. This is an improvement because

> **Question 77.2.2.** Check that $(x-3, y-x-1) = (x-3, y-4)$.

Needless to say, this pattern holds in general.

> **Question 77.2.3.** Let $\{f_i\}$ be a set of polynomials, and consider the ideal I generated by these $\{f_i\}$. Show that $\mathcal{V}(\{f_i\}) = \mathcal{V}(I)$.

Thus we will only consider $\mathcal{V}(I)$ when I is an ideal. Of course, frequently our ideals are generated by one or two polynomials, which leads to:

Abuse of Notation 77.2.4. Given a set of polynomials $f_1, \dots, f_m$ we let $\mathcal{V}(f_1, \dots, f_m)$ be shorthand for $\mathcal{V}((f_1, \dots, f_m))$. In other words we let $\mathcal{V}(f_1, \dots, f_m)$ abbreviate $\mathcal{V}(I)$, where I is the *ideal* $I = (f_1, \dots, f_m)$.

This is where the Noetherian condition really shines: it guarantees that every ideal $I \subseteq \mathbb{C}[x_1, \dots, x_n]$ can be written in the form above with *finitely* many polynomials, because it is *finitely generated*. (The fact that $\mathbb{C}[x_1, \dots, x_n]$ is Noetherian follows from the Hilbert basis theorem, which is Theorem 4.9.5). This is a relief, because dealing with infinite sets of polynomials is not much fun.

[1] Pun not intended but left for amusement value.

§77.3 Radical ideals and Hilbert's Nullstellensatz

Prototypical example for this section: $\sqrt{(x^2)} = (x)$ *in* $\mathbb{C}[x]$, $\sqrt{(12)} = (6)$ *in* $\mathbb{Z}$.

You might ask whether the name is unique now: that is, if $\mathcal{V}(I) = \mathcal{V}(J)$, does it follow that $I = J$? The answer is unfortunately no: the counterexample can be found in just $\mathbb{A}^1$. It is

$$\mathcal{V}(x) = \mathcal{V}(x^2).$$

In other words, the set of solutions to $x = 0$ is the same as the set of solutions to $x^2 = 0$.

Well, that's stupid. We want an operation which takes the ideal (x^2) and makes it into the ideal (x). The way to do so is using the radical of an ideal.

Definition 77.3.1. Let R be a ring. The **radical** of an ideal $I \subseteq R$, denoted $\sqrt{I}$, is defined by

$$\sqrt{I} = \{r \in R \mid r^m \in I \text{ for some integer } m \geq 1\}.$$

If $I = \sqrt{I}$, we say the ideal I itself is **radical**.

For example, $\sqrt{(x^2)} = (x)$. You may like to take the time to verify that $\sqrt{I}$ is actually an ideal.

Remark 77.3.2 (Number theoretic motivation) — This is actually the same as the notion of "radical" in number theory. In $\mathbb{Z}$, the radical of an ideal (n) corresponds to just removing all the duplicate prime factors, so for example

$$\sqrt{(12)} = (6).$$

In particular, if you try to take $\sqrt{(6)}$, you just get (6) back; you don't squeeze out any new prime factors.

This is actually true more generally, and there is a nice corresponding alternate definition: for any ideal I, we have

$$\sqrt{I} = \bigcap_{I \subseteq \mathfrak{p} \text{ prime}} \mathfrak{p}.$$

Although we could prove this now, it will be proved later in Theorem 84.4.2, when we first need it.

Here are the immediate properties you should know.

Proposition 77.3.3 (Properties of radical)

In any ring:

- If I is an ideal, then $\sqrt{I}$ is always a radical ideal.
- Prime ideals are radical.
- For $I \subseteq \mathbb{C}[x_1, \dots, x_n]$ we have $\mathcal{V}(I) = \mathcal{V}(\sqrt{I})$.

Proof. These are all obvious.

- If $f^m \in \sqrt{I}$ then $f^{mn} \in I$, so $f \in \sqrt{I}$.

- If $f^n \in \mathfrak{p}$ for a prime $\mathfrak{p}$, then either $f \in \mathfrak{p}$ or $f^{n-1} \in \mathfrak{p}$, and in the latter case we may continue by induction.
- We have $f(x_1, \dots, x_n) = 0$ if and only if $f(x_1, \dots, x_n)^m = 0$ for some integer m. □

The last bit makes sense: you would never refer to $x = 0$ as $x^2 = 0$, and hence we would always want to call $\mathcal{V}(x^2)$ just $\mathcal{V}(x)$. With this, we obtain a theorem called Hilbert's Nullstellensatz.

Theorem 77.3.4 (Hilbert's Nullstellensatz)
Given an affine variety $V = \mathcal{V}(I)$, the set of polynomials which vanish on all points of V is precisely $\sqrt{I}$. Thus if I and J are ideals in $\mathbb{C}[x_1, \dots, x_n]$, then

$$\mathcal{V}(I) = \mathcal{V}(J) \text{ if and only if } \sqrt{I} = \sqrt{J}.$$

In other words

Radical ideals in $\mathbb{C}[x_1, \dots, x_n]$ correspond exactly to n-dimensional affine varieties.

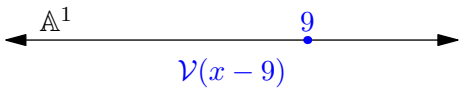
The proof of Hilbert's Nullstellensatz will be given in Problem 77D; for now it is worth remarking that it relies essentially on the fact that $\mathbb{C}$ is *algebraically closed*. For example, it is false in $\mathbb{R}[x]$, with $(x^2 + 1)$ being a maximal ideal with empty vanishing set.

§77.4 Pictures of varieties in $\mathbb{A}^1$

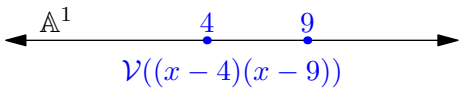
Prototypical example for this section: Finite sets of points (in fact these are the only nontrivial examples).

Let's first draw some pictures. In what follows I'll draw $\mathbb{C}$ as a straight line... sorry.

First of all, let's look at just the complex line $\mathbb{A}^1$. What are the various varieties on it? For starters, we have a single point $9 \in \mathbb{C}$, generated by $(x - 9)$.



Another example is the point 4. And in fact, if we like we can get an ideal consisting of just these two points; consider $\mathcal{V}((x-4)(x-9))$.



In general, in $\mathbb{A}^1$ you can get finitely many points $\{a_1, \dots, a_n\}$ by just taking

$$\mathcal{V}\left((x - a_1)(x - a_2)\dots(x - a_n)\right).$$

On the other hand, you can't get the set $\{0, 1, 2, \dots\}$ as an affine variety; the only polynomial vanishing on all those points is the zero polynomial. In fact, you can convince yourself that these are the only affine varieties, with two exceptions:

- The entire line $\mathbb{A}^1$ is given by $\mathcal{V}(0)$, and
- The empty set is given by $\mathcal{V}(1)$.

Exercise 77.4.1. Show that these are the only varieties of $\mathbb{A}^1$. (Let $\mathcal{V}(I)$ be the variety and pick a $0 \neq f \in I$.)

As you might correctly guess, we have:

Theorem 77.4.2 (Intersections and unions of varieties)

(a) The intersection of affine varieties (even infinitely many) is an affine variety.

(b) The union of finitely many affine varieties is an affine variety.

In fact we have

$$\bigcap_{\alpha} \mathcal{V}(I_\alpha) = \mathcal{V}\left(\sum_{\alpha} I_\alpha\right) \qquad \text{and} \qquad \bigcup_{k=1}^{n} \mathcal{V}(I_k) = \mathcal{V}\left(\bigcap_{k=1}^{n} I_k\right).$$

You are welcome to prove this easy result yourself.

Remark 77.4.3 — Part (a) is a little misleading in that the sum $I + J$ need not be radical: take for example $I = (y - x^2)$ and $J = (y)$ in $\mathbb{C}[x, y]$, where $x \in \sqrt{I + J}$ and $x \notin I + J$. But in part (b) for radical ideals I and J, the intersection $I \cap J$ is radical.

As another easy result concerning the relation between the ideal and variety, we have:

Proposition 77.4.4 ($\mathcal{V}(-)$ is inclusion reversing)

If $I \subseteq J$ then $\mathcal{V}(I) \supseteq \mathcal{V}(J)$. Thus $\mathcal{V}(-)$ is *inclusion-reversing*.

Question 77.4.5. Verify this.

Thus, bigger ideals correspond to smaller varieties.

These results will be used a lot throughout the chapter, so it would be useful for you to be comfortable with the inclusion-reversing nature of $\mathcal{V}$.

Exercise 77.4.6. Some quick exercises to help you be more familiar with the concepts.

1. Let $I = (y - x^2)$ and $J = (x + 1, y + 2)$. What is $\mathcal{V}(I)$ and $\mathcal{V}(J)$?
2. What is the ideal K such that $\mathcal{V}(K)$ is the union of the parabola $y = x^2$ and the point $(-1, -2)$?
3. Let $L = (y - 1)$. What is $\mathcal{V}(L)$?
4. The intersection $\mathcal{V}(I) \cap \mathcal{V}(L)$ consist of two points $(1, 1)$ and $(-1, 1)$. What's the ideal corresponding to it, in terms of I and L?
5. What is $\mathcal{V}(I \cap L)$? What about $\mathcal{V}(IL)$?

Question 77.4.7. Note that the intersection of infinitely many ideals is still an ideal, but the union of infinitely many affine varieties may not be an affine variety.

Consider $I_k = (x-k)$ in $\mathbb{C}[x]$, and take the infinite intersection $I = \bigcap_{k\in\mathbb{N}} I_k$. What is $\mathcal{V}(I)$ and $\bigcup_{k\in\mathbb{N}} \mathcal{V}(I_k)$?

§77.5 Prime ideals correspond to irreducible affine varieties

Prototypical example for this section: (xy) *corresponds to the union of two lines in* $\mathbb{A}^2$.

Note that most of the affine varieties of $\mathbb{A}^1$, like $\{4, 9\}$, are just unions of the simplest "one-point" ideals. To ease our classification, we can restrict our attention to the case of *irreducible* varieties:

Definition 77.5.1. A variety V is **irreducible** if it cannot be written as the union of two proper sub-varieties $V = V_1 \cup V_2$.

Abuse of Notation 77.5.2. Warning: in other literature, irreducible is part of the definition of variety.

Example 77.5.3 (Irreducible varieties of $\mathbb{A}^1$)
The irreducible varieties of $\mathbb{A}^1$ are:

- the empty set $\mathcal{V}(1)$,
- a single point $\mathcal{V}(x-a)$, and
- the entire line $\mathbb{A}^1 = \mathcal{V}(0)$.

Example 77.5.4 (The union of two axes)
Let's take a non-prime ideal in $\mathbb{C}[x,y]$, such as $I = (xy)$. Its vanishing set $\mathcal{V}(I)$ is the union of two lines $x = 0$ and $y = 0$. So $\mathcal{V}(I)$ is reducible.

In general:

Theorem 77.5.5 (Prime $\iff$ irreducible)
Let I be a radical ideal, and $V = \mathcal{V}(I)$ a nonempty variety. Then I is prime if and only if V is irreducible.

Proof. First, assume V is irreducible; we'll show I is prime. Let $f, g \in \mathbb{C}[x_1, \dots, x_n]$ so that $fg \in I$. Then V is a subset of the union $\mathcal{V}(f) \cup \mathcal{V}(g)$; actually, $V = (V \cap \mathcal{V}(f)) \cup (V \cap \mathcal{V}(g))$. Since V is irreducible, we may assume $V = V \cap \mathcal{V}(f)$, hence f vanishes on all of V. So $f \in I$.

The reverse direction is similar. □

§77.6 Pictures in $\mathbb{A}^2$ and $\mathbb{A}^3$

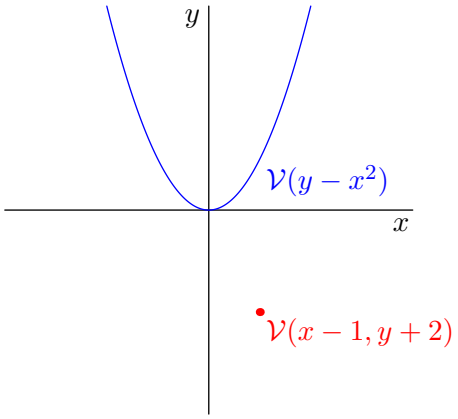
Prototypical example for this section: Various curves and hypersurfaces.

With this notion, we can now draw pictures in "complex affine plane", $\mathbb{A}^2$. What are the irreducible affine varieties in it?

As we saw in the previous discussion, naming irreducible affine varieties in $\mathbb{A}^2$ amounts to naming the prime ideals of $\mathbb{C}[x,y]$. Here are a few.

- The ideal (0) is prime. $\mathcal{V}(0)$ as usual corresponds to the entire plane.
- The ideal $(x-a, y-b)$ is prime, since $\mathbb{C}[x,y]/(x-a,y-b) \cong \mathbb{C}$ is an integral domain. (In fact, since $\mathbb{C}$ is a field, the ideal $(x-a,y-b)$ is *maximal*). The vanishing set of this is $\mathcal{V}(x-a,y-b) = \{(a,b)\} \in \mathbb{C}^2$, so these ideals correspond to a single point.
- Let $f(x,y)$ be an irreducible polynomial, like $y - x^2$. Then (f) is a prime ideal! Here $\mathcal{V}(I)$ is a "degree one curve".

By using some polynomial algebra (again you're welcome to check this; Euclidean algorithm), these are in fact the only prime ideals of $\mathbb{C}[x,y]$. Here's a picture.



As usual, you can make varieties which are just unions of these irreducible ones. For example, if you wanted the variety consisting of a parabola $y = x^2$ plus the point $(20, 15)$ you would write

$$\mathcal{V}\left((y-x^2)(x-20), (y-x^2)(y-15)\right).$$

The picture in $\mathbb{A}^3$ is harder to describe. Again, you have points $\mathcal{V}(x-a,y-b,z-c)$ corresponding to be zero-dimensional points (a,b,c), and two-dimensional surfaces $\mathcal{V}(f)$ for each irreducible polynomial f (for example, $x+y+z=0$ is a plane). But there are more prime ideals, like $\mathcal{V}(x,y)$, which corresponds to the intersection of the planes $x=0$ and $y=0$: this is the one-dimensional z-axis. It turns out there is no reasonable way to classify the "one-dimensional" varieties; they correspond to "irreducible curves".

Thus, as Ravi Vakil [Va17] says: the purely algebraic question of determining the prime ideals of $\mathbb{C}[x,y,z]$ has a fundamentally geometric answer.

§77.7 Maximal ideals

Prototypical example for this section: All maximal ideals are $(x_1 - a_1, \dots, x_n - a_n)$.

Recall that bigger ideals correspond to smaller varieties.

As the above pictures might have indicated, the smallest varieties are *single points*. Moreover, as you might guess from the name, the biggest ideals are the *maximal ideals*. As an example, all ideals of the form

$$(x_1 - a_1, \dots, x_n - a_n)$$

are maximal, since the quotient

$$\mathbb{C}[x_1, \dots, x_n] / (x_1 - a_1, \dots, x_n - a_n) \cong \mathbb{C}$$

is a field. The question is: are all maximal ideals of this form?

The answer is in the affirmative.

Theorem 77.7.1 (Weak Nullstellensatz, phrased with maximal ideals)
Every maximal ideal of $\mathbb{C}[x_1, \dots, x_n]$ is of the form $(x_1 - a_1, \dots, x_n - a_n)$.

The proof of this is surprisingly nontrivial, so we won't include it here yet; see [Va17, §7.4.3]. Again this uses the fact that $\mathbb{C}$ is algebraically closed. (For example $(x^2 + 1)$ is a maximal ideal of $\mathbb{R}[x]$.) Thus:

Over $\mathbb{C}$, maximal ideals correspond to single points.

Consequently, our various ideals over $\mathbb{C}$ correspond to various flavors of affine varieties:

Algebraic flavor	Geometric flavor
radical ideal	affine variety
prime ideal	irreducible variety
maximal ideal	single point
any ideal	(scheme?)

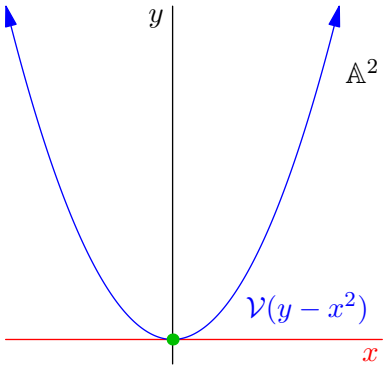
There's one thing I haven't talked about: what's the last entry?

§77.8 Motivating schemes with non-radical ideals

One of the most elementary motivations for the scheme is that we would like to use them to count multiplicity. That is, consider the intersection

$$\mathcal{V}(y - x^2) \cap \mathcal{V}(y) \subseteq \mathbb{A}^2$$

This is the intersection of the parabola with the tangent x-axis, this is the green dot below.



Unfortunately, as a variety, it is just a single point! However, we want to think of this as a "double point": after all, in some sense it has multiplicity 2. You can detect this when you look at the ideals:

$$(y - x^2) + (y) = (x^2, y)$$

and thus, if we blithely ignore taking the radical, we get

$$\mathbb{C}[x,y]/(x^2,y) \cong \mathbb{C}[\varepsilon]/(\varepsilon^2).$$

So the ideals in question are noticing the presence of a double point.

In order to encapsulate this, we need a more refined object than a variety, which (at the end of the day) is just a set of points; it's not possible using topology along to encode more information (there is only one topology on a single point!). This refined object is the *scheme*.

§77.9 A few harder problems to think about

some actual computation here would be good

Problem 77A. Show that $I \subseteq \sqrt{I}$ and $\sqrt{I} \cap \sqrt{J} = \sqrt{IJ} \subseteq \sqrt{I \cap J}$, for two ideals I and J.

Problem 77B. Show that a *real* affine variety $V \subseteq \mathbb{A}^n_{\mathbb{R}}$ can always be written in the form $\mathcal{V}(f)$.

Problem 77C (Complex varieties can't be empty)**.** Prove that if I is a proper ideal in $\mathbb{C}[x_1, \dots, x_n]$ then $\mathcal{V}(I) \neq \varnothing$.

Problem 77D. Show that Hilbert's Nullstellensatz in n dimensions follows from the Weak Nullstellensatz. (This solution is called the **Rabinowitsch Trick**.)

78 Affine varieties as ringed spaces

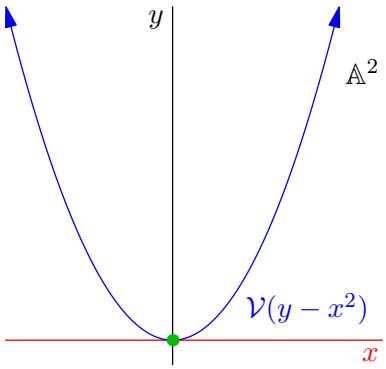
As in the previous chapter, we are working only over affine varieties in $\mathbb{C}$ for simplicity.

§78.1 Synopsis

Group theory was a strange creature in the early 19th century. During the 19th century, a group was literally defined as a subset of $\mathrm{GL}(n)$ or of S_n. Indeed, the word "group" hadn't been invented yet. This may sound ludicrous, but it was true – Sylow developed his theorems without this notion. Only much later was the abstract definition of a group given, an abstract set G which was *independent* of any embedding into S_n, and an object in its own right.

We are about to make the same type of change for our affine varieties. Rather than thinking of them as an object locked into an ambient space $\mathbb{A}^n$ we are instead going to try to make them into an object in their own right. Specifically, for us an affine variety will become a *topological space* equipped with a *ring of functions* for each of its open sets: this is why we call it a **ringed space**.

The bit about the topological space is not too drastic. The key insight is the addition of the ring of functions. For example, consider the double point from last chapter.



As a set, it is a single point, and thus it can have only one possible topology. But the addition of the function ring will let us tell it apart from just a single point.

This construction is quite involved, so we'll proceed as follows: we'll define the structure bit by bit onto our existing affine varieties in $\mathbb{A}^n$, until we have all the data of a ringed space. In later chapters, these ideas will grow up to become the core of modern algebraic geometry: the *scheme*.

§78.2 The Zariski topology on $\mathbb{A}^n$

Prototypical example for this section: In $\mathbb{A}^1$, closed sets are finite collections of points. In $\mathbb{A}^2$, a nonempty open set is the whole space minus some finite collection of curves/points.

We begin by endowing a topological structure on every variety V. Since our affine varieties (for now) all live in $\mathbb{A}^n$, all we have to do is put a suitable topology on $\mathbb{A}^n$, and then just view V as a subspace.

However, rather than putting the standard Euclidean topology on $\mathbb{A}^n$, we put a much more bizarre topology.

Definition 78.2.1. In the **Zariski topology** on $\mathbb{A}^n$, the *closed sets* are those of the form

$$\mathcal{V}(I) \qquad \text{where} \quad I \subseteq \mathbb{C}[x_1, \dots, x_n].$$

Of course, the open sets are complements of such sets.

Example 78.2.2 (Zariski topology on $\mathbb{A}^1$)

Let us determine the open sets of $\mathbb{A}^1$, which as usual we picture as a straight line (ignoring the fact that $\mathbb{C}$ is two-dimensional).
Since $\mathbb{C}[x]$ is a principal ideal domain, rather than looking at $\mathcal{V}(I)$ for every $I \subseteq \mathbb{C}[x]$, we just have to look at $\mathcal{V}(f)$ for a single f. There are a few flavors of polynomials f:

- The zero polynomial 0 which vanishes everywhere: this implies that the entire space $\mathbb{A}^1$ is a closed set.
- The constant polynomial 1 which vanishes nowhere. This implies that $\varnothing$ is a closed set.
- A polynomial $c(x-t_1)(x-t_2)\dots(x-t_n)$ of degree n. It has n roots, and so $\{t_1, \dots, t_n\}$ is a closed set.

Hence the closed sets of $\mathbb{A}^1$ are exactly all of $\mathbb{A}^1$ and finite sets of points (including $\varnothing$). Consequently, the *open* sets of $\mathbb{A}^1$ are

- $\varnothing$, and
- $\mathbb{A}^1$ minus a finite collection (possibly empty) of points.

Thus, the picture of a "typical" open set $\mathbb{A}^1$ might be

It's everything except a few marked points!

Example 78.2.3 (Zariski topology on $\mathbb{A}^2$)

Similarly, in $\mathbb{A}^2$, the interesting closed sets are going to consist of finite unions (possibly empty) of

- Closed curves, like $\mathcal{V}(y-x^2)$ (which is a parabola), and
- Single points, like $\mathcal{V}(x-3, y-4)$ (which is the point $(3,4)$).

Of course, the entire space $\mathbb{A}^2 = \mathcal{V}(0)$ and the empty set $\varnothing = \mathcal{V}(1)$ are closed sets. Thus the nonempty open sets in $\mathbb{A}^2$ consist of the *entire* plane, minus a finite collection of points and one-dimensional curves.

Question 78.2.4. Draw a picture (to the best of your artistic ability) of a "typical" open set in $\mathbb{A}^2$.

All this is to say

The nonempty Zariski open sets are *huge*.

This is an important difference than what you're used to in topology. To be very clear:

- In the past, if I said something like "has so-and-so property in an open neighborhood of point p", one thought of this as saying "is true in a small region around p".
- In the Zariski topology, "has so-and-so property in an open neighborhood of point p" should be thought of as saying "is true for virtually all points, other than those on certain curves".

Indeed, "open neighborhood" is no longer really a accurate description. Nonetheless, in many pictures to follow, it will still be helpful to draw open neighborhoods as circles.

It remains to verify that as I've stated it, the closed sets actually form a topology. That is, I need to verify briefly that

- $\varnothing$ and $\mathbb{A}^n$ are both closed.
- Intersections of closed sets (even infinite) are still closed.
- Finite unions of closed sets are still closed.

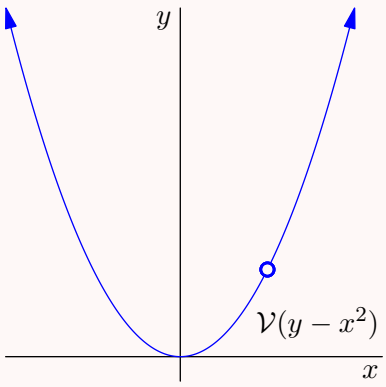
Well, closed sets are the same as affine varieties, so we already know this!

§78.3 The Zariski topology on affine varieties

Prototypical example for this section: If $V = \mathcal{V}(y - x^2)$ is a parabola, then V minus $(1,1)$ is open in V. Also, the plane minus the origin is $D(x) \cup D(y)$.

As we said before, by considering a variety V as a subspace of $\mathbb{A}^n$ it inherits the Zariski topology. One should think of an open subset of V as "V minus a few Zariski-closed sets". For example:

Example 78.3.1 (Open set of a variety)
Let $V = \mathcal{V}(y - x^2) \subseteq \mathbb{A}^2$ be a parabola, and let $U = V \setminus \{(1,1)\}$. We claim U is open in V.



Indeed, $\tilde{U} = \mathbb{A}^2 \setminus \{(1,1)\}$ is open in $\mathbb{A}^2$ (since it is the complement of the closed set $\mathcal{V}(x-1, y-1)$), so $U = \tilde{U} \cap V$ is open in V. Note that on the other hand the set U is *not* open in $\mathbb{A}^2$.

We will go ahead and introduce now a definition that will be very useful later.

Definition 78.3.2. Given $V \subseteq \mathbb{A}^n$ an affine variety and $f \in \mathbb{C}[x_1, \dots, x_n]$, we define the **distinguished open set** $D(f)$ to be the open set in V of points not vanishing on f:

$$D(f) = \{p \in V \mid f(p) \neq 0\} = V \setminus \mathcal{V}(f).$$

In [Va17], Vakil suggests remembering the notation $D(f)$ as "doesn't-vanish set".

Example 78.3.3 (Examples of (unions of) distinguished open sets)

(a) If $V = \mathbb{A}^1$ then $D(x)$ corresponds to a line minus a point.

(b) If $V = \mathcal{V}(y - x^2) \subseteq \mathbb{A}^2$, then $D(x-1)$ corresponds to the parabola minus $(1,1)$.

(c) If $V = \mathbb{A}^2$, then $D(x) \cup D(y) = \mathbb{A}^2 \setminus \{(0,0)\}$ is the punctured plane. You can show that this set is *not* distinguished open.

You can think of the concept as an analog to principal ideal: all open sets can be written in the form $V \setminus \mathcal{V}(I)$ for some ideal I, but if $I = (f)$ is principal then the set can be written as a distinguished open set $D(f)$. Similarly, the intersection of two distinguished open sets is distinguished, just as the product (not intersection!) of two principal ideals is principal.

Proposition 78.3.4 (Properties of distinguished open set)

Recall that $\mathcal{V}$ is inclusion-reversing, so being the complement of $\mathcal{V}$, we would expect D to be "inclusion-preserving". Indeed:

- If $(f) \subseteq (g)$ (that is, $g \mid f$), then $D(f) \subseteq D(g)$.
- Recall that $(fg) \subseteq (f) \cap (g)$. For distinguished open set, we have $D(fg) = D(f) \cap D(g)$.

It is useful to be familiar with the behavior of D.

Question 78.3.5. If $V = \mathcal{A}^2$, then $D(x)$ is the plane minus the y-axis, and $D(y)$ is the plane minus the x-axis. What is $D(xy)$?

§78.4 Coordinate rings

Prototypical example for this section: If $V = \mathcal{V}(y - x^2)$ then $\mathbb{C}[V] = \mathbb{C}[x,y]/(y-x^2)$.

The next thing we do is consider the functions from V to the base field $\mathbb{C}$. We restrict our attention to algebraic (polynomial) functions on a variety V: they should take every point $(a_1, \dots, a_n)$ on V to some complex number $P(a_1, \dots, a_n) \in \mathbb{C}$. For example, a valid function on a three-dimensional affine variety might be $(a,b,c) \mapsto a$; we just call this projection "x". Similarly we have a canonical projection y and z, and we can create polynomials by combining them, say $x^2y + 2xyz$.

Definition 78.4.1. The **coordinate ring** $\mathbb{C}[V]$ of a variety V is the ring of polynomial functions on V. (Notation explained next section.)

Remark 78.4.2 (Meaning of the name "coordinate ring") — We call the functions x, y and z above as the **coordinate functions**, as they maps each point in the variety V to its coordinate. So, the coordinate ring $\mathbb{C}[V]$ is simply the ring generated by $\mathbb{C}$ and the coordinate functions.

At first glance, we might think this is just $\mathbb{C}[x_1, \dots, x_n]$. But on closer inspection we realize that *on a given variety*, some of these functions are the same. For example, consider in $\mathbb{A}^2$ the parabola $V = \mathcal{V}(y - x^2)$. Then the two functions

$$\begin{aligned} V &\to \mathbb{C} \\ (x, y) &\mapsto x^2 \\ (x, y) &\mapsto y \end{aligned}$$

are actually the same function! We have to "mod out" by the ideal I which generates V. This leads us naturally to:

Theorem 78.4.3 (Coordinate rings correspond to ideal)
Let I be a radical ideal, and $V = \mathcal{V}(I) \subseteq \mathbb{A}^n$. Then

$$\mathbb{C}[V] \cong \mathbb{C}[x_1, \dots, x_n]/I.$$

Proof. There's a natural surjection as above

$$\mathbb{C}[x_1, \dots, x_n] \twoheadrightarrow \mathbb{C}[V]$$

and the kernel is I. □

Thus properties of a variety V correspond to properties of the ring $\mathbb{C}[V]$.

§78.5 The sheaf of regular functions

Prototypical example for this section: Let $V = \mathbb{A}^1$, $U = V \setminus \{0\}$. Then $1/x \in \mathcal{O}_V(U)$ is regular on U.

Let V be an affine variety and let $\mathbb{C}[V]$ be its coordinate ring. As mentioned in the start of the chapter, we want to define a variety based on its intrinsic properties only, which is done by studying the collection of algebraic functions on it.

In [Va17] "Motivating example: The sheaf of differentiable functions" section, you can see a comparison of how a differentiable manifold can be studied by studying the differentiable functions on it.

Denote the set of all rational functions on V by $\mathcal{O}_V$ (as will be seen later, this terminology is not quite accurate as we need to allow multiple representations). We can view this as a set, however this does not capture the full structure of the rational functions:

Question 78.5.1. For any two elements f and g in $\mathbb{C}[V]$, show that the set where $\frac{f(x)}{g(x)}$ is well-defined is open in the Zariski topology. (Hint: $g^{\text{pre}}(0)$ is closed.)

So, we want to define a notion of $\mathcal{O}_V(U)$ for any open set U: the "nice" functions on any open subset. Obviously, any function in $\mathbb{C}[V]$ will work as a function on $\mathcal{O}_V(U)$.

However, to capture more of the structure we want to loosen our definition of "nice" function slightly by allowing *rational* functions.

The chief example is that $1/x$ should be a regular function on $\mathbb{A}^1 \setminus \{0\}$. The first natural guess is:

Definition 78.5.2. Let $U \subseteq V$ be an open set of the variety V. A **rational function** on U is a quotient $f(x)/g(x)$ of two elements f and g in $\mathbb{C}[V]$, where we require that $g(x) \neq 0$ for $x \in U$.

However, the definition is slightly too restrictive; we have to allow for multiple representations:

Definition 78.5.3. Let $U \subseteq V$ be open. We say a function $\phi \colon U \to \mathbb{C}$ is a **regular function** if for every point $p \in U$, we can find an open set $U_p \subseteq U$ containing p and a rational function f_p/g_p on U_p such that

$$\phi(x) = \frac{f_p(x)}{g_p(x)} \qquad \forall x \in U_p.$$

In particular, we require $g_p(x) \neq 0$ on the set U_p. We denote the set of all regular functions on U by $\mathcal{O}_V(U)$.

Thus,

ϕ is regular on U if it is locally a rational function.

This definition is misleadingly complicated, and the examples should illuminate it significantly. Firstly, in practice, most of the time we will be able to find a "global" representation of a regular function as a quotient, and we will not need to fuss with the p's. For example:

Example 78.5.4 (Regular functions)

(a) Any function in $f \in \mathbb{C}[V]$ is clearly regular, since we can take $g_p = 1$, $f_p = f$ for every p. So $\mathbb{C}[V] \subseteq \mathcal{O}_V(U)$ for any open set U.

(b) Let $V = \mathbb{A}^1$, $U_0 = V \setminus \{0\}$. Then $1/x \in \mathcal{O}_V(U_0)$ is regular on U_0.

(c) Let $V = \mathbb{A}^1$, $U_{12} = V \setminus \{1, 2\}$. Then

$$\frac{1}{(x-1)(x-2)} \in \mathcal{O}_V(U_{12})$$

is regular on U_{12}.

The "local" clause with p's is still necessary, though.

Example 78.5.5 (Requiring local representations)

Consider the variety

$$V = \mathcal{V}(ab - cd) \subseteq \mathbb{A}^4$$

and the open set $U = V \setminus \mathcal{V}(b,d)$. There is a regular function on U given by

$$(a,b,c,d) \mapsto \begin{cases} a/d & d \neq 0 \\ c/b & b \neq 0. \end{cases}$$

Clearly these are the "same function" (since $ab = cd$), but we cannot write "a/d" or "c/b" to express it because we run into divide-by-zero issues. That's why in the definition of a regular function, we have to allow multiple representations.

In fact, we will see later on that the definition of a regular function is a special case of a more general construction called *sheafification*, in which "presheaves of functions which are P" are transformed into "sheaves of functions which are *locally* P".

§78.6 Regular functions on distinguished open sets

Prototypical example for this section: Regular functions on $\mathbb{A}^1 \setminus \{0\}$ are $P(x)/x^n$.

The division-by-zero, as one would expect, essentially prohibits regular functions on the entire space V; i.e. there are no regular functions in $\mathcal{O}_V(V)$ that were not already in $\mathbb{C}[V]$. Actually, we have a more general result which computes the regular functions on distinguished open sets:

Theorem 78.6.1 (Regular functions on distinguished open sets)
Let $V \subseteq \mathbb{A}^n$ be an affine variety and $D(g)$ a distinguished open subset of it. Then

$$\mathcal{O}_V(D(g)) = \left\{ \frac{f}{g^k} \mid f \in \mathbb{C}[V] \text{ and } k \in \mathbb{Z} \right\}.$$

In particular, $\mathcal{O}_V(V) = \mathcal{O}_V(D(1)) \cong \mathbb{C}[V]$.

The proof of this theorem requires the Nullstellensatz, so it relies on $\mathbb{C}$ being algebraically closed. In fact, a counter-example is easy to find if we replace $\mathbb{C}$ by $\mathbb{R}$: consider $\frac{1}{x^2+1}$.

Proof. Obviously, every function of the form f/g^n works, so we want the reverse direction. This is long, and perhaps should be omitted on a first reading.

Here's the situation. Let $U = D(g)$. We're given a regular function ϕ, meaning at every point $p \in D(g)$, there is an open neighborhood U_p on which ϕ can be expressed as f_p/g_p (where $f_p, g_p \in \mathbb{C}[V]$). Then, we want to construct an $f \in \mathbb{C}[V]$ and an integer n such that $\phi = f/g^n$.

First, look at a particular U_p and f_p/g_p. Shrink U_p to a distinguished open set $D(h_p)$. Then, let $\widetilde{f}_p = f_p h_p$ and $\widetilde{g}_p = g_p h_p$. Thus we have that

$$\frac{\widetilde{f}_p}{\widetilde{g}_p} \text{ is correct on } D(h_p) \subseteq U \subseteq X.$$

The upshot of using the modified f_p and g_p is that:

$$\widetilde{f}_p \widetilde{g}_q = \widetilde{f}_q \widetilde{g}_p \qquad \forall p, q \in U.$$

Indeed, it is correct on $D(h_p) \cap D(h_q)$ by definition, and outside this set both the left-hand side and right-hand side are zero.

Now, we know that $D(g) = \bigcup_{p \in U} D(\widetilde{g}_p)$, i.e.

$$\mathcal{V}(g) = \bigcap_{p \in U} \mathcal{V}(\widetilde{g}_p).$$

So by the Nullstellensatz we know that

$$g \in \sqrt{(\widetilde{g}_p : p \in U)} \implies \exists n : g^n \in (\widetilde{g}_p : p \in U).$$

In other words, for some n and $k_p \in \mathbb{C}[V]$ we have

$$g^n = \sum_p k_p \widetilde{g}_p$$

where only finitely many k_p are not zero. Now, we claim that

$$f \coloneqq \sum_p k_p \widetilde{f}_p$$

works. This just observes by noting that for any $q \in U$, we have

$$f\widetilde{g}_q - g^n \widetilde{f}_q = \sum_p k_p (\widetilde{f}_p \widetilde{g}_q - \widetilde{g}_p \widetilde{f}_q) = 0. \qquad \square$$

This means that the *global* regular functions are just the same as those in the coordinate ring: you don't gain anything new by allowing it to be locally a quotient. (The same goes for distinguished open sets.)

Example 78.6.2 (Regular functions on distinguished open sets)

(a) As said already, taking $g = 1$ we recover $\mathcal{O}_V(V) \cong \mathbb{C}[V]$ for any affine variety V.

(b) Let $V = \mathbb{A}^1$, $U_0 = V \setminus \{0\}$. Then

$$\mathcal{O}_V(U_0) = \left\{ \frac{P(x)}{x^n} \mid P \in \mathbb{C}[x], \quad n \in \mathbb{Z} \right\}.$$

So more examples are $1/x$ and $(x+1)/x^3$.

Question 78.6.3. Why doesn't our theorem on regular functions apply to Example 78.5.5?

The regular functions will become of crucial importance once we define a scheme in the next chapter.

§78.7 Baby ringed spaces

In summary, given an affine variety V we have:

- A structure of a set of points,
- A structure of a topological space V on these points, and

- For every open set $U \subseteq V$, a ring $\mathcal{O}_V(U)$. Elements of the rings are functions $U \to \mathbb{C}$.

Let us agree that:

Definition 78.7.1. A **baby ringed space** is a topological space X equipped with a ring $\mathcal{O}_X(U)$ for every open set U. It is required that elements of the ring $\mathcal{O}_X(U)$ are functions $f \colon U \to \mathbb{C}$; we call these the *regular functions* of X on U.

Therefore, affine varieties are baby ringed spaces.

Remark 78.7.2 — This is not a standard definition. Hehe.

The reason this is called a "baby ringed space" is that in a *ringed space*, the rings $\mathcal{O}_V(U)$ can actually be *any rings*, but they have to satisfy a set of fairly technical conditions. When this happens, it's the $\mathcal{O}_V$ that does all the work; we think of $\mathcal{O}_V$ as a type of functor called a *sheaf*.

Since we are only studying affine/projective/quasi-projective varieties for the next chapters, we will just refer to these as baby ringed spaces so that we don't have to deal with the entire definition. The key concept is that we want to think of these varieties as *intrinsic objects*, free of any embedding. A baby ringed space is philosophically the correct thing to do.

Anyways, affine varieties are baby ringed spaces $(V, \mathcal{O}_V)$. In the next chapter we'll meet projective and quasi-projective varieties, which give more such examples of (baby) ringed spaces. With these examples in mind, we will finally lay down the complete definition of a ringed space, and use this to define a scheme.

§78.8 A few harder problems to think about

Problem 78A†. Show that for any $n \geq 1$ the Zariski topology of $\mathbb{A}^n$ is *not* Hausdorff.

Problem 78B†. Let V be an affine variety, and consider its Zariski topology.

(a) Show that the Zariski topology is **Noetherian**, meaning there is no infinite descending chain $Z_1 \supsetneq Z_2 \supsetneq Z_3 \supsetneq \dots$ of closed subsets.

(b) Prove that a Noetherian topological space is compact. Hence varieties are topologically compact.

Problem 78C* (Punctured Plane)**.** Let $V = \mathbb{A}^2$ and let $X = \mathbb{A}^2 \setminus \{(0,0)\}$ be the punctured plane (which is an open set of V). Compute $\mathcal{O}_V(X)$.

79 Projective varieties

Having studied affine varieties in $\mathbb{A}^n$, we now consider $\mathbb{CP}^n$. We will also make it into a baby ringed space in the same way as with $\mathbb{A}^n$.

§79.1 Graded rings

Prototypical example for this section: $\mathbb{C}[x_0, \dots, x_n]$ *is a graded ring.*

We first take the time to state what a graded ring is, just so that we have this language to use (now and later).

This definition is the same as Definition 76.3.2.

Definition 79.1.1. A **graded ring** R is a ring with the following additional structure: as an abelian group, it decomposes as

$$R = \bigoplus_{d \geq 0} R^d$$

where R^0, R^1, ..., are abelian groups. The ring multiplication has the property that if $r \in R^d$ and $s \in R^e$, we have $rs \in R^{d+e}$. Elements of an R^d are called **homogeneous elements**; we write "$d = \deg r$" to mean "$r \in R^d$".

We denote by R^+ the ideal $R \setminus R^0$ generated by the homogeneous elements of nonzero degree, and call it the **irrelevant ideal**.

Remark 79.1.2 — For experts: all our graded rings are commutative with 1.

Example 79.1.3 (Examples of graded rings)

(a) The ring $\mathbb{C}[x]$ is graded by degree: as abelian groups, $\mathbb{C}[x] \cong \mathbb{C} \oplus x\mathbb{C} \oplus x^2\mathbb{C} \oplus \dots$.

(b) More generally, the polynomial ring $\mathbb{C}[x_0, \dots, x_n]$ is graded by degree.

Abuse of Notation 79.1.4. The notation $\deg r$ is abusive in the case $r = 0$; note that $0 \in R^d$ for every d. So it makes sense to talk about "the" degree of r except when $r = 0$.

We will frequently refer to homogeneous ideals:

Definition 79.1.5. An ideal $I \subseteq \mathbb{C}[x_0, \dots, x_n]$ is **homogeneous** if it can be written as $I = (f_1, \dots, f_m)$ where each f_i is a homogeneous polynomial.

Remark 79.1.6 — If I and J are homogeneous, then so are $I + J$, IJ, $I \cap J$, $\sqrt{I}$.

Lemma 79.1.7 (Graded quotients are graded too)

Let I be a homogeneous ideal of a graded ring R. Then

$$R/I = \bigoplus_{d \geq 0} R^d/(R^d \cap I)$$

realizes R/I as a graded ring.

Since these assertions are just algebra, we omit their proofs here.

Remark 79.1.8 — In some other books, a homogeneous ideal (or **graded ideal**) is sometimes equivalently defined as an ideal I such that $I = \bigoplus_{d \ge 0} (R^d \cap I)$ as abelian group. In fact, we can verify that graded ideals are precisely the ones such that the quotient is naturally graded.

Example 79.1.9 (Example of a graded quotient ring)
Let $R = \mathbb{C}[x,y]$ and set $I = (x^3, y^2)$. Let $S = R/I$. Then

$$\begin{aligned} S^0 &= \mathbb{C} \\ S^1 &= \mathbb{C}x \oplus \mathbb{C}y \\ S^2 &= \mathbb{C}x^2 \oplus \mathbb{C}xy \\ S^3 &= \mathbb{C}x^2y \\ S^d &= 0 \qquad \forall d \ge 4. \end{aligned}$$

So in fact $S = R/I$ is graded, and is a six-dimensional $\mathbb{C}$-vector space.

§79.2 The ambient space

Prototypical example for this section: Perhaps $\mathcal{V}_{pr}(x^2 + y^2 - z^2)$.

The set of points we choose to work with is $\mathbb{CP}^n$ this time, which for us can be thought of as the set of n-tuples

$$(x_0 : x_1 : \cdots : x_n)$$

not all zero, up to scaling. Equivalently, it is the set of lines through the origin in $\mathbb{C}^{n+1}$. Projective space is defined in full in Section 64.6, and you should refer there if you aren't familiar with projective space.

The right way to think about it is "$\mathbb{A}^n$ plus points at infinity":

Definition 79.2.1. We define the set

$$U_i = \{(x_0 : \cdots : x_n) \mid x_i \neq 0\} \subseteq \mathbb{CP}^n.$$

These are called the **standard affine charts**.

The name comes from:

Exercise 79.2.2 (Mandatory)**.** Give a natural bijection from U_i to $\mathbb{A}^n$. Thus we can think of $\mathbb{CP}^n$ as the affine set U_i plus "points at infinity".

Remark 79.2.3 — In fact, these charts U_i make $\mathbb{CP}^n$ with its usual topology into a complex manifold with holomorphic transition functions.

Example 79.2.4 (Colloquially, $\mathbb{CP}^1 = \mathbb{A}^1 \cup \{\infty\}$)
The space $\mathbb{CP}^1$ consists of pairs $(s : t)$, which you can think of as representing the complex number $z/1$. In particular $U_1 = \{(z : 1)\}$ is basically another copy of $\mathbb{A}^1$.

There is only one new point, $(1:0)$.

However, like before we want to impose a Zariski topology on it. For concreteness, let's consider $\mathbb{CP}^2 = \{(x_0 : x_1 : x_2)\}$. We wish to consider zero loci in $\mathbb{CP}^2$, just like we did in affine space, and hence obtain a notion of a projective variety.

But this isn't so easy: for example, the function "x_0" is not a well-defined function on points in $\mathbb{CP}^2$ because $(x_0 : x_1 : x_2) = (5x_0 : 5x_1 : 5x_2)$! So we'd love to consider these "pseudo-functions" that still have zero loci. These are just the homogeneous polynomials f, because f is homogeneous of degree d if and only if

$$f(\lambda x_0, \dots, \lambda x_n) = \lambda^d f(x_0, \dots, x_n).$$

In particular, the relation "$f(x_0, \dots, x_n) = 0$" is well-defined if F is homogeneous. Thus, we can say:

Definition 79.2.5. If f is homogeneous, we can then define its **vanishing locus** as

$$\mathcal{V}_{\text{pr}}(f) = \{(x_0 : \cdots : x_n) \mid f(x_0, \dots, x_n) = 0\}.$$

The homogeneous condition is really necessary. For example, to require "$x_0 - 1 = 0$" makes no sense, since the points $(1:1:1)$ and $(2015:2015:2015)$ are the same.

It's trivial to verify that homogeneous polynomials do exactly what we want; hence we can now define:

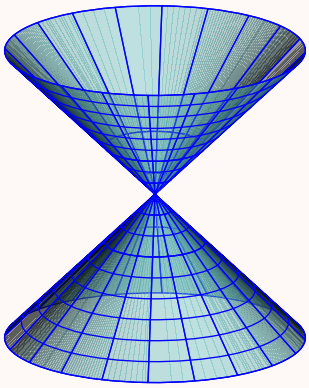
Definition 79.2.6. A **projective variety** in $\mathbb{CP}^n$ is the common zero locus of an arbitrary collection of homogeneous polynomials in $n+1$ variables.

Example 79.2.7 (A conic in $\mathbb{CP}^2$, or a cone in $\mathbb{C}^3$)

Let's try to picture the variety

$$\mathcal{V}_{\text{pr}}(x^2 + y^2 - z^2) \subseteq \mathbb{CP}^2$$

which consists of the points $[x : y : z]$ such that $x^2 + y^2 = z^2$. If we view this as subspace of $\mathbb{C}^3$ (i.e. by thinking of $\mathbb{CP}^2$ as the set of lines through the origin), then we get a "cone":



If we take the standard affine charts now, we obtain:

- At $x = 1$, we get a hyperbola $\mathcal{V}(1 + y^2 - z^2)$.
- At $y = 1$, we get a hyperbola $\mathcal{V}(1 + x^2 - z^2)$.

- At $z = 1$, we get a circle $\mathcal{V}(x^2 + y^2 - 1)$.

That said, over $\mathbb{C}$ a hyperbola and circle are the same thing; I'm cheating a little by drawing $\mathbb{C}$ as one-dimensional, just like last chapter.

Question 79.2.8. Draw the intersection of the cone above with the $z = 1$ plane, and check that you do in fact get a circle. (This geometric picture will be crucial later.)

§79.3 Homogeneous ideals

Now, the next thing we want to do is define $\mathcal{V}_{\text{pr}}(I)$ for an ideal I. Of course, we again run into an issue with things like $x_0 - 1$ not making sense.

The way out of this is to use only *homogeneous* ideals.

Definition 79.3.1. If I is a homogeneous ideal, we define

$$\mathcal{V}_{\text{pr}}(I) = \{x \mid f(x) = 0 \ \forall f \in I\}.$$

Exercise 79.3.2. Show that the notion "$f(x) = 0 \ \forall f \in I$" is well-defined for a homogeneous ideal I.

So, we would hope for a Nullstellensatz-like theorem which bijects the homogeneous radical ideals to projective varieties. Unfortunately:

Example 79.3.3 (Irrelevant ideal)

To crush some dreams and hopes, consider the ideal

$$I = (x_0, x_1, \dots, x_n).$$

This is called the **irrelevant ideal**; it is a homogeneous radical yet $\mathcal{V}_{\text{pr}}(I) = \varnothing$.

However, other than the irrelevant ideal:

Theorem 79.3.4 (Homogeneous Nullstellensatz)

Let I and J be homogeneous ideals.

(a) If $\mathcal{V}_{\text{pr}}(I) = \mathcal{V}_{\text{pr}}(J) \neq \varnothing$ then $\sqrt{I} = \sqrt{J}$.

(b) If $\mathcal{V}_{\text{pr}}(I) = \varnothing$, then either $I = (1)$ or $\sqrt{I} = (x_0, x_1, \dots, x_n)$.

Thus there is a natural bijection between:

- projective varieties in $\mathbb{CP}^n$, and
- homogeneous radical ideals of $\mathbb{C}[x_0, \dots, x_n]$ except for the irrelevant ideal.

Proof. For the first part, let $V = \mathcal{V}_{\text{pr}}(I)$ and $W = \mathcal{V}_{\text{pr}}(J)$ be projective varieties in $\mathbb{CP}^n$. We can consider them as *affine varieties* in $\mathbb{A}^{n+1}$ by using the interpretation of $\mathbb{CP}^n$ as lines through the origin in $\mathbb{C}^n$.

Algebraically, this is done by taking the homogeneous ideals $I, J \subseteq \mathbb{C}[x_0, \dots, x_n]$ and using the same ideals to cut out *affine* varieties $V_{\text{aff}} = \mathcal{V}(I)$ and $W_{\text{aff}} = \mathcal{V}(J)$ in $\mathbb{A}^{n+1}$. For example, the cone $x^2 + y^2 - z^2 = 0$ is a conic (a one-dimensional curve) in $\mathbb{CP}^2$, but can also be thought of as a cone (which is a two-dimensional surface) in $\mathbb{A}^3$.

Then for (a), we have $V_{\text{aff}} = W_{\text{aff}}$, so $\sqrt{I} = \sqrt{J}$.

For (b), either V_{aff} is empty or it is just the origin of $\mathbb{A}^{n+1}$, so the Nullstellensatz implies either $I = (1)$ or $\sqrt{I} = (x_0, \dots, x_n)$ as desired. □

Projective analogues of Theorem 77.4.2 (on intersections and unions of varieties) hold verbatim for projective varieties as well.

§79.4 As ringed spaces

Prototypical example for this section: The regular functions on $\mathbb{CP}^1$ minus a point are exactly those of the form $P(s/t)$.

Now, let us make every projective variety V into a baby ringed space. We already have the set of points, a subset of $\mathbb{CP}^n$.

The topology is defined as follows.

Definition 79.4.1. We endow $\mathbb{CP}^n$ with the **Zariski topology** by declaring the sets of the form $\mathcal{V}_{\text{pr}}(I)$, where I is a homogeneous ideal, to be the closed sets.

Every projective variety V then inherits the Zariski topology from its parent $\mathbb{CP}^n$. The **distinguished open sets** $D(f)$ are $V \setminus \mathcal{V}_{\text{pr}}(f)$.

Thus every projective variety V is now a topological space. It remains to endow it with a sheaf of regular functions $\mathcal{O}_V$. To do this we have to be a little careful. In the affine case we had a nice little ring of functions, the coordinate ring $\mathbb{C}[x_0, \dots, x_n]/I$, that we could use to provide the numerator and denominators. So, it seems natural to then define:

Definition 79.4.2. The **homogeneous coordinate ring** of a projective variety $V = \mathcal{V}_{\text{pr}}(I) \subseteq \mathbb{CP}^n$, where I is homogeneous radical, is defined as the ring

$$\mathbb{C}[V] = \mathbb{C}[x_0, \dots, x_n]/I.$$

Remark 79.4.3 — Unlike the case of Remark 78.4.2, an element of $\mathbb{C}[V]$ no longer correspond to a function from V to $\mathbb{C}$; nevertheless, it is a function from $\mathcal{V}(I) \subseteq \mathbb{A}^{n+1}$ to $\mathbb{C}$.

However, when we define a rational function we must impose a new requirement that the numerator and denominator are the same degree.

Definition 79.4.4. Let $U \subseteq V$ be an open set of a projective variety V. A **rational function** ϕ on a projective variety V is a quotient f/g, where $f, g \in \mathbb{C}[V]$, and f and g are homogeneous of the same degree, and $\mathcal{V}_{\text{pr}}(g) \cap U = \varnothing$. In this way we obtain a function $\phi \colon U \to \mathbb{C}$.

Example 79.4.5 (Examples of rational functions)
Let $V = \mathbb{CP}^1$ have coordinates $(s : t)$.

(a) If $U = V$, then constant functions $c/1$ are the only rational functions on U.

(b) Now let $U_1 = V \setminus \{(1:0)\}$. Then, an example of a regular function is

$$\frac{s^2+9t^2}{t^2} = \left(\frac{s}{t}\right)^2 + 9.$$

If we think of U_1 as $\mathbb{C}$ (i.e. $\mathbb{CP}^1$ minus an infinity point, hence like $\mathbb{A}^1$) then really this is just the function x^2+9.

Then we can repeat the same definition as before:

Definition 79.4.6. Let $U \subseteq V$ be an open set of a projective variety V. We say a function $\phi\colon U \to \mathbb{C}$ is a **regular function** if for every point p, we can find an open set U_p containing p and a rational function f_p/g_p on U_p such that

$$\phi(x) = \frac{f_p(x)}{g_p(x)} \qquad \forall x \in U_p.$$

In particular, we require $U_p \cap \mathcal{V}_{\text{pr}}(g_p) = \varnothing$. We denote the set of all regular functions on U by $\mathcal{O}_V(U)$.

Of course, the rational functions from the previous example are examples of regular functions as well. This completes the definition of a projective variety V as a baby ringed space.

§79.5 Examples of regular functions

Naturally, I ought to tell you what the regular functions on distinguished open sets are; this is an analog to Theorem 78.6.1 from last time.

Theorem 79.5.1 (Regular functions on distinguished open sets for projective varieties)
Let V be a projective variety, and let $g \in \mathbb{C}[V]$ be homogeneous of *positive degree* (thus g is nonconstant). Then

$$\mathcal{O}_V(D(g)) = \left\{ \frac{f}{g^r} \mid f \in \mathbb{C}[V] \text{ homogeneous of degree } r \deg g \right\}.$$

What about the case $g = 1$? A similar result holds, but we need the assumption that V is irreducible.

Definition 79.5.2. A projective variety V is irreducible if it can't be written as the union of two proper (projective) sub-varieties.

Theorem 79.5.3 (Only constant regular functions on projective space)
Let V be an *irreducible* projective variety. Then the only regular functions on V are constant, thus we have

$$\mathcal{O}_V(V) \cong \mathbb{C}.$$

This relies on the fact that $\mathbb{C}$ is algebraically closed.

Proofs of these are omitted for now.

Example 79.5.4 (Irreducibility is needed above)
The reason we need V irreducible is otherwise we could, for example, take V to be the union of two points; in this case $\mathcal{O}_V(V) \cong \mathbb{C}^{\oplus 2}$.

Remark 79.5.5 — It might seem strange that $\mathcal{O}_V(D(g))$ behaves so differently when $g = 1$. One vague explanation is that in a projective variety, a distinguished open $D(g)$ looks much like an affine variety if $\deg g > 0$. For example, in $\mathbb{CP}^1$ we have $\mathbb{CP}^1 \setminus \{0\} \cong \mathbb{A}^1$ (where $\cong$ is used in a sense that I haven't made precise). Thus the claim becomes related to the corresponding affine result. But if $\deg g = 0$ and $g \neq 0$, then $D(g)$ is the entire projective variety, which does not look affine, and thus the analogy breaks down.

Example 79.5.6 (Regular functions on $\mathbb{CP}^1$)
Let $V = \mathbb{CP}^1$, with coordinates $(s : t)$.

(a) By Theorem 79.5.1, if U_1 is the standard affine chart omitting the point $(1 : 0)$, we have $\mathcal{O}_V(U_1) = \left\{ \frac{f}{t^n} \mid \deg f = n \right\}$. One can write this as

$$\mathcal{O}_V(U_1) \cong \{P(s/t) \mid P \in \mathbb{C}[x]\} \cong \mathcal{O}_{\mathbb{A}^1}(\mathbb{A}^1).$$

This conforms with our knowledge that U_1 "looks very much like $\mathbb{A}^1$".

(b) As V is irreducible, $\mathcal{O}_V(V) = \mathbb{C}$: there are no nonconstant functions on $\mathbb{CP}^1$.

Example 79.5.7 (Regular functions on $\mathbb{CP}^2$)
Let $\mathbb{CP}^2$ have coordinates $(x : y : z)$ and let $U_0 = \left\{(x : y : 1) \in \mathbb{CP}^2\right\}$ be the distinguished open set $D(z)$. Then in the same vein,

$$\mathcal{O}_{\mathbb{CP}^2}(U_0) = \left\{ \frac{P(x,y)}{z^n} \mid \deg P = n \right\} \cong \{P(x/z, y/z) \mid P \in \mathbb{C}[x,y]\}.$$

§79.6 A few harder problems to think about

Problems:

80 Bonus: Bézout's theorem

In this chapter we discuss Bézout's theorem. It makes precise the idea that two degree d and e curves in $\mathbb{CP}^2$ should intersect at "exactly" de points. (We work in projective space so e.g. any two lines intersect.)

§80.1 Non-radical ideals

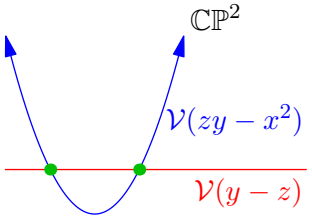
Prototypical example for this section: Tangent to the parabola.

We need to account for multiplicities. So we will whenever possible work with homogeneous ideals I, rather than varieties V, because we want to allow the possibility that I is not radical. Let's see how we might do so.

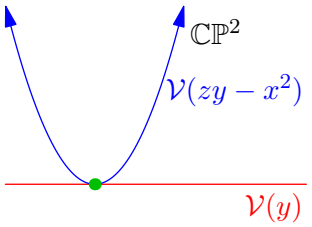
For a first example, suppose we intersect $y = x^2$ with the line $y = 1$; or more accurately, in projective coordinates of $\mathbb{CP}^2$, the parabola $zy = x^2$ and $y = z$. The ideal of the intersection is

$$(zy - x^2, y - z) = (x^2 - z^2, y - z) \subseteq \mathbb{C}[x, y, z].$$

So this corresponds to having two points; this gives two intersection points: $(1 : 1 : 1)$ and $(-1 : 1 : 1)$. Here is a picture of the two varieties in the affine $z = 1$ chart:



That's fine, but now suppose we intersect $zy = x^2$ with the line $y = 0$ instead. Then we instead get a "double point":



The corresponding ideal is this time

$$(zy - x^2, y) = (x^2, y) \subseteq \mathbb{C}[x, y, z].$$

This ideal is *not* radical, and when we take $\sqrt{(x^2, y)} = (x, y)$ we get the ideal which corresponds to a single projective point $(0 : 0 : 1)$ of $\mathbb{CP}^2$. This is why we work with ideals rather than varieties: we need to tell the difference between (x^2, y) and (x, y).

§80.2 Hilbert functions of finitely many points

Prototypical example for this section: The Hilbert function attached to the double point (x^2, y) is eventually the constant 2.

Definition 80.2.1. Given a nonempty projective variety V, there is a unique radical ideal I such that $V = \mathcal{V}_{\text{pr}}(I)$. In this chapter we denote it by $\mathcal{I}_{\text{rad}}(V)$. For an empty variety we set $\mathcal{I}_{\text{rad}}(\varnothing) = (1)$, rather than choosing the irrelevant ideal.

Definition 80.2.2. Let $I \subseteq \mathbb{C}[x_0, \dots, x_n]$ be homogeneous. We define the **Hilbert function** of I, denoted $h_I \colon \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}$ by

$$h_I(d) = \dim_{\mathbb{C}} \left(\mathbb{C}[x_0, \dots, x_n]/I\right)^d$$

i.e. $h_I(d)$ is the dimension of the dth graded part of $\mathbb{C}[x_0, \dots, x_n]/I$.

Definition 80.2.3. If V is a projective variety, we set $h_V = h_{\mathcal{I}_{\text{rad}}(V)}$, where $I = \mathcal{I}_{\text{rad}}(V)$ is the *radical* ideal satisfying $V = \mathcal{V}_{\text{pr}}(I)$ as defined above.

In this case, $\mathbb{C}[x_0, \dots, x_n]/I$ is just $\mathbb{C}[V]$.

Example 80.2.4 (Examples of Hilbert functions in zero dimensions)
For concreteness, let us use $\mathbb{CP}^2$.

(a) If V is the single point $(0:0:1)$, with ideal $\mathcal{I}_{\text{rad}}(V) = (x, y)$, then

$$\mathbb{C}[V] = \mathbb{C}[x, y, z]/(x, y) \cong \mathbb{C}[z] \cong \mathbb{C} \oplus z\mathbb{C} \oplus z^2\mathbb{C} \oplus z^3\mathbb{C} \dots$$

which has dimension 1 in all degrees. Consequently, we have

$$h_I(d) \equiv 1.$$

(b) Now suppose we use the "double point" ideal $I = (x^2, y)$. This time, we have

$$\begin{aligned} \mathbb{C}[x, y, z]/(x^2, y) &\cong \mathbb{C}[z] \oplus x\mathbb{C}[z] \\ &\cong \mathbb{C} \oplus (x\mathbb{C} \oplus z\mathbb{C}) \oplus (xz\mathbb{C} \oplus z^2\mathbb{C}) \oplus (xz^2\mathbb{C} \oplus z^3\mathbb{C}) \oplus \dots. \end{aligned}$$

From this we deduce that

$$h_I(d) = \begin{cases} 2 & d = 1, 2, 3, \dots \\ 1 & d = 0. \end{cases}$$

(c) Let's now take the variety $V = \{(1:1:1), (-1:1:1)\}$ consisting of two points, with $\mathcal{I}_{\text{rad}}(V) = (x^2 - z^2, y - z)$. Then

$$\begin{aligned} \mathbb{C}[x, y, z]/(x^2 - z^2, y - z) &\cong \mathbb{C}[x, z]/(x^2 - z^2) \\ &\cong \mathbb{C}[z] \oplus x\mathbb{C}[z]. \end{aligned}$$

So this example has the same Hilbert function as the previous one.

Abuse of Notation 80.2.5. I'm abusing the isomorphism symbol $\mathbb{C}[z] \cong \mathbb{C} \oplus z\mathbb{C} \oplus z^2\mathbb{C}$ and similarly in other examples. This is an isomorphism only on the level of $\mathbb{C}$-vector

spaces. However, in computing Hilbert functions of other examples I will continue using this abuse of notation.

Example 80.2.6 (Hilbert functions for empty varieties)
Suppose $I \subsetneq \mathbb{C}[x_0, \dots, x_n]$ is an ideal, possibly not radical but such that

$$\mathcal{V}_{\text{pr}}(I) = \varnothing$$

hence $\sqrt{I} = (x_0, \dots, x_n)$ is the irrelevant ideal. Thus there are integers d_i for $i = 0, \dots, n$ such that $x_i^{d_i} \in I$ for every i; consequently, $h_I(d) = 0$ for any $d > d_0 + \cdots + d_n$. We summarize this by saying that

$$h_I(d) = 0 \text{ for all } d \gg 0.$$

Here the notation $d \gg 0$ means "all sufficiently large d".

From these examples we see that if I is an ideal, then the Hilbert function appears to eventually be constant, with the desired constant equal to the size of $\mathcal{V}_{\text{pr}}(I)$, "with multiplicity" in the case that I is not radical.

Let's prove this. Before proceeding we briefly remind the reader of short exact sequences: a sequence of maps of $0 \to V \hookrightarrow W \twoheadrightarrow X \to 0$ is one such that the $\text{im}(V \hookrightarrow W) = \ker(W \twoheadrightarrow X)$ (and of course the maps $V \hookrightarrow W$ and $W \twoheadrightarrow X$ are injective and surjective). If V, W, X are finite-dimensional vector spaces over $\mathbb{C}$ this implies that $\dim W = \dim V + \dim X$.

Proposition 80.2.7 (Hilbert functions of $I \cap J$ and $I + J$)
Let I and J be homogeneous ideals in $\mathbb{C}[x_0, \dots, x_n]$. Then

$$h_{I \cap J} + h_{I+J} = h_I + h_J.$$

Proof. Consider any $d \geq 0$. Let $S = \mathbb{C}[x_0, \dots, x_n]$ for brevity. Then

$$0 \longrightarrow [S/(I \cap J)]^d \hookrightarrow [S/I]^d \oplus [S/J]^d \twoheadrightarrow [S/(I+J)]^d \longrightarrow 0$$

$$f \longmapsto (f, f)$$

$$(f, g) \longmapsto f - g$$

is a short exact sequence of vector spaces. Therefore, for every $d \geq 0$ we have that

$$\dim [S/I]^d \oplus [S/J]^d = \dim [S/(I \cap J)]^d + \dim [S/(I+J)]^d$$

which gives the conclusion. □

Example 80.2.8 (Hilbert function of two points in $\mathbb{CP}^1$)
In $\mathbb{CP}^1$ with coordinate ring $\mathbb{C}[s, t]$, consider $I = (s)$ the ideal corresponding to the point $(0 : 1)$ and $J = (t)$ the ideal corresponding to the point $(1 : 0)$. Then $I \cap J = (st)$ is the ideal corresponding to the disjoint union of these two points,

while $I + J = (s, t)$ is the irrelevant ideal. Consequently $h_{I+J}(d) = 0$ for $d \gg 0$. Therefore, we get

$$h_{I \cap J}(d) = h_I(d) + h_J(d) \text{ for } d \gg 0$$

so the Hilbert function of a two-point projective variety is the constant 2 for $d \gg 0$.

This example illustrates the content of the main result:

Theorem 80.2.9 (Hilbert functions of zero-dimensional varieties)
Let V be a projective variety consisting of m points (where $m \geq 0$ is an integer). Then

$$h_V(d) = m \text{ for } d \gg 0.$$

Proof. We already did $m = 0$, so assume $m \geq 1$. Let $I = \mathcal{I}_{\text{rad}}(V)$ and for $k = 1, \dots, m$ let $I_k = \mathcal{I}_{\text{rad}}(k\text{th point of } V)$.

Exercise 80.2.10. Show that $h_{I_k}(d) = 1$ for every d. (Modify Example 80.2.4(a).)

Hence we can proceed by induction on $m \geq 2$, with the base case $m = 1$ already done above. For the inductive step, we use the projective analogues of Theorem 77.4.2. We know that $h_{I_1 \cap \dots \cap I_{m-1}}(d) = m - 1$ for $d \gg 0$ (this is the first $m - 1$ points; note that $I_1 \cap \dots \cap I_{m-1}$ is radical). To add in the mth point we note that

$$h_{I_1 \cap \dots \cap I_m}(d) = h_{I_1 \cap \dots I_{m-1}}(d) + h_{I_m}(d) - h_J(d)$$

where $J = (I_1 \cap \dots \cap I_{m-1}) + I_m$. The ideal J may not be radical, but satisfies $\mathcal{V}_{\text{pr}}(J) = \varnothing$ by an earlier example, hence $h_J = 0$ for $d \gg 0$. This completes the proof. □

In exactly the same way we can prove that:

Corollary 80.2.11 (h_I eventually constant when $\dim \mathcal{V}_{\text{pr}}(I) = 0$)
Let I be an ideal, not necessarily radical, such that $\mathcal{V}_{\text{pr}}(I)$ consists of finitely many points. Then the Hilbert h_I is eventually constant.

Proof. Induction on the number of points, $m \geq 1$. The base case $m = 1$ was essentially done in Example 80.2.4(b) and Exercise 80.2.10. The inductive step is literally the same as in the proof above, except no fuss about radical ideals. □

§80.3 Hilbert polynomials

So far we have only talked about Hilbert functions of zero-dimensional varieties, and showed that they are eventually constant. Let's look at some more examples.

Example 80.3.1 (Hilbert function of $\mathbb{CP}^n$)
The Hilbert function of $\mathbb{CP}^n$ is

$$h_{\mathbb{CP}^n}(d) = \binom{d+n}{n} = \frac{1}{n!}(d+n)(d+n-1)\dots(d+1)$$

by a "balls and urns" argument. This is a polynomial of degree n.

Example 80.3.2 (Hilbert function of the parabola)
Consider the parabola $zy - x^2$ in $\mathbb{CP}^2$ with coordinates $\mathbb{C}[x, y, z]$. Then

$$\mathbb{C}[x, y, z]/(zy - x^2) \cong \mathbb{C}[y, z] \oplus x\mathbb{C}[y, z].$$

A combinatorial computation gives that

$$\begin{aligned} h_{(zy-x^2)}(0) &= 1 && \text{Basis } 1 \\ h_{(zy-x^2)}(1) &= 3 && \text{Basis } x, y, z \\ h_{(zy-x^2)}(2) &= 5 && \text{Basis } xy,\ xz,\ y^2,\ yz,\ z^2. \end{aligned}$$

We thus in fact see that $h_{(zy-x^2)}(d) = 2d - 1$.

In fact, this behavior of "eventually polynomial" always works.

Theorem 80.3.3 (Hilbert polynomial)
Let $I \subseteq \mathbb{C}[x_0, \dots, x_n]$ be a homogeneous ideal, not necessarily radical. Then

(a) There exists a polynomial χ_I such that $h_I(d) = \chi_I(d)$ for all $d \gg 0$.

(b) $\deg \chi_I = \dim \mathcal{V}_{\text{pr}}(I)$ (if $\mathcal{V}_{\text{pr}}(I) = \varnothing$ then $\chi_I = 0$).

(c) The polynomial $m! \cdot \chi_I$ has integer coefficients.

Proof. The base case was addressed in the previous section.

For the inductive step, consider $\mathcal{V}_{\text{pr}}(I)$ with dimension m. Consider a hyperplane H such that no irreducible component of $\mathcal{V}_{\text{pr}}(I)$ is contained inside H (we quote this fact without proof, as it is geometrically obvious, but the last time I tried to write the proof I messed up). For simplicity, assume WLOG that $H = \mathcal{V}_{\text{pr}}(x_0)$.

Let $S = \mathbb{C}[x_0, \dots, x_n]$ again. Now, consider the short exact sequence

$$0 \longrightarrow [S/I]^{d-1} \hookrightarrow [S/I]^d \twoheadrightarrow [S/(I + (x_0))]^d \longrightarrow 0$$

$$f \longmapsto f \cdot x_0$$

$$f \longmapsto f.$$

(The injectivity of the first map follows from the assumption about irreducible components of $\mathcal{V}_{\text{pr}}(I)$.) Now exactness implies that

$$h_I(d) - h_I(d-1) = h_{I+(x_0)}(d).$$

The last term geometrically corresponds to $\mathcal{V}_{\text{pr}}(I) \cap H$; it has dimension $m - 1$, so by the inductive hypothesis we know that

$$h_I(d) - h_I(d-1) = \frac{c_0 d^{m-1} + c_1 d^{m-2} + \cdots + c_{m-1}}{(m-1)!} \qquad d \gg 0$$

for some integers $c_0, \dots, c_{m-1}$. Then we are done by the theory of **finite differences** of polynomials. □

§80.4 Bézout's theorem

Definition 80.4.1. We call χ_I the **Hilbert polynomial** of I. If χ_I is nonzero, we call the leading coefficient of $m!\chi_I$ the **degree** of I, which is an integer, denoted $\deg I$.

Of course for projective varieties V we let $h_V = h_{\mathcal{I}_{\text{rad}}(V)}$, and $\deg V = \deg \mathcal{I}_{\text{rad}}(V)$.

> **Remark 80.4.2 —** Note that the degree of an ideal $\deg I$ is not the same as $\deg h_I$!

Let us show some properties of the degrees, which will allow us to compute the degree of any projective variety from its irreducible components.

Proposition 80.4.3 (Properties of degrees)

For two varieties V and W, we have the following:

- If V and W are disjoint and have the same dimension, then $\deg(V \cup W) = \deg V + \deg W$.
- If $\dim V < \dim W$, then $\deg(V \cup W) = \deg W$.

So,

The degree is additive over components, and it measures the "degree" of the highest-dimensional component.

Proof. Follows from the properties of Hilbert polynomial in Theorem 80.3.3 and Proposition 80.2.7, and that the leading coefficient only depends on the largest-degree summand. □

Example 80.4.4 (Examples of degrees)

(a) If V is a finite set of $n \geq 1$ points, it has degree n.

(b) If I corresponds to a double point, it has degree 2.

(c) $\mathbb{CP}^n$ has degree 1.

(d) Any line or plane, being "isomorphic" to $\mathbb{CP}^1$ and $\mathbb{CP}^2$ respectively, has degree 1.

(e) The parabola has degree 2. (Note that, as an algebraic variety, the parabola is isomorphic to a line!)

(f) The union of the parabola and a point has degree 2.

Now, you might guess that if f is a homogeneous quadratic polynomial then the degree of the principal ideal (f) is 2, and so on. (Thus for example we expect a circle to have degree 2.) This is true:

Theorem 80.4.5 (Bézout's theorem)
Let I be a homogeneous ideal of $\mathbb{C}[x_0, \dots, x_n]$, such that $\dim \mathcal{V}_{\text{pr}}(I) \geq 1$. Let $f \in \mathbb{C}[x_0, \dots, x_n]$ be a homogeneous polynomial of degree k which does not vanish on any irreducible component of $\mathcal{V}_{\text{pr}}(I)$. Then

$$\deg (I + (f)) = k \deg I.$$

Geometrically,

If V is any projective variety, $\mathcal{V}(f)$ is a hyperplane of degree k, then their intersection $V \cap \mathcal{V}(f)$ has degree $k \deg V$ — unless some irreducible component of V is contained inside $\mathcal{V}(f)$.

This is what we mentioned at the beginning of the chapter.

Because the ideal I may not be radical, the geometric interpretation statement is not the most general possible — the problem will be rectified later with the generalization to schemes.

Proof. Let $S = \mathbb{C}[x_0, \dots, x_n]$ again. This time the exact sequence is

$$0 \longrightarrow [S/I]^{d-k} \hookrightarrow [S/I]^d \twoheadrightarrow [S/(I + (f))]^d \longrightarrow 0.$$

We leave this olympiad-esque exercise as Problem 80A. □

§80.5 Applications

First, we show that the notion of degree is what we expect.

Corollary 80.5.1 (Hypersurfaces: the degree deserves its name)
Let V be a hypersurface, i.e. $\mathcal{I}_{\text{rad}}(V) = (f)$ for f a homogeneous polynomial of degree k. Then $\deg V = k$.

Proof. Recall $\deg(0) = \deg \mathbb{CP}^n = 1$. Take $I = (0)$ in Bézout's theorem. □

The common special case in $\mathbb{CP}^2$ is:

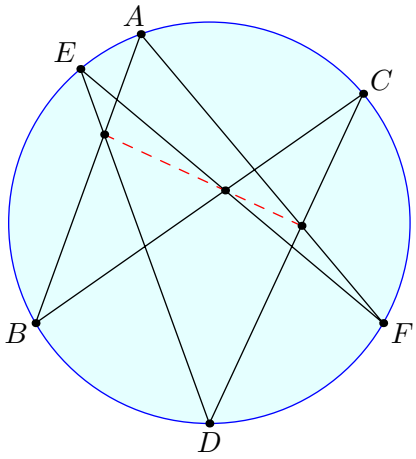
Corollary 80.5.2 (Bézout's theorem for curves)
For any two curves X and Y in $\mathbb{CP}^2$ without a common irreducible component,

$$|X \cap Y| \leq \deg X \cdot \deg Y.$$

Now, we use this to prove Pascal's theorem.

Theorem 80.5.3 (Pascal's theorem)
Let A, B, C, D, E, F be six distinct points which lie on a conic $\mathscr{C}$ in $\mathbb{CP}^2$. Then the points $AB \cap DE$, $BC \cap EF$, $CD \cap FA$ are collinear.

Proof. Let X be the variety equal to the union of the three lines AB, CD, EF, hence $X = \mathcal{V}_{\text{pr}}(f)$ for some cubic polynomial f (which is the product of three linear ones). Similarly, let $Y = \mathcal{V}_{\text{pr}}(g)$ be the variety equal to the union of the three lines BC, DE, FA.



Now let P be an arbitrary point on the conic on $\mathscr{C}$, distinct from the six points A, B, C, D, E, F. Consider the projective variety

$$V = \mathcal{V}_{\text{pr}}(\alpha f + \beta g)$$

where the constants α and β are chosen such that $P \in V$.

> **Question 80.5.4.** Show that V also contains the six points A, B, C, D, E, F as well as the three points $AB \cap DE$, $BC \cap EF$, $CD \cap FA$ regardless of which α and β are chosen.

Now, note that $|V \cap \mathscr{C}| \geq 7$. But $\deg V = 3$ and $\deg \mathscr{C} = 2$. This contradicts Bézout's theorem unless V and $\mathscr{C}$ share an irreducible component. This can only happen if V is the union of a line and conic, for degree reasons; i.e. we must have that

$$V = \mathscr{C} \cup \text{line}.$$

Finally note that the three intersection points $AB \cap DE$, $BC \cap EF$ and $CD \cap FA$ do not lie on $\mathscr{C}$, so they must lie on this line. □

We'd like to remark that the Pascal's theorem is just a special case of the Cayley-Bacharach theorem, which can be used to prove that the addition operation on an elliptic curve is associative. Interested readers may want to try proving the Cayley-Bacharach theorem using the same technique.

§80.6 A few harder problems to think about

Problem 80A. Complete the proof of Bézout's theorem from before.

Problem 80B (USA TST 2016/6)**.** Let ABC be an acute scalene triangle and let P be a point in its interior. Let A_1, B_1, C_1 be projections of P onto triangle sides BC, CA, AB, respectively. Find the locus of points P such that AA_1, BB_1, CC_1 are concurrent and $\angle PAB + \angle PBC + \angle PCA = 90°$.

81 Morphisms of varieties

In preparation for our work with schemes, we will finish this part by talking about *morphisms* between affine and projective varieties, given that we have taken the time to define them.

Idea: we know both affine and projective varieties are special cases of baby ringed spaces, so in fact we will just define a morphism between *any* two baby ringed spaces.

§81.1 Defining morphisms of baby ringed spaces

Prototypical example for this section: See next section.

Let $(X, \mathcal{O}_X)$ and $(Y, \mathcal{O}_Y)$ be baby ringed spaces, and think about how to define a morphism between them.

The guiding principle in algebra is that we want morphisms to be functions on underlying structure, but also respect the enriched additional data on top. To give some examples from the very beginning of time:

Example 81.1.1 (How to define a morphism)

- Consider groups. A group G has an underlying set (of elements), which we then enrich with a multiplication operation. So a homomorphism is a map of the underlying sets, plus it has to respect the group multiplication.
- Consider R-modules. Each R-module has an underlying abelian group, which we then enrich with scalar multiplication. So we require that a linear map respects the scalar multiplication as well, in addition to being a homomorphism of abelian groups.
- Consider topological spaces. A space X has an underlying set (of points), which we then enrich with a topology of open sets. So we consider maps of the set of points which respect the topology (pre-images of open sets are open).

This time, the ringed spaces $(X, \mathcal{O}_X)$ have an underlying *topological space*, which we have enriched with a structure sheaf. So, we want a continuous map $f \colon X \to Y$ of these topological spaces, which we then need to respect the sheaf of regular functions.

How might we do this? Well, if we let $\psi \colon Y \to \mathbb{C}$ be a regular function, then composition gives a natural way to write a map $X \to Y \to \mathbb{C}$. We then want to require that this is also a regular function.

More generally, we can take any regular function on Y and obtain some function on X, which we call a pullback. We then require that all the pullbacks are regular on X.

Definition 81.1.2. Let $(X, \mathcal{O}_X)$ and $(Y, \mathcal{O}_Y)$ be baby ringed spaces. Given a map $f \colon X \to Y$ and a regular function $\phi \in \mathcal{O}_Y(U)$, we define the **pullback** of ϕ, denoted $f^\sharp \phi$, to be the composed function

$$f^{\text{pre}}(U) \xrightarrow{f} U \xrightarrow{\phi} \mathbb{C}.$$

The use of the word "pullback" is the same as in our study of differential forms.

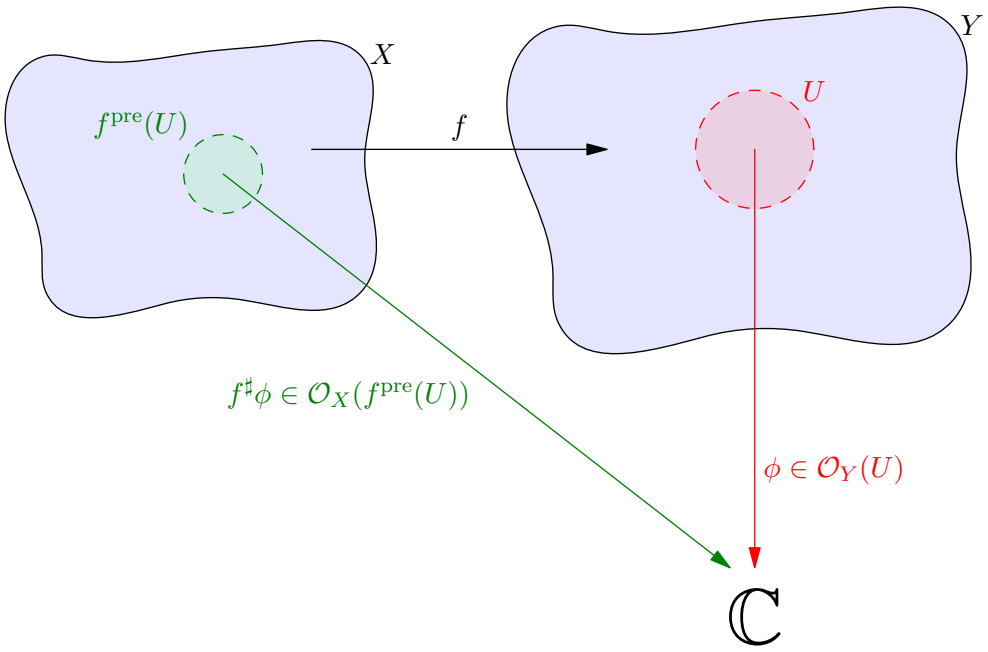
Definition 81.1.3. Let $(X, \mathcal{O}_X)$ and $(Y, \mathcal{O}_Y)$ be baby ringed spaces. A continuous map of topological spaces $f\colon X \to Y$ is a **morphism** if every pullback of a regular function on Y is a regular function on X.

Two baby ringed spaces are **isomorphic** if there are mutually inverse morphisms between them, which we then call **isomorphisms**.

In particular, the pullback gives us a (reversed) *ring homomorphism*

$$f^\sharp\colon \mathcal{O}_Y(U) \to \mathcal{O}_X(f^{\text{pre}}(U))$$

for *every* U; thus our morphisms package a lot of information. Here's a picture of a morphism f, and the pullback of $\phi\colon U \to \mathbb{C}$ (where $U \subseteq Y$).



Example 81.1.4 (The pullback of $\frac{1}{y-25}$ under $t \mapsto t^2$)

The map

$$f\colon X = \mathbb{A}^1 \to Y = \mathbb{A}^1 \quad \text{by} \quad t \mapsto t^2$$

is a morphism of varieties. For example, consider the regular function $\varphi = \frac{1}{y-25}$ on the open set $Y \setminus \{25\} \subseteq Y$. The f-inverse image is $X \setminus \{\pm 5\}$. Thus the pullback is

$$f^\sharp\varphi\colon X \setminus \{\pm 5\} \to Y \setminus \{25\}$$
$$\text{by} \quad x \mapsto \frac{1}{x^2 - 25}$$

which is regular on $X \setminus \{\pm 5\}$.

§81.2 Classifying the simplest examples

Prototypical example for this section: Theorem 81.2.2; they're just polynomials.

On a philosophical point, we like the earlier definition because it adheres to our philosophy of treating our varieties as intrinsic objects, rather than embedded ones. However, it is somewhat of a nuisance to actually verify it.

So in this section, we will

- classify all the morphisms from $\mathbb{A}^m \to \mathbb{A}^n$, and
- classify all the morphisms from $\mathbb{CP}^m \to \mathbb{CP}^n$.

It what follows I will wave my hands a lot in claiming that something is a morphism, since doing so is mostly detail checking. The theorems which follow will give us alternative definitions of morphism which are more coordinate-based and easier to use for actual computations.

§81.2.i Affine classification

Earlier we saw how $t \mapsto t^2$ gives us a map. More generally, given any polynomial $P(t)$, the map $t \mapsto P(t)$ will work. And in fact, that's all:

> **Exercise 81.2.1.** Let $X = \mathbb{A}^1$, $Y = \mathbb{A}^1$. By considering $\text{id} \in \mathcal{O}_Y(Y)$, show that no other regular functions exist.

In fact, let's generalize the previous exercise:

Theorem 81.2.2 (Regular maps of affine varieties are globally polynomials)
Let $X \subseteq \mathbb{A}^m$ and $Y \subseteq \mathbb{A}^n$ be affine varieties. Every morphism $f \colon X \to Y$ of varieties is given by

$$x = (x_1, \dots, x_m) \overset{f}{\mapsto} (P_1(x), \dots, P_n(x))$$

where $P_1, \dots, P_n$ are polynomials.

Proof. It's not too hard to see that all such functions work, so let's go the other way. Let $f \colon X \to Y$ be a morphism.

First, remark that $f^{\text{pre}}(Y) = X$. Now consider the regular function $\pi_1 \in \mathcal{O}_Y(Y)$, given by the projection $(y_1, \dots, y_n) \mapsto y_1$. Thus we need $f \circ \pi_1$ to be regular on X.

But for affine varieties $\mathcal{O}_X(X)$ is just the coordinate ring $\mathbb{C}[X]$ and so we know there is a polynomial P_1 such that $f \circ \pi_1 = P_1$. Similarly for the other coordinates. □

§81.2.ii Projective classification

Unfortunately, the situation is a little weirder in the projective setting. If $X \subseteq \mathbb{CP}^m$ and $Y \subseteq \mathbb{CP}^n$ are projective varieties, then every function

$$x = (x_0 : x_1 : \cdots : x_m) \mapsto (P_0(x) : P_1(x) : \cdots : P_n(x))$$

is a valid morphism, provided the P_i are homogeneous of the same degree and don't all vanish simultaneously. However if we try to repeat the proof for affine varieties we run into an issue: there is no π_1 morphism. (Would we send $(1:1) = (2:2)$ to 1 or 2?)

And unfortunately, there is no way to repair this. Counterexample:

Example 81.2.3 (Projective map which is not globally polynomial)

Let $V = \mathcal{V}_{\text{pr}}(xy - z^2) \subseteq \mathbb{CP}^2$. Then the map

$$V \to \mathbb{CP}^1 \quad \text{by} \quad (x:y:z) \mapsto \begin{cases} (x:z) & x \neq 0 \\ (z:y) & y \neq 0 \end{cases}$$

turns out to be a morphism of projective varieties. This is well defined just because $(x:z) = (z:y)$ if $x, y \neq 0$; this should feel reminiscent of the definition of regular function.

The good news is that "local" issues are the only limiting factor.

Theorem 81.2.4 (Regular maps of projective varieties are locally polynomials)
Let $X \subseteq \mathbb{CP}^m$ and $Y \subseteq \mathbb{CP}^n$ be projective varieties and let $f\colon X \to Y$ be a morphism. Then at every point $p \in X$ there exists an open neighborhood $U_p \ni p$ and polynomials P_0, P_1, …, P_n (which depend on U) so that

$$f(x) = (P_0(x) : P_1(x) : \cdots : P_n(x)) \quad \forall x = (x_0 : \cdots : x_n) \in U_p.$$

Of course the polynomials P_i must be homogeneous of the same degree and cannot vanish simultaneously on any point of U_p.

Example 81.2.5 (Example of an isomorphism)
In fact, the map $V = \mathcal{V}_{\text{pr}}(xy - z^2) \to \mathbb{CP}^1$ is an isomorphism. The inverse map $\mathbb{CP}^1 \to V$ is given by

$$(s:t) \mapsto (s^2 : t^2 : st).$$

Thus actually $V \cong \mathbb{CP}^1$.

§81.3 Some more applications and examples

Prototypical example for this section: $\mathbb{A}^1 \hookrightarrow \mathbb{CP}^1$ *is a good one.*

The previous section complete settles affine varieties to affine varieties, and projective varieties to projective varieties. However, the definition we gave at the start of the chapter works for *any* baby ringed spaces, and therefore there is still a lot of room to explore.

For example, **we can have affine spaces talk to projective ones**. Why not? The power of our pullback-based definition is that you enable any baby ringed spaces to communicate, even if they live in different places.

Example 81.3.1 (Embedding $\mathbb{A}^1 \hookrightarrow \mathbb{CP}^1$)
Consider a morphism

$$f\colon \mathbb{A}^1 \hookrightarrow \mathbb{CP}^1 \quad \text{by} \quad t \mapsto (t:1).$$

This is also a morphism of varieties. (Can you see what the pullbacks look like?) This reflects the fact that $\mathbb{CP}^1$ is "$\mathbb{A}^1$ plus a point at infinity".

Here is another way you can generate more baby ringed spaces. Given any projective variety, you can take an open subset of it, and that will itself be a baby ringed space. We give this a name:

Definition 81.3.2. A **quasi-projective variety** is an open set X of a projective variety V. It is a baby ringed space $(X, \mathcal{O}_X)$ too, because for any open set $U \subseteq X$ we simply define $\mathcal{O}_X(U) = \mathcal{O}_V(U)$.

We chose to take open subsets of projective varieties because this will subsume the affine ones, for example:

Example 81.3.3 (The parabola is quasi-projective)
Consider the parabola $V = \mathcal{V}(y - x^2) \subset \mathbb{A}^2$. We take the projective variety $W = \mathcal{V}_{\text{pr}}(zy - x^2)$ and look at the standard affine chart $D(z)$. Then there is an isomorphism

$$\begin{aligned} V &\to D(z) \subseteq W \\ (x, y) &\mapsto (x : y : 1) \\ (x/z, y/z) &\hookleftarrow (x : y : z). \end{aligned}$$

Consequently, V is (isomorphic to) an open subset of W, thus we regard it as quasi-projective.

In general this proof can be readily adapted:

Proposition 81.3.4 (Affine $\subseteq$ quasi-projective)
Every affine variety is isomorphic to a quasi-projective one (i.e. every affine variety is an open subset of a projective variety).

So quasi-projective varieties generalize both types of varieties we have seen.

§81.4 The hyperbola effect

Prototypical example for this section: $\mathbb{A}^1 \setminus \{0\}$ *is even affine*

So here is a natural question: are there quasi-projective varieties which are neither affine nor projective? The answer is yes, but for the sake of narrative I'm going to play dumb and find a *non-example*, with the actual example being given in the problems.

Our first guess might be to take the simplest projective variety, say $\mathbb{CP}^1$, and delete a point (to get an open set). This is quasi-projective, but it's isomorphic to $\mathbb{A}^1$. So instead we start with the simplest affine variety, say $\mathbb{A}^1$, and try to delete a point.

Surprisingly, this doesn't work.

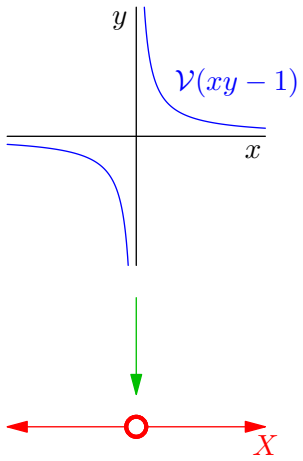
Example 81.4.1 (Crucial example: punctured line is isomorphic to hyperbola)
Let $X = \mathbb{A}^1 \setminus \{0\}$ be an quasi-projective variety. We claim that in fact we have an isomorphism

$$X \cong V = \mathcal{V}(xy - 1) \subseteq \mathbb{A}^2$$

which shows that X is still isomorphic to an affine variety. The maps are

$$\begin{aligned} X &\leftrightarrow V \\ t &\mapsto (t, 1/t) \\ x &\leftarrow\!\shortmid (x, y). \end{aligned}$$

Intuitively, the "hyperbola $y = 1/x$" in $\mathbb{A}^2$ can be projected onto the x-axis. Here is the relevant picture.



Actually, deleting any number of points from $\mathbb{A}^1$ fails. If we delete $\{1, 2, 3\}$, the resulting open set is isomorphic as a baby ringed space to $\mathcal{V}(y(x-1)(x-2)(x-3)-1)$, which colloquially might be called $y = \frac{1}{(x-1)(x-2)(x-3)}$.

The truth is more general.

Distinguished open sets of affine varieties are affine.

Here is the exact isomorphism.

Theorem 81.4.2 (Distinguished open subsets of affines are affine)
Consider $X = D(f) \subseteq V = \mathcal{V}(f_1, \dots, f_m) \subseteq \mathbb{A}^n$, where V is an affine variety, and the distinguished open set X is thought of as a quasi-projective variety. Define

$$W = \mathcal{V}(f_1, \dots, f_m, y \cdot f - 1) \subseteq \mathbb{A}^{n+1}$$

where y is the $(n+1)$st coordinate of $\mathbb{A}^{n+1}$.
Then $X \cong W$.

For lack of a better name, I will dub this the **hyperbola effect**, and it will play a significant role later on.

Therefore, if we wish to find an example of a quasi-projective variety which is not affine, one good place to look would be an open set of an affine space which is not distinguished open. If you are ambitious now, you can try to prove the punctured plane (that is, $\mathbb{A}^2$ minus the origin) works. We will see that example once again later in the next chapter, so you will have a second chance to do so.

§81.5 A few harder problems to think about

Problem 81A. Consider the map

$$\mathbb{A}^1 \to \mathcal{V}(y^2 - x^3) \subseteq \mathbb{A}^2 \quad \text{by} \quad t \mapsto (t^2, t^3).$$

Show that it is a morphism of varieties, but it is not an isomorphism.

Problem 81B†. Show that every projective variety has an open neighborhood which is isomorphic to an affine variety. In this way, "projective varieties are locally affine".

Problem 81C. Let V be a affine variety and let W be a irreducible projective variety. Prove that $V \cong W$ if and only if V and W are a single point.

Problem 81D (Punctured plane is not affine)**.** Let $X = \mathbb{A}^2 \setminus \{(0,0)\}$ be an open set of $\mathbb{A}^2$. Let V be any affine variety and let $f \colon X \to V$ be a morphism. Show that f is not an isomorphism.

XX

Algebraic Geometry II: Affine Schemes

Part XX: Contents

82 Sheaves and ringed spaces

Most of the complexity of the affine variety V earlier comes from $\mathcal{O}_V$. This is a type of object called a "sheaf". The purpose of this chapter is to completely define what this sheaf is, and just what it is doing.

§82.1 Motivation and warnings

The typical example to keep in mind is a sheaf of "functions with property P" on a topological space X: for every open set U, $\mathscr{F}(U)$ gives us the ring of functions on X. However, we will work very abstractly and only assume $\mathscr{F}(U)$ is a ring, without an interpretation as "functions".

Throughout this chapter, I will not only be using algebraic geometry examples, but also those with X a topological space and $\mathscr{F}$ being a sheaf of differentiable/analytic/etc functions. One of the nice things about sheaves is that the same abstraction works fine, so you can train your intuition with both algebraic and analytic examples. In particular, we can keep drawing open sets U as ovals, even though in the Zariski topology that's not what they look like.

The payoff for this abstraction is that it will allow us to define an arbitrary scheme in Chapter 84. Varieties use $\mathbb{C}[x_1, x_2, \dots, x_n]/I$ as their "ring of functions", and by using the fully general sheaf we replace this with *any* commutative ring. In particular, we could choose $\mathbb{C}[x]/(x^2)$ and this will give the "multiplicity" behavior that we sought all along.

§82.2 Pre-sheaves

Prototypical example for this section: The sheaf of holomorphic (or regular, continuous, differentiable, constant, whatever) functions.

The proper generalization of our $\mathcal{O}_V$ is a so-called sheaf of rings. Recall that $\mathcal{O}_V$ took *open sets of* V to *rings*, with the interpretation that $\mathcal{O}_V(U)$ was a "ring of functions".

Recall from Section 78.5 that $\mathcal{O}_V$, as a set, consist of simply the algebraic functions. However, if we view $\mathcal{O}_V$ purely as a set, the structure of the functions is essentially thrown way.

Let us see how the functions in $\mathcal{O}_V$ are related to each other:

- Each function in $\mathcal{O}_V$ is defined on a open set $U \subseteq V$.
- If two functions are defined on the same open set, you can add and multiply them together. In other words, $\mathcal{O}_V(U)$ is a ring.
- Given a function $f \in \mathcal{O}_V(U)$, we can restrict it to a smaller open subset $W \subseteq U$.

These are the operations that we will impose on a pre-sheaf.

§82.2.i Usual definition

So here is the official definition of a pre-sheaf. We will only define a pre-sheaf of rings, however it's possible to define a pre-sheaf of sets, pre-sheaf of abelian groups, etc.

Definition 82.2.1. For a topological space X let $\text{OpenSets}(X)$ denote the open sets of X.

Definition 82.2.2. A **pre-sheaf** of rings on a space X is a function

$$\mathscr{F} \colon \text{OpenSets}(X) \to \mathsf{Rings}$$

meaning each open set gets associated with a ring $\mathscr{F}(U)$. Each individual element of $\mathscr{F}(U)$ is called a **section**.

It is also equipped with a **restriction map** for any $U_1 \subseteq U_2$; this is a map

$$\text{res}_{U_1,U_2} \colon \mathscr{F}(U_2) \to \mathscr{F}(U_1).$$

The map satisfies two axioms:

- The map $\text{res}_{U,U}$ is the identity, and
- Whenever we have nested subsets

  $$U_{\text{small}} \subseteq U_{\text{med}} \subseteq U_{\text{big}}$$

  the diagram

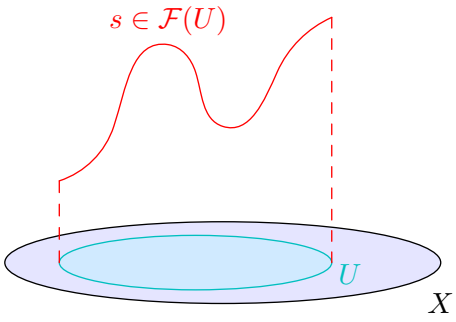
  $$\begin{array}{ccc} \mathscr{F}(U_{\text{big}}) & \xrightarrow{\text{res}} & \mathscr{F}(U_{\text{med}}) \\ & \searrow^{\text{res}} & \downarrow{\scriptstyle \text{res}} \\ & & \mathscr{F}(U_{\text{small}}) \end{array}$$

  commutes.

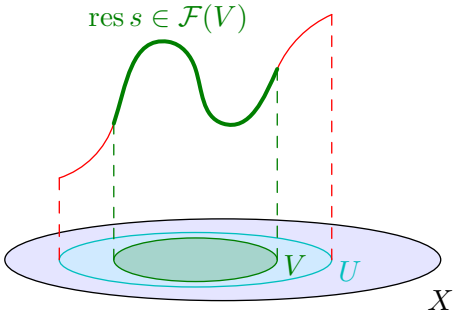
Definition 82.2.3. An element of $\mathscr{F}(X)$ is called a **global section**.

Abuse of Notation 82.2.4. If $s \in \mathscr{F}(U_2)$ is some section and $U_1 \subseteq U_2$, then rather than write $\text{res}_{U_1,U_2}(s)$ I will write $s\restriction_{U_1}$ instead: "s restricted to U_1". This is abuse of notation because the section s is just an element of some ring, and in the most abstract of cases may not have a natural interpretation as function.

Here is a way you can picture sections. In all our usual examples, sheaves return functions an open set U. So, we draw a space X, and an open set U, and then we want to draw a "function on U" to represent a section s. Crudely, we will illustrate s by drawing an xy-plot of a curve, since that is how we were told to draw functions in grade school.



Then, the restriction corresponds to, well, taking just a chunk of the section.



All of this is still a dream, since s in reality is an element of a ring. However, by the end of this chapter we will be able to make our dream into a reality.

Example 82.2.5 (Examples of pre-sheaves)

(a) For an affine variety V, $\mathcal{O}_V$ is of course a sheaf, with $\mathcal{O}_V(U)$ being the ring of regular functions on U. The restriction map just says that if $U_1 \subseteq U_2$, then a function $s \in \mathcal{O}_V(U_2)$ can also be thought of as a function $s\restriction_{U_1} \in \mathcal{O}_V(U_1)$, hence the name "restriction". The commutativity of the diagram then follows.

(b) Let $X \subseteq \mathbb{R}^n$ be an open set. Then there is a sheaf of smooth/differentiable/etc. functions on X. In fact, one can do the same construction for any manifold M.

(c) Similarly, if $X \subseteq \mathbb{C}$ is open, we can construct a sheaf of holomorphic functions on X.

In all these examples, the sections $s \in \mathscr{F}(U)$ are really functions on the space, but in general they need not be.

In practice, thinking about the restriction maps might be more confusing than helpful; it is better to say:

Pre-sheaves should be thought of as "returning the ring of functions with a property P".

§82.2.ii Categorical definition

If you really like category theory, we can give a second equivalent and shorter definition. Despite being a category lover myself, I find this definition less intuitive, but its brevity helps with remembering the first one.

Abuse of Notation 82.2.6. By abuse of notation, $\text{OpenSets}(X)$ will also be thought of as a posetal category by inclusion. Thus $\varnothing$ is an initial object and the entire space X is a terminal object.

Definition 82.2.7. A **pre-sheaf** of rings on X is a contravariant functor

$$\mathscr{F} \colon \text{OpenSets}(X)^{\text{op}} \to \mathsf{Rings}.$$

Exercise 82.2.8. Check that these definitions are equivalent.

In particular, it is possible to replace Rings with any category we want. We will not need to do so any time soon, but it's worth mentioning.

§82.3 Stalks and germs

Prototypical example for this section: Germs of real smooth functions tell you the derivatives, but germs of holomorphic functions determine the entire function.

As we mentioned, the helpful pictures from the previous section are still just metaphors, because there is no notion of "value". With the addition of the words "stalk" and "germ", we can actually change that.

Definition 82.3.1. Let $\mathscr{F}$ be a pre-sheaf (of rings). For every point p we define the **stalk** $\mathscr{F}_p$ to be the set

$$\{(s, U) \mid s \in \mathscr{F}(U), p \in U\}$$

modulo the equivalence relation $\sim$ that

$$(s_1, U_1) \sim (s_2, U_2) \quad \text{if} \quad s_1 \restriction_V = s_2 \restriction_V$$

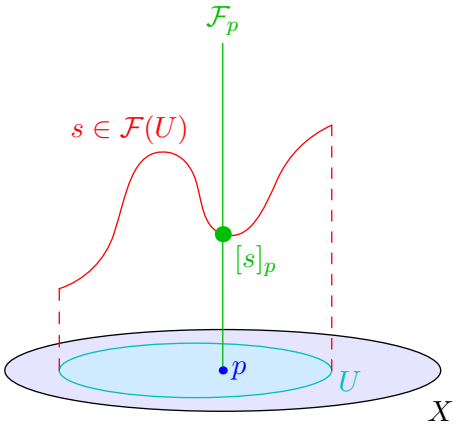
for some open set V with $V \ni p$ and $V \subseteq U_1 \cap U_2$. The equivalence classes themselves are called **germs**.

Definition 82.3.2. The germ of a given $s \in \mathscr{F}(U)$ at a point p is the equivalence class for $(s, U) \in \mathscr{F}_p$. We denote this by $[s]_p$.

It is rarely useful to think of a germ as an ordered pair, since the set U can get arbitrarily small. Instead, one should think of a germ as a "shred" of some section near p. A nice summary for the right mindset might be:

A germ is an "enriched value"; the stalk is the set of possible germs.

Let's add this to our picture from before. If we insist on continuing to draw our sections as xy-plots, then above each point p a good metaphor would be a vertical line out from p. The germ would then be the "enriched value of s at p". We just draw that as a big dot in our plot. The main difference is that the germ is enriched in the sense that the germ carries information in a small region around p as well, rather than literally just the point p itself. So accordingly we draw a large dot for $[s]_p$, rather than a small dot at p.



Before going on, we might as well note that the stalks are themselves rings, not just sets: we can certainly add or subtract enriched values.

Definition 82.3.3. The stalk $\mathscr{F}_p$ can itself be regarded as a ring: for example, addition is done by

$$(s_1, U_1) + (s_2, U_2) = (s_1 \restriction_{U_1 \cap U_2} + s_2 \restriction_{U_1 \cap U_2}, U_1 \cap U_2)\,.$$

Example 82.3.4 (Germs of real smooth functions)
Let $X = \mathbb{R}$ and let $\mathscr{F}$ be the pre-sheaf on X of smooth functions (i.e. $\mathscr{F}(U)$ is the set of smooth real-valued functions on U).
Consider a global section, $s\colon \mathbb{R} \to \mathbb{R}$ (thus $s \in \mathscr{F}(X)$) and its germ at 0.

(a) From the germ we can read off $s(0)$, obviously.

(b) We can also find $s'(0)$, because the germ carries enough information to compute the limit $\lim_{h\to 0} \frac{1}{h}[s(h) - s(0)]$.

(c) Similarly, we can compute the second derivative and so on.

(d) However, we can't read off, say, $s(3)$ from the germ. For example, consider the function from Example 29.4.4,

$$s(x) = \begin{cases} e^{-\frac{1}{x-1}} & x > 1 \\ 0 & x \le 1. \end{cases}$$

Note $s(3) = e^{-\frac{1}{2}}$, but $[\text{zero function}]_0 = [s]_0$. So germs can't distinguish between the zero function and s.

Example 82.3.5 (Germs of holomorphic functions)
Holomorphic functions are surprising in this respect. Consider the sheaf $\mathscr{F}$ on $\mathbb{C}$ of *holomorphic* functions.
Take $s\colon \mathbb{C} \to \mathbb{C}$ a global section. Given the germ of s at 0, we can read off $s(0)$, $s'(0)$, et cetera. The miracle of complex analysis is that just knowing the derivatives of s at zero is enough to reconstruct all of s: we can compute the Taylor series of s now. **Thus germs of holomorphic functions determine the entire function**; they "carry more information" than their real counterparts.
In particular, we can concretely describe the stalks of the pre-sheaf:

$$\mathscr{F}_p = \left\{ \sum_{k\ge 0} c_k (z-p)^k \text{ convergent near } p \right\}.$$

For example, this includes germs of meromorphic functions, so long as there is no pole at p itself.

And of course, our algebraic geometry example. This example will matter a lot later, so we do it carefully now.

Abuse of Notation 82.3.6. Rather than writing $(\mathcal{O}_X)_p$ we will write $\mathcal{O}_{X,p}$.

Theorem 82.3.7 (Stalks of $\mathcal{O}_V$)
Let $V \subseteq \mathbb{A}^n$ be a variety, and assume $p \in V$ is a point. Then

$$\mathcal{O}_{V,p} \cong \left\{ \frac{f}{g} \mid f, g \in \mathbb{C}[V],\ g(p) \neq 0 \right\}.$$

Proof. A regular function φ on $U \subseteq V$ is supposed to be a function on U that "locally" is a quotient of two functions in $\mathbb{C}[V]$. Since we are looking at the stalk near p, though, the germ only cares up to the choice of representation at p, and so we can go ahead and write

$$\mathcal{O}_{V,p} = \left\{ \left(\frac{f}{g}, U \right) \mid U \ni p,\ f,g \in \mathbb{C}[V],\ g \neq 0 \text{ on } U \right\}$$

modulo the same relation.

Now we claim that the map

$$\mathcal{O}_{V,p} \to \text{desired RHS} \qquad \text{by} \qquad \left(\frac{f}{g}, U \right) \mapsto \frac{f}{g}$$

is an isomorphism.

- Injectivity: We are working with complex polynomials, so we know that a rational function is determined by its behavior on any open neighborhood of p; thus two germ representatives $(\frac{f_1}{g_1}, U_1)$ and $(\frac{f_2}{g_2}, U_2)$ agree on $U_1 \cap U_2$ if and only if they are actually the same quotient.
- Surjectivity: take $U = D(g)$. □

Example 82.3.8 (Stalks of your favorite varieties at the origin)

(a) Let $V = \mathbb{A}^1$; then the stalk of $\mathcal{O}_V$ at each point $p \in V$ is

$$\mathcal{O}_{V,p} = \left\{ \frac{f(x)}{g(x)} \mid g(p) \neq 0 \right\}.$$

Examples of elements are $x^2 + 5$, $\frac{1}{x-1}$ if $p \neq 1$, $\frac{x+7}{x^2-9}$ if $p \neq \pm 3$, and so on.

(b) Let $V = \mathbb{A}^2$; then the stalk of $\mathcal{O}_V$ at the origin is

$$\mathcal{O}_{V,(0,0)} = \left\{ \frac{f(x,y)}{g(x,y)} \mid g(0,0) \neq 0 \right\}.$$

Examples of elements are $x^2 + y^2$, $\frac{x^3}{xy+1}$, $\frac{13x+37y}{x^2+8y+2}$.

(c) Let $V = \mathcal{V}(y - x^2) \subseteq \mathbb{A}^2$; then the stalk of $\mathcal{O}_V$ at the origin is

$$\mathcal{O}_{V,(0,0)} = \left\{ \frac{f(x,y)}{g(x,y)} \mid f,g \in \mathbb{C}[x,y]/(y-x^2),\ g(0,0) \neq 0 \right\}.$$

For example, $\frac{y}{1+x} = \frac{x^2}{1+x}$ denote the same element in the stalk. Actually, you could give a canonical choice of representative by replacing y with x^2 everywhere, so it would also be correct to write

$$\mathcal{O}_{V,(0,0)} = \left\{ \frac{f(x)}{g(x)} \mid\ g(0) \neq 0 \right\}$$

which is the same as the first example.

Remark 82.3.9 (Aside for category lovers) — You may notice that $\mathscr{F}_p$ seems to be "all the $\mathscr{F}_p(U)$ coming together", where $p \in U$. And in fact, $\mathscr{F}_p(U)$ is the

categorical *colimit* of the diagram formed by all the $\mathscr{F}(U)$ such that $p \in U$. This is often written

$$\mathscr{F}_p = \varinjlim_{U \ni p} \mathscr{F}(U)$$

Thus we can define stalks in any category with colimits, though to be able to talk about germs the category needs to be concrete.

§82.4 Sheaves

Prototypical example for this section: Constant functions aren't sheaves, but locally constant ones are.

Since we care so much about stalks, which study local behavior, we will impose additional local conditions on our pre-sheaves. One way to think about this is:

Sheaves are pre-sheaves for which P is a *local* property.

The formal definition doesn't illuminate this as much as the examples do, but sadly I have to give the definition first for the examples to make sense.

Definition 82.4.1. A **sheaf** $\mathscr{F}$ on a topological space X is a pre-sheaf obeying two additional axioms: Suppose U is an open set in X, and U is covered by open sets $U_\alpha \subseteq U$. Then:

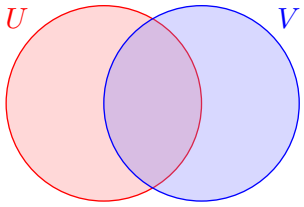
1. (Identity) If $s, t \in \mathscr{F}(U)$ are sections, and $s\restriction_{U_\alpha} = t\restriction_{U_\alpha}$ for all α, then $s = t$.
2. (Gluing) Consider sections $s_\alpha \in \mathscr{F}(U_\alpha)$ for each α. Suppose that

$$s_\alpha\restriction_{U_\alpha \cap U_\beta} = s_\beta\restriction_{U_\alpha \cap U_\beta}$$

   for each U_α and U_β. Then we can find $s \in \mathscr{F}(U)$ such that $s\restriction_{U_\alpha} = s_\alpha$.

Remark 82.4.2 (For keepers of the empty set) — The above axioms imply $\mathscr{F}(\varnothing) = 0$ (the zero ring), when $\mathscr{F}$ is a sheaf of rings. This is not worth worrying about until you actually need it, so you can forget I said that.

This is best illustrated by picture in the case of just two open sets: consider two open sets U and V. Then the sheaf axioms are saying something about $\mathscr{F}(U \cup V)$, $\mathscr{F}(U \cap V)$, $\mathscr{F}(U)$ and $\mathscr{F}(V)$.



Then for a sheaf of functions, the axioms are saying that:

- If s and t are functions (with property P) on $U \cup V$ and $s\restriction_U = t\restriction_U$, $s\restriction_V = t\restriction_V$, then $s = t$ on the entire union. This is clear.

- If s_1 is a function with property P on U and s_2 is a function with property P on V, and the two functions agree on the overlap, then one can glue them to obtain a function s on the whole space: this is obvious, but **the catch is that the collated function needs to have property P as well** (i.e. needs to be an element of $\mathscr{F}(U \cup V)$). That's why it matters that P is local.

So you can summarize both of these as saying: any two functions on U and V which agree on the overlap glue to a *unique* function on $U \cup V$. If you like category theory, you might remember we alluded to this in Example 69.2.1.

Exercise 82.4.3 (For the categorically inclined)**.** Show that the diagram

$$\begin{array}{ccc} \mathscr{F}(U \cup V) & \longrightarrow & \mathscr{F}(U) \\ \downarrow & & \downarrow \\ \mathscr{F}(V) & \longrightarrow & \mathscr{F}(U \cap V) \end{array}$$

is a pullback square.

Now for the examples.

Example 82.4.4 (Examples and non-examples of sheaves)

Note that every example of a stalk we computed in the previous section was of a sheaf. Here are more details:

(a) Pre-sheaves of arbitrary / continuous / differentiable / smooth / holomorphic functions are still sheaves. This is because to verify a function is continuous, one only needs to look at small open neighborhoods at once.

(b) Let $X = \mathbb{R}$, and define the presheaf of rings $\mathscr{F}$ by:

$$\mathscr{F}(U) = \{f \colon U \to \mathbb{R} \mid \text{there exists continuous } g \colon \mathbb{R} \to \mathbb{R} \text{ such that } g|_U = f\}.$$

Then $\mathscr{F}$ is not a sheaf. Indeed, $s_1(x) = 0$ in $\mathscr{F}((-1,0))$ and $s_2(x) = 1$ in $\mathscr{F}((0,1))$ agrees on the (empty) overlap, but they cannot be glued together to an element in $\mathscr{F}((-1,0) \cup (0,1))$.

(c) For a complex variety V, $\mathcal{O}_V$ is a sheaf, precisely because our definition was *locally* quotients of polynomials.

(d) The pre-sheaf of *constant* real functions on a space X is *not* a sheaf in general, because it fails the gluing axiom. Namely, suppose that $U_1 \cap U_2 = \varnothing$ are disjoint open sets of X. Then if s_1 is the constant function 1 on U_1 while s_2 is the constant function 2 on U_2, then we cannot glue these to a constant function on $U_1 \cup U_2$.

(e) On the other hand, *locally constant* functions do produce a sheaf. (A function is locally constant if for every point it is constant on some open neighborhood.)

In fact, the sheaf in e is what is called a *sheafification* of the pre-sheaf constant functions, which we define momentarily.

§82.5 For sheaves, sections "are" sequences of germs

Prototypical example for this section: A real function on U is a sequence of real numbers $f(p)$ for each $p \in U$ satisfying some local condition. Analogously, a section $s \in \mathscr{F}(U)$ is a sequence of germs satisfying some local compatibility condition.

Once we impose the sheaf axioms, our metaphorical picture will actually be more or less complete. Just as a function was supposed to be a choice of value at each point, a section will be a choice of germ at each stalk.

Example 82.5.1 (Real functions vs. germs)

Let X be a space and let $\mathscr{F}$ be the sheaf of smooth functions. Take a section $f \in \mathscr{F}(U)$.

- As a function, f is just a choice of value $f(p) \in \mathbb{R}$ at every point p, subject to a local "smooth" condition.
- Let's now think of f as a sequence of germs. At every point p the germ $[f]_p \in \mathscr{F}_p$ gives us the value $f(p)$ as we described above. The germ packages even more data than this: from the germ $[f]_p$ alone we can for example compute $f'(p)$. Nonetheless we stretch the analogy and think of f as a choice of germ $[f]_p \in \mathscr{F}_p$ at each point p.

Thus we can replace the notion of the value $f(p)$ with germ $[f]_p$. This is useful because in a general sheaf $\mathscr{F}$, the notion $s(p)$ is not defined while the notion $[s]_p$ is.

From the above example it's obvious that if we know each germ $[s]_p$, this should let us reconstruct the entire section s. Let's check this from the sheaf axioms:

Exercise 82.5.2 (Sections are determined by stalks)**.** Let $\mathscr{F}$ be a sheaf. Consider the natural map

$$\mathscr{F}(U) \to \prod_{p \in U} \mathscr{F}_p$$

described above. Show that this map is injective, i.e. the germs of s at every point $p \in U$ determine the section s. (You will need the "identity" sheaf axiom, but not "gluing".)

However, this map is clearly not surjective! Nonetheless we can describe the image: we want a sequence of germs $(g_p)_{p \in U}$ such that near every germ g_p, the germs g_q are "compatible" with g_p. We make this precise:

Definition 82.5.3. Let $\mathscr{F}$ be pre-sheaf and let U be an open set. A sequence $(g_p)_{p \in U}$ of germs (with $g_p \in \mathscr{F}_p$ for each p) is said to be **compatible** if they can be "locally collated":

> For any $p \in U$ there exists an open neighborhood $U_p \ni p$ and a section $s \in \mathscr{F}(U_p)$ on it such that $[s]_q = g_q$ for each $q \in U_p$.

Intuitively, the germs should "collate together" to some section near each *individual* point q (but not necessarily to a section on all of U).

We let the reader check this definition is what we want:

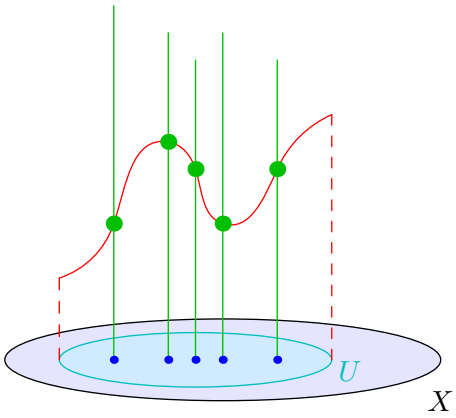
Exercise 82.5.4. Prove that any choice of compatible germs over U collates together to a section of U. (You will need the "gluing" sheaf axiom, but not "identity".)

Putting together the previous two exercise gives:

Theorem 82.5.5 (Sections "are" just compatible germs)
Let $\mathscr{F}$ be a sheaf. There is a natural bijection between

- sections of $\mathscr{F}(U)$, and
- sequences of compatible germs over U.

We draw this in a picture below by drawing several stalks, rather than just one, with the germs above. The stalks at different points need not be related to each other, so I have drawn the stalks to have different heights to signal this. And, the caveat is that the germs are large enough that germs which are "close to each other" should be "compatible".



This is in exact analogy to the way that e.g. a smooth real-valued function on U is a choice of real number $f(p) \in \mathbb{R}$ at each point $p \in U$ satisfying a local smoothness condition.

Thus the notion of stalks is what lets us recover the viewpoint that sections are "functions". Therefore for theoretical purposes,

With sheaf axioms, sections are sequences of compatible germs.

In particular, this makes restriction morphisms easy to deal with: just truncate the sequence of germs!

§82.6 Sheafification (optional)

Prototypical example for this section: The pre-sheaf of constant functions becomes the sheaf of locally constant functions.

The idea is that if $\mathscr{F}$ is the pre-sheaf of "functions with property P" then we want to associate a sheaf $\mathscr{F}^{\text{sh}}$ of "functions which are locally P", which makes them into a sheaf. We have already seen two examples of this:

Example 82.6.1 (Sheafification)

(a) If X is a topological space, and $\mathscr{F}$ is the pre-sheaf of constant functions on open sets of X, then $\mathscr{F}^{\text{sh}}$ is the sheaf of locally constant functions.

(b) If V is an affine variety, and $\mathscr{F}$ is the pre-sheaf of rational functions, then $\mathscr{F}^{\text{sh}}$ is the sheaf of regular functions (which are locally rational).

The procedure is based on stalks and germs. We saw that for a sheaf, sections correspond to sequences of compatible germs. For a pre-sheaf, we can still define stalks and germs, but their properties will be less nice. But given our initial pre-sheaf $\mathscr{F}$, we *define* the sections of $\mathscr{F}^{\text{sh}}$ to be sequences of compatible $\mathscr{F}$-germs.

Definition 82.6.2. The **sheafification** $\mathscr{F}^{\text{sh}}$ of a pre-sheaf $\mathscr{F}$ is defined by

$$\mathscr{F}^{\text{sh}}(U) = \{\text{sequences of compatible } \mathscr{F}\text{-germs } (g_p)_{p \in U}\}.$$

Question 82.6.3. Complete the definition by describing the restriction morphisms of $\mathscr{F}^{\text{sh}}$.

Abuse of Notation 82.6.4. I'll usually be equally sloppy in the future: when defining a sheaf $\mathscr{F}$, I'll only say what $\mathscr{F}(U)$ is, with the restriction morphisms $\mathscr{F}(U_2) \to \mathscr{F}(U_1)$ being implicit.

The construction is contrived so that given a section $(g_p)_{p \in U} \in \mathscr{F}^{\text{sh}}(U)$ the germ at a point p is g_p:

Lemma 82.6.5 (Stalks preserved by sheafification)

Let $\mathscr{F}$ be a pre-sheaf and $\mathscr{F}^{\text{sh}}$ its sheafification. Then for any point q, there is an isomorphism

$$(\mathscr{F}^{\text{sh}})_q \cong \mathscr{F}_q.$$

Proof. A germ in $(\mathscr{F}^{\text{sh}})_q$ looks like $((g_p)_{p \in U}, U)$, where $g_p = (s_p, U_p)$ are themselves germs of $\mathscr{F}_p$, and $q \in U$. Then the isomorphism is given by

$$((g_p)_{p \in U}, U) \mapsto g_q \in \mathscr{F}_q.$$

The inverse map is given by for each $g = (s, U) \in \mathscr{F}_q$ by

$$g \mapsto ((g)_{p \in U}, U) \in (\mathscr{F}^{\text{sh}})_q$$

i.e. the sequence of germs is the constant sequence. □

We will use sheafification in the future to economically construct sheaves. However, in practice, the details of the construction will often not matter.

§82.7 A few harder problems to think about

Problem 82A. Prove that if $\mathscr{F}$ is already a sheaf, then $\mathscr{F}(U) \cong \mathscr{F}^{\text{sh}}(U)$ for every open set U.

Problem 82B. Let X be a space with two points $\{p, q\}$ and let $\mathscr{F}$ be a sheaf on it. Suppose $\mathscr{F}_p = \mathbb{Z}/5\mathbb{Z}$ and $\mathscr{F}_q = \mathbb{Z}$. Describe $\mathscr{F}(U)$ for each open set U of X, where

(a) X is equipped with the discrete topology.

(b) X is equipped $\varnothing$, $\{p\}$, $\{p, q\}$ as the only open sets.

Problem 82C (Skyscraper sheaf)**.** Let Y be a topological space. Fix $p \in Y$ a point, and R a ring. The **skyscraper sheaf** is defined by

$$\mathscr{F}(U) = \begin{cases} R & p \in U \\ 0 & \text{otherwise} \end{cases}$$

with restriction maps in the obvious manner. Compute all the stalks of $\mathscr{F}$.

(Possible suggestion: first do the case where Y is Hausdorff, where your intuition will give the right answer. Then do the pathological case where every open set of Y contains p. Then try to work out the general answer.)

Problem 82D. Let $\mathscr{F}$ be a sheaf of rings on a space X and let $s \in \mathscr{F}(X)$ be a global section. Define the **support** of s as

$$Z = \{p \in X \mid [s]_p \neq 0 \in \mathscr{F}_p\}.$$

Show that Z is a closed set of X.

83 Localization

Before we proceed on to defining an affine scheme, we will take the time to properly cover one more algebraic construction that of a *localization*. This is mandatory because when we define a scheme, we will find that all the sections and stalks are actually obtained using this construction.

One silly slogan might be:

Localization is the art of adding denominators.

You may remember that when we were working with affine varieties, there were constantly expressions of the form $\left\{ \frac{f}{g} \mid g(p) \neq 0 \right\}$ and the like. The point is that we introduced a lot of denominators. Localization will give us a concise way of doing this in general.

This thus also explain why the operation is called "localization": we start from a set of "global" functions, and get a (larger) set of functions that are well-defined on a "smaller" open set, or in an open neighborhood of point p.

Of course, this is the Zariski topology, so "small" means "everywhere except certain curves".

Notational note: moving forward we'll prefer to denote rings by A, B, ..., rather than R, S, ...

§83.1 Spoilers

Here is a preview of things to come, so that you know what you are expecting. Some things here won't make sense, but that's okay, it is just foreshadowing.

Let $V \subseteq \mathbb{A}^n$, and for brevity let $R = \mathbb{C}[V]$ be its coordinate ring. We saw in previous sections how to compute $\mathcal{O}_V(D(g))$ and $\mathcal{O}_{V,p}$ for $p \in V$ a point. For example, if we take $\mathbb{A}^1$ and consider a point p, then $\mathcal{O}_{\mathbb{A}^1}(D(x-p)) = \left\{ \frac{f(x)}{(x-p)^n} \right\}$ and $\mathcal{O}_{\mathbb{A}^1,p} = \left\{ \frac{f(x)}{g(x)} \mid g(p) \neq 0 \right\}$. More generally, we had

$$\mathcal{O}_V(D(g)) = \left\{ \frac{f}{g^n} \mid f \in R \right\} \quad \text{by Theorem 78.6.1}$$
$$\mathcal{O}_{V,p} = \left\{ \frac{f}{g} \mid f, g \in R, g(p) \neq 0 \right\} \quad \text{by Theorem 82.3.7.}$$

We will soon define something called a localization, which will give us a nice way of expressing the above: if $R = \mathbb{C}[V]$ is the coordinate ring, then the above will become abbreviated to just

$$\mathcal{O}_V(D(g)) = R[g^{-1}]$$
$$\mathcal{O}_{V,p} = R_{\mathfrak{m}} \quad \text{where } \{p\} = \mathcal{V}(\mathfrak{m}).$$

The former will be pronounced "R localized away from g" while the latter will be pronounced "R localized at $\mathfrak{m}$".

Even more generally, next chapter we will throw out the coordinate ring R altogether and replace it with a general commutative ring A (which are still viewed as functions).

We will construct a ringed space called $X = \operatorname{Spec} A$, whose elements are *prime ideals* of A and is equipped with the Zariski topology and a sheaf $\mathcal{O}_X$. It will turn out that, the right way to define the sheaf $\mathcal{O}_X$ is to use localization,

$$\mathcal{O}_X(D(f)) = A[f^{-1}]$$
$$\mathcal{O}_{X,\mathfrak{p}} = A_{\mathfrak{p}}$$

for any element $f \in A$ and prime ideal $\mathfrak{p} \in \operatorname{Spec} A$. Thus just as with complex affine varieties, localizations will give us a way to more or less describe the sheaf $\mathcal{O}_X$ completely.

In other words,

Localization is the purely algebraic way to *define* the ring of regular functions on a smaller open set from the ring of "global" regular functions.

§83.2 The definition

Definition 83.2.1. A subset $S \subseteq A$ is a **multiplicative set** if $1 \in S$ and S is closed under multiplication.

Definition 83.2.2. Let A be a ring and $S \subset A$ a multiplicative set. Then the **localization of A at S**, denoted $S^{-1}A$, is defined as the set of fractions

$$\{a/s \mid a \in A, s \in S\}$$

where we declare two fractions $a_1/s_1 = a_2/s_2$ to be equal if $s(a_1s_2 - a_2s_1) = 0$ for some $s \in S$. Addition and multiplication in this ring are defined in the obvious way.

In particular, if $0 \in S$ then $S^{-1}A$ is the zero ring. So we usually only take situations where $0 \notin S$.

We give in brief now two examples which will be motivating forces for the construction of the affine scheme.

Example 83.2.3 (Localizations of $\mathbb{C}[x]$)
Let $A = \mathbb{C}[x]$.

(a) Suppose we let $S = \{1, x, x^2, x^3, \dots\}$ be the powers of x. Then

$$S^{-1}A = \left\{ \frac{f(x)}{x^n} \mid f \in \mathbb{C}[x], n \in \mathbb{Z}_{\geq 0} \right\}.$$

In other words, we get the Laurent polynomials in x.

You might recognize this as

$$\mathcal{O}_V(U) \text{ where } V = \mathbb{A}^1,\ U = V \setminus \{0\}.$$

i.e. the sections of the punctured line. In line with the "hyperbola effect", this is also expressible as $\mathbb{C}[x,y]/(xy-1)$.

(b) Let $p \in \mathbb{C}$. Suppose we let $S = \{g(x) \mid g(p) \neq 0\}$, i.e. we allow any denominators where $g(p) \neq 0$. Then

$$S^{-1}A = \left\{ \frac{f(x)}{g(x)} \mid f, g \in \mathbb{C}[x], g(p) \neq 0 \right\}.$$

You might recognize this is as the stalk $\mathcal{O}_{\mathbb{A}^1,p}$. This will be important later on.

Remark 83.2.4 (Why the extra s?) — We cannot use the simpler $a_1s_2 - a_2s_1 = 0$ since otherwise the equivalence relation may fail to be transitive. Here is a counterexample: take

$$A = \mathbb{Z}/12\mathbb{Z} \qquad S = \{2, 4, 8\}.$$

Then we have for example $\frac{0}{1} = \frac{0}{2} = \frac{12}{2} = \frac{6}{1}$. So we need to have $\frac{0}{1} = \frac{6}{1}$ which is only true with the first definition. Of course, if A is an integral domain (and $0 \notin S$) then this is a moot point.
Alternatively, one can start with this simpler relation, and take the transitive closure; this yields an equivalent definition.

Example 83.2.5 (Field of fractions)
Let A be an integral domain and $S = A \setminus \{0\}$. Then $S^{-1}A = \operatorname{Frac}(A)$.

§83.3 Localization away from an element

Prototypical example for this section: $\mathbb{Z}$ *localized away from* 6 *has fractions* $\frac{m}{2^x3^y}$.

We now focus on the two special cases of localization we will need the most; one in this section, the other in the next section.

Definition 83.3.1. For $f \in A$, we define the **localization of A away from f**, denoted $A[1/f]$ or $A[f^{-1}]$, to be $\{1, f, f^2, f^3, \dots\}^{-1}A$. (Note that $\{1, f, f^2, \dots\}$ is multiplicative.)

Remark 83.3.2 — In the literature it is more common to see the notation A_f instead of $A[1/f]$. This is confusing, because in the next section we define $A_{\mathfrak{p}}$ which is almost the opposite. So I prefer this more suggestive (but longer) notation.

Example 83.3.3 (Some arithmetic examples of localizations)
(a) We localize $\mathbb{Z}$ away from 6:

$$\mathbb{Z}[1/6] = \left\{ \frac{m}{6^n} \mid m \in \mathbb{Z}, n \in \mathbb{Z}_{\geq 0} \right\}.$$

So $A[1/6]$ consist of those rational numbers whose denominators have only powers of 2 and 3. For example, it contains $\frac{5}{12} = \frac{15}{36}$.

(b) Here is a more confusing example: if we localize $\mathbb{Z}/60\mathbb{Z}$ away from the element 5, we get $(\mathbb{Z}/60\mathbb{Z})[1/5] \cong \mathbb{Z}/12\mathbb{Z}$. You should try to think about why this is the case. We will see a "geometric" reason later, in Section 86.9.

Example 83.3.4 (Localization at an element, algebraic geometry flavored)
We saw that if A is the coordinate ring of a variety, then $A[1/g]$ is interpreted geometrically as $\mathcal{O}_V(D(g))$. Here are some special cases:

(a) As we saw, if $A = \mathbb{C}[x]$, then $A[1/x] = \left\{ \frac{f(x)}{x^n} \right\}$ consists of Laurent polynomials.

(b) Let $A = \mathbb{C}[x,y,z]$. Then

$$A[1/x] = \left\{ \frac{f(x,y,z)}{x^n} \mid f \in \mathbb{C}[x,y,z],\ n \geq 0 \right\}$$

is rational functions whose denominators are powers of x.

(c) Let $A = \mathbb{C}[x,y]$. If we localize away from $y - x^2$ we get

$$A[(y-x^2)^{-1}] = \left\{ \frac{f(x,y)}{(y-x^2)^n} \mid f \in \mathbb{C}[x,y],\ n \geq 0 \right\}$$

By now you should recognize this as $\mathcal{O}_{\mathbb{A}^2}(D(y-x^2))$.

Example 83.3.5 (An example with zero-divisors)
Let $A = \mathbb{C}[x,y]/(xy)$ (which intuitively is the coordinate ring of two axes). Suppose we localize at x: equivalently, allowing denominators of x. Since $xy = 0$ in A, we now have $0 = x^{-1}(xy) = y$, so $y = 0$ in A, and thus y just goes away completely. From this we get a ring isomorphism

$$A[1/x] \cong \mathbb{C}[x, 1/x].$$

Later, we will be able to use our geometric intuition to "see" this at Section 86.14, once we have defined the affine scheme.

§83.4 Localization at a prime ideal

Prototypical example for this section: $\mathbb{Z}$ *localized at* (5) *has fractions* $\frac{m}{n}$ *with* $5 \nmid n$.

Definition 83.4.1. If A is a ring and $\mathfrak{p}$ is a prime ideal, then we define

$$A_{\mathfrak{p}} \coloneqq (A \setminus \mathfrak{p})^{-1} A.$$

This is called the **localization at $\mathfrak{p}$**.

Question 83.4.2. Why is $S = A \setminus \mathfrak{p}$ multiplicative in the above definition?

This special case is important because we will see that stalks of schemes will all be of this shape. In fact, the same was true for affine varieties too.

Example 83.4.3 (Relation to affine varieties)
Let $V \subseteq \mathbb{A}^n$, let $A = \mathbb{C}[V]$ and let $p = (a_1, \dots, a_n)$ be a point. Consider the maximal (hence prime) ideal

$$\mathfrak{m} = (x_1 - a_1, \dots, x_n - a_n).$$

Observe that a function $f \in A$ vanishes at p if and only if $f \pmod{\mathfrak{m}} = 0$,

equivalently $f \in \mathfrak{m}$. Thus, by Theorem 82.3.7 we can write

$$\begin{aligned}\mathcal{O}_{V,p} &= \left\{ \frac{f}{g} \mid f, g \in A, g(p) \neq 0 \right\} \\ &= \left\{ \frac{f}{g} \mid f \in A, g \in A \setminus \mathfrak{m} \right\} \\ &= (A \setminus \mathfrak{m})^{-1} A = A_{\mathfrak{m}}.\end{aligned}$$

So, we can also express $\mathcal{O}_{V,p}$ concisely as a localization.

Consequently, we give several examples in this vein.

Example 83.4.4 (Geometric examples of localizing at a prime)

(a) We let $\mathfrak{m}$ be the maximal ideal (x) of $A = \mathbb{C}[x]$. Then

$$A_{\mathfrak{m}} = \left\{ \frac{f(x)}{g(x)} \mid g(0) \neq 0 \right\}$$

consists of the Laurent series.

(b) We let $\mathfrak{m}$ be the maximal ideal (x, y) of $A = \mathbb{C}[x, y]$. Then

$$A_{\mathfrak{m}} = \left\{ \frac{f(x,y)}{g(x,y)} \mid g(0,0) \neq 0 \right\}.$$

(c) Let $\mathfrak{p}$ be the prime ideal $(y - x^2)$ of $A = \mathbb{C}[x, y]$. Then

$$A_{\mathfrak{p}} = \left\{ \frac{f(x,y)}{g(x,y)} \mid g \notin (y - x^2) \right\}.$$

This is a bit different from what we've seen before: the polynomials in the denominator are allowed to vanish at a point like $(1, 1)$, as long as they don't vanish on *every* point on the parabola. This doesn't correspond to any stalk we're familiar with right now, but it will later (it will be the "stalk at the generic point of the parabola").

(d) Let $A = \mathbb{C}[x]$ and localize at the prime ideal (0). This gives

$$A_{(0)} = \left\{ \frac{f(x)}{g(x)} \mid g(x) \neq 0 \right\}.$$

This is all rational functions, period.

Remark 83.4.5 (Notational philosophy) To reiterate:

- when localizing away from an element, you allow the functions to blow up at (the vanishing set of) that element;
- when localizing at a prime, you allow the functions to blow up everywhere *except* at (the whole vanishing set of) that prime.

Thus we see why we say the 2 notations are opposites.
Thinking of functions that "may not blow up at the whole vanishing set" can be confusing, so another (hopefully) more intuitive way to think about localizing at a prime is that the function must not blow up at the point corresponding to the prime ideal. For example, if $\mathfrak{p}$ is the ideal $(y-x^2)$ in $A = \mathbb{C}[x,y]$, then $A_{\mathfrak{p}}$ is the set of functions that do not blow up at the generic point on the parabola.

Example 83.4.6 (Arithmetic examples)
We localize $\mathbb{Z}$ at a few different primes.

(a) If we localize $\mathbb{Z}$ at (0):

$$\mathbb{Z}_{(0)} = \left\{ \frac{m}{n} \mid n \neq 0 \right\} \cong \mathbb{Q}.$$

(b) If we localize $\mathbb{Z}$ at (3), we get

$$\mathbb{Z}_{(3)} = \left\{ \frac{m}{n} \mid \gcd(n,3) = 1 \right\}$$

which is the ring of rational numbers whose denominators are relatively prime to 3.

Example 83.4.7 (Field of fractions)
If A is an integral domain, the localization $A_{(0)}$ is the field of fractions of A.

§83.5 Prime ideals of localizations

Prototypical example for this section: The examples with $A = \mathbb{Z}$.

We take the time now to mention how you can think about prime ideals of localized rings.

Proposition 83.5.1 (The prime ideals of $S^{-1}A$)
Let A be a ring and $S \subseteq A$ a multiplicative set. Then there is a natural inclusion-preserving bijection between:

- The set of prime ideals of $S^{-1}A$, and
- The set of prime ideals of A not intersecting S.

Proof. Consider the homomorphism $\iota \colon A \to S^{-1}A$. For any prime ideal $\mathfrak{q} \subseteq S^{-1}A$, its pre-image $\iota^{\text{pre}}(\mathfrak{q})$ is a prime ideal of A (by Problem 5C*). Conversely, for any prime ideal $\mathfrak{p} \subseteq A$ not meeting S, $S^{-1}\mathfrak{p} = \left\{ \frac{a}{s} \mid a \in \mathfrak{p}, s \in S \right\}$ is a prime ideal of $S^{-1}A$. An annoying check shows that this produces the required bijection. □

In practice, we will almost always use the corollary where S is one of the two special cases we discussed at length:

Corollary 83.5.2 (Spectrums of localizations)
Let A be a ring.

(a) If f is an element of A, then the prime ideals of $A[1/f]$ are naturally in bijection with prime ideals of A **do not contain the element** f.

(b) If $\mathfrak{p}$ is a prime ideal of A, then the prime ideals of $A_{\mathfrak{p}}$ are naturally in bijection with prime ideals of A which are **subsets of** $\mathfrak{p}$.

Proof. Part (b) is immediate; a prime ideal doesn't meet $A \setminus \mathfrak{p}$ exactly if it is contained in $\mathfrak{p}$. For part (a), we want prime ideals of A not containing any *power* of f. But if the ideal is prime and contains f^n, then it should contain either f or f^{n-1}, and so at least for prime ideals these are equivalent. □

Notice again how the notation is a bit of a nuisance. Anyways, here are some examples, to help cement the picture.

Example 83.5.3 (Prime ideals of $\mathbb{Z}[1/6]$)
Suppose we localize $\mathbb{Z}$ away from the element 6, i.e. consider $\mathbb{Z}[1/6]$. As we saw,

$$\mathbb{Z}[1/6] = \left\{ \frac{n}{2^x 3^y} \mid n \in \mathbb{Z}, x, y \in \mathbb{Z}_{\geq 0} \right\}.$$

consist of those rational numbers whose denominators have only powers of 2 and 3. Note that $(5) \subset \mathbb{Z}[1/6]$ is a prime ideal: those elements of $\mathbb{Z}[1/6]$ with 5 dividing the numerator. Similarly, (7), (11), (13), ... and even (0) give prime ideals of $\mathbb{Z}[1/6]$. But (2) and (3) no longer correspond to prime ideals; in fact in $\mathbb{Z}[1/6]$ we have $(2) = (3) = (1)$, the whole ring.

Example 83.5.4 (Prime ideals of $\mathbb{Z}_{(5)}$)
Suppose we localize $\mathbb{Z}$ at the prime (5). As we saw,

$$\mathbb{Z}_{(5)} = \left\{ \frac{m}{n} \mid m, n \in \mathbb{Z}, 5 \nmid n \right\}.$$

consist of those rational numbers whose denominators are not divisible by 5. This is an integral domain, so (0) is still a prime ideal. There is one other prime ideal: (5), i.e. those elements whose numerators are divisible by 5.
There are no other prime ideals: if $p \neq 5$ is a rational prime, then $(p) = (1)$, the whole ring, again.

§83.6 Prime ideals of quotients

While we are here, we mention that the prime ideals of quotients A/I can be interpreted in terms of those of A (as in the previous section for localization). You may remember this from Problem 4D* a long time ago, if you did that problem; but for our purposes we actually only care about the prime ideals.

Proposition 83.6.1 (The prime ideals of A/I)

If A is a ring and I is any ideal (not necessarily prime) then the prime (resp. maximal) ideals of A/I are in bijection with prime (resp. maximal) ideals of A which are **supersets of** I. This bijection is inclusion-preserving.

Proof. Consider the quotient homomorphism $\psi\colon A \twoheadrightarrow A/I$. For any prime ideal $\mathfrak{q} \subseteq A/I$, its pre-image $\psi^{\text{pre}}(\mathfrak{q})$ is a prime ideal (by Problem 5C*). Conversely, for any prime ideal $\mathfrak{p}$ with $I \subseteq \mathfrak{p} \subseteq A$, we get a prime ideal of A/I by looking at $\mathfrak{p} \pmod I$. An annoying check shows that this produces the required bijection. It is also inclusion-preserving — from which the same statement holds for maximal ideals. □

Example 83.6.2 (Prime ideals of $\mathbb{Z}/60\mathbb{Z}$)

The ring $\mathbb{Z}/60\mathbb{Z}$ has three prime ideals:

$$\begin{aligned}(2) &= \{0, 2, 4, \dots, 58\} \\ (3) &= \{0, 3, 6, \dots, 57\} \\ (5) &= \{0, 5, 10, \dots, 55\}.\end{aligned}$$

Back in $\mathbb{Z}$, these correspond to the three prime ideals which are supersets of $60\mathbb{Z} = \{\dots, -60, 0, 60, 120, \dots\}$.

§83.7 Localization commutes with quotients

Prototypical example for this section: $(\mathbb{C}[x,y]/(xy))[1/x] \cong \mathbb{C}[x, x^{-1}]$.

While we are here, we mention a useful result from commutative algebra which lets us compute localizations in quotient rings, which are surprisingly unintuitive. You will *not* have a reason to care about this until we reach Section 85.4.ii, and so this is only placed earlier to emphasize that it's a purely algebraic fact that we can (and do) state this early, even though we will not need it anytime soon.

Let's say we have a quotient ring like

$$A/I = \mathbb{C}[x,y]/(xy)$$

and want to compute the localization of this ring away from the element x. (To be pedantic, we are actually localizing away from $x \pmod{xy}$, the element of the quotient ring, but we will just call it x.) You will quickly find that even the notation becomes clumsy: it is

$$(\mathbb{C}[x,y]/(xy))\,[1/x] \tag{83.1}$$

which is hard to think about, because the elements in play are part of the *quotient*: how are we supposed to think about

$$\frac{1 \pmod{xy}}{x \pmod{xy}}$$

for example? The zero-divisors in play may already make you feel uneasy.

However, it turns out that we can actually do the localization *first*, meaning the answer is just

$$\mathbb{C}[x, y, 1/x]/(xy) \tag{83.2}$$

which then becomes $\mathbb{C}[x, x^{-1}, y]/(y) \cong \mathbb{C}[x, x^{-1}]$.

This might look like it should be trivial, but it's not as obvious as you might expect. There is a sleight of hand present here with the notation:

- In (83.1), the notation (xy) stands for an ideal of $\mathbb{C}[x,y]$ — that is, the set $xy\mathbb{C}[x,y]$.
- In (83.2) the notation (xy) now stands for an ideal of $\mathbb{C}[x,x^{-1},y]$ — that is, the set $xy\mathbb{C}[x,x^{-1},y]$.

So even writing down the *statement* of the theorem is actually going to look terrible.

In general, what we want to say is that if we have our ring A with ideal I and S is some multiplicative subset of A, then

$$\text{Colloquially: } \text{“}S^{-1}(A/I) = (S^{-1}A)/I\text{”}.$$

But there are two things wrong with this:

- The main one is that I is not an ideal of $S^{-1}A$, as we saw above. This is remedied by instead using $S^{-1}I$, which consists of those elements of those elements $\frac{x}{s}$ for $x \in I$ and $s \in S$. As we saw this distinction is usually masked in practice, because we will usually write $I = (a_1, \dots, a_n) \subseteq A$ in which case the new ideal $S^{-1}I \subseteq A$ can be denoted in exactly the same way: $(a_1, \dots, a_n)$, just regarded as a subset of $S^{-1}A$ now.
- The second is that S is not, strictly speaking, a subset of A/I, either. But this is easily remedied by instead using the image of S under the quotient map $A \twoheadrightarrow A/I$.

  We actually already saw this in the previous example: when trying to localize $\mathbb{C}[x,y]/(xy)$, we were really localizing at the element $x \pmod{xy}$, but (as always) we just denoted it by x anyways.

And so after all those words, words, words, we have the hideous:

Theorem 83.7.1 (Localization commutes with quotients)

Let S be a multiplicative set of a ring A, and I an ideal of A. Let $\overline{S}$ be the image of S under the projection map $A \twoheadrightarrow A/I$. Then

$$\overline{S}^{-1}(A/I) \cong S^{-1}A/S^{-1}I$$

where $S^{-1}I = \left\{\frac{x}{s} \mid x \in I, s \in S\right\}$.

Proof. Omitted; Atiyah-Macdonald is the right reference for these type of things in the event that you do care. □

The notation is a hot mess. But when we do calculations in practice, we instead write

$$\left(\mathbb{C}[x,y,z]/(x^2+y^2-z^2)\right)[1/x] \cong \mathbb{C}[x,y,z,1/x]/(x^2+y^2-z^2)$$

or (for an example where we localize at a prime ideal)

$$\left(\mathbb{Z}[x,y,z]/(x^2+yz)\right)_{(x,y)} \cong \mathbb{Z}[x,y,z]_{(x,y)}/(x^2+yz)$$

and so on — the pragmatism of our "real-life" notation which hides some details actually guides our intuition (rather than misleading us). So maybe the moral of this section is that whenever you compute the localization of the quotient ring, if you just suspend belief for a bit, then you will probably get the right answer.

We will later see geometric interpretations of these facts when we work with $\operatorname{Spec} A/I$, at which point they will become more natural.

§83.8 A few harder problems to think about

Problem 83A. Let $A = \mathbb{Z}/2016\mathbb{Z}$, and consider the element $60 \in A$. Compute $A[1/60]$, the localization of A away from 60.

Problem 83B (Injectivity of localizations)**.** Let A be a ring and $S \subseteq A$ a multiplicative set. Find necessary and sufficient conditions for the map $A \to S^{-1}A$ to be injective.

Problem 83C* (Alluding to local rings)**.** Let A be a ring, and $\mathfrak{p}$ a prime ideal. How many maximal ideals does $A_{\mathfrak{p}}$ have?

Problem 83D. Let A be a ring such that $A_{\mathfrak{p}}$ is an integral domain for every prime ideal $\mathfrak{p}$ of A. Must A be an integral domain?

84 Affine schemes: the Zariski topology

Now that we understand sheaves well, we can define an affine scheme. It will be a ringed space, so we need to define

- The set of points,
- The topology on it, and
- The structure sheaf on it.

In this chapter, we handle the first two parts; Chapter 85 does the last one.

Quick note: Chapter 86 contains a long list of examples of affine schemes. So if something written in this chapter is not making sense, one thing worth trying is skimming through Chapter 86 to see if any of the examples there are more helpful.

§84.1 Some more advertising

Let me describe what the construction of $\operatorname{Spec} A$ is going to do.

In the case of $\mathbb{A}^n$, we used $\mathbb{C}^n$ as the set of points and $\mathbb{C}[x_1, \dots, x_n]$ as the ring of functions but then remarked that the set of points of $\mathbb{C}^n$ corresponded to the maximal ideals of $\mathbb{C}[x_1, \dots, x_n]$. In an *affine scheme*, we will take an *arbitrary* ring A, and generate the entire structure from just A itself. The final result is called $\operatorname{Spec} A$, the **spectrum** of A. The affine varieties $\mathcal{V}(I)$ we met earlier will just be $\operatorname{Spec} \mathbb{C}[\mathcal{V}(I)] = \operatorname{Spec} \mathbb{C}[x_1, \dots, x_n]/I$, but now we will be able to take *any* ideal I, thus finally completing the table at the end of the "affine variety" chapter.

To emphasize the point:

For affine varieties V, the spectrum of the coordinate ring $\mathbb{C}[V]$ is V.

Thus, we may also think of Spec as the opposite operation of taking the ring of global sections, defined purely-algebraically in order to depend only on the intrinsic properties of the affine variety itself (the ring $\mathcal{O}_V$) and not the embedding.

The construction of the affine scheme in this way will have three big generalizations:

1. We no longer have to work over an algebraically closed field $\mathbb{C}$, or even a field at all. This will be the most painless generalization: you won't have to adjust your current picture much for this to work.
2. We allow non-radical ideals: $\operatorname{Spec} \mathbb{C}[x]/(x^2)$ will be the double point we sought for so long. This will let us formalize the notion of a "fat" or "fuzzy" point.
3. Our affine schemes will have so-called *non-closed points*: points which you can visualize as floating around, somewhere in the space but nowhere in particular. (They'll correspond to prime non-maximal ideals.) These will take the longest to get used to, but as we progress we will begin to see that these non-closed points actually make life *easier*, once you get a sense of what they look like.

§84.2 The set of points

Prototypical example for this section: $\operatorname{Spec} \mathbb{C}[x_1, \dots, x_n]/I$.

First surprise, for a ring A:

Definition 84.2.1. The set $\operatorname{Spec} A$ is defined as the set of prime ideals of A.

This might be a little surprising, since we might have guessed that $\operatorname{Spec} A$ should just have the maximal ideals. What do the remaining ideals correspond to? The answer is that they will be so-called *non-closed points* or *generic points* which are "somewhere" in the space, but nowhere in particular. (The name "non-closed" is explained next chapter.)

Remark 84.2.2 — As usual A itself is not a prime ideal, but (0) is prime if and only if A is an integral domain.

Example 84.2.3 (Examples of spectrums)

(a) $\operatorname{Spec} \mathbb{C}[x]$ consists of a point $(x-a)$ for every $a \in \mathbb{C}$, which correspond to what we geometrically think of as $\mathbb{A}^1$. It additionally consists of a point (0), which we think of as a "non-closed point", nowhere in particular.

(b) $\operatorname{Spec} \mathbb{C}[x,y]$ consists of points $(x-a, y-b)$ (which are the maximal ideals) as well as (0) again, a non-closed point that is thought of as "somewhere in $\mathbb{C}^2$, but nowhere in particular". It also consists of non-closed points corresponding to irreducible polynomials $f(x,y)$, for example $(y-x^2)$, which is a "generic point on the parabola".

(c) If k is a field, $\operatorname{Spec} k$ is a single point, since the only maximal ideal of k is (0).

Example 84.2.4 (Complex affine varieties)

Let $I \subseteq \mathbb{C}[x_1, \dots, x_n]$ be an ideal. By Proposition 83.6.1, the set

$$\operatorname{Spec} \mathbb{C}[x_1, \dots, x_n]/I$$

consists of those prime ideals of $\mathbb{C}[x_1, \dots, x_n]$ which contain I: in other words, it has a point for every closed irreducible subvariety of $\mathcal{V}(I)$. So in addition to the "geometric points" (corresponding to the maximal ideals $(x_1 - a_1, \dots, x_n - a_n)$ we have non-closed points along each of the varieties).

The non-closed points are the ones you are not used to: there is one for each non-maximal prime ideal (visualized as "irreducible subvariety"). I like to visualize them in my head like a fly: you can hear it, so you know it is floating *somewhere* in the room, but as it always moving, you never know exactly where. So the generic point of $\operatorname{Spec} \mathbb{C}[x,y]$ corresponding to the prime ideal (0) is floating everywhere in the plane, the one for the ideal $(y-x^2)$ floats along the parabola, etc.

Image from [Wa].

Remark 84.2.5 (Why don't the prime non-maximal ideals correspond to the whole parabola?) — We have already seen a geometric reason in Section 83.4 earlier: localizing a ring at a prime non-maximal ideal gives the functions that may blow up somewhere in the parabola, but not *generically*.

Example 84.2.6 (More examples of spectrums)

(a) $\operatorname{Spec}\mathbb{Z}$ consists of a point for every prime p, plus a generic point that is somewhere, but no where in particular.

(b) $\operatorname{Spec}\mathbb{C}[x]/(x^2)$ has only (x) as a prime ideal. The ideal (0) is not prime since $0 = x \cdot x$. Thus as a *topological space*, $\operatorname{Spec}\mathbb{C}[x]/(x^2)$ is a single point.

(c) $\operatorname{Spec}\mathbb{Z}/60\mathbb{Z}$ consists of three points. What are they?

§84.3 The Zariski topology on the spectrum

Prototypical example for this section: Still $\operatorname{Spec}\mathbb{C}[x_1, \dots, x_n]/I$.

Now, we endow a topology on $\operatorname{Spec} A$. Since the points on $\operatorname{Spec} A$ are the prime ideals, we continue the analogy by thinking of the points f as functions on $\operatorname{Spec} A$. That is:

Definition 84.3.1. Let $f \in A$ and $\mathfrak{p} \in \operatorname{Spec} A$. Then the **value** of f at $\mathfrak{p}$ is defined to be $f \pmod{\mathfrak{p}}$, an element of $A/\mathfrak{p}$. We denote it $f(\mathfrak{p})$.

Example 84.3.2 (Vanishing locii in $\mathbb{A}^n$)

Suppose $A = \mathbb{C}[x_1, \dots, x_n]$, and $\mathfrak{m} = (x_1 - a_1, x_2 - a_2, \dots, x_n - a_n)$ is a maximal ideal of A. Then for a polynomial $f \in \mathbb{C}$,

$$f \pmod{\mathfrak{m}} = f(a_1, \dots, a_n)$$

with the identification that $A/\mathfrak{m} \cong \mathbb{C}$.

Example 84.3.3 (Functions on $\operatorname{Spec}\mathbb{Z}$)

Consider $A = \operatorname{Spec}\mathbb{Z}$. Then 2019 is a function on A. Its value at the point (5) is $4 \pmod 5$; its value at the point (7) is $3 \pmod 7$.

Indeed if you replace A with $\mathbb{C}[x_1, \dots, x_n]$ and $\operatorname{Spec} A$ with $\mathbb{A}^n$ in everything that follows, then everything will become quite familiar.

Definition 84.3.4. Let $f \in A$. We define the **vanishing locus** of f to be

$$\mathcal{V}(f) = \{\mathfrak{p} \in \operatorname{Spec} A \mid f(\mathfrak{p}) = 0\} = \{\mathfrak{p} \in \operatorname{Spec} A \mid f \in \mathfrak{p}\}.$$

More generally, just as in the affine case, we define the vanishing locus for an ideal I as

$$\begin{aligned}\mathcal{V}(I) &= \{\mathfrak{p} \in \operatorname{Spec} A \mid f(\mathfrak{p}) = 0 \ \forall f \in I\} \\ &= \{\mathfrak{p} \in \operatorname{Spec} A \mid f \in \mathfrak{p} \ \forall f \in I\} \\ &= \{\mathfrak{p} \in \operatorname{Spec} A \mid I \subseteq \mathfrak{p}\}.\end{aligned}$$

Finally, we define the **Zariski topology** on $\operatorname{Spec} A$ by declaring that the sets of the form $\mathcal{V}(I)$ are closed.

We now define a few useful topological notions:

Definition 84.3.5. Let X be a topological space. A point $p \in X$ is a **closed point** if the set $\{p\}$ is closed.

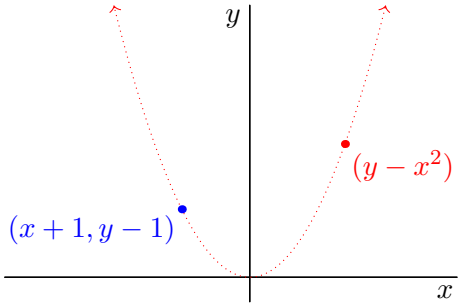
> **Question 84.3.6** (Mandatory)**.** Show that a point (i.e. prime ideal) $\mathfrak{m} \in \operatorname{Spec} A$ is a closed point if and only if $\mathfrak{m}$ is a maximal ideal.

Recall also in Definition 7.2.4 we denote by $\overline{S}$ the closure of a set S (i.e. the smallest closed set containing S); so you can think of a closed point p also as one whose closure is just $\{p\}$. Therefore the Zariski topology lets us refer back to the old "geometric" as just the closed points.

Example 84.3.7 (Non-closed points, continued)
Let $A = \mathbb{C}[x, y]$ and let $\mathfrak{p} = (y - x^2) \in \operatorname{Spec} A$; this is the "generic point" on a parabola. It is not closed, but we can compute its closure:

$$\overline{\{\mathfrak{p}\}} = \mathcal{V}(\mathfrak{p}) = \{\mathfrak{q} \in \operatorname{Spec} A \mid \mathfrak{q} \supseteq \mathfrak{p}\}.$$

This closure contains the point $\mathfrak{p}$ as well as several maximal ideals $\mathfrak{q}$, such as $(x - 2, y - 4)$ and $(x - 3, y - 9)$. In other words, the closure of the "generic point" of the parabola is literally the set of all points that are actually on the parabola (including generic points).
That means the way to picture $\mathfrak{p}$ is a point that is floating "somewhere on the parabola", but nowhere in particular. It makes sense then that if we take the closure, we get the entire parabola, since $\mathfrak{p}$ "could have been" any of those points.



Example 84.3.8 (The generic point of the y-axis isn't on the x-axis)
Let $A = \mathbb{C}[x,y]$ again. Consider $\mathcal{V}(y)$, which is the x-axis of $\operatorname{Spec} A$. Then consider $\mathfrak{p} = (x)$, which is the generic point on the y-axis. Observe that

$$\mathfrak{p} \notin \mathcal{V}(y).$$

The geometric way of saying this is that a *generic point* on the y-axis does not lie on the x-axis.

We now also introduce one more word:

Definition 84.3.9. A topological space X is **irreducible** if either of the following two conditions hold:

- The space X cannot be written as the union of two proper closed subsets.
- Any two nonempty open sets of X intersect.

A subset Z of X (usually closed) is irreducible if it is irreducible as a subspace.

Exercise 84.3.10. Show that the two conditions above are indeed equivalent. Also, show that the closure of a point is always irreducible.

This is the analog of the "irreducible" we defined for affine varieties, but it is now a topological definition, although in practice this definition is only useful for spaces with the Zariski topology. Indeed, if any two nonempty open sets intersect (and there is more than one point), the space is certainly not Hausdorff! As with our old affine varieties, the intuition is that $\mathcal{V}(xy)$ (the union of two lines) should not be irreducible.

Example 84.3.11 (Reducible and irreducible spaces)
(a) The closed set $\mathcal{V}(xy) = \mathcal{V}(x) \cup \mathcal{V}(y)$ is reducible.

(b) The entire plane $\operatorname{Spec} \mathbb{C}[x,y]$ is irreducible. There is actually a simple (but counter-intuitive, since you are just getting used to generic points) reason why this is true: the generic point (0) is in *every* open set, ergo, any two open sets intersect.

So actually, the generic points kind of let us cheat our way through the following bit:

Proposition 84.3.12 (Spectrums of integral domains are irreducible)
If A is an integral domain, then $\operatorname{Spec} A$ is irreducible.

Proof. Just note (0) is a prime ideal, and is in every open set. □

You should compare this with our old classical result that $\mathbb{C}[x_1, \dots, x_n]/I$ was irreducible as an affine variety exactly when I was prime. This time, the generic point actually takes care of the work for us: the fact that it is *allowed* to float anywhere in the plane lets us capture the idea that $\mathbb{A}^2$ should be irreducible without having to expend any additional effort.

Remark 84.3.13 — Surprisingly, the converse of this proposition is false: we have seen $\operatorname{Spec} \mathbb{C}[x]/(x^2)$ has only one point, so is certainly irreducible. But $A = \mathbb{C}[x]/(x^2)$ is not an integral domain. So this is one weird-ness introduced by allowing "non-radical" behavior.

At this point you might notice something:

Theorem 84.3.14 (Points are in bijection with irreducible closed sets)
Consider $X = \operatorname{Spec} A$. For every irreducible closed set Z, there is exactly one point $\mathfrak{p}$ such that $Z = \overline{\{\mathfrak{p}\}}$. (In particular points of X are in bijection with closed subsets of X.)

Idea of proof. The point $\mathfrak{p}$ corresponds to the closed set $\mathcal{V}(\mathfrak{p})$, which one can show is irreducible. □

This gives you a better way to draw non-closed points: they are the generic points lying along any irreducible closed set (consisting of more than just one point).

At this point,[1] I may as well give you the real definition of generic point.

Definition 84.3.15. Given a topological space X, a **generic point** η is a point whose closure is the entire space X.

So for us, when A is an integral domain, $\operatorname{Spec} A$ has generic point (0).

Abuse of Notation 84.3.16. Very careful readers might note I am being a little careless with referring to $(y-x^2)$ as "the generic point along the parabola" in $\operatorname{Spec} \mathbb{C}[x,y]$. What's happening is that $\mathcal{V}(y-x^2)$ is a closed set, and as a topological subspace, it has generic point $(y-x^2)$.

§84.4 On radicals

Back when we studied classical algebraic geometry in $\mathbb{C}^n$, we saw Hilbert's Nullstellensatz (Theorem 77.3.4) show up to give bijections between radical ideals and affine varieties; we omitted the proof, because it was nontrivial.

However, for a *scheme*, where the points *are* prime ideals (rather than tuples in $\mathbb{C}^n$), the corresponding results will actually be *easy*: even in the case where $A = \mathbb{C}[x_1, \dots, x_n]$, the addition of prime ideals (instead of just maximal ideals) will actually *simplify* the proof, because radicals play well with prime ideals.

We still have the following result.

Proposition 84.4.1 ($\mathcal{V}(\sqrt{I}) = \mathcal{V}(I)$)
For any ideal I of a ring A we have $\mathcal{V}(\sqrt{I}) = \mathcal{V}(I)$.

Proof. We have $\sqrt{I} \supseteq I$. Hence automatically $\mathcal{V}(\sqrt{I}) \subseteq \mathcal{V}(I)$.

Conversely, if $\mathfrak{p} \in \mathcal{V}(I)$, then $I \subseteq \mathfrak{p}$, so $\sqrt{I} \subseteq \sqrt{\mathfrak{p}} = \mathfrak{p}$ (by Proposition 77.3.3). □

[1] Pun not intended

We hinted the key result in an earlier remark, and we now prove it.

Theorem 84.4.2 (Radical is intersection of primes)
Let I be an ideal of a ring A. Then

$$\sqrt{I} = \bigcap_{\mathfrak{p} \supseteq I} \mathfrak{p}.$$

Proof. This is a famous statement from commutative algebra, and we prove it here only for completeness. It is "doing most of the work".

Note that if $I \subseteq \mathfrak{p}$, then $\sqrt{I} \subseteq \sqrt{\mathfrak{p}} = \mathfrak{p}$; thus $\sqrt{I} \subseteq \bigcap_{\mathfrak{p} \supseteq I} \mathfrak{p}$.

Conversely, suppose $x \notin \sqrt{I}$, meaning $1, x, x^2, x^3, \ldots \notin I$. Then, consider the localization $(A/I)[1/x]$, which is not the zero ring. Like any ring, it has some maximal ideal (Krull's theorem). This means our usual bijection between prime ideals of $(A/I)[1/x]$, prime ideals of A/I and prime ideals of A gives some prime ideal $\mathfrak{p}$ of A containing I but not containing x. Thus $x \notin \bigcap_{\mathfrak{p} \supseteq I} \mathfrak{p}$, as desired.

The key idea here is, for $x \in A$, $x^n = 0$ for some positive *finite* integer n if and only if $A[1/x] = 0$.

So, in other words,

$$\begin{aligned}
& x \in \sqrt{(0)} \\
\iff & x^n = 0 \text{ for some positive integer } n \\
\iff & A[1/x] = 0 \\
\iff & \text{for all prime ideals } \mathfrak{p},\ x \in \mathfrak{p} \\
\iff & x \in \bigcap_{\mathfrak{p}} \mathfrak{p}.
\end{aligned}$$

When I is not (0), consider the ring A/I instead. □

Geometrically speaking, this theorem states:

> For any f a regular function on $\operatorname{Spec} A/I$, then
>
> $f^n = 0$ for some positive integer n $\iff$ f vanishes at all points in $\operatorname{Spec} A/I$.

To which, the proof above reads:

$$\begin{aligned}
& f \in \sqrt{(I)} \\
\iff & f^n \in I \text{ for some positive integer } n \\
\iff & (A/I)[1/f] = 0 \\
\iff & \operatorname{Spec}(A/I)[1/f] \text{ is empty} \\
\iff & \text{for all } \mathfrak{p} \in \operatorname{Spec} A/I,\ f \text{ vanishes at } \mathfrak{p} \\
\iff & f \in \bigcap_{\mathfrak{p} \in \operatorname{Spec} A/I} \mathfrak{p}.
\end{aligned}$$

You may want to run through the proof with the example $A = k[x]$, $I = (x^2)$ and $f = x$ in Section 86.7, keeping in mind the image of $\operatorname{Spec} A/I$ as a "fuzzy" point and f being a nonzero function that takes value zero at every point.

Remark 84.4.3 (A variant of Krull's theorem) — The longer direction of this proof is essentially saying that for any $x \in A$, there is a maximal ideal of A not containing x. The "short" proof is to use Krull's theorem on $(A/I)[1/x]$ as above, but one can also still prove it directly using Zorn's lemma (by copying the proof of the original Krull's theorem).

Example 84.4.4 ($\sqrt{(2016)} = (42)$ in $\mathbb{Z}$)
In the ring $\mathbb{Z}$, we see that $\sqrt{(2016)} = (42)$, since the distinct primes containing (2016) are (2), (3), (7).

Geometrically, this gives us a good way to describe $\sqrt{I}$: it is the *set of all functions vanishing on all of* $\mathcal{V}(I)$. Indeed, we may write

$$\sqrt{I} = \bigcap_{\mathfrak{p} \supseteq I} \mathfrak{p} = \bigcap_{\mathfrak{p} \in \mathcal{V}(I)} \mathfrak{p} = \bigcap_{\mathfrak{p} \in \mathcal{V}(I)} \{f \in A \mid f(\mathfrak{p}) = 0\}.$$

We can now state:

Theorem 84.4.5 (Radical ideals correspond to closed sets)
Let I and J be ideals of A, and considering the space $\operatorname{Spec} A$. Then

$$\mathcal{V}(I) = \mathcal{V}(J) \iff \sqrt{I} = \sqrt{J}.$$

In particular, radical ideals exactly correspond to closed subsets of $\operatorname{Spec} A$.

Proof. If $\mathcal{V}(I) = \mathcal{V}(J)$, then $\sqrt{I} = \bigcap_{\mathfrak{p} \in \mathcal{V}(I)} \mathfrak{p} = \bigcap_{\mathfrak{p} \in \mathcal{V}(J)} \mathfrak{p} = \sqrt{J}$ as needed.
Conversely, suppose $\sqrt{I} = \sqrt{J}$. Then $\mathcal{V}(I) = \mathcal{V}(\sqrt{I}) = \mathcal{V}(\sqrt{J}) = \mathcal{V}(J)$. □

Compare this to the theorem we had earlier that the *irreducible* closed subsets correspond to *prime* ideals!

§84.5 A few harder problems to think about

As Chapter 86 contains many examples of affine schemes to train your intuition, it's possibly worth reading even before attempting these problems, even though there will be some parts that won't make sense yet.

Problem 84A ($\operatorname{Spec} \mathbb{Q}[x]$)**.** Describe the points and topology of $\operatorname{Spec} \mathbb{Q}[x]$.

Problem 84B (Product rings)**.** Describe the points and topology of $\operatorname{Spec} A \times B$ in terms of $\operatorname{Spec} A$ and $\operatorname{Spec} B$.

85 Affine schemes: the sheaf

We now complete our definition of $X = \operatorname{Spec} A$ by defining the sheaf $\mathcal{O}_X$ on it, making it into a ringed space. This is done quickly in the first section.

As before, our goal is:

> **The sheaf $\mathcal{O}_X$ coincides with the sheaf of regular functions on affine varieties, so that we can apply our geometric intuition to $\operatorname{Spec} A$ when A is an arbitrary ring.**

However, we will then spend the next several chapters trying to convince the reader to *forget* the definition we gave, in practice. This is because practically, the sections of the sheaves are best computed by not using the definition directly, but by using some other results.

Along the way we'll develop some related theory: in computing the stalks we'll find out the definition of a local ring, and in computing the sections we'll find out about distinguished open sets.

A reminder once again: Chapter 86 has many more concrete examples. It's not a bad idea to look through there for more examples if anything in this chapter trips you up.

§85.1 A useless definition of the structure sheaf

Prototypical example for this section: Still $\mathbb{C}[x_1, \dots, x_n]/I$.

We have now endowed $\operatorname{Spec} A$ with the Zariski topology, and so all that remains is to put a sheaf $\mathcal{O}_{\operatorname{Spec} A}$ on it. To do this we want a notion of "regular functions" as before.

This is easy to do since we have localizations on hand.

Definition 85.1.1. First, let $\mathscr{F}$ be the pre-sheaf of "globally rational" functions: i.e. we define $\mathscr{F}(U)$ to be the localization

$$\mathscr{F}(U) = \left\{ \frac{f}{g} \mid f, g \in A \text{ and } g(\mathfrak{p}) \neq 0 \ \forall \mathfrak{p} \in U \right\} = \left(A \setminus \bigcup_{\mathfrak{p} \in U} \mathfrak{p} \right)^{-1} A.$$

We now define the structure sheaf on $\operatorname{Spec} A$. It is

$$\mathcal{O}_{\operatorname{Spec} A} = \mathscr{F}^{\text{sh}}$$

i.e. the sheafification of the $\mathscr{F}$ we just defined.

> **Exercise 85.1.2.** Compare this with the definition for $\mathcal{O}_V$ with V a complex variety, and check that they essentially match.

And thus, we have completed the transition to adulthood, with a complete definition of the affine scheme.

If you really like compatible germs, you can write out the definition:

Definition 85.1.3. Let A be a ring. Then $\operatorname{Spec} A$ is made into a ringed space by setting

$$\mathcal{O}_{\operatorname{Spec} A}(U) = \{(f_{\mathfrak{p}} \in A_{\mathfrak{p}})_{\mathfrak{p}\in U} \text{ which are locally quotients}\}.$$

That is, it consists of sequence $(f_{\mathfrak{p}})_{\mathfrak{p}\in U}$, with each $f_{\mathfrak{p}} \in A_{\mathfrak{p}}$, such that for every point $\mathfrak{p}$ there is an open neighborhood $U_{\mathfrak{p}}$ and an $f, g \in A$ such that $f_{\mathfrak{q}} = \frac{f}{g} \in A_{\mathfrak{q}}$ for all $\mathfrak{q} \in U_{\mathfrak{p}}$.

We will now **basically forget about this definition**, because we will never use it in practice. In the next two sections, we will show you:

- that the stalks $\mathcal{O}_{\operatorname{Spec} A,\mathfrak{p}}$ are just $A_{\mathfrak{p}}$, and
- that the sections $\mathcal{O}_{\operatorname{Spec} A}(U)$ can be computed, for any open set U, by focusing only on the special case where $U = D(f)$ is a distinguished open set.

These two results will be good enough for all of our purposes, so we will be able to not use this definition. (Hence the lack of examples in this section.)

§85.2 The value of distinguished open sets (or: how to actually compute sections)

Prototypical example for this section: $D(x)$ *in* $\operatorname{Spec} \mathbb{C}[x]$ *is the punctured line.*

We will now really hammer in the importance of the distinguished open sets. The definition is analogous to before:

Definition 85.2.1. Let $f \in \operatorname{Spec} A$. Then $D(f)$ is the set of $\mathfrak{p}$ such that $f(\mathfrak{p}) \neq 0$, a **distinguished open set**.

Distinguished open sets will have three absolutely crucial properties, which build on each other.

§85.2.i A basis of the Zariski topology

The first is a topological observation:

Theorem 85.2.2 (Distinguished open sets form a base)
The distinguished open sets $D(f)$ form a basis for the Zariski topology: any open set U is a union of distinguished open sets.

Proof. Let U be an open set; suppose it is the complement of closed set $V(I)$. Then verify that

$$U = \bigcup_{f\in I} D(f). \qquad \square$$

§85.2.ii Sections are computable

The second critical fact is that the sections on distinguished open sets can be computed explicitly.

Theorem 85.2.3 (Sections of $D(f)$ are localizations away from f)
Let A be a ring and $f \in A$. Then

$$\mathcal{O}_{\operatorname{Spec} A}(D(f)) \cong A[1/f].$$

Proof. Omitted, but similar to Theorem 78.6.1. □

Example 85.2.4 (The punctured line is isomorphic to a hyperbola)
The "hyperbola effect" appears again:

$$\mathcal{O}_{\operatorname{Spec}\mathbb{C}[x]}(D(x)) = \mathbb{C}[x, x^{-1}] \cong \mathbb{C}[x, y]/(xy - 1).$$

On a tangential note, we had better also note somewhere that $\operatorname{Spec} A = D(1)$ is itself distinguished open, so the global sections can be recovered.

Corollary 85.2.5 (A is the ring of global sections)
The ring of global sections of $\operatorname{Spec} A$ is A.

Proof. By previous theorem, $\mathcal{O}_{\operatorname{Spec} A}(\operatorname{Spec} A) = \mathcal{O}_{\operatorname{Spec} A}(D(1)) = A[1/1] = A$. □

§85.2.iii They are affine

We know $\mathcal{O}_X(D(f)) = A[1/f]$. In fact, if you draw $\operatorname{Spec} A[1/f]$, you will find that it looks exactly like $D(f)$. So the third final important fact is that $D(f)$ will actually be *isomorphic* to $\operatorname{Spec} A[1/f]$ (just like the line minus the origin is isomorphic to the hyperbola). We can't make this precise yet, because we have not yet discussed morphisms of schemes, but it will be handy later (though not right away).

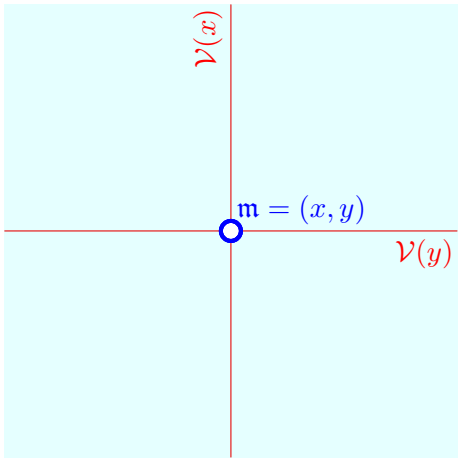
§85.2.iv Classic example: the punctured plane

We now give the classical example of a computation which shows how you can forget about sheafification, if you never liked it.[1] The idea is that:

We can compute any section $\mathcal{O}_X(U)$ in practice by using distinguished open sets and sheaf axioms.

Let $X = \operatorname{Spec}\mathbb{C}[x, y]$, and consider the origin, i.e. the point $\mathfrak{m} = (x, y)$. This ideal is maximal, so it corresponds to a closed point, and we can consider the open set U consisting of all the points other than $\mathfrak{m}$. We wish to compute $\mathcal{O}_X(U)$.

[1]This perspective is so useful that some sources, like Vakil [Va17, §4.1] will *define* $\mathcal{O}_{\operatorname{Spec} A}$ by requiring $\mathcal{O}_{\operatorname{Spec} A}(D(f)) = A[1/f]$, rather than use sheafification as we did.



Unfortunately, U is not distinguished open. But, we can compute it anyways by writing $U = D(x) \cup D(y)$: conveniently, $D(x) \cap D(y) = D(xy)$. By the sheaf axioms, we have a pullback square

$$\begin{array}{ccc} \mathcal{O}_X(U) & \longrightarrow & \mathcal{O}_X(D(x)) = \mathbb{C}[x, y, x^{-1}] \\ \downarrow & & \downarrow \\ \mathcal{O}_X(D(y)) = \mathbb{C}[x, y, y^{-1}] & \longrightarrow & \mathcal{O}_X(D(xy)) = \mathbb{C}[x, y, x^{-1}, y^{-1}]. \end{array}$$

In other words, $\mathcal{O}_X(U)$ consists of pairs

$$f \in \mathbb{C}[x, y, x^{-1}]$$
$$g \in \mathbb{C}[x, y, y^{-1}]$$

which agree on the overlap: $f = g$ on $D(x) \cap D(y)$. Well, we can describe f as a polynomial with some x's in the denominator, and g as a polynomial with some y's in the denominator. If they match, the denominator is actually constant. Put crudely,

$$\mathbb{C}[x, y, x^{-1}] \cap \mathbb{C}[x, y, y^{-1}] = \mathbb{C}[x, y].$$

In conclusion,

$$\mathcal{O}_X(U) = \mathbb{C}[x, y].$$

That is, we get no additional functions.

§85.3 The stalks of the structure sheaf

Prototypical example for this section: The stalk of $\operatorname{Spec} \mathbb{C}[x, y]$ *at* $\mathfrak{m} = (x, y)$ *are rational functions defined at the origin.*

Don't worry, this one is easier than last section.

§85.3.i They are localizations

Theorem 85.3.1 (Stalks of $\operatorname{Spec} A$ are $A_{\mathfrak{p}}$)
Let A be a ring and let $\mathfrak{p} \in \operatorname{Spec} A$. Then

$$\mathcal{O}_{\operatorname{Spec} A, \mathfrak{p}} \cong A_{\mathfrak{p}}.$$

In particular $\operatorname{Spec} A$ is a locally ringed space.

Proof. Since sheafification preserved stalks, it's enough to check it for $\mathscr{F}$ the pre-sheaf of globally rational functions in our definition. The proof is basically the same as Theorem 82.3.7: there is an obvious map $\mathscr{F}_{\mathfrak{p}} \to A_{\mathfrak{p}}$ on germs by

$$(U, f/g \in \mathscr{F}(U)) \mapsto f/g \in A_{\mathfrak{p}}.$$

(Note the f/g on the left lives in $\mathscr{F}(U)$ but the one on the right lives in $A_{\mathfrak{p}}$). We show injectivity and surjectivity:

- Injective: suppose $(U_1, f_1/g_1)$ and $(U_2, f_2/g_2)$ are two germs with $f_1/g_1 = f_2/g_2 \in A_{\mathfrak{p}}$. This means $h(g_1 f_2 - f_2 g_1) = 0$ in A, for some nonzero h. Then both germs identify with the germ $(U_1 \cap U_2 \cap D(h), f_1/g_1)$.
- Surjective: let $U = D(g)$. □

Example 85.3.2 (Denominators not divisible by x)
We have seen this example so many times that I will only write it in the new notation, and make no further comment: if $X = \operatorname{Spec} \mathbb{C}[x]$ then

$$\mathcal{O}_{\operatorname{Spec} X,(x)} = \mathbb{C}[x]_{(x)} = \left\{ \frac{f}{g} \mid g(0) \neq 0 \right\}.$$

Example 85.3.3 (Denominators not divisible by x or y)
Let $X = \operatorname{Spec} \mathbb{C}[x, y]$ and let $\mathfrak{m} = (x, y)$ be the origin. Then

$$\mathbb{C}[x, y]_{(x,y)} = \left\{ \frac{f(x,y)}{g(x,y)} \mid g(0,0) \neq 0 \right\}.$$

If you want more examples, take any of the ones from Section 83.4, and try to think about what they mean geometrically.

§85.3.ii Motivating local rings: germs should package values

Let's return to our well-worn example $X = \operatorname{Spec} \mathbb{C}[x, y]$ and consider $\mathfrak{m} = (x, y)$ the origin. The stalk was

$$\mathcal{O}_{X,\mathfrak{m}} = \mathbb{C}[x, y]_{(x,y)} = \left\{ \frac{f(x,y)}{g(x,y)} \mid g(0,0) \neq 0 \right\}.$$

So let's take some section like $f = \frac{1}{xy+4}$, which is a section of $U = D(xy + 4)$ (or some smaller open set, but we'll just use this one for simplicity). We also have $U \ni m$, and so f gives a germ at $\mathfrak{m}$.

On the other hand, f also has a value at $\mathfrak{m}$: it is $f \pmod{\mathfrak{m}} = \frac{1}{4}$. And in general, the ring of possible values of a section at the origin $\mathfrak{m}$ is $\mathbb{C}[x, y]/\mathfrak{m} \cong \mathbb{C}$.

Now, you might recall that I pressed the point of view that a germ might be thought of as an "enriched value". Then it makes sense that if you know the germ of a section f at a point $\mathfrak{m}$ — i.e., you know the "enriched value" — then you should be able to compute its value as well. What this means is that we ought to have some map

$$A_{\mathfrak{m}} \to A/\mathfrak{m}$$

sending germs to their associated values.

Indeed you can, and this leads us to...

§85.4 Local rings and residue fields: linking germs to values

Prototypical example for this section: The residue field of $\operatorname{Spec}\mathbb{C}[x,y]$ *at* $\mathfrak{m}=(x,y)$ *is* $\mathbb{C}$.

§85.4.i Localizations give local rings

This notation is about to get really terrible, but bear with me.

Theorem 85.4.1 (Stalks are local rings)
Let A be a ring and $\mathfrak{p}$ any prime ideal. Then the localization $A_{\mathfrak{p}}$ has exactly one maximal ideal, given explicitly by

$$\mathfrak{p}A_{\mathfrak{p}} = \left\{ \frac{f}{g} \mid f \in \mathfrak{p},\ g \notin \mathfrak{p} \right\}.$$

The ideal $\mathfrak{p}A_{\mathfrak{p}}$ thus captures the idea of "germs vanishing at $\mathfrak{p}$".[2]

Proof in a moment; for now let's introduce some words so we can give our examples in the proper language.

Definition 85.4.2. A ring R with exactly one maximal ideal $\mathfrak{m}$ will be called a **local ring**. The **residue field** is the quotient $A/\mathfrak{m}$.

Question 85.4.3. Are fields local rings?

Thus what we find is that:

The stalks consist of the possible enriched values (germs); the residue field is the set of (un-enriched) values.

Example 85.4.4 (The stalk at the origin of $\operatorname{Spec}\mathbb{C}[x,y]$)
Again set $A=\mathbb{C}[x,y]$, $X=\operatorname{Spec}A$ and $\mathfrak{p}=(x,y)$ so that $\mathcal{O}_{X,\mathfrak{p}}=A_{\mathfrak{p}}$. (I switched to $\mathfrak{p}$ for the origin, to avoid confusion with the maximal ideal $\mathfrak{p}A_{\mathfrak{p}}$ of the local ring $A_{\mathfrak{p}}$.) As we said many times already, $A_{\mathfrak{p}}$ consists of rational functions not vanishing at the origin, such as $f=\frac{1}{xy+4}$.
What is the unique maximal ideal $\mathfrak{p}A_{\mathfrak{p}}$? Answer: it consists of the rational functions which *vanish* at the origin: for example, $\frac{x}{x^2+3y}$, or $\frac{3x+5y}{2}$, or $\frac{-xy}{4(xy+4)}$. If we allow ourselves to mod out by such functions, we get the residue field $\mathbb{C}$, and f will have the value $\frac{1}{4}$, since

$$\frac{1}{xy+4} - \underbrace{\frac{-xy}{4(xy+4)}}_{\text{vanishes at origin}} = \frac{1}{4}.$$

More generally, suppose f is any section of some open set containing $\mathfrak{p}$. Let $c\in\mathbb{C}$ be the value $f(\mathfrak{p})$, that is, $f \pmod{\mathfrak{p}}$. Then $f-c$ is going to be another section which vanishes at the origin $\mathfrak{p}$, so as promised, $f \equiv c \pmod{\mathfrak{p}A_{\mathfrak{p}}}$.

[2]The notation $\mathfrak{p}A_{\mathfrak{p}}$ really means the set of $f\cdot h$ where $f\in\mathfrak{p}$ (viewed as a subset of $A_{\mathfrak{p}}$ by $f\mapsto\frac{f}{1}$) and $h\in A_{\mathfrak{p}}$. I personally find this is more confusing than helpful, so I'm footnoting it.

Okay, we can write down a proof of the theorem now.

Proof of Theorem 85.4.1. One may check that the set $I = \mathfrak{p}A_{\mathfrak{p}}$ is an ideal of $A_{\mathfrak{p}}$. Moreover, $1 \notin I$, so I is proper.

To prove it is maximal and unique, it suffices to prove that any $f \in A_{\mathfrak{p}}$ with $f \notin I$ is a *unit* of $A_{\mathfrak{p}}$. This will imply I is maximal: there are no more non-units to add. It will also imply I is the only maximal ideal: because any proper ideal can't contain units, so is contained in I.

This is actually easy. An element of $A_{\mathfrak{p}}$ not in I must be $x = \frac{f}{g}$ for $f, g \in A$ and $f, g \notin \mathfrak{p}$. For such an element, $x^{-1} = \frac{g}{f} \notin \mathfrak{p}$ too. So x is a unit. End proof. □

Even more generally:

If a sheaf $\mathscr{F}$ consists of "field-valued functions", the stalk $\mathscr{F}_p$ probably has a maximal ideal consisting of the germs vanishing at p.

Example 85.4.5 (Local rings in non-algebraic geometry sheaves)

Let's go back to the example of $X = \mathbb{R}$ and $\mathscr{F}(U)$ the smooth functions, and consider the stalk $\mathscr{F}_p$, where $p \in X$. Define the ideal $\mathfrak{m}_p$ to be the set of germs (s, U) for which $s(p) = 0$.

Then $\mathfrak{m}_p$ is maximal: we have an exact sequence

$$0 \to \mathfrak{m}_p \to \mathscr{F}_p \xrightarrow{(s,U) \mapsto s(p)} \mathbb{R} \to 0$$

and so $\mathscr{F}_p / \mathfrak{m}_p \cong \mathbb{R}$, which is a field.

It remains to check there are no nonzero maximal ideals. Now note that if $s \notin \mathfrak{m}_p$, then s is nonzero in some open neighborhood of p, and one can construct the function $1/s$ on it. So **every element of $\mathscr{F}_p \setminus \mathfrak{m}_p$ is a unit**; and again $\mathfrak{m}_p$ is in fact the only maximal ideal!

Thus the stalks of each of the following types of sheaves are local rings, too.

- Sheaves of continuous real/complex functions on a topological space
- Sheaves of smooth functions on any manifold
- etc.

§85.4.ii Computing values: a convenient square

Very careful readers might have noticed something a little uncomfortable in our extended example with $\operatorname{Spec} A$ with $A = \mathbb{C}[x, y]$ and $\mathfrak{p} = (x, y)$ the origin. Let's consider $f = \frac{1}{xy+4}$. We took $f \pmod{x, y}$ in the original ring A in order to decide the value "should" be $\frac{1}{4}$. However, all our calculations actually took place not in the ring A, but instead in the ring $A_{\mathfrak{p}}$. Does this cause issues?

Thankfully, no, nothing goes wrong, even in a general ring A.

Definition 85.4.6. We let the quotient $A_{\mathfrak{p}}/\mathfrak{p}A_{\mathfrak{p}}$, i.e. the **residue field** of the stalk of $\operatorname{Spec} A$ at $\mathfrak{p}$, be denoted by $\kappa(\mathfrak{p})$.

Then the following is a special case of Theorem 83.7.1 (localization commutes with quotients):

Theorem 85.4.7 (The germ-to-value square)

Let A be a ring and $\mathfrak{p}$ a prime ideal. The following diagram commutes:

$$\begin{array}{ccc} A & \xrightarrow{\text{localize}} & A_{\mathfrak{p}} \\ {\scriptstyle \text{mod } \mathfrak{p}}\downarrow & & \downarrow{\scriptstyle \text{mod } \mathfrak{p}} \\ A/\mathfrak{p} & \xrightarrow{\text{Frac}(-)} & \kappa(\mathfrak{p}) \end{array}$$

In particular, $\kappa(\mathfrak{p})$ can also be described as $\text{Frac}(A/\mathfrak{p})$.

So for example, if $A = \mathbb{C}[x,y]$ and $\mathfrak{p} = (x,y)$, then $A/\mathfrak{p} = \mathbb{C}$ and $\text{Frac}(A/\mathfrak{p}) = \text{Frac}(\mathbb{C}) = \mathbb{C}$, as we expected. In practice, $\text{Frac}(A/\mathfrak{p})$ is probably the easier way to compute $\kappa(\mathfrak{p})$ for any prime ideal $\mathfrak{p}$.

§85.5 Recap

To recap the last two chapters, let A be a ring.

- We define $X = \operatorname{Spec} A$ to be the set of prime ideals of A.
  - The maximal ideals are the "closed points" we are used to, but the prime ideals are "generic points".
- We equip $\operatorname{Spec} A$ with the Zariski topology by declaring $\mathcal{V}(I)$ to be the closed sets, for ideals $I \subseteq A$.
  - The distinguished open sets $D(f)$, form a topological basis.
  - The irreducible closed sets are exactly the closures of points.
- Finally, we defined a sheaf $\mathcal{O}_X$. We set up the definition such that
  - $\mathcal{O}_X(D(f)) = A[1/f]$: at distinguished open sets $D(f)$, we get localizations too.
  - $\mathcal{O}_{X,\mathfrak{p}} = A_{\mathfrak{p}}$: the stalks are localizations at a prime.

  Since $D(f)$ is a basis, these two properties lets us explicitly compute $\mathcal{O}_X(U)$ for any open set U, so we don't have to resort to the definition using sheafification.

§85.6 Functions are determined by germs, not values

Prototypical example for this section: The functions 0 *and* x *on* $\operatorname{Spec} \mathbb{C}[x]/(x^2)$.

We close the chapter with a word of warning. In any ringed space, a section is determined by its germs; so that on $\operatorname{Spec} A$ a function $f \in A$ is determined by its germ in each stalk $A_{\mathfrak{p}}$. However, we now will mention that an $f \in A$ is *not* determined by its value $f(\mathfrak{p}) = f \pmod{\mathfrak{p}}$ at each point.

The famous example is:

Example 85.6.1 (On the double point, all multiples of x are zero at all points)

The space $\operatorname{Spec} \mathbb{C}[x]/(x^2)$ has only one point, (x). The functions 0 and x (and for that matter $2x$, $3x$, ...) all vanish on it. This shows that functions are not determined uniquely by values in general.

Fortunately, we can explicitly characterize when this sort of "bad" behavior happens. Indeed, we want to see when $f(\mathfrak{p}) = g(\mathfrak{p})$ for every $\mathfrak{p}$, or equivalently, $h = f - g$ vanishes on every prime ideal $\mathfrak{p}$. This is equivalent to having

$$h \in \bigcap_{\mathfrak{p}} \mathfrak{p} = \sqrt{(0)}$$

the radical of the *zero* ideal. Thus in the prototype, the failure was caused by the fact that $x^n = 0$ for some large n.

Definition 85.6.2. For a ring A, the radical of the zero ideal, $\sqrt{(0)}$, is called the **nilradical** of A. Elements of the nilradical are called **nilpotents**. We say A is **reduced** if 0 is the only nilpotent, i.e. $\sqrt{(0)} = (0)$.

Question 85.6.3. Are integral domains reduced?

Then our above discussion gives:

Theorem 85.6.4 (Nilpotents are the only issue)
Two functions f and g have the same value on all points of $\operatorname{Spec} A$ if and only if $f - g$ is nilpotent.

In particular, when A is a reduced ring, even the values $f(\mathfrak{p})$ as $\mathfrak{p} \in \operatorname{Spec} A$ are enough to determine $f \in A$.

§85.7 A few harder problems to think about

As Chapter 86 contains many examples of affine schemes to train your intuition; it's likely to be worth reading even before attempting these problems.

Problem 85A† (Spectrums are quasicompact)**.** Show that $\operatorname{Spec} A$ is quasicompact for any ring A.

Problem 85B (Punctured gyrotop, communicated by Aaron Pixton)**.** The gyrotop is the scheme $X = \operatorname{Spec} \mathbb{C}[x, y, z]/(xy, z)$. We let U denote the open subset obtained by deleting the closed point $\mathfrak{m} = (x, y, z)$. Compute $\mathcal{O}_X(U)$.

Problem 85C. Show that a ring R is a local ring if and only of the following property is true: for any $x \in R$, either x or $1 - x$ is a unit.

Problem 85D. Let R be a local ring, and $\mathfrak{m}$ be its maximal ideal. Describe $R_{\mathfrak{m}}$.

Problem 85E. Let A be a ring, and $\mathfrak{m}$ a maximal ideal. Consider $\mathfrak{m}$ as a point of $\operatorname{Spec} A$. Show that $\kappa(\mathfrak{m}) \cong A/\mathfrak{m}$.

86 Interlude: eighteen examples of affine schemes

To cement in the previous two chapters, we now give an enormous list of examples. Each example gets its own section, rather than having page-long orange boxes.

One common theme you will find as you wade through the examples is that your geometric intuition may be better than your algebraic one. For example, while studying $k[x,y]/(xy)$ you will say "geometrically, I expect so-and-so to look like other thing", but when you write down the algebraic statements you find two expressions that are don't look equal to you. However, if you then do some calculation you will find that they were isomorphic after all. So in that sense, in this chapter you will learn to begin drawing pictures of algebraic statements — which is great!

As another example, all the lemmas about prime ideals from our study of localizations will begin to now take concrete forms: you will see many examples that

- $\operatorname{Spec} A/I$ looks like $\mathcal{V}(I)$ of $\operatorname{Spec} A$,
- $\operatorname{Spec} A[1/f]$ looks like $D(f)$ of $\operatorname{Spec} A$,
- $\operatorname{Spec} A_{\mathfrak{p}}$ looks like $\mathcal{O}_{\operatorname{Spec} A,\mathfrak{p}}$ of $\operatorname{Spec} A$.

In everything that follows, k is any field. We will also use the following color connotations:

- The closed points of the scheme are drawn in blue.
- Non-closed points are drawn in red, with their "trails" dotted red.
- Stalks are drawn in green, when they appear.

§86.1 Example: $\operatorname{Spec} k$, a single point

This one is easy: for any field k, $X = \operatorname{Spec} k$ has a single point, corresponding to the only proper ideal (0). There is only way to put a topology on it.

As for the sheaf,

$$\mathcal{O}_X(X) = \mathcal{O}_{X,(0)} = k.$$

So the space is remembering what field it wants to be over. If we are complex analysts, the set of functions on a single point is $\mathbb{C}$; if we are number theorists, maybe the set of functions on a single point is $\mathbb{Q}$.

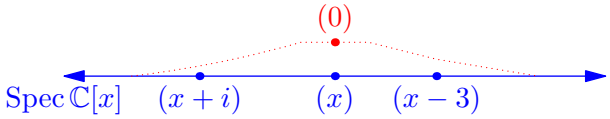
§86.2 $\operatorname{Spec} \mathbb{C}[x]$, a one-dimensional line

The scheme $X = \operatorname{Spec} \mathbb{C}[x]$ is our beloved one-dimensional line. It consists of two types of points:

- The closed points $(x-a)$, corresponding to each complex number $a \in \mathbb{C}$, and
- The *generic* point (0).

As for the Zariski topology, every open set contains (0), which captures the idea it is close to everywhere: no matter where you stand, you can still hear the buzzing of the fly! True to the irreducibility of this space, the open sets are huge: the proper *closed sets* consist of finitely many closed points.

Here is a picture: for lack of better place to put it, the point (0) is floating around just above the line in red.



The notion of "value at $\mathfrak{p}$" works as expected. For example, $f = x^2 + 5$ is a global section of $\mathbb{C}[x]$. If we evaluate it at $\mathfrak{p} = x - 3$, we find $f(\mathfrak{p}) = f \pmod{\mathfrak{p}} = x^2 + 5 \pmod{x-3} = 14 \pmod{x-3}$. Indeed,

$$\kappa(\mathfrak{p}) \cong \mathbb{C}$$

meaning the stalks all have residue field $\mathbb{C}$. As

$$\mathbb{C}[x]/\mathfrak{p} \cong \mathbb{C} \qquad \text{by } x \mapsto 3$$

we see we are just plugging $x = 3$.

Of course, the stalk at $(x-3)$ carries more information. In this case it is $\mathbb{C}[x]_{(x-3)}$. Which means that if we stand near the point $(x-3)$, rational functions are all fine as long as no $x - 3$ appears in the denominator. So, $\frac{x^2+8}{(x-1)(x-5)}$ is a fine example of a germ near $x = 3$.

Things get more interesting if we consider the generic point $\eta = (0)$.

What is the stalk $\mathcal{O}_{X,\eta}$? Well, it should be $\mathbb{C}[x]_{(0)} = \mathbb{C}(x)$, which is the again the set of *rational* functions. And that's what you expect. For example, $\frac{x^2+8}{(x-1)(x-5)}$ certainly describes a rational function on "most" complex numbers.

What happens if we evaluate the global section $f = x^2 + 5$ at η? Well, we just get $f(\eta) = x^2 + 5$ — taking modulo 0 doesn't do much. Fitting, it means that if you want to be able to evaluate a polynomial f at a general complex number, you actually just need the whole polynomial (or rational function). We can think of this in terms of the residue field being $\mathbb{C}(x)$:

$$\kappa((0)) = \operatorname{Frac}(\mathbb{C}[x]/(0)) \cong \operatorname{Frac}\mathbb{C}[x] = \mathbb{C}(x).$$

§86.3 $\operatorname{Spec}\mathbb{R}[x]$, a one-dimensional line with complex conjugates glued (no fear nullstellensatz)

Despite appearances, this actually looks almost exactly like $\operatorname{Spec}\mathbb{C}[x]$, even more than you expect. The main thing to keep in mind is that now (x^2+1) is a point, which you can loosely think of as $\pm i$. So it almost didn't matter that $\mathbb{R}$ is not algebraically closed; the $\mathbb{C}$ is showing through anyways. But this time, because we only consider real coefficient polynomials, we do not distinguish between "conjugate" $+i$ and $-i$. Put another way, we have folded $a+bi$ and $a-bi$ into a single point: $(x+i)$ and $(x-i)$ merge to form x^2+1.

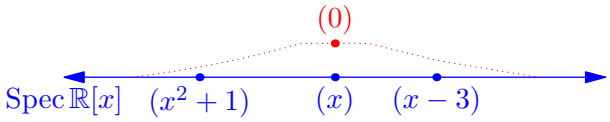
To be explicit, there are three types of points:

- $(x-a)$ for each real number a
- $(x^2 - ax + b)$ if $a^2 < 4b$, and

- the generic point (0), again.

The ideals $(x-a)$ and (x^2-ax+b) are each closed points: the quotients with $\mathbb{R}[x]$ are both fields ($\mathbb{R}$ and $\mathbb{C}$, respectively).

We have been drawing $\operatorname{Spec}\mathbb{C}[x]$ as a one-dimensional line, so $\operatorname{Spec}\mathbb{R}[x]$ will be drawn the same way.



One nice thing about this is that the nullstellensatz is less scary than it was with classical varieties. The short version is that the function x^2+1 vanishes at a point of $\operatorname{Spec}\mathbb{R}[x]$, namely (x^2+1) itself! (So in some ways we're sort of automatically working with the algebraic closure.)

You might remember a long time ago we made a big fuss about the weak nullstellensatz, for example in Problem 77C: if I was a proper ideal in $\mathbb{C}[x_1,\dots,x_n]$ there was *some* point $(a_1,\dots,a_n) \in \mathbb{C}^n$ such that $f(a_1,\dots,a_n)=0$ for all $f \in I$. With schemes, it doesn't matter anymore: if I is a proper ideal of a ring A, then some maximal ideal contains it, and so $\mathcal{V}(I)$ is nonempty in $\operatorname{Spec} A$.

We better mention that the stalks this time look different than expected. Here are some examples:

$$\begin{aligned} \kappa\left((x^2+1)\right) &\cong \mathbb{R}[x]/(x^2+1) \cong \mathbb{C} \\ \kappa\left((x-3)\right) &\cong \mathbb{R}[x]/(x-3) \cong \mathbb{R} \\ \kappa\left((0)\right) &\cong \operatorname{Frac}(\mathbb{R}[x]/(0)) \cong \mathbb{R}(x) \end{aligned}$$

Notice the residue fields above the "complex" points are bigger: functions on them take values in $\mathbb{C}$.

§86.4 $\operatorname{Spec} k[x]$, over any ground field

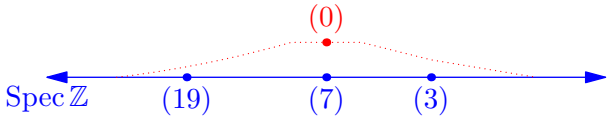
In general, if $\overline{k}$ is the algebraic closure of k, then $\operatorname{Spec} k[x]$ looks like $\operatorname{Spec}\overline{k}[x]$ with all the Galois conjugates glued together. So we will almost never need "algebraically closed" hypotheses anymore: we're working with polynomial ideals, so all the elements are implicitly there, anyways.

§86.5 $\operatorname{Spec}\mathbb{Z}$, a one-dimensional scheme

The great thing about $\operatorname{Spec}\mathbb{Z}$ is that it basically looks like $\operatorname{Spec} k[x]$, too, being a one-dimensional scheme. It has two types of prime ideals:

- (p), for every rational prime p,
- and the generic point (0).

So the picture almost does not change.



This time $\eta = (0)$ has stalk $\mathbb{Z}_{(0)} = \mathbb{Q}$, so a "rational function" is literally a rational number! Thus, $\frac{20}{19}$ is a function with a double root at (2), a root at (5), and a simple pole at (19). If we evaluate it at $\mathfrak{p} = (7)$, we get 3 (mod 7). In general, the residue fields are what you'd guess:

$$\kappa((p)) = \mathbb{Z}/(p) \cong \mathbb{F}_p$$

for each prime p, and $\kappa((0)) \cong \mathbb{Q}$.

The stalk is bigger than the residue field at the closed points: for example

$$\mathcal{O}_{\operatorname{Spec}\mathbb{Z},(3)} \cong \left\{ \frac{m}{n} \mid 3 \nmid n \right\}$$

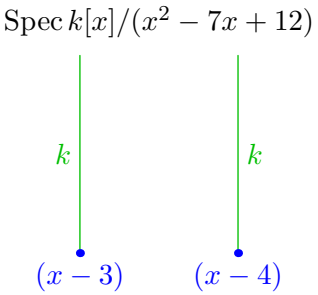
consists of rational numbers with no pole at 3. The stalk at the generic point is $\mathbb{Z}_{(0)} \cong \operatorname{Frac}\mathbb{Z} = \mathbb{Q}$.

§86.6 $\operatorname{Spec} k[x]/(x^2 - 7x + 12)$, two points

If we were working with affine varieties, you would already know what the answer is: $x^2 - 7x + 12 = 0$ has solutions $x = 3$ and $x = 4$, so this should be a scheme with two points.

To see this come true, we use Proposition 83.6.1: the points of $\operatorname{Spec} k[x]/(x^2 - 7x + 12)$ should correspond to prime ideals of $k[x]$ containing $(x^2 - 7x + 12)$. As $k[x]$ is a PID, there are only two, $(x-3)$ and $(x-4)$. They are each maximal, since their quotient with $k[x]$ is a field (namely k), so as promised $\operatorname{Spec} k[x]/(x^2 - 7x + 12)$ has just two closed points.

Each point has a stalk above it isomorphic to k. A section on the whole space X is just a choice of two values, one at $(x-3)$ and one at $(x-4)$.



So actually, this is a geometric way of thinking about the ring-theoretic fact that

$$k[x]/\left(x^2 - 7x + 12\right) \cong k \times k \quad \text{by } f \mapsto (f(3), f(4)).$$

Also, this is the first example of a reducible space in this chapter: in fact X is even disconnected. Accordingly there is no generic point floating around: as the space is discrete, all points are closed.

§86.7 $\operatorname{Spec} k[x]/(x^2)$, the double point

We can now elaborate on the "double point" scheme

$$X_2 = \operatorname{Spec} k[x]/(x^2)$$

since it is such an important motivating example. How it does differ from the "one-point" scheme $X_1 = \operatorname{Spec} k[x]/(x) = \operatorname{Spec} k$? Both X_2 and X_1 have exactly one point, and so obviously the topologies are the same too.

The difference is that the stalk (equivalently, the section, since we have only one point) is larger:

$$\mathcal{O}_{X_2,(x)} = \mathcal{O}_{X_2}(X_2) = k[x]/(x^2).$$

So to specify a function on a double point, you need to specify two parameters, not just one: if we take a polynomial

$$f = a_0 + a_1 x + \cdots \in k[x]$$

then evaluating it at the double point will remember both a_0 and the "first derivative" say.

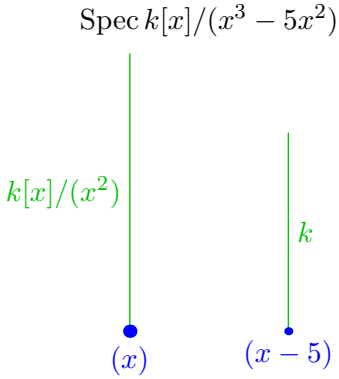
I should mention that if you drop all the way to the residue fields, you can't tell the difference between the double point and the single point anymore. For the residue field of $\operatorname{Spec} k[x]/(x^2)$ at (x) is

$$\operatorname{Frac}(A/(x)) = \operatorname{Frac} k = k.$$

Thus the set of *values* is still just k (leading to the "nilpotent" discussion at the end of last chapter); but the stalk, having "enriched" values, can tell the difference.

§86.8 $\operatorname{Spec} k[x]/(x^3 - 5x^2)$, a double point and a single point

There is no problem putting the previous two examples side by side: the scheme $X = \operatorname{Spec} k[x]/(x^3 - 5x^2)$ consists of a double point next to a single point. Note that the stalks are different: the one above the double point is larger.



This time, we implicitly have the ring isomorphism

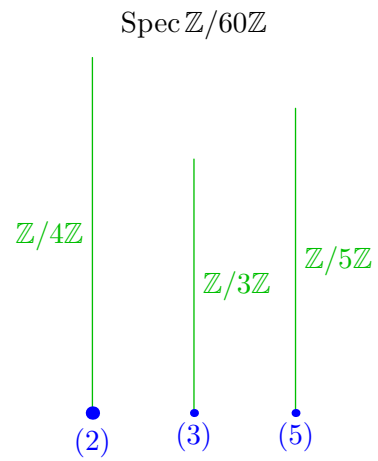
$$k[x]/(x^3 - 5x^2) \cong k[x]/(x^2) \times k$$

by $f \mapsto (f(0) + f'(0)x, f(5))$. The derivative is meant formally here!

§86.9 $\operatorname{Spec} \mathbb{Z}/60\mathbb{Z}$, a scheme with three points

We've being seeing geometric examples of ring products coming up, but actually the Chinese remainder theorem you are used to with integers is no different. (This example $X = \operatorname{Spec} \mathbb{Z}/60\mathbb{Z}$ is taken from [Va17, §4.4.11].)

By Proposition 83.6.1, the prime ideals of $\mathbb{Z}/60\mathbb{Z}$ are (2), (3), (5). But you can think of this also as coming out of $\operatorname{Spec} \mathbb{Z}$: as 60 was a function with a double root at (2), and single roots at (3) and (5).



Actually, although I have been claiming the ring isomorphisms, the sheaves really actually give us a full proof. Let me phrase it in terms of global sections:

$$\begin{aligned}\mathbb{Z}/60\mathbb{Z} &= \mathcal{O}_X(X) \\ &= \mathcal{O}_X(\{(2)\}) \times \mathcal{O}_X(\{(3)\}) \times \mathcal{O}_X(\{(5)\}) \\ &= \mathcal{O}_{X,(2)} \times \mathcal{O}_{X,(3)} \times \mathcal{O}_{X,(5)} \\ &= \mathbb{Z}/4\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/5\mathbb{Z}.\end{aligned}$$

So the theorem that $\mathcal{O}_X(X) = A$ for $X = \operatorname{Spec} A$ is doing the "work" here; the sheaf axioms then give us the Chinese remainder theorem from here.

On that note, this gives us a way of thinking about the earlier example that

$$(\mathbb{Z}/60\mathbb{Z})[1/5] \cong \mathbb{Z}/12\mathbb{Z}.$$

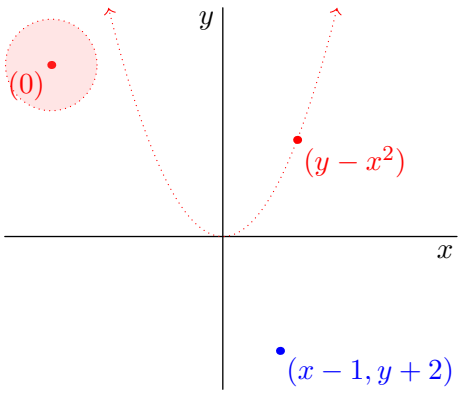
Indeed, $\operatorname{Spec} \mathbb{Z}/60\mathbb{Z}[1/5]$ is supposed to look like the distinguished open set $D(5)$: which means we delete the point (5) from the picture above. That leaves us with $\mathbb{Z}/12\mathbb{Z}$.

§86.10 $\operatorname{Spec} k[x,y]$, the two-dimensional plane

We have seen this scheme already: it is visualized as a plane. There are three types of points:

- The closed points $(x-a, y-b)$, which consists of single points of the plane.
- A non-closed point $(f(x,y))$ for any irreducible polynomial f, which floats along some irreducible curve. We illustrate this by drawing the dotted curve along which the point is floating.
- The generic point (0), floating along the entire plane. I don't know a good place to put it in the picture, so I'll just put it somewhere and draw a dotted circle around it.

Here is an illustration of all three types of points.



We also go ahead and compute the stalks above each point.

- The stalk above $(x - 1, y + 2)$ is the set of rational functions $\frac{f(x,y)}{g(x,y)}$ such that $g(1, -2) \neq 0$.

- The stalk above the non-closed point $(y - x^2)$ is the set of rational functions $\frac{f(x,y)}{g(x,y)}$ such that $g(t, t^2) \neq 0$. For example the function $\frac{xy}{x+y-2}$ is still fine; despite the fact that the denominator vanishes at the point $(1, 1)$ and $(-2, 4)$ on the parabola, it is a function on a "generic point" (crudely, "most points") of the parabola.

- The stalk above (0) is the entire fraction field $k(x, y)$ of rational functions.

Let's consider the global section $f = x^2 + y^2$ and also take the value at each of the above points.

- $f \pmod{x - 1, y - 2} = 5$, so f has value 5 at $(x - 1, y + 2)$.

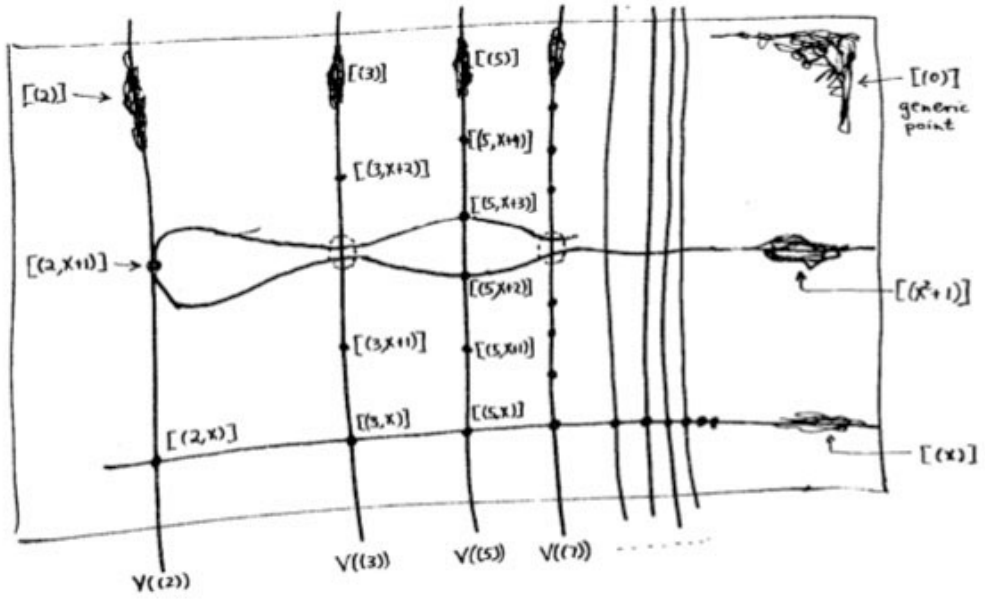
- The new bit is that we can think of evaluating f along the parabola too — it is given a particular value in the quotient $k[x, y]/(y - x^2)$. We can think of it as $f = x^2 + y^2 \equiv x^2 + x^4 \pmod{y - x^2}$ for example. Note that if we know the value of f at the generic point of the parabola, we can therefore also evaluate it at any closed point on the parabola.

- At the generic point (0), $f \pmod 0 = f$. So "evaluating at the generic point" does nothing, as in any other scheme.

§86.11 $\operatorname{Spec} \mathbb{Z}[x]$, a two-dimensional scheme, and Mumford's picture

We saw $\operatorname{Spec} \mathbb{Z}$ looked a lot like $\operatorname{Spec} k[x]$, and we will now see that $\operatorname{Spec} \mathbb{Z}[x]$ looks a lot like $\operatorname{Spec} k[x, y]$.

There is a famous picture of this scheme in Mumford's "red book", which I will produce here for culture-preservation reasons, even though there are some discrepancies between the pictures that we previously drew.



Mumford uses $[\mathfrak{p}]$ to denote the point $\mathfrak{p}$, which we don't, so you can ignore the square brackets that appear everywhere. The non-closed points are illustrated as balls of fuzz.

As before, there are three types of prime ideals, but they will look somewhat more different:

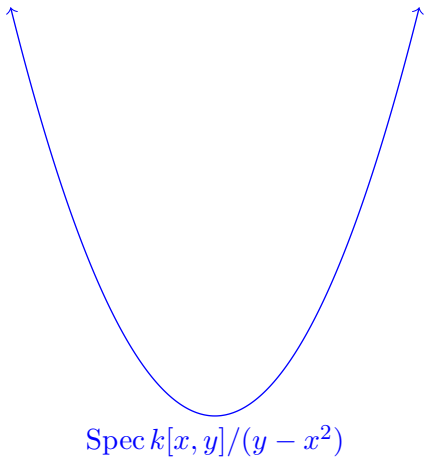
- The closed points are now pairs $(p, f(x))$ where p is a prime and f is an irreducible polynomial modulo p. Indeed, these are the maximal ideals: the quotient $\mathbb{Z}[x]/(p, f)$ becomes some finite extension of $\mathbb{F}_p$.
- There are now two different "one-dimensional" non-closed points:
  - Each rational prime gives a point (p) and
  - Each irreducible polynomial f gives a point (f).

  Indeed, note that the quotients of $\mathbb{Z}[x]$ by each are integral domains.
- $\mathbb{Z}[x]$ is an integral domain, so as always (0) is our generic point for the entire space.

There is one bit that I would do differently, in $\mathcal{V}(3)$ and $\mathcal{V}(7)$, there ought to be a point $(3, x^2+1)$, which is not drawn as a closed point in the picture, but rather as dashed oval. This is not right in the topological sense: as $\mathfrak{m} = (3, x^2+1)$ is a maximal ideal, so it really is one closed point in the scheme. But the reason it might be thought of as "doubled", is that $\mathbb{Z}[x]/(3, x^2+1)$, the residue field at $\mathfrak{m}$, is a two-dimensional $\mathbb{F}_3$ vector space.

§86.12 $\operatorname{Spec} k[x,y]/(y-x^2)$, the parabola

By Proposition 83.6.1, the prime ideals of $k[x,y]/(y-x^2)$ correspond to the prime ideals of $k[x,y]$ which are supersets of $(y-x^2)$, or equivalently the points of $\operatorname{Spec} k[x,y]$ contained inside the closed set $\mathcal{V}(y-x^2)$. Moreover, the subspace topology on $\mathcal{V}(y-x^2)$ coincides with the topology on $\operatorname{Spec} k[x,y]/(y-x^2)$.



This holds much more generally:

Exercise 86.12.1 (Boring check)**.** Show that if I is an ideal of a ring A, then $\operatorname{Spec} A/I$ is homeomorphic as a topological space to the closed subset $\mathcal{V}(I)$ of $\operatorname{Spec} A$.

So this is the notion of "closed embedding": the parabola, which was a closed subset of $\operatorname{Spec} k[x,y]$, is itself a scheme. It will be possible to say more about this, once we actually define the notion of a morphism.

The sheaf on this scheme only remembers the functions on the parabola, though: the stalks are not "inherited", so to speak. To see this, let's compute the stalk at the origin: Theorem 83.7.1 tells us it is

$$k[x,y]_{(x,y)}/(y-x^2) \cong k[x,x^2]_{(x,x^2)} \cong k[x]_{(x)}$$

which is the same as the stalk of the affine line $\operatorname{Spec} k[x]$ at the origin. Intuitively, not surprising; if one looks at any point of the parabola near the origin, it looks essentially like a line, as do the functions on it.

The stalk above the generic point is $\operatorname{Frac}(k[x,y]/(y-x^2))$: so rational functions, with the identification that $y = x^2$. Also unsurprising.

Finally, we expect the parabola is actually isomorphic to $\operatorname{Spec} k[x]$, since there is an isomorphism $k[x,y]/(y-x^2) \cong k[x]$ by sending $y \mapsto x^2$. Pictorially, this looks like "un-bending" the hyperbola. In general, we would hope that when two rings A and B are isomorphic, then $\operatorname{Spec} A$ and $\operatorname{Spec} B$ should be "the same" (otherwise we would be horrified), and we'll see later this is indeed the case.

§86.13 $\operatorname{Spec}\mathbb{Z}[i]$, the Gaussian integers (one-dimensional)

You can play on this idea some more in the integer case. Note that

$$\mathbb{Z}[i] \cong \mathbb{Z}[x]/(x^2+1)$$

which means this is a "dimension-one" closed set within $\operatorname{Spec}\mathbb{Z}[x]$. In this way, we get a scheme whose elements are *Gaussian primes*.

You can tell which closed points are "bigger" than others by looking at the residue fields. For example the residue field of the point $(2+i)$ is

$$\kappa\left((2+i)\right) = \mathbb{Z}[i]/(2+i) \cong \mathbb{F}_5$$

but the residue field of the point (3) is

$$\kappa\left((3)\right) \cong \mathbb{Z}[i]/(3) \cong \mathbb{F}_9$$

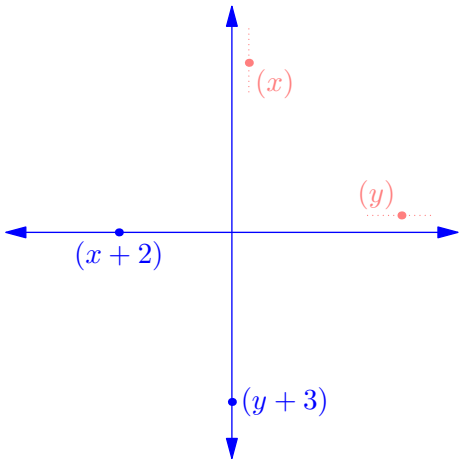
which is a degree two $\mathbb{F}_3$-extension.

§86.14 Long example: $\operatorname{Spec} k[x,y]/(xy)$, two axes

This is going to be our first example of a non-irreducible scheme.

§86.14.i Picture

Like before, topologically it looks like the closed set $\mathcal{V}(xy)$ of $\operatorname{Spec} k[x,y]$. Here is a picture:



To make sure things are making sense:

> **Question 86.14.1** (Sanity check)**.** Verify that $(y+3)$ is really a maximal ideal of $\operatorname{Spec} k[x,y]/(xy)$ lying in $\mathcal{V}(x)$.

The ideal (0) is longer prime, so it is not a point of this space. Rather, there are two non-closed points this time: the ideals (x) and (y), which can be visualized as floating around each of the two axes. This space is reducible, since it can be written as the union of two proper closed sets, $\mathcal{V}(x) \cup \mathcal{V}(y)$. (It is still *connected*, as a topological space.)

§86.14.ii Throwing out the y-axis

Consider the distinguished open set $U = D(x)$. This corresponds to deleting $\mathcal{V}(x)$, the y-axis. Therefore we expect that $D(x)$ "is" just $\operatorname{Spec} k[x]$ with the origin deleted, and in particular that we should get $k[x,x^{-1}]$ for the sections. Indeed,

$$\begin{aligned}\mathcal{O}_{\operatorname{Spec} k[x,y]/(xy)}(D(x)) &\cong (k[x,y]/(xy))[1/x] \\ &\cong k[x,x^{-1},y]/(xy) \cong k[x,x^{-1},y]/(y) \cong k[x,x^{-1}].\end{aligned}$$

where $(xy) = (y)$ follows from x being a unit. Everything as planned.

§86.14.iii Stalks above some points

Let's compute the stalk above the point $\mathfrak{m} = (x+2)$, which we think of as the point $(-2, 0)$ on the x-axis. (If it makes you more comfortable, note that $\mathfrak{m} \ni y(x+2) = 2y$ and hence $y \in \mathfrak{m}$, so we could also write $\mathfrak{m} = (x+2, y)$.) The stalk is

$$\mathcal{O}_{\operatorname{Spec} k[x,y]/(xy),\mathfrak{m}} = (k[x,y]/(xy))_{(x+2)}.$$

But I claim that y becomes the zero element with this localization. Indeed, we have $\frac{y}{1} = \frac{0}{x} = 0$. Hence the entire thing collapses to just

$$\mathcal{O}_{\operatorname{Spec} k[x,y]/(xy),\mathfrak{m}} = k[x]_{(x+2)}$$

which anyways is the stalk of $(x+2)$ in $\operatorname{Spec} k[x]$. That's expected. If we have a space with two lines but we're standing away from the origin, then the stalk is not going to pick up the weird behavior at that far-away point; it only cares about what happens near $\mathfrak{m}$, and so it looks just like an affine line there.

Remark 86.14.2 — Note that $(k[x,y]/(xy))_{(x+2)}$ is *not* the same as $k[x,y]_{(x+2)}/(xy)$; the order matters here. In fact, the latter is the zero ring, since both x and y, and hence xy, are units.

The generic point (y) (which floats around the x-axis) will tell a similar tale: if we look at the stalk above it, we ought to find that it doesn't recognize the presence of the y-axis, because "nearly all" points don't recognize it either. To actually compute the stalk:

$$\mathcal{O}_{\operatorname{Spec} k[x,y]/(xy),(y)} = (k[x,y]/(xy))_{(y)}.$$

Again $\frac{y}{1} = \frac{0}{x} = 0$, so this is just

$$\mathcal{O}_{\operatorname{Spec} k[x,y]/(xy),(y)} \cong k[x]_{(0)} \cong k(x).$$

which is what we expected (it is the same as the stalk above (0) in $\operatorname{Spec} k[x]$).

§86.14.iv Stalk above the origin (tricky)

The stalk above the origin (x, y) is interesting, and has some ideas in it we won't be able to explore fully without talking about localizations of modules. The localization is given by

$$(k[x,y]/(xy))_{(x,y)}$$

and hence the elements should be

$$\frac{c + (a_1 x + a_2 x^2 + \dots) + (b_1 y + b_2 y^2 + \dots)}{c' + (a_1' x + a_2' x^2 + \dots) + (b_1' y + b_2' y^2 + \dots)}$$

where $c' \neq 0$.

You might feel unsatisfied with this characterization. Here is some geometric intuition. You can write the global section ring as

$$k[x,y]/(xy) = c + (a_1 x + a_2 x^2 + \dots) + (b_1 y + b_2 y^2 + \dots)$$

meaning any global section is the sum of an x-polynomial and a y-polynomial. This is *not* just the ring product $k[x] \times k[y]$, though; the constant term is shared. So it's better thought of as pairs of polynomials in $k[x]$ and $k[y]$ which agree on the constant term.

If you like category theory, it is thus a fibered product

$$k[x,y]/(xy) \cong k[x] \times_k k[y]$$

with morphism $k[x] \to k$ and $k[y] \to k$ by sending x and y to zero. In that way, we can mostly decompose $k[x,y]/(xy)$ into its two components.

We really ought to be able to do the same as the stalk: we wish to say that

$$\mathcal{O}_{\operatorname{Spec} k[x,y]/(xy),(x,y)} \cong k[x]_{(x)} \times_k k[y]_{(y)}.$$

English translation: a "typical" germ ought to look like $\frac{3+x}{x^2+7} + \frac{4+y^3}{y^2+y+7}$, with the x and y parts decoupled. Equivalently, the stalk should consist of pairs of x-germs and y-germs that agree at the origin.

In fact, this is true! This might come as a surprise, but let's see why we expect this. Suppose we take the germ

$$\frac{1}{1-(x+y)}.$$

If we hold our breath, we could imagine expanding it as a geometric series: $1 + (x+y) + (x+y)^2 + \dots$. As $xy = 0$, this just becomes $1 + x + x^2 + x^3 + \dots + y + y^2 + y^3 + \dots$. This is nonsense (as written), but nonetheless it suggests the conjecture

$$\frac{1}{1-(x+y)} = \frac{1}{1-x} + \frac{1}{1-y} - 1$$

which you can actually verify is true.

> **Question 86.14.3.** Check this identity holds.

Of course, this is a lot of computation just for one simple example. Is there a way to make it general? Yes: the key claim is that "localization commutes with *limits*". You can try to work out the statement now if you want, but we won't do so.

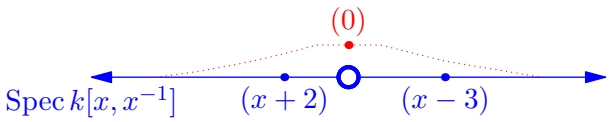
§86.15 $\operatorname{Spec} k[x,x^{-1}]$, the punctured line (or hyperbola)

This is supposed to look like $D(x)$ of $\operatorname{Spec} k[x]$, or the line with the origin deleted it. Alternatively, we could also write

$$k[x,x^{-1}] \cong k[x,y]/(xy-1)$$

so that the scheme could also be drawn as a hyperbola.

First, here's the 1D illustration.

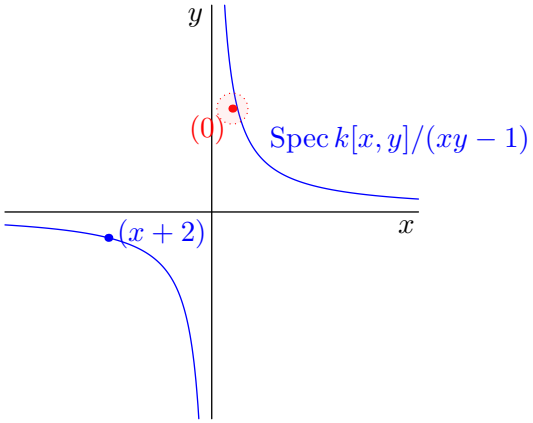


We actually saw this scheme already when we took $\operatorname{Spec} k[x,y]/(xy)$ and looked at $D(y)$, too. Anyways, let us compute the stalk at $(x-3)$ now; it is

$$\mathcal{O}_{\operatorname{Spec} k[x,x^{-1}],(x-3)} \cong k[x,x^{-1}]_{(x-3)} \cong k[x]_{(x-3)}$$

since x^{-1} is in $k[x]_{(x-3)}$ anyways. So again, we see that the deletion of the origin doesn't affect the stalk at the farther away point $(x-3)$.

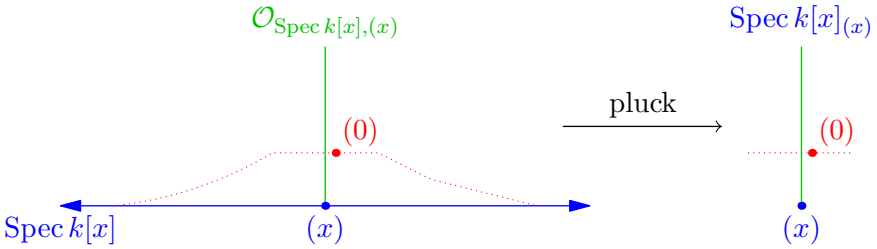
As mentioned, since $k[x, x^{-1}]$ is isomorphic to $k[x, y]/(xy - 1)$, another way to draw the visualize the same curve would be to draw the hyperbola (which you can imagine as flattening to give the punctured line.) There is one generic point (0) since $k[x, y]/(xy-1)$ really is an integral domain, as well as points like $(x + 2, y + 1/2) = (x + 2) = (y + 1/2)$.



§86.16 Spec $k[x]_{(x)}$, zooming in to the origin of the line

We know already that $\mathcal{O}_{\operatorname{Spec} A, \mathfrak{p}} \cong A_{\mathfrak{p}}$: so $A_{\mathfrak{p}}$ should be the stalk at $\mathfrak{p}$. In this example we will see that $\operatorname{Spec} A_{\mathfrak{p}}$ should be drawn sort of as this stalk, too.

We saw earlier how to draw a picture of $\operatorname{Spec} k[x]$. You can also draw a picture of the stalk above the origin (x), which you might visualize as a grass or other plant growing above (x) if you like agriculture. In that case, $\operatorname{Spec} k[x]_{(x)}$ might look like what happens if you pluck out that stalk from the affine line.



Since $k[x]_{(x)}$ is a local ring (it is the localization of a prime ideal), this point has only one closed point: the maximal ideal (x). However, surprisingly, it has one more point: a "generic" point (0). So $\operatorname{Spec} k[x]_{(x)}$ is a *two-point space*, but it does not have the discrete topology: (x) is a closed point, but (0) is not. (This makes it a nice counter-example for exercises of various sorts.)

So, topologically what's happening is that when we zoom in to (x), the generic point (0) (which was "close to every point") remains, floating above the point (x).

Note that the stalk above our single closed point (x) is the same as it was before:

$$\left(k[x]_{(x)}\right)_{(x)} \cong k[x]_{(x)}.$$

Indeed, in general if R is a local ring with maximal ideal $\mathfrak{m}$, then $R_{\mathfrak{m}} \cong R$: since every element $x \notin \mathfrak{m}$ was invertible anyways. Thus in the picture, the stalk is drawn the same.

Similarly, the stalk above (0) is the same as it was before we plucked it out:

$$\left(k[x]_{(x)}\right)_{(0)} = \operatorname{Frac} k[x]_{(x)} = k(x).$$

More generally:

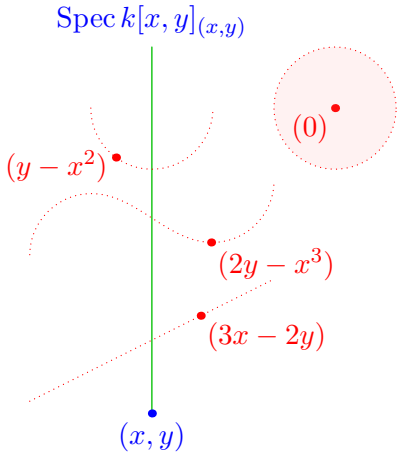
Exercise 86.16.1. Let A be a ring, and $\mathfrak{q} \subseteq \mathfrak{p}$ prime ideals. Check that $A_{\mathfrak{q}} \cong (A_{\mathfrak{p}})_{\mathfrak{q}}$, where we view $\mathfrak{q}$ as a prime ideal of $A_{\mathfrak{p}}$.

So when we zoom in like this, all the stalks stay the same, even above the non-closed points.

§86.17 $\operatorname{Spec} k[x,y]_{(x,y)}$, zooming in to the origin of the plane

The situation is more surprising if we pluck the stalk above the origin of $\operatorname{Spec} k[x,y]$, the two-dimensional plane. The points of $\operatorname{Spec} k[x,y]_{(x,y)}$ are supposed to be the prime ideals of $k[x,y]$ which are contained in (x,y); geometrically these are (x,y) and the generic points passing through the origin. For example, there will be a generic point for the parabola $(y-x^2)$ contained in $k[x,y]_{(x,y)}$, and another one $(y-x)$ corresponding to a straight line, etc.

So we have the single closed point (x,y) sitting at the bottom, and all sorts of "one-dimensional" generic points floating above it: lines, parabolas, you name it. Finally, we have (0), a generic point floating in two dimensions, whose closure equals the entire space.



§86.18 $\operatorname{Spec} k[x,y]_{(0)} = \operatorname{Spec} k(x,y)$, the stalk above the generic point

The generic point of the plane just has stalk $\operatorname{Spec} k(x,y)$: which is the spectrum of a field, hence a single point. The stalk remains intact as compared to when planted in $\operatorname{Spec} k[x,y]$; the functions are exactly rational functions in x and y.

§86.19 A few harder problems to think about

Problem 86A. Draw a picture of $\operatorname{Spec} \mathbb{Z}[1/55]$, describe the topology, and compute the stalk at each point.

Problem 86B. Draw a picture of $\operatorname{Spec} \mathbb{Z}_{(5)}$, describe the topology, and compute the stalk at each point.

Problem 86C. Let $A = (k[x,y]/(xy))[(x+y)^{-1}]$. Draw a picture of $\operatorname{Spec} A$. Show that it is not connected as a topological space.

Problem 86D. Let $A = k[x, y]_{(y-x^2)}$. Draw a picture of Spec A.

87 Morphisms of locally ringed spaces

Having set up the definition of a locally ringed space, we are almost ready to define morphisms between them. Throughout this chapter, you should imagine your ringed spaces are the affine schemes we have so painstakingly defined; but it will not change anything to work in the generality of arbitrary locally ringed spaces.

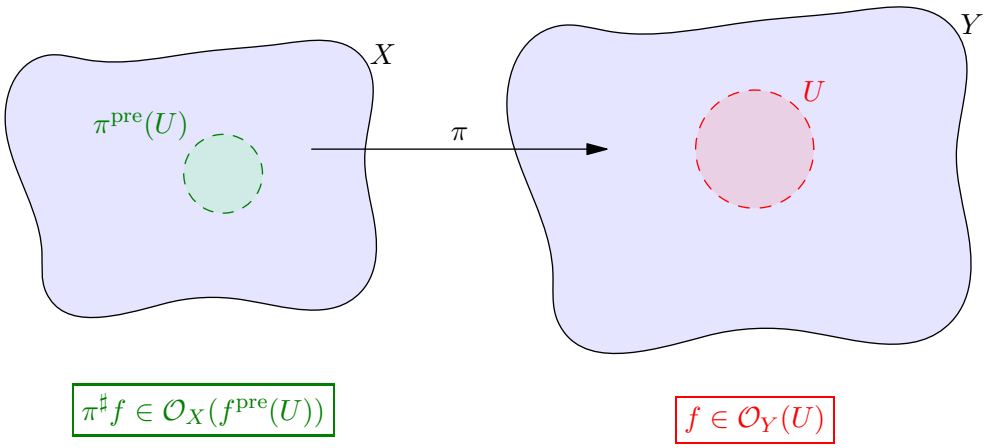
§87.1 Morphisms of ringed spaces via sections

Let $(X, \mathcal{O}_X)$ and $(Y, \mathcal{O}_Y)$ be ringed spaces. We want to give a define what it means to have a function $\pi\colon X \to Y$ between them.[1] We start by requiring the map to be continuous, but this is not enough: there is a sheaf on it!

Well, you might remember what we did for baby ringed spaces: any time we had a function on an open set of $U \subseteq Y$, we wanted there to be an analogous function on $\pi^{\text{pre}}(U) \subseteq X$. For baby ringed spaces, this was done by composition, since the elements of the sheaf *were* really complex valued functions:

$$\pi^\sharp\phi \text{ was defined as } \phi \circ \pi.$$

The upshot was that we got a map $\mathcal{O}_Y(U) \to \mathcal{O}_X(\pi^{\text{pre}}(U))$ for every open set U.



Now, for general locally ringed spaces, the sections are just random rings, which may not be so well-behaved. So the solution is that we *include* the data of $f^\sharp$ as part of the definition of a morphism.

> **Remark 87.1.1 —** As we will see in Example 87.4.2, unlike the situation in algebraic varieties where the morphism is uniquely determined by the map of topological space, here $\pi^\sharp$ is not necessarily uniquely determined by the map π. Thus, including the $\pi^\sharp$ is necessary.

Definition 87.1.2. A **morphism of ringed spaces** $(X, \mathcal{O}_X) \to (Y, \mathcal{O}_Y)$ consists of a pair $(\pi, \pi^\sharp)$ where $\pi\colon X \to Y$ is a continuous map (of topological spaces), and $\pi^\sharp$ consists of a choice of ring homomorphism

$$\pi^\sharp_U\colon \mathcal{O}_Y(U) \to \mathcal{O}_X(\pi^{\text{pre}}(U))$$

[1] Notational connotations: for ringed spaces, π will be used for maps, since f is often used for sections.

for every open set $U \subseteq Y$, such that the restriction diagram

$$\begin{array}{ccc} \mathcal{O}_Y(U) & \longrightarrow & \mathcal{O}_X(\pi^{\text{pre}}(U)) \\ \downarrow & & \downarrow \\ \mathcal{O}_Y(V) & \longrightarrow & \mathcal{O}_X(\pi^{\text{pre}}(V)) \end{array}$$

commutes for $V \subseteq U$.

Abuse of Notation 87.1.3. We will abbreviate $(\pi, \pi^\sharp) \colon (X, \mathcal{O}_X) \to (Y, \mathcal{O}_Y)$ to just $\pi \colon X \to Y$, despite the fact this notation is exactly the same as that for topological spaces.

There is an obvious identity map, and so we can also define isomorphism etc. in the categorical way.

§87.2 Morphisms of ringed spaces via stalks

Unsurprisingly, the sections are clumsier to work with than the stalks, now that we have grown to love localization. So rather than specifying $\pi^\sharp_U$ on every open set U, it seems better if we could do it by stalks (there are fewer stalks than open sets, so this saves us a lot of work!).

We start out by observing that we *do* get a morphism of stalks.

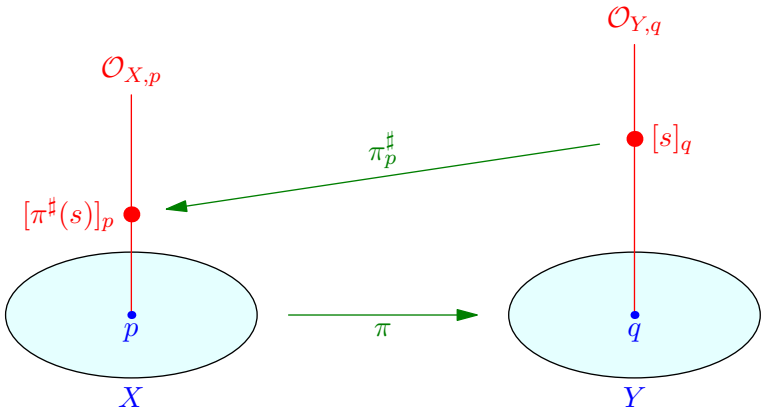
Proposition 87.2.1 (Induced stalk morphisms)

If $\pi \colon X \to Y$ is a map of ringed spaces sending $\pi(p) = q$, then we get a map

$$\pi^\sharp_p \colon \mathcal{O}_{Y,q} \to \mathcal{O}_{X,p}$$

whenever $\pi(p) = q$.

This means you can draw a morphism of locally ringed spaces as a continuous map on the topological space, plus for each $\pi(p) = q$, an assignment of each germ at q to a germ at p.



Again, compare this to the pullback picture: this is roughly saying that if a function f has some enriched value at q, then $\pi^\sharp(f)$ should be assigned a corresponding enriched value at p. The analogy is not perfect since the stalks at q and p may be different rings in general, but there should at least be a ring homomorphism (the assignment).

Proof. If (s, U) is a germ at q, then $(\pi^\sharp(s), \pi^{\text{pre}}(U))$ is a germ at p, and this is a well-defined morphism because of compatibility with restrictions. □

We already obviously have uniqueness in the following senes.

Proposition 87.2.2 (Uniqueness of morphisms via stalks)
Consider a map of ringed spaces $(\pi, \pi^\sharp)\colon (X, \mathcal{O}_X) \to (Y, \mathcal{O}_Y)$ and the corresponding map $\pi_p^\sharp$ of stalks. Then $\pi^\sharp$ is uniquely determined by $\pi_p^\sharp$.

Proof. Given a section $s \in \mathcal{O}_Y(U)$, let

$$t = \pi_U^\sharp(s) \in \mathcal{O}_X(\pi^{\text{pre}}(U))$$

denote the image under $\pi^\sharp$.

We know t_p for each $p \in \pi^{\text{pre}}(U)$, since it equals $\pi_p^\sharp(t)$ by definition. That is, we know all the germs of t. So we know t. □

However, it seems clear that not every choice of stalk morphisms will lead to $\pi_U^\sharp$: some sort of "continuity" or "compatibility" is needed. You can actually write down the explicit statement: each sequence of compatible germs over U should get mapped to a sequence of compatible germs over $\pi^{\text{pre}}(U)$. We avoid putting up a formal statement of this for now, because the statement is clumsy, and you're about to see it in practice (where it will make more sense).

Remark 87.2.3 (Isomorphisms are determined by stalks) — One fact worth mentioning, that we won't prove, but good to know: a map of ringed spaces $(\pi, \pi^\sharp)\colon (X, \mathcal{O}_X) \to (Y, \mathcal{O}_Y)$ is an isomorphism if and only if π is a homeomorphism, and moreover $\pi_p^\sharp$ is an isomorphism for each $p \in X$.

§87.3 Morphisms of locally ringed spaces

On the other hand, we've seen that our stalks are local rings, which enable us to actually talk about *values*. And so we want to add one more compatibility condition to ensure that our notion of value is preserved. Now the stalks at p and q in the previous picture might be different, so $\kappa(p)$ and $\kappa(q)$ might even be different fields.

Definition 87.3.1. A **morphism of locally ringed spaces** is a morphism of ringed spaces $\pi\colon X \to Y$ with the following additional property: whenever $\pi(p) = q$, the map at the stalks also induces a well-defined ring homomorphism

$$\pi_p^\sharp\colon \kappa(q) \to \kappa(p).$$

So we require $\pi_p^\sharp$ induces a field homomorphism[2] on the *residue fields*. In particular, since $\pi^\sharp(0) = 0$, this means something very important:

In a morphism of locally ringed spaces, a germ vanishes at q if and only if the corresponding germ vanishes at p.

[2]Which means it is automatically injective, by Problem 5B.

Exercise 87.3.2 (So-called "local ring homomorphism")**.** Show that this is equivalent to requiring

$$(\pi_p^\sharp)^{\text{img}}(\mathfrak{m}_{Y,q}) \subseteq \mathfrak{m}_{X,p}$$

or in English, a germ at q has value zero iff the corresponding germ at p has value zero.

I don't like this formulation $(\pi^\sharp)^{\text{img}}(\mathfrak{m}_{Y,q}) \subseteq \mathfrak{m}_{X,p}$ as much since it hides the geometric intuition behind a lot of symbols: that we want the notion of "value at a point" to be preserved in some way.

At this point, we can state the definition of a scheme, and we do so, although we won't really use it for a few more sections.

Definition 87.3.3. A **scheme** is a locally ringed space for which every point has an open neighborhood isomorphic to an affine scheme. A morphism of schemes is just a morphism of locally ringed spaces.

In particular, $\operatorname{Spec} A$ is a scheme (the open neighborhood being the entire space!). And so let's start by looking at those.

§87.4 A few examples of morphisms between affine schemes

Okay, sorry for lack of examples in previous few sections. Let's make amends now, where you can see all the moving parts in action.

§87.4.i One-point schemes

Example 87.4.1 ($\operatorname{Spec}\mathbb{R}$ is well-behaved)

There is only one map $X = \operatorname{Spec}\mathbb{R} \to \operatorname{Spec}\mathbb{R} = Y$. Indeed, these are spaces with one point, and specifying the map $\mathbb{R} = \mathcal{O}_Y(Y) \to \mathcal{O}_X(X) = \mathbb{R}$ can only be done in one way, since there is only one field automorphism of $\mathbb{R}$ (the identity).

Example 87.4.2 ($\operatorname{Spec}\mathbb{C}$ horror story)

There are multiple maps $X = \operatorname{Spec}\mathbb{C} \to \operatorname{Spec}\mathbb{C} = Y$, horribly enough! Indeed, these are spaces with one point, so again we're just reduced to specifying a map $\mathbb{C} = \mathcal{O}_Y(Y) \to \mathcal{O}_X(X) = \mathbb{C}$. However, in addition to the identity map, complex conjugation also works, as well as some so-called "wild automorphisms" of $\mathbb{C}$.

This behavior is obviously terrible, so for illustration reasons, some of the examples use $\mathbb{R}$ instead of $\mathbb{C}$ to avoid the horror story we just saw. However, there is an easy fix using "scheme over $\mathbb{C}$" which will force the ring homomorphisms to fix $\mathbb{C}$, later.

Example 87.4.3 ($\operatorname{Spec} k$ and $\operatorname{Spec} k'$)

In general, if k and k' are fields, we see that maps $\operatorname{Spec} k \to \operatorname{Spec} k'$ are in bijection with field homomorphism $k' \to k$, since that's all there is left to specify.

§87.4.ii Examples of constant maps

Example 87.4.4 (Constant map to $(y-3)$)
We analyze scheme morphisms

$$X = \operatorname{Spec} \mathbb{R}[x] \xrightarrow{\pi} \operatorname{Spec} \mathbb{R}[y] = Y$$

which send all points of X to $\mathfrak{m} = (y-3) \in Y$. Constant maps are continuous no matter how bizarre your topology is, so this lets us just focus our attention on the sections.
This example is simple enough that we can even do it by sections, as much as I think stalks are simpler. Let U be any open subset of Y, then we need to specify a map

$$\pi^\sharp_U \colon \mathcal{O}_Y(U) \to \mathcal{O}_X(\pi^{\text{pre}}(U)).$$

If U does not contain $(y-3)$, then $\pi^{\text{pre}}(U) = \varnothing$, so $\mathcal{O}_X(\varnothing) = 0$ is the zero ring and there is nothing to do.
Conversely, if U does contain $(y-3)$ then $\pi^{\text{pre}}(U) = X$, so this time we want to specify a map

$$\pi^\sharp_U \colon \mathcal{O}_Y(U) \to \mathcal{O}_X(X) = \mathbb{R}[x]$$

which satisfies restriction maps. Note that for any U, the element y must be mapped to a unit in $\mathbb{R}[x]$; since $1/y$ is a section too for a subset of U not containing (y). In more detail, let $W = U \cap D(y)$ so that $(y) \notin W$, then

$$\pi^\sharp_W(y) = \pi^\sharp_U(y) \quad \text{and} \quad \pi^\sharp_W(y)\pi^\sharp_W(1/y) = 1.$$

Actually for any real number $c \neq 3$, $y - c$ must be mapped to a unit in $\mathbb{R}[x]$. This can only happen if $y \mapsto 3 \in \mathbb{R}[x]$.
As we have specified $\mathbb{R}[y] \to \mathbb{R}[x]$ with $y \mapsto 3$, that determines all the ring homomorphisms we needed.

But we could have used stalks, too. We wanted to specify a morphism

$$\mathbb{R}[y]_{(y-3)} = \mathcal{O}_{\operatorname{Spec} Y,(y-3)} \to \mathcal{O}_{\operatorname{Spec} X,\mathfrak{p}}$$

for every prime ideal $\mathfrak{p}$, sending compatible germs to compatible germs... but wait, $(y-3)$ is spitting out all the germs. So every *individual* germ in $\mathcal{O}_{\operatorname{Spec} Y,(y-3)}$ needs to yield a (compatible) germ above every point of $\operatorname{Spec} X$, which is the data of an entire global section. So we're actually trying to specify

$$\mathbb{R}[y]_{(y-3)} = \mathcal{O}_{\operatorname{Spec} Y,(y-3)} \to \mathcal{O}_{\operatorname{Spec} X}(\operatorname{Spec} X) = \mathbb{R}[x].$$

This requires $y \mapsto 3$, as we saw, since $y - c$ is a unit of $\mathbb{R}[x]$ for any $c \neq 3$.

Example 87.4.5 (Constant map to (y^2+1) does not exist)
Let's see if there are constant maps $X = \operatorname{Spec} \mathbb{R}[x] \to \operatorname{Spec} \mathbb{R}[y] = Y$ which send everything to (y^2+1). Copying the previous example, we see that we want

$$\mathcal{O}_Y(U) \to \mathcal{O}_X(X) = \mathbb{R}[x].$$

We find that y and $1/y$ have nowhere to go: the same argument as last time shows

that $y - c$ should be a unit of $\mathbb{R}[x]$; this time for any real number c.
Like this time, stalks show this too, even with just residue fields. We would for example need a field homomorphism

$$\mathbb{C} = \kappa((y^2+1)) \to \kappa((x)) = \mathbb{R}$$

which does not exist.

You might already notice the following:

Example 87.4.6 (The generic point repels smaller points)
Changing the tune, consider maps $\operatorname{Spec}\mathbb{C}[x] \to \operatorname{Spec}\mathbb{C}[y]$. We claim that if $\mathfrak{m}$ is a maximal ideal (closed point) of $\mathbb{C}[x]$, then it can never be mapped to the generic point (0) of $\mathbb{C}[y]$.
For otherwise, we would get a local ring homomorphism

$$\mathbb{C}(y) \cong \mathcal{O}_{\operatorname{Spec} k[y],(0)} \to \mathcal{O}_{\operatorname{Spec}\mathbb{C}[x],\mathfrak{m}} \cong \mathbb{C}[x]_{\mathfrak{m}}$$

which in particular means we have a map on the residue fields

$$\mathbb{C}(y) \to \mathbb{C}[x]/\mathfrak{m} \cong \mathbb{C}$$

which is impossible, there is no such field homomorphism at all (why?).

The last example gives some nice intuition in general: "more generic" points tend to have bigger stalks than "less generic" points, hence repel them.

§87.4.iii The map $t \mapsto t^2$

We now consider what we would think of as the map $t \mapsto t^2$.

Example 87.4.7 (The map $t \mapsto t^2$)
We consider a map

$$\pi \colon X = \operatorname{Spec}\mathbb{C}[x] \to \operatorname{Spec}\mathbb{C}[y] = Y$$

defined on points as follows:

$$\pi((0)) = (0)$$
$$\pi((x-a)) = (y-a^2).$$

You may check if you wish this map is continuous. I claim that, surprisingly, you can actually read off $\pi^\sharp$ from just this behavior at points. The reason is that we imposed the requirement that a section s can vanish at $\mathfrak{q} \in Y$ if and only if $\pi_X^\sharp(s)$ vanishes at $\mathfrak{p} \in X$, where $\pi(\mathfrak{p}) = \mathfrak{q}$. So, now:

- Consider the section $y \in \mathcal{O}_Y(Y)$, which vanishes only at $(y) \in \operatorname{Spec}\mathbb{C}[y]$; then its image $\pi_Y^\sharp(y) \in \mathcal{O}_X(X)$ must vanish at exactly $(x) \in \operatorname{Spec}\mathbb{C}[x]$, so $\pi_Y^\sharp(y) = x^n$ for some integer $n \geq 1$.

- Consider the section $y-4 \in \mathcal{O}_Y(Y)$, which vanishes only at $(y-4) \in \operatorname{Spec} \mathbb{C}[y]$; then its image $\pi_Y^\sharp(y-4) \in \mathcal{O}_X(X)$ must vanish at exactly $(x-2) \in \operatorname{Spec} \mathbb{C}[x]$ and $(x+2) \in \operatorname{Spec} \mathbb{C}[x]$. So $\pi_Y^\sharp(y)-4$ is divisible by $(x-2)^a(x+2)^b$ for some $a \geq 1$ and $b \geq 1$.

Thus $y \mapsto x^2$ in the top level map of sections $\pi_Y^\sharp$: and hence also in all the maps of sections (as well as at all the stalks).

The above example works equally well if t^2 is replaced by some polynomial $f(t)$, so that $(x-a)$ maps to $(y-f(y))$. The image of y must be a polynomial $g(x)$ with the property that $g(x)-c$ has the same roots as $f(x)-c$ for any $c \in \mathbb{C}$. Put another way, f and g have the same values, so $f=g$.

Remark 87.4.8 (Generic point stalk overpowered) — I want to also point out that you can read off the polynomial just from the stalk at the generic point: for example, the previous example has

$$\mathbb{C}(y) \cong \mathcal{O}_{\operatorname{Spec} \mathbb{C}[y],(0)} \to \mathcal{O}_{\operatorname{Spec} \mathbb{C}[x],(0)} \cong \mathbb{C}(x) \qquad y \mapsto x^2.$$

This is part of the reason why generic points are so powerful. We expect that with polynomials, if you know what happens to a "generic" point, you can figure out the entire map. This intuition is true: knowing where each germ at the generic point goes is enough to tell us the whole map.

§87.4.iv An arithmetic example

Example 87.4.9 ($\operatorname{Spec} \mathbb{Z}[i] \to \operatorname{Spec} \mathbb{Z}$)

We now construct a morphism of schemes $\pi\colon \operatorname{Spec} \mathbb{Z}[i] \to \operatorname{Spec} \mathbb{Z}$. On points it behaves by

$$\begin{aligned} \pi((0)) &= (0) \\ \pi((p)) &= (p) \\ \pi((a+bi)) &= (a^2+b^2) \end{aligned}$$

where $a+bi$ is a Gaussian prime: so for example $\pi((2+i)) = (5)$ and $\pi((1+i)) = (2)$. We could figure out the induced map on stalks now, much like before, but in a moment we'll have a big theorem that spares us the trouble.

§87.5 The big theorem

We did a few examples of $\operatorname{Spec} A \to \operatorname{Spec} B$ by hand, specifying the full data of a map of locally ringed spaces. It turns out that in fact, we didn't to specify that much data, and much of the process can be automated:

Proposition 87.5.1 (Affine reconstruction)

Let $\pi\colon \operatorname{Spec} A \to \operatorname{Spec} B$ be a map of schemes. Let $\psi\colon B \to A$ be the ring homomorphism obtained by taking global sections, i.e.

$$\psi = \pi_B^\sharp\colon \mathcal{O}_{\operatorname{Spec} B}(\operatorname{Spec} B) \to \mathcal{O}_{\operatorname{Spec} A}(\operatorname{Spec} A).$$

Then we can recover π given only ψ; in fact, π is given explicitly by

$$\pi(\mathfrak{p}) = \psi^{\text{pre}}(\mathfrak{p})$$

and

$$\pi_{\mathfrak{p}}^\sharp\colon \mathcal{O}_{Y,\pi(\mathfrak{p})} \to \mathcal{O}_{X,\mathfrak{p}} \quad \text{by} \quad f/g \mapsto \psi(f)/\psi(g).$$

This is the big miracle of affine schemes. Despite the enormous amount of data packaged into the definition, we can compress maps between affine schemes into just the single ring homomorphism on the top level.

Proof. This requires two parts.

- We need to check that the maps agree on *points*; surprisingly this is the harder half. To see how this works, let $\mathfrak{q} = \pi(\mathfrak{p})$. The key fact is that a function $f \in B$ vanishes on $\mathfrak{q}$ if and only if $\pi_B^\sharp(f)$ vanishes on $\mathfrak{p}$ (because $\pi_B^\sharp$ is supposed to be a homomorphism of *local* rings). Therefore,

$$\begin{aligned}
\pi(\mathfrak{p}) = \mathfrak{q} &= \{f \in B \mid f \in \mathfrak{q}\} \\
&= \{f \in B \mid f \text{ vanishes on } \mathfrak{q}\} \\
&= \left\{f \in B \mid \pi_B^\sharp(f) \text{ vanishes on } \mathfrak{p}\right\} \\
&= \left\{f \in B \mid \pi_B^\sharp(f) \in \mathfrak{p}\right\} = \{f \in B \mid \psi(f) \in \mathfrak{p}\} \\
&= \psi^{\text{pre}}(\mathfrak{p}).
\end{aligned}$$

- We also want to check the maps on the stalks is the same. Suppose $\mathfrak{p} \in \operatorname{Spec} A$, $\mathfrak{q} \in \operatorname{Spec} B$, and $\mathfrak{p} \mapsto \mathfrak{q}$ (under both of the above).

  In our original π, consider the map $\pi_{\mathfrak{p}}^\sharp\colon B_{\mathfrak{q}} \to A_{\mathfrak{p}}$. We know that it sends each $f \in B$ to $\psi(f) \in A$, by taking the germ of each global section $f \in B$ at $\mathfrak{q}$. Thus it must send f/g to $\psi(f)/\psi(g)$, being a ring homomorphism, as needed. □

All of this suggests a great idea: if $\psi\colon B \to A$ is *any* ring homomorphism, we ought to be able to construct a map of schemes by using fragments of the proof we just found. The only extra work we have to do is verify that we get a continuous map in the Zariski topology, and that we can get a suitable $\pi^\sharp$.

We thus get the huge important theorem about affine schemes.

Theorem 87.5.2 ($\operatorname{Spec} A \to \operatorname{Spec} B$ is just $B \to A$)

These two construction gives a bijection between ring homomorphisms $B \to A$ and $\operatorname{Spec} A \to \operatorname{Spec} B$.

Proof. We have seen how to take each $\pi\colon \operatorname{Spec} A \to \operatorname{Spec} B$ and get a ring homomorphism ψ. Proposition 87.5.1 shows this map is injective. So we just need to check it is surjective — that every ψ arises from some π.

Given $\psi\colon B \to A$, we define $(\pi, \pi^\sharp)\colon \operatorname{Spec} A \to \operatorname{Spec} B$ by copying Proposition 87.5.1 and checking that everything is well-definced. The details are:

- For each prime ideal $\mathfrak{p} \in \operatorname{Spec} A$, we let $\pi(\mathfrak{p}) = \psi^{\text{pre}}(\mathfrak{p}) \in \operatorname{Spec} B$ (which by Problem 5C* is also prime).

  > **Exercise 87.5.3.** Show that the resulting map π is continuous in the Zariski topology.

- Now we want to also define maps on the stalks, and so for each $\pi(\mathfrak{p}) = \mathfrak{q}$ we set
$$B_{\mathfrak{q}} \ni \frac{f}{g} \mapsto \frac{\psi(f)}{\psi(g)} \in A_{\mathfrak{p}}.$$
This makes sense since $g \notin \mathfrak{q} \implies \psi(g) \notin \mathfrak{p}$. Also $f \in \mathfrak{q} \implies \psi(f) \in \mathfrak{p}$, we find this really is a local ring homomorphism (sending the maximal ideal of $B_{\mathfrak{q}}$ into the one of $A_{\mathfrak{p}}$).

  Observe that if f/g is a *section* over an open set $U \subseteq B$ (meaning g does not vanish at the primes in U), then $\psi(f)/\psi(g)$ is a section over $\pi^{\text{pre}}(U)$ (meaning $\psi(g)$ does not vanish at the primes in $\pi^{\text{pre}}(U)$). Therefore, compatible germs over B get sent to compatible germs over A, as needed.

Finally, the resulting π has $\pi^\sharp_B = \psi$ on global sections, completing the proof. □

This can be summarized succinctly using category theory:

Corollary 87.5.4 (Categorical interpretation)
The opposite category of rings $\mathsf{CRing}^{\text{op}}$, is "equivalent" to the category of affine schemes, AffSch, with Spec as a functor.

This means for example that $\operatorname{Spec} A \cong \operatorname{Spec} B$, naturally, whenever $A \cong B$.

To make sure you realize that this theorem is important, here is an amusing comment I found on MathOverflow while reading about algebraic geometry references[3]:

> *He [Hartshorne] never mentions that the category of affine schemes is dual to the category of rings, as far as I can see. I'd expect to see that in huge letters near the definition of scheme. How could you miss that out?*

§87.6 More examples of scheme morphisms

Now that we have the big hammer, we can talk about examples much more briefly than we did a few sections ago. Before throwing things around, I want to give another definition that will eliminate the weird behavior we saw with $\mathbb{C} \to \mathbb{C}$ having nontrivial field automorphisms:

Definition 87.6.1. Let S be a scheme. A **scheme over S** or **S-scheme** is a scheme X together with a map $X \to S$. A morphism of S-schemes is a scheme morphism $X \to Y$ such that the diagram
$$\begin{array}{ccc} X & \longrightarrow & Y \\ & \searrow & \downarrow \\ & & S \end{array}$$
commutes. Often, if $S = \operatorname{Spec} k$, we will refer to X by schemes over k or k-schemes for short.

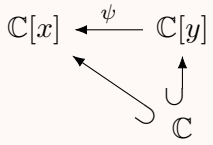
[3] From https://mathoverflow.net/q/2446/70654.

Example 87.6.2 (Spec $k[\dots]$)

If $X = \operatorname{Spec} k[x_1, \dots, x_n]/I$ for some ideal I, then X is a k-scheme in a natural way; since we have an obvious homomorphism $k \hookrightarrow k[x_1, \dots, x_n]/I$ which gives a map $X \to \operatorname{Spec} k$.

Example 87.6.3 (Spec $\mathbb{C}[x] \to$ Spec $\mathbb{C}[y]$)

As $\mathbb{C}$-schemes, maps $\operatorname{Spec} \mathbb{C}[x] \to \operatorname{Spec} \mathbb{C}[y]$ coincide with ring homomorphisms from $\psi \colon \mathbb{C}[y] \to \mathbb{C}[x]$ such that the diagram



commutes. We see that the "over $\mathbb{C}$" condition is eliminating the pathology from before: the ψ is required to preserve $\mathbb{C}$. So the morphism is determined by the image of y, i.e. the choice of a polynomial in $\mathbb{C}[x]$. For example, if $\psi(y) = x^2$ we recover the first example we saw. This matches our intuition that these maps should correspond to polynomials.

Example 87.6.4 (Spec $\mathcal{O}_K$)

This generalizes Spec $\mathbb{Z}[i]$ from before. If K is a number field and $\mathcal{O}_K$ is the ring of integers, then there is a natural morphism $\operatorname{Spec} \mathcal{O}_K \to \operatorname{Spec} \mathbb{Z}$ from the (unique) ring homomorphism $\mathbb{Z} \hookrightarrow \mathcal{O}_K$. Above each rational prime $(p) \in \mathbb{Z}$, one obtains the prime ideals that p splits as. (We don't have a way of capturing ramification yet, alas.)

§87.7 A little bit on non-affine schemes

We can finally state the isomorphism that we wanted for a long time (first mentioned in Section 85.2.iii):

Theorem 87.7.1 (Distinguished open sets are isomorphic to affine schemes)

Let A be a ring and f an element. Then

$$\operatorname{Spec} A[1/f] \cong D(f) \subseteq \operatorname{Spec} A.$$

Proof. Annoying check, not included yet. (We have already seen the bijection of prime ideals, at the level of points.) □

Corollary 87.7.2 (Open subsets are schemes)

(a) Any nonempty open subset of an affine scheme is itself a scheme.

(b) Any nonempty open subset of any scheme (affine or not) is itself a scheme.

Proof. Part (a) has essentially been done already:

> **Question 87.7.3.** Combine Theorem 85.2.2 with the previous proposition to deduce (a).

Part (b) then follows by noting that if U is an open set, and p is a point in U, then we can take an affine open neighborhood $\operatorname{Spec} A$ at p, and then cover $U \cap \operatorname{Spec} A$ with distinguished open subsets of $\operatorname{Spec} A$ as in (a). □

We now reprise Section 85.2.iv (except $\mathbb{C}$ will be replaced by k). We have seen it is an open subset U of $\operatorname{Spec} k[x,y]$, so it is a scheme.

> **Question 87.7.4.** Show that in fact U can be covered by two open sets which are both affine.

However, we show now that you really do need two distinguished open sets.

Proposition 87.7.5 (Famous example: punctured plane isn't affine)
The punctured plane $U = (U, \mathcal{O}_U)$, obtained by deleting (x,y) from $\operatorname{Spec} k[x,y]$, is not isomorphic to any affine scheme $\operatorname{Spec} B$.

The intuition is that $\mathcal{O}_U(U) = k[x,y]$, but U is not the plane.

Proof. We already know $\mathcal{O}_U(U) = k[x,y]$ and we have a good handle on it. For example, $y \in \mathcal{O}_U(U)$ is a global section which vanishes on what looks like the y-axis. Similarly, $x \in \mathcal{O}_X(X)$ is a global section which vanishes on what looks like the y-axis. In particular, no point of U vanishes at both.

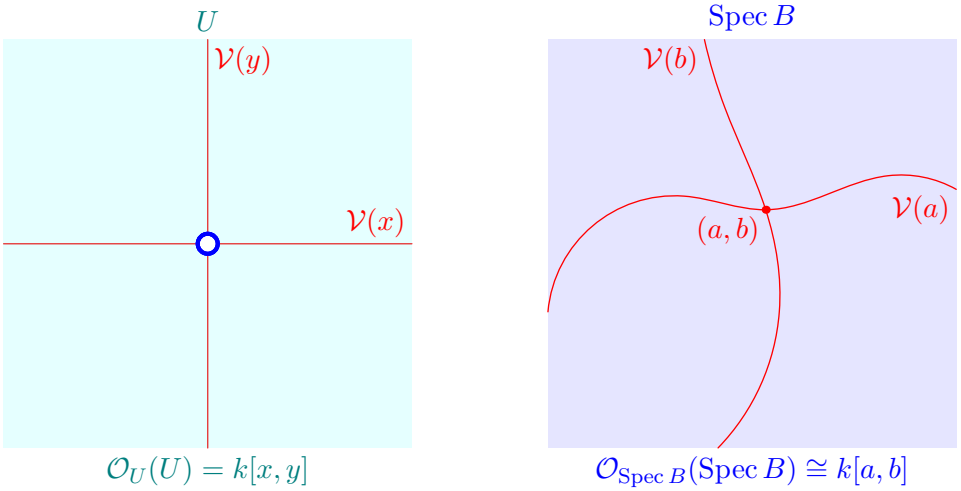
Now assume for contradiction that we have an isomorphism

$$\psi \colon \operatorname{Spec} B \to U.$$

By taking the map on global sections (part of the definition),

$$k[x,y] = \mathcal{O}_U(U) \xrightarrow{\psi^\sharp} \mathcal{O}_{\operatorname{Spec} B}(\operatorname{Spec} B) \cong B.$$

The global sections x and y in $\mathcal{O}_U(U)$ should then have images a and b in B; and it follows we have a ring isomorphism $B \cong k[a,b]$.



Now in $\operatorname{Spec} B$, $\mathcal{V}(a) \cap \mathcal{V}(b)$ is a closed set containing a single point, the maximal ideal $\mathfrak{m} = (a, b)$. Thus in $\operatorname{Spec} B$ there is exactly one point vanishing at both a and b. *Because we required morphisms of schemes to preserve values* (hence the big fuss about locally ringed spaces), that means there should be a single point of U vanishing at both x and y. But there isn't — it was the origin we deleted. □

§87.8 Where to go from here

This chapter concludes the long setup for the definition of a scheme. For now, this unfortunately is as far as I have time to go. So, if you want to actually see how schemes are used in "real life", you'll have to turn elsewhere.

A good reference I happily recommend is [Ga03]; a more difficult (and famous) one is [Va17]. See Appendix A for further remarks.

§87.9 A few harder problems to think about

Problem 87A. Given an affine scheme $X = \operatorname{Spec} R$, show that there is a unique morphism of schemes $X \to \operatorname{Spec} \mathbb{Z}$, and describe where it sends points of X.

Problem 87B. Is the open subset of $\operatorname{Spec} \mathbb{Z}[x]$ obtained by deleting the point $\mathfrak{m} = (2, x)$ isomorphic to some affine scheme?

XXI

Set Theory I: ZFC, Ordinals, and Cardinals

Part XXI: Contents

88 Interlude: Cauchy's functional equation and Zorn's lemma

This is an informal chapter on Zorn's lemma, which will give an overview of what's going to come in the last parts of the Napkin. It can be omitted without loss of continuity.

In the world of olympiad math, there's a famous functional equation that goes as follows:

$$f\colon \mathbb{R} \to \mathbb{R} \qquad f(x+y) = f(x) + f(y).$$

Everyone knows what its solutions are! There's an obvious family of solutions $f(x) = cx$. Then there's also this family of... uh... noncontinuous solutions (mumble grumble) pathological (mumble mumble) Axiom of Choice (grumble).

There's also this thing called Zorn's lemma. It sounds terrifying, because it's equivalent to the Axiom of Choice, which is also terrifying because why not.

In this post I will try to de-terrify these things, because they're really not as terrifying as they sound.

§88.1 Let's construct a monster

Let us just see if we can try and construct a "bad" f and see what happens.

By scaling, let's assume WLOG that $f(1) = 1$. Thus $f(n) = n$ for every integer n, and you can easily show from here that

$$f\left(\frac{m}{n}\right) = \frac{m}{n}.$$

So f is determined for all rationals. And then you get stuck.

None of this is useful for determining, say, $f(\sqrt{2})$. You could add and subtract rational numbers all day and, say, $\sqrt{2}$ isn't going to show up at all.

Well, we're trying to set things on fire anyways, so let's set

$$f(\sqrt{2}) = 2015$$

because why not? By the same induction, we get $f(n\sqrt{2}) = 2015n$, and then that

$$f\left(a + b\sqrt{2}\right) = a + 2015b.$$

Here a and b are rationals. Well, so far so good – as written, this is a perfectly good solution, other than the fact that we've only defined f on a tiny portion of the real numbers.

Well, we can do this all day:

$$f\left(a + b\sqrt{2} + c\sqrt{3} + d\pi\right) = a + 2015b + 1337c - 999d.$$

Perfectly consistent.

You can kind of see how we should keep going now. Just keep throwing in new real numbers which are "independent" to the previous few, assigning them to whatever junk we want. It feels like it *should* be workable. . .

In a moment I'll explain what "independent" means (though you might be able to guess already), but at the moment there's a bigger issue: no matter how many numbers we throw, it seems like we'll never finish. Let's address the second issue first.

§88.2 Review of finite induction

When you do induction, you get to count off 1, 2, 3, ... and so on. So for example, suppose we had a "problem" such as:

> Prove that the intersection of n open intervals is either $\varnothing$ or an open interval.

You can do this by induction easily: it's true for $n = 2$, and for the larger cases it's similarly easy.

But you can't conclude from this that *infinitely* many open intervals intersect at some open interval. Indeed, this is false: consider the intervals

$$(-1, 1), \quad \left(-\frac{1}{2}, \frac{1}{2}\right), \quad \left(-\frac{1}{3}, \frac{1}{3}\right), \quad \left(-\frac{1}{4}, \frac{1}{4}\right), \quad \dots$$

This *infinite* set of intervals intersects at a single point $\{0\}$!

The moral of the story is that induction doesn't let us reach infinity. Too bad, because we'd have loved to use induction to help us construct a monster. That's what we're doing, after all – adding things in one by one.

§88.3 Transfinite induction

Well, it turns out we can, but we need a new notion of number, the so-called *ordinal number*. I define these in their full glory in the first two sections of Chapter 90 (and curious readers are even invited to jump ahead to those two sections), but for this chapter I won't need that full definition yet.

Here's what I want to say: after all the natural numbers

$$0,\ 1,\ \dots,$$

I'll put a *new number* called ω, the first ordinal greater than all the natural numbers. After that there's more numbers called

$$\omega + 1,\ \omega + 2,\ \dots$$

and eventually a number called $\omega \cdot 2$.

The list goes on:

$$\begin{aligned}
&0, 1, 2, 3, \dots, \omega \\
&\quad \omega + 1, \omega + 2, \dots, \omega + \omega \\
&\quad \omega \cdot 2 + 1, \omega \cdot 2 + 2, \dots, \omega \cdot 3 \\
&\quad \vdots \\
&\quad \omega^2 + 1, \omega^2 + 2, \dots \\
&\quad \vdots \\
&\quad \omega^3, \dots, \omega^4, \dots, \omega^\omega \dots, \omega^{\omega^{\omega^\cdots}}
\end{aligned}$$

Pictorially, it kind of looks like this:

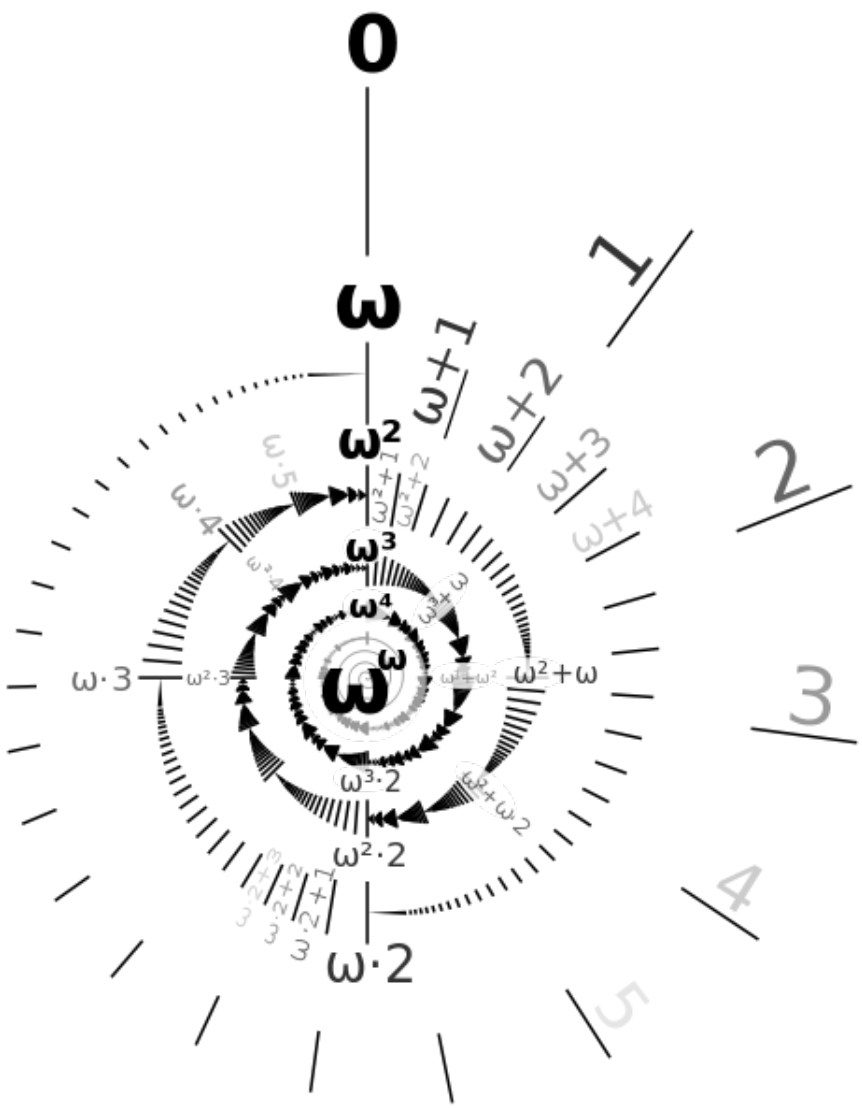


Image from [ca]

(Note that the diagram only shows an initial segment; there are still larger ordinals like $\omega^{\omega^\omega} + 1000$ and so on).

Anyways, in the same way that natural numbers "dominate" all finite sets, the ordinals dominate *all the sets*, in the following sense. Essentially, assuming the Axiom of Choice, it follows that for every set S there's some ordinal α which is larger than S (in a sense I won't make precise until later chapters).

But it turns out (and you can intuitively see) that as large as the ordinals grow, there is no *infinite descending chain*. Meaning: if I start at an ordinal (like $2\omega + 4$) and jump down, I can only take finitely many jumps before I hit 0. (To see this, try writing down a chain starting at $2\omega + 4$ yourself.) Hence, induction and recursion still work verbatim:

Theorem 88.3.1 (Transfinite induction)
Given a statement $P(-)$, suppose that

- $P(0)$ is true, and
- If $P(\alpha)$ is true for all $\alpha < \beta$, then $P(\beta)$ is true.

Then $P(\beta)$ is true.

Similarly, you're allowed to do recursion to define x_β if you know the value of x_α for all $\alpha < \beta$.

The difference from normal induction or recursion is that we'll often only do things like "define $x_{n+1} = \dots$". But this is not enough to define x_α for all α. To see this, try using our normal induction and see how far we can climb up the ladder.

Answer: you can't get ω! It's not of the form $n+1$ for any of our natural numbers n – our finite induction only lets us get up to the ordinals less than ω. Similarly, the simple $+1$ doesn't let us hit the ordinal $\omega \cdot 2$, even if we already have $\omega + n$ for all n. Such ordinals are called **limit ordinals**. The ordinal that *are* of the form $\alpha + 1$ are called **successor ordinals**.

So a transfinite induction or recursion is very often broken up into three cases. In the induction phrasing, it looks like

- (Zero Case) First, resolve $P(0)$.
- (Successor Case) Show that from $P(\alpha)$ we can get $P(\alpha+1)$.
- (Limit Case) Show that $P(\lambda)$ holds given $P(\alpha)$ for all $\alpha < \lambda$, where λ is a limit ordinal.

Similarly, transfinite recursion often is split into cases too.

- (Zero Case) First, define x_0.
- (Successor Case) Define $x_{\alpha+1}$ from x_α.
- (Limit Case) Define x_λ from x_α for all $\alpha < \lambda$, where λ is a limit ordinal.

In both situations, finite induction only does the first two cases, but if we're able to do the third case we can climb far above the barrier ω.

§88.4 Wrapping up functional equations

Let's return to solving our problem.

Let S_n denote the set of "base" numbers we have at the nth step. In our example, we might have

$$S_1 = \{1\}, \quad S_2 = \left\{1, \sqrt{2}\right\}, \quad S_3 = \left\{1, \sqrt{2}, \sqrt{3}\right\}, \quad S_4 = \left\{1, \sqrt{2}, \sqrt{3}, \pi\right\}, \quad \dots$$

and we'd like to keep building up S_i until we can express all real numbers. For completeness, let me declare $S_0 = \varnothing$.

First, I need to be more precise about "independent". Intuitively, this construction is working because

$$a + b\sqrt{2} + c\sqrt{3} + d\pi$$

is never going to equal zero for rational numbers a, b, c, d (other than all zeros). In general, a set X of numbers is "independent" if the combination

$$c_1x_1 + c_2x_2 + \cdots + c_mx_m = 0$$

never occurs for rational numbers $\mathbb{Q}$ unless $c_1 = c_2 = \cdots = c_m = 0$. Here $x_i \in X$ are distinct. Note that even if X is infinite, I can only take finite sums! (This notion has a name: we want X to be **linearly independent** over $\mathbb{Q}$; see the chapter on vector spaces for more on this!)

When do we stop? We'd like to stop when we have a set $S_{\text{something}}$ that's so big, every real number can be written in terms of the independent numbers. (This notion also has a name: it's called a $\mathbb{Q}$-basis.) Let's call such a set **spanning**; we stop once we hit a spanning set.

The idea that we can induct still seems okay: suppose S_α isn't spanning. Then there's some number that is independent of S_α, say $\sqrt{2015}\pi$ or something. Then we just add it to get $S_{\alpha+1}$. And we keep going.

Unfortunately, as I said before it's not enough to be able to go from S_α to $S_{\alpha+1}$ (successor case); we need to handle the limit case as well. But it turns out there's a trick we can do. Suppose we've constructed *all* the sets S_0, S_1, S_2, ..., one for each positive integer n, and none of them are spanning. The next thing I want to construct is S_ω; somehow I have to "jump". To do this, I now take the infinite union

$$S_\omega \stackrel{\text{def}}{=} S_0 \cup S_1 \cup S_2 \cup \ldots.$$

The elements of this set are also independent (why?).

Ta-da! With the simple trick of "union all the existing sets", we've just jumped the hurdle to the first limit ordinal ω. Then we can construct $S_{\omega+1}$, $S_{\omega+2}$, ..., once again – just keep throwing in elements. Then when we need to jump the next hurdle to $S_{2\omega}$, we just do the same trick of "union-ing" all the previous sets.

So we can formalize the process as follows:

1. Let $S_0 = \varnothing$.
2. For a successor stage $S_{\alpha+1}$, add any element to S_α to obtain $S_{\alpha+1}$.
3. For a limit stage S_λ, take the union $\bigcup_{\gamma<\lambda} S_\gamma$.

How do we know that we'll stop eventually? Well, the thing is that this process consumes a lot of real numbers. In particular, the ordinals get larger than the size of $\mathbb{R}$ (assuming Choice). Hence if we don't stop we will quite literally reach a point where we have used up every single real number. Clearly that's impossible, because by then the elements can't possibly be independent!

So by transfinite recursion, we eventually hit some S_γ which is spanning: the elements are all independent, but every real number can be expressed using it. Done!

§88.5 Zorn's lemma

Now I can tell you what Zorn's lemma is: it lets us do the same thing in any poset.

We can think of the above example as follows: consider all sets of independent elements. These form a partially ordered set by inclusion, and what we did was quite literally climb up a chain

$$S_0 \subsetneq S_1 \subsetneq S_2 \subsetneq \ldots.$$

It's not quite climbing since we weren't just going one step at a time: we had to do "jumps" to get up to S_ω and resume climbing. But the main idea is to climb up a poset until we're at the very top; in the previous case, when we reached the spanning set.

The same thing works verbatim with any partially ordered set $\mathbb{P}$. Let's define some terminology. A **local maximum** of the entire poset $\mathbb{P}$ is an element which has no other elements strictly greater than it. (Most authors refer to this as "maximal element", but I think "local maximum" is a more accurate term.)

Now a **chain of length** γ is a set of elements p_α for every $\alpha < \gamma$ such that $p_0 < p_1 < p_2 < \dots$. (Observe that a chain has a last element if and only if γ is a successor ordinal, like $\omega + 3$.) An **upper bound** to a chain is an element $\tilde{p}$ which is greater than or equal to all elements of the chain; In particular, if γ is a successor ordinal, then just taking the last element of the chain works.

In this language, Zorn's lemma states that

Theorem 88.5.1 (Zorn's lemma)

Let $\mathbb{P}$ be a nonempty partially ordered set. If every chain has an upper bound, then $\mathbb{P}$ has a local maximum.

Chains with length equal to a successor ordinal always have upper bounds, but this is not true in the limit case. So the hypothesis of Zorn's lemma is exactly what lets us "jump" up to define p_ω and other limit ordinals. And the proof of Zorn's lemma is straightforward: keep climbing up the poset at successor stages, using Zorn's condition to jump up at limit stages, and thus building a really long chain. But we have to eventually stop, or we literally run out of elements of $\mathbb{P}$. And the only possible stopping point is a local maximum.

If we want to phrase our previous solution in terms of Zorn's lemma, we'd say:

Proof. Look at the poset whose elements are sets of independent real numbers. Every chain $S_0 \subsetneq S_1 \subsetneq \dots$ has an upper bound $\bigcup S_\alpha$ (which you have to check is actually an element of the poset). Thus by Zorn, there is a local maximum S. Then S must be spanning, because otherwise we could add an element to it. □

So really, Zorn's lemma is encoding all of the work of climbing that I argued earlier. It's a neat little package that captures all the boilerplate, and tells you exactly what you need to check.

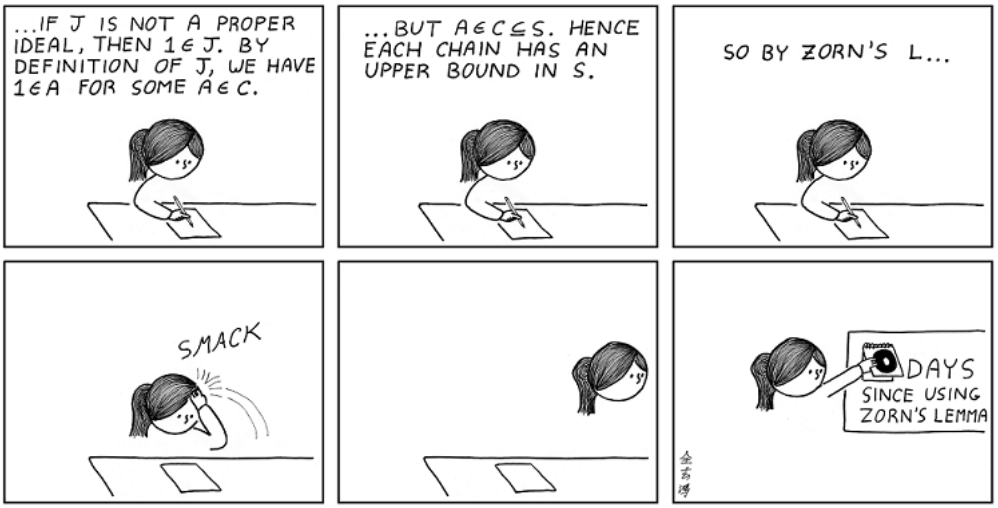


Image from [Go09]

One last thing you might ask: where is the Axiom of Choice used? Well, the idea is that for any chain there could be lots of $\tilde{p}$'s, and you need to pick one of them. Since you are making arbitrary choices infinitely many times, you need the Axiom of Choice. (Actually, you also need Choice to talk about cardinalities as in Theorem 88.3.1.) But really, it's nothing special.

§88.6 A few harder problems to think about

Problem 88A. Suppose $f\colon (0,\infty) \to \mathbb{R}$ satisfies $f(1) = 1$, $f(\pi) = \frac{22}{7}$, and

$$f(x+y) = f(x) + f(y)$$

for all $x, y > 0$. Given that $\pi \approx 3.14159265358979$, find, with proof and without a calculator, an example of a real number r such that $0 < r < 1$ and $f(r) > 9000$.

Problem 88B. Suppose $f\colon (0,\infty) \to \mathbb{R}$ satisfies

$$f(x+y) = f(x) + f(y)$$

for all $x, y > 0$. Label each of the following statements as true or false.

1. The function f can be extended to $f\colon \mathbb{R} \to \mathbb{R}$ while still satisfying Cauchy's functional equation.
2. If f is extended as in the previous statement, then f must be odd.
3. If $f(x) \geq 0$ for all $x > 0$, then f is linear.
4. If f is strictly increasing, then f is linear.
5. The function f is not a bijection.
6. It's possible for f to be injective but not linear.
7. It's possible for f to be surjective.
8. It's possible for f to be nonconstant and only take rational values.
9. It's possible for f to only take irrational values.

Problem 88C (Tukey's lemma)**.** Let $\mathcal{F}$ be a nonempty family of sets. Assume that for any set A, the set A is in $\mathcal{F}$ if and only if all its finite subsets are in $\mathcal{F}$.

Prove that there exists a maximal set $Y \in \mathcal{F}$ (i.e. Y not contained in any other set of $\mathcal{F}$).

89 Zermelo-Fraenkel with choice

Chapter 3.1 of [Le14] has a nice description of this.

§89.1 The ultimate functional equation

In abstract mathematics, we often define structures by what *properties* they should have; for example, a group is a set and a binary operation satisfying so-and-so axioms, while a metric space is a set and a distance function satisfying so-and-so axioms.

Nevertheless, these definitions rely on previous definitions. The colorful illustration of [Le14] on this:

- A *vector space* is an abelian group with...
- An *abelian group* has a binary operation such that...
- A *binary operation* on a set is...
- A *set* is ...

and so on.

We have to stop at some point, because infinite lists of definitions are bad. The stopping turns out to be a set, "defined" by properties. The trick is that we never actually define what a set is, but nonetheless postulate that these sets satisfy certain properties: these are the ZFC axioms. Loosely, ZFC can be thought of as the *ultimate functional equation*.

Before talking about what these axioms are, I should talk about the caveats.

§89.2 Cantor's paradox

Intuitively, a set is an unordered collection of elements. Two sets are equal if they share the same elements:

$$\{x \mid x \text{ is a featherless biped}\} = \{x \mid x \text{ is a human}\}$$

(let's put aside the issue of dinosaurs).

As another example, we have our empty set $\varnothing$ that contains no objects. We can have a set $\{1, 2, 3\}$, or maybe the set of natural numbers $\mathbb{N} = \{0, 1, 2, \dots\}$. (For the purposes of set theory, 0 is usually considered a natural number.) Sets can even contain other sets, like $\{\mathbb{Z}, \mathbb{Q}, \mathbb{N}\}$. Fine and dandy, right?

The trouble is that this definition actually isn't good enough, and here's why. If we just say "a set is any collection of objects", then we can consider a really big set V, the set of all sets. So far no problem, right? We would have the oddity that $V \in V$, but oh well, no big deal.

Unfortunately, this existence of this V leads immediately to a paradox. The classical one is Russell's Paradox. I will instead present a somewhat simpler one: not only does V contain itself, *every subset* $S \subseteq V$ is itself an element of V (i.e. $S \in V$). If we let $\mathcal{P}(V)$ denote the **power set** of V (i.e. all the subsets of V), then we have an inclusion

$$\mathcal{P}(V) \hookrightarrow V.$$

This is bad, since:

Lemma 89.2.1 (Cantor's diagonal argument)
For *any* set X, it's impossible to construct an injective map $\iota\colon \mathcal{P}(X) \hookrightarrow X$.

Proof. Assume for contradiction ι exists.

Exercise 89.2.2. Show that if, ι exists, then there exists a surjective map $j\colon X \twoheadrightarrow \mathcal{P}(X)$. (This is easier than it appears, just "invert ι").

We now claim that j can't exist.

Let me draw a picture for j to give the idea first:

		x_1	x_2	x_3	x_4	x_5	$\dots$
x_1	$\overset{j}{\mapsto}$	**0**	1	1	0	1	$\dots$
x_2	$\overset{j}{\mapsto}$	1	**1**	0	1	1	$\dots$
x_3	$\overset{j}{\mapsto}$	0	1	**0**	0	1	$\dots$
x_4	$\overset{j}{\mapsto}$	1	0	0	**1**	0	$\dots$
x_5	$\overset{j}{\mapsto}$	0	1	1	1	**1**	$\dots$
$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Here, for each $j(x) \subseteq X$, I'm writing "1" to mean that the element is inside $j(x)$, and "0" otherwise. So $j(x_1) = \{x_2, x_3, x_5 \dots\}$. (Here the indices are ordinals rather than integers as X may be uncountable. Experts may notice I've tacitly assumed a well-ordering of X; but this picture is for motivation only so I won't dwell on the point.) Then we can read off the diagonal to get a new set. In our example, the diagonal specifies a set $A = \{x_2, x_4, x_5 \dots\}$. Then we "invert" it to get a set $B = \{x_1, x_3, \dots\}$.

Back to the formal proof. As motivated above, we define

$$B = \{x \mid x \notin j(x)\}.$$

By construction, $B \subseteq X$ is not in the image of j, which is a contradiction since j was supposed to be surjective. □

Now if you're not a set theorist, you could probably just brush this off, saying "oh well, I guess you can't look at certain sets". But if you're a set theorist, this worries you, because you realize it means that you can't just define a set as "a collection of objects", because then everything would blow up. Something more is necessary.

§89.3 The language of set theory

We need a way to refer to sets other than the informal description of "collection of objects".

So here's what we're going to do. We'll start by defining a formal *language of set theory*, a way of writing logical statements. First of all we can throw in our usual logical operators:

- $\forall$ means "for all"
- $\exists$ means "exists"

- $=$ means "equal"
- $X \implies Y$ means "if X then Y"
- $A \wedge B$ means "A and B"
- $A \vee B$ means "A or B"
- $\neg A$ means "not A".

Since we're doing set theory, there's only one more operator we add in: the inclusion $\in$. And that's all we're going to use (for now).

So how do we express something like "the set $\{1, 2\}$"? The trick is that we're not going to actually "construct" any sets, but rather refer to them indirectly, like so:

$$\exists S : x \in S \iff ((x = 1) \vee (x = 2)) .$$

This reads: "there exists an S such that x is in S if and only if either $x = 1$ or $x = 2$". We don't have to refer to sets as objects in and of themselves anymore — we now have a way to "create" our sets, by writing formulas for exactly what they contain. This is something a machine can parse.

Well, what are we going to do with things like 1 and 2, which are not sets? Answer:

Elements of sets are themselves sets.

We're going to make **everything** into a set. Natural numbers will be sets. Ordered pairs will be sets. Functions will be sets. Later, I'll tell you exactly how we manage to do something like encode 1 as a set. For now, all you need to know is that sets don't just hold objects; they hold other sets.

So now it makes sense to talk about whether something is a set or not: $\exists x$ means "x is a set", while $\nexists x$ means "x is not a set". In other words, we've rephrased the problem of deciding whether something is a set to whether it exists, which makes it easier to deal with in our formal language. That means that our axiom system had better find some way to let us show a lot of things exist, without letting us prove

$$\exists S \forall x : x \in S.$$

For if we prove this formula, then we have our "bad" set that caused us to go down the rabbit hole in the first place.

§89.4 The axioms of ZFC

I don't especially want to get into details about these axioms; if you're interested, read:

- https://blog.evanchen.cc/2014/11/13/set-theory-an-intro-to-zfc-part-1/
- https://blog.evanchen.cc/2014/11/18/set-theory-part-2-constructing-the-ordinals/

Here is a much terser description of the axioms, which also includes the corresponding sentence in the language of set theory. It is worth the time to get some practice parsing $\forall$, $\exists$, etc. and you can do so by comparing the formal sentences with the natural statement of the axiom.

First, the two easiest axioms:

- Extensionality is the sentence $\forall x \forall y \left((\forall a \left(a \in x \iff a \in y \right)) \implies x = y \right)$, which says that if two sets x and y have the same elements, then $x = y$.
- EmptySet is the sentence $\exists a : \forall x \ \neg(x \in a)$; it says there exists a set with no elements. By Extensionality this set is unique, so we denote it $\varnothing$.

The next two axioms give us basic ways of building new sets.

- Given two elements x and y, there exists a set a containing only those two elements. In machine code, this is the sentence Pairing, written
$$\forall x \forall y \exists a \quad \forall z, \ z \in a \iff \left((z = x) \vee (z = y) \right).$$
By Extensionality this set a is unique, so we write $a = \{x, y\}$.
- Given a set a, we can create the union of the elements of a. For example, if $a = \{\{1,2\},\{3,4\}\}$, then $U = \{1,2,3,4\}$ is a set. Formally, this is the sentence Union:
$$\forall a \exists U \quad \forall x \ \left[(x \in U) \iff (\exists y : x \in y \in a) \right].$$
Since U is unique by Extensionality, we denote it $\cup a$.
- We can construct the **power set** $\mathcal{P}(x)$. Formally, the sentence PowerSet says that
$$\forall x \exists P \forall y (y \in P \iff y \subseteq x)$$
where $y \subseteq x$ is short for $\forall z (z \in y \implies z \in x)$. As Extensionality gives us uniqueness of P, we denote it $\mathcal{P}(x)$.
- Foundation says there are no infinite descending chains
$$x_0 \ni x_1 \ni x_2 \ni \dots.$$
This is important, because it lets us induct. In particular, **no set contains itself**.
- Infinity implies that $\omega = \{0, 1, \dots\}$ is a set.

These are all things you are already used to, so keep your intuition there. The next one is less intuitive:

- The **schema of restricted comprehension** says: if we are *given a set* X, and some formula $\phi(x)$ then we can *filter* through the elements of X to get a subset
$$Y = \{x \in X \mid \phi(x)\}.$$
Formally, given a formula ϕ:
$$\forall X \quad \exists Y \quad \forall y (y \in Y \iff y \in X \wedge \phi(y)).$$

Notice that we may *only* do this filtering over an already given set. So it is not valid to create $\{x \mid x \text{ is a set}\}$. We are thankful for this, because this lets us evade Cantor's paradox.

Abuse of Notation 89.4.1. Note that technically, there are infinitely many sentences, a Comprehension$_\phi$ for every possible formula ϕ. By abuse of notation, we let Comprehension abbreviate the infinitely many axioms Comprehension$_\phi$ for every ϕ.

There is one last schema called Replacement$_\phi$. Suppose X is a set and $\phi(x,y)$ is some formula such that for every $x \in X$, there is a *unique* y in the universe such that $\phi(x,y)$ is true: for example "$y = x \cup \{x\}$" works. (In effect, ϕ is defining a function f on X.) Then there exists a set Y consisting exactly of these images: (i.e. $f^{\text{img}}(X)$ is a set).

Abuse of Notation 89.4.2. By abuse of notation, we let Replacement abbreviate the infinitely many axioms Replacement$_\phi$ for every ϕ.

Remark 89.4.3 — What do we mean here that "for every $x \in X$, there is a *unique* y in the universe such that $\phi(x,y)$ is true"? How can we decide, given a formula ϕ, whether that statement is true, for Replacement$_\phi$ to be an axiom?
Turns out we cannot in general. But we don't need it! To circumvent the problem, for every $\phi(x,y)$, the axiom Replacement$_\phi$ states that

$$\text{“}\phi \text{ defines a function”} \implies \forall X \quad \exists Y \quad \text{“}Y = f^{\text{img}}(X)\text{”}.$$

In other words, the hypothesis that ϕ is a function is "folded in" the axiom Replacement$_\phi$ itself.
This will not really matter to us for now, but later on, it will matter in model theory, where we will state in Lemma 92.5.1 what it means for a model M to satisfy Replacement.

We postpone discussion of the Axiom of Choice momentarily.

§89.5 Encoding

Now that we have this rickety universe of sets, we can start re-building math. You'll get to see this more in the next chapter on ordinal numbers.

Definition 89.5.1. An **ordered pair** (x,y) is a set of the form

$$(x,y) \coloneqq \{\{x\}, \{x,y\}\}.$$

Note that $(x,y) = (a,b)$ if and only if $x = a$ and $y = b$. Ordered k-tuples can be defined recursively: a three-tuple (a,b,c) means $(a,(b,c))$.

Definition 89.5.2. A **function** $f\colon X \to Y$ is defined as a collection of ordered pairs such that

- If $(x,y) \in f$, then $x \in X$ and $y \in Y$.
- For every $x \in X$, there is a unique $y \in Y$ such that $(x,y) \in f$. We denote this y by $f(x)$.

Definition 89.5.3. The **natural numbers** are defined inductively as

$$\begin{aligned} 0 &= \varnothing \\ 1 &= \{0\} \\ 2 &= \{0,1\} \\ 3 &= \{0,1,2\} \\ &\vdots \end{aligned}$$

The set of all natural numbers is denoted ω.

Abuse of Notation 89.5.4. Yes, I'm sorry, in set theory 0 is considered a natural number. For this reason I'm using ω and not $\mathbb{N}$ since I explicitly have $0 \notin \mathbb{N}$ in all other parts of this book.

Et cetera, et cetera.

§89.6 Choice and well-ordering

The Axiom of Choice states that given a collection Y of nonempty sets, there is a function $g\colon Y \to \cup Y$ which "picks" an element of each member of Y. That means $g(y) \in y$ for every $y \in Y$. (The typical illustration is that Y contains infinitely many drawers, and each drawer (a y) has some sock in it.)

Formally, it is the sentence

$$\forall Y \left(\varnothing \notin Y \implies \exists g\colon Y \to \cup Y \text{ such that } \forall y \in Y \left(g(y) \in y\right).\right)$$

The tricky part is not that we can conceive of such a function g, but that in fact this function g is *actually a set*.

There is an equivalent formulation which is often useful.

Definition 89.6.1. A **well-ordering** $<$ of X is a strict, total order on X which has no infinite descending chains.

Well-orderings on a set are very nice, because we can pick minimal elements: this lets us do induction, for example.

Example 89.6.2 (Examples and non-examples of well-orderings)

(a) The natural numbers $\omega = \{0, 1, 2, \dots\}$ are well-ordered by $<$.

(b) The integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ are not well-ordered by $<$, because there are infinite descending chains (take $-1 > -2 > -3 > \dots$).

(c) The positive real numbers are not well-ordered by $<$, again because of the descending chain $\frac{1}{1} > \frac{1}{2} > \frac{1}{3} > \dots$.

(d) The positive integers are not well-ordered by the divisibility operation $|$. While there are no descending chains, there are elements which cannot be compared (for example $3 \nmid 5$, $5 \nmid 3$ and $3 \neq 5$).

Theorem 89.6.3 (Well-ordering theorem)

Assuming Choice, for every set we can place some well-ordering on it.

In fact, the well-ordering theorem is actually equivalent to the axiom of choice.

§89.7 Sets vs classes

Prototypical example for this section: The set of all sets is the standard example of a proper class.

We close the discussion of ZFC by mentioning "classes".

Roughly, the "bad thing" that happened was that we considered a set S, the "set of all sets", and it was *too big*. That is,

$$\{x \mid x \text{ is a set}\}$$

is not good. Similarly, we cannot construct a set

$$\{x \mid x \text{ is an ordered pair}\}.$$

The lesson of Cantor's Paradox is that we cannot create any sets we want; we have to be more careful than that.

Nonetheless, if we are given a set we can still tell whether or not it is an ordered pair. So for convenience, we will define a **class** to be a "concept" like the "class of all ordered pairs". Formally, a class is defined by some formula ϕ: it consists of the sets which satisfy the formula.

In particular:

Definition 89.7.1. The class of all sets is denoted V, defined by $V = \{x \mid x = x\}$. It is called the **von Neumann universe**.

A class is a **proper class** if it is not a set, so for example we have:

Theorem 89.7.2 (There is no set of all sets)
V is a proper class.

Proof. Assume not, and V is a set. Then $V \in V$, which violates Foundation. (In fact, V cannot be a set even without Foundation, as we saw earlier). □

Abuse of Notation 89.7.3. Given a class C, we will write $x \in C$ to mean that x has the defining property of C. For example, $x \in V$ means "x is a set".

It does not mean x is an element of V – this doesn't make sense as V is not a set.

§89.8 A few harder problems to think about

Problem 89A. Let A and B be sets. Show that $A \cap B$ and $A \times B$ are sets.

Problem 89B. Show that the class of all groups is a proper class. (You can take the definition of a group as a pair $(G, \cdot)$ where $\cdot$ is a function $G \times G \to G$.)

Problem 89C. Show that the axiom of choice follows from the well-ordering theorem.

Problem 89D†. Prove that actually, Replacement $\implies$ Comprehension.

Problem 89E (From Taiwan IMO training camp)**.** Consider infinitely many people each wearing a hat, which is either red, green, or blue. Each person can see the hat color of everyone except themselves. Simultaneously each person guesses the color of their hat. Show that they can form a strategy such that at most finitely many people guess their color incorrectly.

90 Ordinals

§90.1 Counting for preschoolers

In preschool, we were told to count as follows. We defined a set of symbols 1, 2, 3, 4, ... Then the teacher would hold up three apples and say:

> "One . . . two . . . three! There are three apples."

Image from [Ho]

The implicit definition is that the *last* number said is the final answer. This raises some obvious problems if we try to count infinite sets, but even in the finite world, this method of counting fails for the simplest set of all: how many apples are in the following picture?

Image from [Kr]

Answer: 0. There is nothing to say, and our method of counting has failed for the simplest set of all: the empty set.

§90.2 Counting for set theorists

Prototypical example for this section: $\omega + 1 = \{0, 1, 2, \dots, \omega\}$ *might work.*

Rather than using the *last* number listed, I propose instead starting with a list of symbols 0, 1, 2, ... and making the final answer the *first* number which was *not* said. Thus to count three apples, we would say

> "Zero . . . one . . . two! There are three apples."

We will call these numbers *ordinal numbers* (rigorous definition later). In particular, we'll *define* each ordinal to be the set of things we say:

$$\begin{aligned} 0 &= \varnothing \\ 1 &= \{0\} \\ 2 &= \{0, 1\} \\ 3 &= \{0, 1, 2\} \\ &\vdots \end{aligned}$$

In this way we can write out the natural numbers. You can have some fun with this, by saying things like

$$4 \coloneqq \{\{\}, \{\{\}\}, \{\{\}, \{\{\}\}\}, \{\{\}, \{\{\}\}, \{\{\}, \{\{\}\}\}\}\}.$$

In this way, we soon write down all the natural numbers. The next ordinal, ω,[1] is defined as

$$\omega = \{0, 1, 2, \dots\}$$

Then comes

$$\begin{aligned} \omega + 1 &= \{0, 1, 2, \dots, \omega\} \\ \omega + 2 &= \{0, 1, 2, \dots, \omega, \omega + 1\} \\ \omega + 3 &= \{0, 1, 2, \dots, \omega, \omega + 1, \omega + 2\} \\ &\vdots \end{aligned}$$

And in this way we define $\omega + n$, and eventually reach

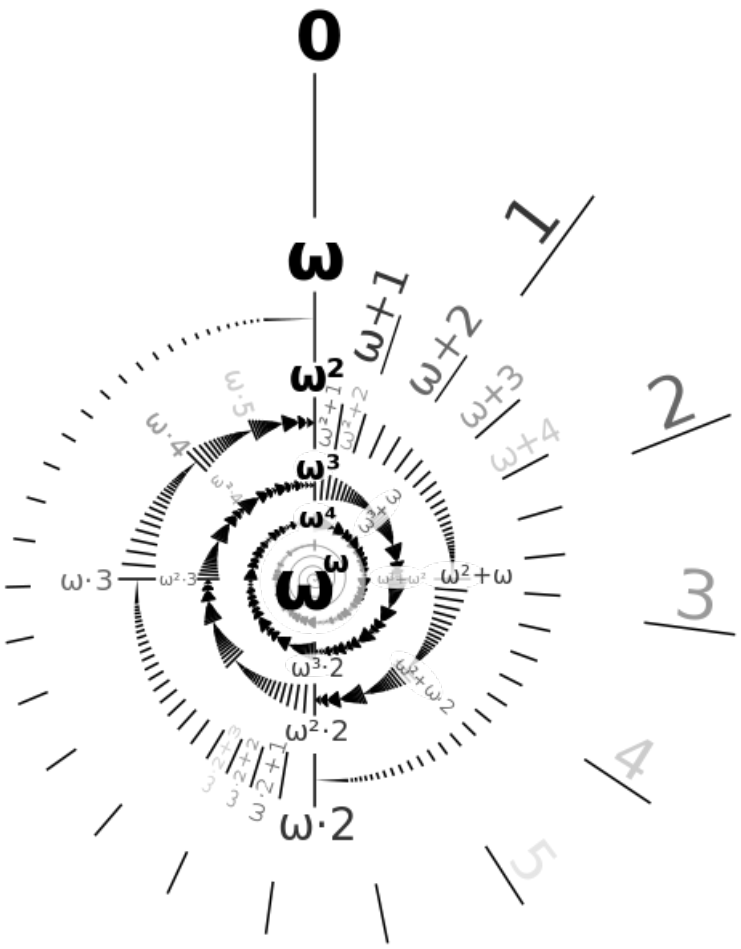
$$\begin{aligned} \omega \cdot 2 = \omega + \omega &= \{0, 1, 2 \dots, \omega, \omega + 1, \omega + 2, \dots\} \\ \omega \cdot 2 + 1 &= \{0, 1, 2 \dots, \omega, \omega + 1, \omega + 2, \dots, \omega \cdot 2\}. \end{aligned}$$

[1] As mentioned in the last chapter, it's not immediate that ω is a set; its existence is generally postulated by the Infinity axiom.

In this way we obtain

$$
\begin{aligned}
&0,\ 1,\ 2,\ 3,\ \ldots,\ \omega \\
&\quad \omega+1,\ \omega+2,\ \ldots,\ \omega+\omega \\
&\quad \omega\cdot 2+1,\ \omega\cdot 2+2,\ \ldots,\ \omega\cdot 3, \\
&\quad \vdots \\
&\quad \omega^2+1,\ \omega^2+2,\ \ldots \\
&\quad \vdots \\
&\quad \omega^3,\ \ldots,\ \omega^4,\ \ldots,\ \omega^\omega \\
&\quad \vdots \\
&\quad \omega^{\omega^{\omega^{\cdots}}}
\end{aligned}
$$

The first several ordinals can be illustrated in a nice spiral.



Remark 90.2.1 — You may think, well, why don't we define the ordinals like this

instead? It certainly looks shorter and simpler!

$$\begin{aligned} 0 &= \varnothing \\ 1 &= \{0\} \\ 2 &= \{1\} \\ 3 &= \{2\} \\ &\vdots \\ \omega &= \{0, 1, 2, \dots\} \\ \omega+1 &= \{\omega\} \\ &\vdots \end{aligned}$$

There are a few reasons why the usual definition is better.

- The "alternative definition" above is not uniform — an ordinal of the form $\alpha+1$ is defined by a singleton set, while the other ordinals are defined by a set that leads up to it.
- Comparison is simpler: for two ordinals α and β, $\alpha < \beta$ if and only if $\alpha \in \beta$.
- Cardinals will be simpler: the size of the set 5 is exactly 5.

Remark 90.2.2 (Digression) — The number $\omega^{\omega^{\omega^{\cdots}}}$ has a name, ε_0; it has the property that $\omega^{\varepsilon_0} = \varepsilon_0$. The reason for using "$\varepsilon$" (which is usually used to denote small quantities) is that, despite how huge it may appear, it is actually a countable set. More on that later.

§90.3 Definition of an ordinal

Our informal description of ordinals gives us a chain

$$0 \in 1 \in 2 \in \cdots \in \omega \in \omega+1 \in \dots.$$

To give the actual definition of an ordinal, I need to define two auxiliary terms first.

Definition 90.3.1. A set x is **transitive** if whenever $z \in y \in x$, we have $z \in x$ also.

Example 90.3.2 (7 is transitive)
The set 7 is transitive: for example, $2 \in 5 \in 7 \implies 2 \in 7$.

Question 90.3.3. Show that this is equivalent to: whenever $y \in x$, $y \subseteq x$.

Moreover, recall the definition of "well-ordering": a strict linear order with no infinite descending chains.

Example 90.3.4 ($\in$ is a well-ordering on $\omega \cdot 3$)

In $\omega \cdot 3$, we have an ordering

$$0 \in 1 \in 2 \in \cdots \in \omega \in \omega + 1 \in \cdots \in \omega \cdot 2 \in \omega \cdot 2 + 1 \in \ldots.$$

which has no infinite descending chains. Indeed, a typical descending chain might look like

$$\omega \cdot 2 + 6 \ni \omega \cdot 2 \ni \omega + 2015 \ni \omega + 3 \ni \omega \ni 1000 \ni 256 \ni 42 \ni 7 \ni 0.$$

Even though there are infinitely many elements, there is no way to make an infinite descending chain.

Exercise 90.3.5. (Important) Convince yourself there are no infinite descending chains of ordinals at all, without using the Foundation axiom.

Definition 90.3.6. An **ordinal** is a transitive set which is well-ordered by $\in$. The class of all ordinals is denoted On.

Question 90.3.7. Satisfy yourself that this definition works.

Example 90.3.8

- All of 0, 1, 2, $\ldots$, ω, $\omega + 1$, $\ldots$ defined above are ordinals.
- $\{3\}$ is not an ordinal — it's not transitive because $2 \in 3$, but $2 \notin \{3\}$.
- $\{0, 1, 2, \{0, 2\}\}$ is not an ordinal — the two elements 1 and $\{0, 2\}$ are not comparable.

We typically use Greek letters α, β, etc. for ordinal numbers.

Definition 90.3.9. We write

- $\alpha < \beta$ to mean $\alpha \in \beta$, and $\alpha > \beta$ to mean $\alpha \ni \beta$.
- $\alpha \le \beta$ to mean $\alpha \in \beta$ or $\alpha = \beta$, and $\alpha \ge \beta$ to mean $\alpha \ni \beta$ or $\alpha = \beta$,

Theorem 90.3.10 (Ordinals are strictly ordered)

Given any two ordinal numbers α and β, either $\alpha < \beta$, $\alpha = \beta$ or $\alpha > \beta$.

Proof. Surprisingly annoying, thus omitted. The key idea is that we can define $\min(\alpha, \beta) = \alpha \cap \beta$, then prove that this must be equal to either α or β. □

Theorem 90.3.11 (Ordinals represent all order types)

Suppose $<$ is a well-ordering on a set X. Then there exists a unique ordinal α such that there is a bijection $\alpha \to X$ which is order preserving.

Thus ordinals represent the possible *equivalence classes* of order types. Any time you have a well-ordered set, it is isomorphic to a unique ordinal.

We now formalize the "+1" operation we were doing:

Definition 90.3.12. Given an ordinal α, we let $\alpha+1 = \alpha \cup \{\alpha\}$. An ordinal of the form $\alpha+1$ is called a **successor ordinal**.

Definition 90.3.13. If λ is an ordinal which is neither zero nor a successor ordinal, then we say λ is a **limit ordinal**.

Example 90.3.14 (Successor and limit ordinals)

7, $\omega+3$, $\omega\cdot 2+2015$ are successor ordinals, but ω and $\omega\cdot 2$ are limit ordinals.

§90.4 Ordinals are "tall"

First, we note that:

Theorem 90.4.1 (There is no set of all ordinals)

On is a proper class.

Proof. Assume for contradiction not. Then On is well-ordered by $\in$ and transitive, so On is an ordinal, i.e. $\text{On} \in \text{On}$, which violates Foundation. □

Exercise 90.4.2 (Unimportant)**.** Give a proof without Foundation by considering $\text{On}+1$.

From this we deduce:

Theorem 90.4.3 (Sets of ordinals are bounded)

Let $A \subseteq \text{On}$. Then there is some ordinal α such that $A \subseteq \alpha$ (i.e. A must be bounded).

Proof. Otherwise, look at $\bigcup A$. It is a set. But if A is unbounded it must equal On, which is a contradiction. □

In light of this, every set of ordinals has a **supremum**, which is the least upper bound. We denote this by $\sup A$.

Question 90.4.4. Show that

(a) $\sup(\alpha+1) = \alpha$ for any ordinal α.

(b) $\sup \lambda = \lambda$ for any limit ordinal λ.

The pictorial "tall" will be explained in a few sections.

§90.5 Transfinite induction and recursion

The fact that $\in$ has no infinite descending chains means that induction and recursion still work verbatim.

Theorem 90.5.1 (Transfinite induction)
Given a statement $P(-)$, suppose that

- $P(0)$ is true, and
- If $P(\alpha)$ is true for all $\alpha < \beta$, then $P(\beta)$ is true.

Then $P(\alpha)$ is true for every ordinal α.

Theorem 90.5.2 (Transfinite recursion)
To define a sequence x_α for every ordinal α, it suffices to

- define x_0, then
- for any β, define x_β using only x_α for any $\alpha < \beta$.

The difference between this and normal induction lies in the *limit ordinals*. In real life, we might only do things like "define $x_{n+1} = \dots$". But this is not enough to define x_α for all α, because we can't hit ω this way. Similarly, the simple $+1$ doesn't let us hit the ordinal $\omega \cdot 2$, even if we already have $\omega + n$ for all n. In other words, simply incrementing by 1 cannot get us past limit stages, but using transfinite induction to jump upwards lets us sidestep this issue.

So a transfinite induction is often broken up into three cases. In the induction phrasing, it looks like

- (Zero Case) First, resolve $P(0)$.
- (Successor Case) Show that from $P(\alpha)$ we can get $P(\alpha + 1)$.
- (Limit Case) For λ a limit ordinal, show that $P(\lambda)$ holds given $P(\alpha)$ for all $\alpha < \lambda$.

Similarly, transfinite recursion is often split into cases too.

- (Zero Case) First, define x_0.
- (Successor Case) Define $x_{\alpha+1}$ from x_α.
- (Limit Case) Define x_λ from x_α for all $\alpha < \lambda$, where λ is a limit ordinal.

In both situations, finite induction only does the first two cases, but if we're able to do the third case we can climb above the barrier ω.

§90.6 Ordinal arithmetic

Prototypical example for this section: $1 + \omega = \omega \neq \omega + 1$.

To give an example of transfinite recursion, let's define addition of ordinals. Recall that we defined $\alpha + 1 = \alpha \cup \{\alpha\}$. By transfinite recursion, let

$$\begin{aligned} \alpha + 0 &= \alpha \\ \alpha + (\beta + 1) &= (\alpha + \beta) + 1 \\ \alpha + \lambda &= \bigcup_{\beta < \lambda} (\alpha + \beta). \end{aligned}$$

Here $\lambda \neq 0$.

We can also do this explicitly: The picture is to just line up α before β. That is, we can consider the set

$$X = (\{0\} \times \alpha) \cup (\{1\} \times \beta)$$

(i.e. we tag each element of α with a 0, and each element of β with a 1). We then impose a well-ordering on X by a lexicographic ordering $<_{\text{lex}}$ (sort by first component, then by second). This well-ordering is isomorphic to a unique ordinal.

Example 90.6.1 $(2+3=5)$

Under the explicit construction for $\alpha = 2$ and $\beta = 3$, we get the set

$$X = \{(0,0) < (0,1) < (1,0) < (1,1) < (1,2)\}$$

which is isomorphic to 5.

Example 90.6.2 (Ordinal arithmetic is not commutative)

Note that $1 + \omega = \omega$! Indeed, under the transfinite definition, we have

$$1 + \omega = \cup_n (1+n) = 1 \cup 2 \cup 3 \cup \dots = \omega.$$

With the explicit construction, we have

$$X = \{(0,0) < (1,0) < (1,1) < (1,2) < \dots\}$$

which is isomorphic to ω.

Exercise 90.6.3. Show that $n + \omega = \omega$ for any $n \in \omega$.

Remark 90.6.4 — Ordinal addition is not commutative. However, from the explicit construction we can see that it is at least associative.
Furthermore, you can see that for small enough $\alpha \neq 0$, then $\alpha + \beta = \beta$ may happen; however, this does not happen on the other side — if $\beta < \gamma$, then $\alpha + \beta < \alpha + \gamma$.

Similarly, we can define multiplication in two ways. By transfinite induction:

$$\begin{aligned} \alpha \cdot 0 &= 0 \\ \alpha \cdot (\beta + 1) &= (\alpha \cdot \beta) + \alpha \\ \alpha \cdot \lambda &= \bigcup_{\beta < \lambda} \alpha \cdot \beta. \end{aligned}$$

We can also do an explicit construction: This time, the picture is to line up β copies, each copy contains α items. That is, $\alpha \cdot \beta$ is the order type of

$$<_{\text{lex}} \text{ applied to } \beta \times \alpha.$$

Example 90.6.5 (Ordinal multiplication is not commutative)
We have $\omega \cdot 2 = \omega + \omega$, but $2 \cdot \omega = \omega$.

Exercise 90.6.6. Prove this.

Exercise 90.6.7. Verify that ordinal multiplication (like addition) is associative but not commutative. (Look at $\gamma \times \beta \times \alpha$.)

Similar to ordinal addition, defining $\alpha \cdot (\beta + 1) = (\alpha \cdot \beta) + \alpha$ makes sure that if $\beta < \gamma$ then $\alpha \cdot \beta < \alpha \cdot \gamma$ — as long as $\alpha > 0$.

Exponentiation can also be so defined, though the explicit construction is less natural — since we will not use this definition in the rest of the book, you may ignore it.

For $\alpha = 0$, define $0^0 = 1$ and $0^\beta = 0$ for all $\beta > 0$. Otherwise:

$$\begin{aligned} \alpha^0 &= 1 \\ \alpha^{\beta+1} &= \alpha^\beta \cdot \alpha \\ \alpha^\lambda &= \bigcup_{\beta<\lambda} \alpha^\beta. \end{aligned}$$

Exercise 90.6.8. Verify that $2^\omega = \omega$.

§90.7 The hierarchy of sets

We now define the **von Neumann Hierarchy** by transfinite recursion.

Definition 90.7.1. By transfinite recursion, we set

$$\begin{aligned} V_0 &= \varnothing \\ V_{\alpha+1} &= \mathcal{P}(V_\alpha) \\ V_\lambda &= \bigcup_{\alpha<\lambda} V_\alpha \end{aligned}$$

By transfinite induction, we see V_α is transitive and that $V_\alpha \subseteq V_\beta$ for all $\alpha < \beta$.

Example 90.7.2 (V_α for $\alpha \leq 3$)
The first few levels of the hierarchy are:

$$\begin{aligned} V_0 &= \varnothing \\ V_1 &= \{0\} \\ V_2 &= \{0, 1\} \\ V_3 &= \{0, 1, 2, \{1\}\}. \end{aligned}$$

Notice that for each n, V_n consists of only finite sets, and each n appears in V_{n+1} for the first time. Observe that

$$V_\omega = \bigcup_{n\in\omega} V_n$$

consists only of finite sets; thus ω appears for the first time in $V_{\omega+1}$.

Question 90.7.3. How many sets are in V_5?

Definition 90.7.4. The **rank** of a set y, denoted $\operatorname{rank}(y)$, is the smallest ordinal α such that $y \in V_{\alpha+1}$.

Example 90.7.5

$\operatorname{rank}(2) = 2$, and actually $\operatorname{rank}(\alpha) = \alpha$ for any ordinal α (problem later). This is the reason for the extra "+1".

Question 90.7.6. Show that $\operatorname{rank}(y)$ is the smallest ordinal α such that $y \subseteq V_\alpha$.

It's not yet clear that the rank of a set actually exists, so we prove:

Theorem 90.7.7 (The von Neumann hierarchy is complete)

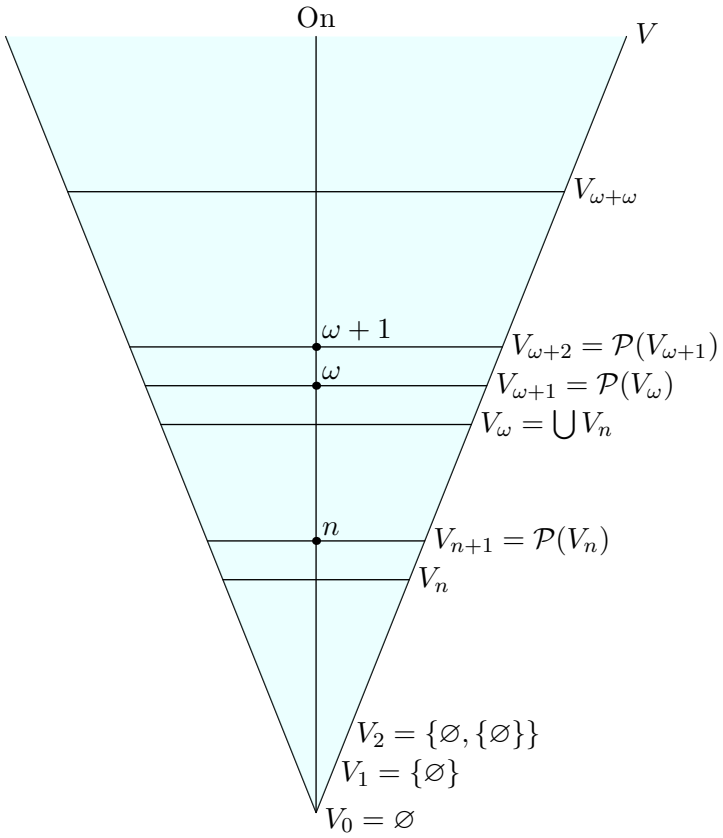
The class V is equal to $\bigcup_{\alpha \in \mathrm{On}} V_\alpha$. In other words, every set appears in some V_α.

Proof. Assume for contradiction this is false. The key is that because $\in$ satisfies Foundation, we can take a $\in$-minimal counterexample x. Thus $\operatorname{rank}(y)$ is defined for every $y \in x$, and we can consider (by Replacement) the set

$$\{\operatorname{rank}(y) \mid y \in x\}.$$

Since it is a set of ordinals, it is bounded. So there is some large ordinal α such that $y \in V_\alpha$ for all $y \in x$, i.e. $x \subseteq V_\alpha$, so $x \in V_{\alpha+1}$. □

This leads us to a picture of the universe V:



We can imagine the universe V as a triangle, built in several stages or layers, $V_0 \subsetneq V_1 \subsetneq V_2 \subsetneq \dots$. This universe doesn't have a top: but each of the V_i do. However, the universe has a very clear bottom. Each stage is substantially wider than the previous one.

In the center of this universe are the ordinals: for every successor V_α, exactly one new ordinal appears, namely α. Thus we can picture the class of ordinals as a thin line that stretches the entire height of the universe. A set has rank α if it appears at the same stage that α does.

All of number theory, the study of the integers, lives inside V_ω. Real analysis, the study of real numbers, lives inside $V_{\omega+1}$, since a real number can be encoded as a subset of $\mathbb{N}$ (by binary expansion). Functional analysis lives one step past that, $V_{\omega+2}$. For all intents and purposes, most mathematics does not go beyond $V_{\omega+\omega}$. This pales in comparison to the true magnitude of the whole universe.

§90.8 A few harder problems to think about

Problem 90A. Prove that $\operatorname{rank}(\alpha) = \alpha$ for any α by transfinite induction.

Problem 90B (Online Math Open)**.** Count the number of transitive sets in V_5.

Problem 90C (Goodstein)**.** Let a_2 be any positive integer. We define the infinite sequence a_2, a_3, ... recursively as follows. If $a_n = 0$, then $a_{n+1} = 0$. Otherwise, we write a_n in base n, then write all exponents in base n, and so on until all numbers in the expression are at most n. Then we replace all instances of n by $n+1$ (including the

exponents!), subtract 1, and set the result to a_{n+1}. For example, if $a_2 = 11$ we have

$$\begin{aligned}
a_2 &= 2^3 + 2 + 1 = 2^{2+1} + 2 + 1 \\
a_3 &= 3^{3+1} + 3 + 1 - 1 = 3^{3+1} + 3 \\
a_4 &= 4^{4+1} + 4 - 1 = 4^{4+1} + 3 \\
a_5 &= 5^{5+1} + 3 - 1 = 5^{5+1} + 2
\end{aligned}$$

and so on. Prove that $a_N = 0$ for some integer $N > 2$.

91 Cardinals

An ordinal measures a total ordering. However, it does not do a fantastic job at measuring size. For example, there is a bijection between the elements of ω and $\omega+1$:

$$\begin{aligned} \omega+1 &= \{\ \omega\ 0\ 1\ 2\ \dots\ \} \\ \omega &= \{\ 0\ 1\ 2\ 3\ \dots\ \}. \end{aligned}$$

In fact, as you likely already know, there is even a bijection between ω and ω^2:

$+$	0	1	2	3	4	$\dots$
0	0	1	3	6	10	$\dots$
ω	2	4	7	11	$\dots$	
$\omega\cdot 2$	5	8	12	$\dots$		
$\omega\cdot 3$	9	13	$\dots$			
$\omega\cdot 4$	14	$\dots$				

So ordinals do not do a good job of keeping track of size. For this, we turn to the notion of a cardinal number.

§91.1 Equinumerous sets and cardinals

Definition 91.1.1. Two sets A and B are **equinumerous**, written $A \approx B$, if there is a bijection between them.

Definition 91.1.2. A **cardinal** is an ordinal κ such that for no $\alpha < \kappa$ do we have $\alpha \approx \kappa$.

Example 91.1.3 (Examples of cardinals)
Every finite number is a cardinal. Moreover, ω is a cardinal. However, $\omega+1$, ω^2, ω^{2015} are not, because they are countable.

Example 91.1.4 (ω^ω is countable)
Even ω^ω is not a cardinal, since it is a countable union

$$\omega^\omega = \bigcup_n \omega^n$$

and each ω^n is countable.

Question 91.1.5. Why must an infinite cardinal be a limit ordinal?

Remark 91.1.6 — There is something fishy about the definition of a cardinal: it relies on an *external* function f. That is, to verify κ is a cardinal I can't just look at κ itself; I need to examine the entire universe V to make sure there does not exist a bijection $f\colon \kappa \to \alpha$ for $\alpha < \kappa$. For now this is no issue, but later in model

theory this will lead to some highly counterintuitive behavior.

§91.2 Cardinalities

Now that we have defined a cardinal, we can discuss the size of a set by linking it to a cardinal.

Definition 91.2.1. The **cardinality** of a set X is the *least* ordinal κ such that $X \approx \kappa$. We denote it by $|X|$.

Question 91.2.2. Why must $|X|$ be a cardinal?

Remark 91.2.3 — One needs the well-ordering theorem (equivalently, choice) in order to establish that such an ordinal κ actually exists.

Since cardinals are ordinals, it makes sense to ask whether $\kappa_1 \le \kappa_2$, and so on. Our usual intuition works well here.

Proposition 91.2.4 (Restatement of cardinality properties)
Let X and Y be sets.

(i) $X \approx Y$ if and only if $|X| = |Y|$, if and only if there's a bijection from X to Y.

(ii) $|X| \le |Y|$ if and only if there is an injective map $X \hookrightarrow Y$.

Diligent readers are invited to try and prove this.

§91.3 Aleph numbers

Prototypical example for this section: $\aleph_0 = \omega$, *and* $\aleph_1$ *is the first uncountable ordinal.*

First, let us check that cardinals can get arbitrarily large:

Proposition 91.3.1
We have $|X| < |\mathcal{P}(X)|$ for every set X.

Proof. There is an injective map $X \hookrightarrow \mathcal{P}(X)$ but there is no injective map $\mathcal{P}(X) \hookrightarrow X$ by Lemma 89.2.1. □

Thus we can define:

Definition 91.3.2. For a cardinal κ, we define κ^+ to be the least cardinal above κ, called the **successor cardinal**.

This κ^+ exists and has $\kappa^+ \le |\mathcal{P}(\kappa)|$.
Next, we claim that:

Exercise 91.3.3. Show that if A is a set of cardinals, then $\cup A$ is a cardinal.

Thus by transfinite induction we obtain that:

Definition 91.3.4. For any $\alpha \in \mathrm{On}$, we define the **aleph numbers** as

$$\begin{aligned} \aleph_0 &= \omega \\ \aleph_{\alpha+1} &= (\aleph_\alpha)^+ \\ \aleph_\lambda &= \bigcup_{\alpha<\lambda} \aleph_\alpha. \end{aligned}$$

Thus we have the sequence of cardinals

$$0 < 1 < 2 < \cdots < \aleph_0 < \aleph_1 < \cdots < \aleph_\omega < \aleph_{\omega+1} < \ldots.$$

By definition, $\aleph_0$ is the cardinality of the natural numbers, $\aleph_1$ is the first uncountable ordinal, ...

We claim the aleph numbers constitute all the cardinals:

Lemma 91.3.5 (Aleph numbers constitute all infinite cardinals)
If κ is a cardinal then either κ is finite (i.e. $\kappa \in \omega$) or $\kappa = \aleph_\alpha$ for some $\alpha \in \mathrm{On}$.

Proof. Assume κ is infinite, and take α minimal with $\aleph_\alpha \geq \kappa$. Suppose for contradiction that we have $\aleph_\alpha > \kappa$. We may assume $\alpha > 0$, since the case $\alpha = 0$ is trivial.

If $\alpha = \overline{\alpha} + 1$ is a successor, then

$$\aleph_{\overline{\alpha}} < \kappa < \aleph_\alpha = (\aleph_{\overline{\alpha}})^+$$

which contradicts the definition of the successor cardinal.

If $\alpha = \lambda$ is a limit ordinal, then $\aleph_\lambda$ is the supremum $\bigcup_{\gamma<\lambda} \aleph_\gamma$. So there must be some $\gamma < \lambda$ with $\aleph_\gamma > \kappa$, which contradicts the minimality of α. □

Definition 91.3.6. An infinite cardinal which is not a successor cardinal is called a **limit cardinal**. It is exactly those cardinals of the form $\aleph_\lambda$, for λ a limit ordinal, plus $\aleph_0$.

§91.4 Cardinal arithmetic

Prototypical example for this section: $\aleph_0 \cdot \aleph_0 = \aleph_0 + \aleph_0 = \aleph_0$

Recall the way we set up ordinal arithmetic. Note that in particular, $\omega + \omega > \omega$ and $\omega^2 > \omega$. Since cardinals count size, this property is undesirable, and we want to have

$$\begin{aligned} \aleph_0 + \aleph_0 &= \aleph_0 \\ \aleph_0 \cdot \aleph_0 &= \aleph_0 \end{aligned}$$

because $\omega + \omega$ and $\omega \cdot \omega$ are countable. In the case of cardinals, we simply "ignore order".

The definition of cardinal arithmetic is as expected:

Definition 91.4.1 (Cardinal arithmetic)**.** Given cardinals κ and μ, define

$$\kappa + \mu \coloneqq |(\{0\} \times \kappa) \cup (\{1\} \times \mu)|$$

and

$$\kappa \cdot \mu \coloneqq |\mu \times \kappa|.$$

Question 91.4.2. Check this agrees with what you learned in pre-school for finite cardinals.

Abuse of Notation 91.4.3. This is a slight abuse of notation since we are using the same symbols as for ordinal arithmetic, even though the results are different ($\omega \cdot \omega = \omega^2$ but $\aleph_0 \cdot \aleph_0 = \aleph_0$). In general, I'll make it abundantly clear whether I am talking about cardinal arithmetic or ordinal arithmetic.

To help combat this confusion, we use separate symbols for ordinals and cardinals. Specifically, ω will always refer to $\{0, 1, \dots\}$ viewed as an ordinal; $\aleph_0$ will always refer to the same set viewed as a cardinal. More generally,

Definition 91.4.4. Let $\omega_\alpha = \aleph_\alpha$ viewed as an ordinal.

However, as we've seen already we have that $\aleph_0 \cdot \aleph_0 = \aleph_0$. In fact, this holds even more generally:

Theorem 91.4.5 (Infinite cardinals squared)
Let κ be an infinite cardinal. Then $\kappa \cdot \kappa = \kappa$.

Proof. Obviously $\kappa \cdot \kappa \geq \kappa$, so we want to show $\kappa \cdot \kappa \leq \kappa$.

The idea is to try to repeat the same proof that we had for $\aleph_0 \cdot \aleph_0 = \aleph_0$, so we re-iterate it here. We took the "square" of elements of $\aleph_0$, and then *re-ordered* it according to the diagonal:

	0	1	2	3	4	…
0	0	1	3	6	10	…
1	2	4	7	11	…	
2	5	8	12	…		
3	9	13	…			
4	14	…				

We'd like to copy this idea for a general κ; however, since addition is less well-behaved for infinite ordinals it will be more convenient to use $\max\{\alpha, \beta\}$ rather than $\alpha + \beta$. Specifically, we put the ordering $<_{\max}$ on $\kappa \times \kappa$ as follows: for (α_1, β_1) and (α_2, β_2) in $\kappa \times \kappa$ we declare $(\alpha_1, \beta_1) <_{\max} (\alpha_2, \beta_2)$ if

- $\max\{\alpha_1, \beta_1\} < \max\{\alpha_2, \beta_2\}$ or
- $\max\{\alpha_1, \beta_1\} = \max\{\alpha_2, \beta_2\}$ and (α_1, β_1) is lexicographically earlier than (α_2, β_2).

This alternate ordering (which deliberately avoids referring to the addition) looks like:

	0	1	2	3	4	…
0	0	1	4	9	16	…
1	2	3	5	10	17	…
2	6	7	8	11	18	…
3	12	13	14	15	19	…
4	20	21	22	23	24	…
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Now we proceed by transfinite induction on κ. The base case is $\kappa = \aleph_0$, done above. Now, $<_{\max}$ is a well-ordering of $\kappa \times \kappa$, so we know it is in order-preserving bijection with

some ordinal γ. Our goal is to show that $|\gamma| \le \kappa$. To do so, it suffices to prove that for any $\overline{\gamma} \in \gamma$, we have $|\overline{\gamma}| < \kappa$.

Suppose $\overline{\gamma}$ corresponds to the point $(\alpha, \beta) \in \kappa \times \kappa$ under this bijection. If α and β are both finite then certainly $\overline{\gamma}$ is finite too. Otherwise, let $\overline{\kappa} = \max\{\alpha, \beta\} < \kappa$; then the number of points below $\overline{\gamma}$ is at most

$$|\alpha| \cdot |\beta| \le \overline{\kappa} \cdot \overline{\kappa} = \overline{\kappa}$$

by the inductive hypothesis. So $|\overline{\gamma}| \le \overline{\kappa} < \kappa$ as desired. □

From this it follows that cardinal addition and multiplication is really boring:

Theorem 91.4.6 (Infinite cardinal arithmetic is trivial)
Given cardinals κ and μ, one of which is infinite, we have

$$\kappa \cdot \mu = \kappa + \mu = \max\{\kappa, \mu\}.$$

Proof. The point is that both of these are less than the square of the maximum. Writing out the details:

$$\begin{aligned} \max\{\kappa, \mu\} &\le \kappa + \mu \\ &\le \kappa \cdot \mu \\ &\le \max\{\kappa, \mu\} \cdot \max\{\kappa, \mu\} \\ &= \max\{\kappa, \mu\}. \end{aligned}$$

□

§91.5 Cardinal exponentiation

Prototypical example for this section: $2^\kappa = |\mathcal{P}(\kappa)|$.

Definition 91.5.1. Suppose κ and λ are cardinals. Then

$$\kappa^\lambda := |\mathscr{F}(\lambda, \kappa)|.$$

Here $\mathscr{F}(A, B)$ is the set of functions from A to B.

Abuse of Notation 91.5.2. As before, we are using the same notation for both cardinal and ordinal arithmetic. Sorry!

In particular, $2^\kappa = |\mathcal{P}(\kappa)| > \kappa$, and so from now on we can use the notation 2^κ freely. (Note that this is totally different from ordinal arithmetic; there we had $2^\omega = \bigcup_{n \in \omega} 2^n = \omega$. In cardinal arithmetic $2^{\aleph_0} > \aleph_0$.)

I have unfortunately not told you what $2^{\aleph_0}$ equals. A natural conjecture is that $2^{\aleph_0} = \aleph_1$; this is called the **Continuum Hypothesis**. It turns out that this is *undecidable* – it is not possible to prove or disprove this from the ZFC axioms.

§91.6 Cofinality

Prototypical example for this section: $\aleph_0$, $\aleph_1$, ... *are all regular, but* $\aleph_\omega$ *has cofinality* ω.

Definition 91.6.1. Let λ be an ordinal (usually a limit ordinal), and α another ordinal. A map $f : \alpha \to \lambda$ of ordinals is called **cofinal** if for every $\overline{\lambda} < \lambda$, there is some $\overline{\alpha} \in \alpha$ such that $f(\overline{\alpha}) \ge \overline{\lambda}$. In other words, the map reaches arbitrarily high into λ.

Example 91.6.2 (Example of a cofinal map)

(a) The map $\omega \to \omega^\omega$ by $n \mapsto \omega^n$ is cofinal.

(b) For any ordinal α, the identity map $\alpha \to \alpha$ is cofinal.

Definition 91.6.3. Let λ be a limit ordinal. The **cofinality** of λ, denoted $\operatorname{cof}(\lambda)$, is the smallest ordinal α such that there is a cofinal map $\alpha \to \lambda$.

Question 91.6.4. Why must α be an infinite cardinal?

Usually, we are interested in taking the cofinality of a cardinal κ.

Pictorially, you can imagine standing at the bottom of the universe and looking up the chain of ordinals to κ. You have a machine gun and are firing bullets upwards, and you want to get arbitrarily high but less than κ. The cofinality is then the number of bullets you need to do this.

We now observe that "most" of the time, the cofinality of a cardinal is itself.[1] Such a cardinal is called **regular**.

Example 91.6.5 ($\aleph_0$ is regular)

$\operatorname{cof}(\aleph_0) = \aleph_0$, because no finite subset of $\aleph_0 = \omega$ can reach arbitrarily high.

Example 91.6.6 ($\aleph_1$ is regular)

$\operatorname{cof}(\aleph_1) = \aleph_1$. Indeed, assume for contradiction that some countable set of ordinals $A = \{\alpha_0, \alpha_1, \dots\} \subseteq \aleph_1$ reaches arbitrarily high inside $\aleph_1$. Then $\Lambda = \cup A$ is a *countable* ordinal, because it is a countable union of countable ordinals. In other words $\Lambda \in \aleph_1$. But Λ is an upper bound for A, contradiction.

On the other hand, there *are* cardinals which are not regular; since these are the "rare" cases we call them **singular**.

Example 91.6.7 ($\aleph_\omega$ is not regular)

Notice that $\aleph_0 < \aleph_1 < \aleph_2 < \dots$ reaches arbitrarily high in $\aleph_\omega$, despite only having $\aleph_0$ terms. It follows that $\operatorname{cof}(\aleph_\omega) = \aleph_0$.

We now confirm a suspicion you may have:

Theorem 91.6.8 (Successor cardinals are regular)

If $\kappa = \overline{\kappa}^+$ is a successor cardinal, then it is regular.

Proof. We copy the proof that $\aleph_1$ was regular.

Assume for contradiction that for some $\mu \le \overline{\kappa}$, there are μ sets reaching arbitrarily high in κ as a cardinal. Observe that each of these sets must have cardinality at most $\overline{\kappa}$.

[1] Be careful — the cofinality of an *ordinal* is usually strictly less than itself. In fact, if the cofinality of an ordinal is itself, then that ordinal must be a cardinal.

We take the union of all μ sets, which gives an ordinal Λ serving as an upper bound.

The number of elements in the union is at most

$$\#\text{sets} \cdot \#\text{elms} \le \mu \cdot \overline{\kappa} = \overline{\kappa}$$

and hence $|\Lambda| \le \overline{\kappa} < \kappa$. □

§91.7 Inaccessible cardinals

So, what about limit cardinals? It seems that most of them are singular: if $\aleph_\lambda \neq \aleph_0$ is a limit cardinal (that is, λ is a limit ordinal), then the sequence $\{\aleph_\alpha\}_{\alpha\in\lambda}$ (of length λ) is certainly cofinal.

Example 91.7.1 (Beth fixed point)

Consider the monstrous cardinal

$$\kappa = \aleph_{\aleph_{\aleph_{\ddots}}}.$$

This might look frighteningly huge, as $\kappa = \aleph_\kappa$, but its cofinality is ω as it is the limit of the sequence

$$\aleph_0, \aleph_{\aleph_0}, \aleph_{\aleph_{\aleph_0}}, \dots$$

More generally, one can in fact prove that

$$\operatorname{cof}(\aleph_\lambda) = \operatorname{cof}(\lambda).$$

But it is actually conceivable that λ is so large that $\lambda = \aleph_\lambda$.

A regular limit cardinal other than $\aleph_0$ has a special name: it is **weakly inaccessible**. Such cardinals are so large that it is impossible to prove or disprove their existence in ZFC. It is the first of many so-called "large cardinals".

An infinite cardinal κ is a **strong limit cardinal** if

$$\forall \overline{\kappa} < \kappa \quad 2^{\overline{\kappa}} < \kappa$$

for any cardinal $\overline{\kappa}$. For example, $\aleph_0$ is a strong limit cardinal.

Question 91.7.2. Why must strong limit cardinals actually be limit cardinals? (This is offensively easy.)

Remark 91.7.3 — A limit cardinal can equivalently be defined as a nonzero cardinal κ such that

$$\forall \overline{\kappa} < \kappa \quad (\overline{\kappa})^+ < \kappa.$$

If you compare it with the definition of strong limit cardinals, you can see the parallel. (This remark also gives an answer to the previous question.)

A regular strong limit cardinal other than $\aleph_0$ is called **strongly inaccessible**.

§91.8 A few harder problems to think about

Problem 91A. Compute $|V_\omega|$.

Problem 91B. Prove that for any limit ordinal α, $\text{cof}(\alpha)$ is a *regular* cardinal.

Problem 91C* (Strongly inaccessible cardinals)**.** Show that for any strongly inaccessible κ, we have $|V_\kappa| = \kappa$.

Problem 91D (König's theorem)**.** Show that

$$\kappa^{\text{cof}(\kappa)} > \kappa$$

for every infinite cardinal κ.

XXII

Set Theory II: Model Theory and Forcing

Part XXII: Contents

92 Inner model theory

Model theory is *really* meta, so you will have to pay attention here.

Roughly, a "model of ZFC" is a set with a binary relation that satisfies the ZFC axioms, just as a group is a set with a binary operation that satisfies the group axioms. Unfortunately, unlike with groups, it is very hard for me to give interesting examples of models, for the simple reason that we are literally trying to model the entire universe.

§92.1 Models

Prototypical example for this section: $(\omega, \in)$ *obeys* PowerSet, V_κ *is a model for* κ *inaccessible (later).*

Definition 92.1.1. A **model** $\mathscr{M}$ consists of a set M and a binary relation $E \subseteq M \times M$. (The E relation is the "$\in$" for the model.)

> **Remark 92.1.2 —** I'm only considering *set-sized* models where M is a set. Experts may be aware that I can actually play with M being a class, but that would require too much care for now.

If you have a model, you can ask certain things about it, such as "does it satisfy EmptySet?". Let me give you an example of what I mean and then make it rigorous.

Example 92.1.3 (A stupid model)

Let's take $\mathscr{M} = (M, E) = (\omega, \in)$. This is not a very good model of ZFC, but let's see if we can make sense of some of the first few axioms.

(a) $\mathscr{M}$ satisfies Extensionality, which is the sentence

$$\forall x \forall y \forall a : (a \in x \iff a \in y) \implies x = y.$$

This just follows from the fact that E is actually $\in$.

(b) $\mathscr{M}$ satisfies EmptySet, which is the sentence

$$\exists a : \forall x \, \neg(x \in a).$$

Namely, take $a = \varnothing \in \omega$.

(c) $\mathscr{M}$ does not satisfy Pairing, since $\{1, 3\}$ is not in ω, even though $1, 3 \in \omega$.

(d) Miraculously, $\mathscr{M}$ satisfies Union, since for any $n \in \omega$, $\cup n$ is $n-1$ (unless $n = 0$). The Union axiom states that

$$\forall a \exists U \quad \forall x \; [(x \in U) \iff (\exists y : x \in y \in a)] \,.$$

An important thing to notice is that the "$\forall a$" ranges only over the sets in the model of the universe, $\mathscr{M}$.

Example 92.1.4 (Important: this stupid model satisfies PowerSet)

Most incredibly of all: $\mathscr{M} = (\omega, \in)$ satisfies PowerSet. This is a really important example.

You might think this is ridiculous. Look at $2 = \{0, 1\}$. The power set of this is $\{0, 1, 2, \{1\}\}$ which is not in the model, right?

Well, let's look more closely at PowerSet. It states that:

$$\forall x \exists a \forall y (y \in a \iff y \subseteq x).$$

What happens if we set $x = 2 = \{0, 1\}$? Well, actually, we claim that $a = 3 = \{0, 1, 2\}$ works. The key point is "for all y" – this *only ranges over the objects in* $\mathscr{M}$. In $\mathscr{M}$, the only subsets of 2 are $0 = \varnothing$, $1 = \{0\}$ and $2 = \{0, 1\}$. The "set" $\{1\}$ in the "real world" (in V) is not a set in the model $\mathscr{M}$.

In particular, you might say that in this strange new world, we have $2^n = n + 1$, since $n = \{0, 1, \dots, n-1\}$ really does have only $n + 1$ subsets.

Example 92.1.5 (Sentences with parameters)

The sentences we ask of our model are allowed to have "parameters" as well. For example, if $\mathscr{M} = (\omega, \in)$ as before then $\mathscr{M}$ satisfies the sentence

$$\forall x \in 3 (x \in 5).$$

§92.2 Sentences and satisfaction

With this intuitive notion, we can define what it means for a model to satisfy a sentence.

Definition 92.2.1. Note that any sentence ϕ can be written in one of five forms:

- $x \in y$
- $x = y$
- $\neg\psi$ ("not ψ") for some shorter sentence ψ
- $\psi_1 \vee \psi_2$ ("ψ_1 or ψ_2") for some shorter sentences ψ_1, ψ_2
- $\exists x \psi$ ("exists x") for some shorter sentence ψ.

Question 92.2.2. What happened to $\wedge$ (and) and $\forall$ (for all)? (Hint: use $\neg$.)

Often (almost always, actually) we will proceed by so-called "induction on formula complexity", meaning that we define or prove something by induction using this. Note that we require all formulas to be finite.

Now suppose we have a sentence ϕ, like $a = b$ or $\exists a \forall x \neg(x \in a)$, plus a model $\mathscr{M} = (M, E)$. We want to ask whether $\mathscr{M}$ satisfies ϕ.

To give meaning to this, we have to designate certain variables as **parameters**. For example, if I asked you

"Does $a = b$?"

the first question you would ask is what a and b are. So a, b would be parameters: I have to give them values for this sentence to make sense.

On the other hand, if I asked you

"Does $\exists a \forall x \neg(x \in a)$?"

then you would just say "yes". In this case, x and a are *not* parameters. In general, parameters are those variables whose meaning is not given by some $\forall$ or $\exists$.

In what follows, we will let $\phi(x_1, \dots, x_n)$ denote a formula ϕ, whose parameters are $x_1, \dots, x_n$. Note that possibly $n = 0$, for example all ZFC axioms have no parameters.

> **Question 92.2.3.** Try to guess the definition of satisfaction before reading it below. (It's not very hard to guess!)

Definition 92.2.4. Let $\mathscr{M} = (M, E)$ be a model. Let $\phi(x_1, \dots, x_n)$ be a sentence, and let $b_1, \dots, b_n \in M$. We will define a relation

$$\mathscr{M} \vDash \phi[b_1, \dots, b_n]$$

and say $\mathscr{M}$ **satisfies** the sentence ϕ with parameters $b_1, \dots, b_n$.

The relationship is defined by induction on formula complexity as follows:

- If ϕ is "$x_1 = x_2$" then $\mathscr{M} \vDash \phi[b_1, b_2] \iff b_1 = b_2$.
- If ϕ is "$x_1 \in x_2$" then $\mathscr{M} \vDash \phi[b_1, b_2] \iff b_1 \; E \; b_2$.
  (This is what we mean by "E interprets $\in$".)
- If ϕ is "$\neg\psi$" then $\mathscr{M} \vDash \phi[b_1, \dots, b_n] \iff \mathscr{M} \nvDash \psi[b_1, \dots, b_n]$.
- If ϕ is "$\psi_1 \vee \psi_2$" then $\mathscr{M} \vDash \phi[b_1, \dots, b_n]$ means $\mathscr{M} \vDash \psi_i[b_1, \dots, b_n]$ for some $i = 1, 2$.
- Most important case: suppose ϕ is $\exists x \psi(x, x_1, \dots, x_n)$. Then $\mathscr{M} \vDash \phi[b_1, \dots, b_n]$ if and only if
  $$\exists b \in M \text{ such that } \mathscr{M} \vDash \psi[b, b_1, \dots, b_n].$$
  Note that ψ has one extra parameter.

Notice where the information of the model actually gets used. We only ever use E in interpreting $x_1 \in x_2$; unsurprising. But we only ever use the set M when we are running over $\exists$ (and hence $\forall$). That's well-worth keeping in mind:

> **The behavior of a model essentially comes from $\exists$ and $\forall$, which search through the entire model M.**

And finally,

Definition 92.2.5. A **model of ZFC** is a model $\mathscr{M} = (M, E)$ satisfying all ZFC axioms.

We are especially interested in models of the form $(M, \in)$, where M is a *transitive* set. (We want our universe to be transitive, otherwise we would have elements of sets which are not themselves in the universe, which is very strange.) Such a model is called a **transitive model**.

Abuse of Notation 92.2.6. If M is a transitive set, the model $(M, \in)$ will be abbreviated to just M.

Definition 92.2.7. An **inner model** of ZFC is a transitive model satisfying ZFC.

Remark 92.2.8 — The definition of a model of ZFC only uses $M \vDash \varphi$ where φ has no parameters; nevertheless, you can see that we define what $M \vDash \varphi$ means when φ has parameters because it's used in the definition of $M \vDash \exists x \psi(x)$.
The extension $\varphi(x_1, \dots, x_n)$ is written with round parentheses, but $M \vDash \varphi[b_1, \dots, b_n]$ is written with square brackets — you can think of it as "formally substitute" the parameters $b_1, \dots, b_n$ into φ, because if $b_1, \dots, b_n$ is "actually" substituted into φ, then $\varphi(b_1, \dots, b_n)$ is just a single boolean value.

§92.3 The Levy hierarchy

Prototypical example for this section: $\texttt{isSubset}(x, y)$ *is absolute. The axiom* EmptySet *is* Σ_1, $\texttt{isPowerSetOf}(X, x)$ *is* Π_1.

A key point to remember is that the behavior of a model is largely determined by $\exists$ and $\forall$. It turns out we can say even more than this.

Consider a formula such as

$$\texttt{isEmpty}(x) : \neg\exists a (a \in x)$$

which checks whether a given set x has an element in it. Technically, this has an "$\exists$" in it. But somehow this $\exists$ does not really search over the entire model, because it is *bounded* to search in x. That is, we might informally rewrite this as

$$\neg(\exists a \in x)$$

which doesn't fit into the strict form, but points out that we are only looking over $a \in x$. We call such a quantifier a **bounded quantifier**.

We like sentences with bounded quantifiers because they designate properties which are **absolute** over transitive models. It doesn't matter how strange your surrounding model M is. As long as M is transitive,

$$M \vDash \texttt{isEmpty}(\varnothing)$$

will always hold. Similarly, the sentence

$$\texttt{isSubset}(x, y) : x \subseteq y \text{ i.e. } \forall a \in x (a \in y)$$

is absolute. Sentences with this property are called Σ_0 or Π_0.

The situation is different with a sentence like

$$\texttt{isPowerSetOf}(y, x) : \forall z \, (z \subseteq x \iff z \in y)$$

which in English means "y is the power set of x", or just $y = \mathcal{P}(x)$. The $\forall z$ is *not* bounded here. This weirdness is what allows things like

$$\omega \vDash \text{“}\{0, 1, 2\} \text{ is the power set of } \{0, 1\}\text{”}$$

and hence

$$\omega \vDash \text{PowerSet}$$

which was our stupid example earlier. The sentence $\texttt{isPowerSetOf}$ consists of an unbounded $\forall$ followed by an absolute sentence, so we say it is Π_1.

More generally, the **Levy hierarchy** keeps track of how bounded our quantifiers are. Specifically,

- Formulas which have only bounded quantifiers are $\Delta_0 = \Sigma_0 = \Pi_0$.
- Formulas of the form $\exists x_1 \dots \exists x_k \psi$ where ψ is Π_n are considered Σ_{n+1}.
- Formulas of the form $\forall x_1 \dots \forall x_k \psi$ where ψ is Σ_n are considered Π_{n+1}.

(A formula which is both Σ_n and Π_n is called Δ_n, but we won't use this except for $n = 0$.)

Example 92.3.1 (Examples of Δ_0 sentences)

(a) The sentences $\texttt{isEmpty}(x)$, $x \subseteq y$, as discussed above.

(b) The formula "x is transitive" can be expanded as a Δ_0 sentence.

(c) The formula "x is an ordinal" can be expanded as a Δ_0 sentence.

Exercise 92.3.2. Write out the expansions for "x is transitive" and "x is an ordinal" in a Δ_0 form.

Example 92.3.3 (More complex formulas)

(a) The axiom EmptySet is Σ_1; it is $\exists a(\texttt{isEmpty}(a))$, and $\texttt{isEmpty}(a)$ is Δ_0.

(b) The formula "$y = \mathcal{P}(x)$" is Π_1, as discussed above.

(c) The formula "x is countable" is Σ_1. One way to phrase it is "$\exists f$ an injective map $x \hookrightarrow \omega$", which necessarily has an unbounded "$\exists f$".

(d) The axiom PowerSet is Π_3:

$$\forall y \exists P \forall x (x \subseteq y \iff x \in P).$$

Remark 92.3.4 (Why only alternating unbounded quantifier count?) — Note that a formula $\exists a \exists b \psi(a, b)$ can alternatively be written as $\exists c(c$ is an ordered pair $(a, b) \wedge \psi(a, b))$, which explains why we only want to consider the formula $\exists a \exists b \psi(a, b)$ as Σ_1.

§92.4 Substructures, and Tarski-Vaught

Let $\mathscr{M}_1 = (M_1, E_1)$ and $\mathscr{M}_2 = (M_2, E_2)$ be models.

Definition 92.4.1. We say that $\mathscr{M}_1 \subseteq \mathscr{M}_2$ if $M_1 \subseteq M_2$ and E_1 agrees with E_2; we say $\mathscr{M}_1$ is a **substructure** of $\mathscr{M}_2$.

That's boring. The good part is:

Definition 92.4.2. We say $\mathscr{M}_1 \prec \mathscr{M}_2$, or $\mathscr{M}_1$ is an **elementary substructure** of $\mathscr{M}_2$, if $\mathscr{M}_1 \subseteq \mathscr{M}_2$ and for *every* sentence $\phi(x_1, \dots, x_n)$ and parameters $b_1, \dots, b_n \in M_1$, we have

$$\mathscr{M}_1 \vDash \phi[b_1, \dots, b_n] \iff \mathscr{M}_2 \vDash \phi[b_1, \dots, b_n].$$

In other words, $\mathscr{M}_1$ and $\mathscr{M}_2$ agree on every sentence possible. Note that the b_i have to come from $\mathscr{M}_1$; if the b_i came from $\mathscr{M}_2$ then asking something of $\mathscr{M}_1$ wouldn't make sense.

Let's ask now: how would $\mathscr{M}_1 \prec \mathscr{M}_2$ fail to be true? If we look at the possible sentences, none of the atomic formulas, nor the "$\wedge$" and "$\neg$", are going to cause issues.

The intuition you should be getting by now is that things go wrong once we hit $\forall$ and $\exists$. They won't go wrong for bounded quantifiers. But unbounded quantifiers search the entire model, and that's where things go wrong.

To give a "concrete example": imagine $\mathscr{M}_1$ is MIT, and $\mathscr{M}_2$ is the state of Massachusetts. If $\mathscr{M}_1$ thinks there exist hackers at MIT, certainly there exist hackers in Massachusetts. Where things go wrong is something like:

$$\mathscr{M}_2 \vDash \text{“}\exists x : x \text{ is a course numbered } > 50\text{”}.$$

This is true for $\mathscr{M}_2$ because we can take the witness $x = $ Math 55, say. But it's false for $\mathscr{M}_1$, because at MIT all courses are numbered 18.701 or something similar.

The issue is that the *witnesses* for statements in $\mathscr{M}_2$ do not necessarily propagate down to witnesses for $\mathscr{M}_1$.

The Tarski-Vaught test says this is the only impediment: if every witness in $\mathscr{M}_2$ can be replaced by one in $\mathscr{M}_1$ then $\mathscr{M}_1 \prec \mathscr{M}_2$.

Lemma 92.4.3 (Tarski-Vaught)
Let $\mathscr{M}_1 \subseteq \mathscr{M}_2$. Then $\mathscr{M}_1 \prec \mathscr{M}_2$ if and only if: For every sentence $\phi(x, x_1, \dots, x_n)$ and parameters $b_1, \dots, b_n \in M_1$: if there is a witness $\tilde{b} \in M_2$ to $\mathscr{M}_2 \vDash \phi(\tilde{b}, b_1 \dots, b_n)$ then there is a witness $b \in M_1$ to $\mathscr{M}_1 \vDash \phi(b, b_1, \dots, b_n)$.

Proof. Easy after the above discussion. To formalize it, use induction on formula complexity. □

§92.5 Obtaining the axioms of ZFC

We now want to write down conditions for M to satisfy ZFC axioms. The idea is that almost all the ZFC axioms are just Σ_1 claims about certain desired sets, and so verifying an axiom reduces to checking some appropriate "closure" condition: that the witness to the axiom is actually in the model.

For example, the EmptySet axiom is "$\exists a(\texttt{isEmpty}(a))$", and so we're happy as long as $\varnothing \in M$, which is of course true for any nonempty transitive set M.

Lemma 92.5.1 (Transitive sets inheriting ZFC)
Let M be a nonempty transitive set. Then

(i) M satisfies Extensionality, Foundation, EmptySet.

(ii) $M \vDash$ Pairing if $x, y \in M \implies \{x, y\} \in M$.

(iii) $M \vDash$ Union if $x \in M \implies \cup x \in M$.

(iv) $M \vDash$ PowerSet if $x \in M \implies \mathcal{P}(x) \cap M \in M$.

(v) $M \vDash$ Replacement if for every $x \in M$ and every function $F \colon x \to M$ which is M-definable with parameters, we have $F^{\text{img}}(x) \in M$ as well.

(vi) $M \vDash$ Infinity as long as $\omega \in M$.

Here, a set $X \subseteq M$ is **M-definable with parameters** if it can be realized as

$$X = \{x \in M \mid \phi[x, b_1, \dots, b_n]\}$$

for some (fixed) choice of parameters $b_1, \dots, b_n \in M$. We allow $n = 0$, in which case we say X is **M-definable without parameters**. Note that X need not itself be in M! As a trivial example, $X = M$ is M-definable without parameters (just take $\phi[x]$ to always be true), and certainly we do not have $X \in M$.

Exercise 92.5.2. Verify (i)-(iv) above.

Remark 92.5.3 — Converses to the statements of Lemma 92.5.1 are true for all claims other than (vi).

§92.6 Mostowski collapse

Up until now I have been only talking about transitive models, because they were easier to think about. Here's a second, better reason we might only care about transitive models.

Lemma 92.6.1 (Mostowski collapse lemma)
Let $X = (X, \in)$ be a model satisfying Extensionality, where X is a set (possibly not transitive). Then there exists an isomorphism $\pi \colon X \to M$ for a transitive model $M = (M, \in)$.

This is also called the *transitive collapse*. In fact, both π and M are unique.

Proof. The idea behind the proof is very simple. Since $\in$ is well-founded and extensional (satisfies Foundation and Extensionality, respectively), we can look at the $\in$-minimal element $x_\varnothing$ of X with respect to $\in$. Clearly, we want to send that to $0 = \varnothing$.

Then we take the next-smallest set under $\in$, and send it to $1 = \{\varnothing\}$. We "keep doing this"; it's not hard to see this does exactly what we want.

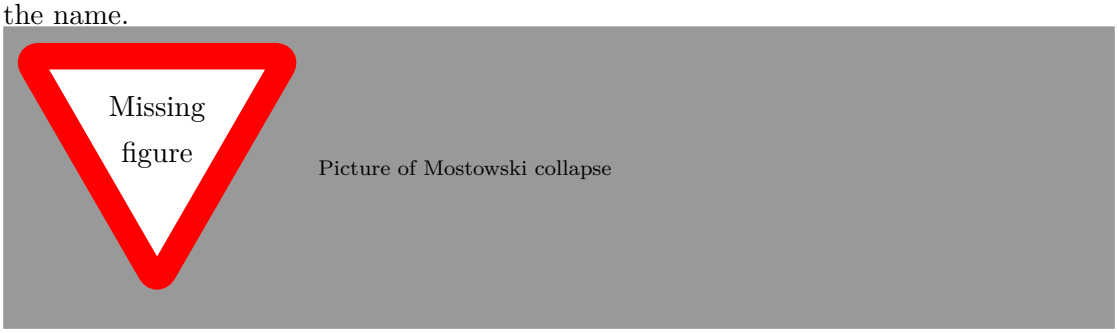
To formalize, define π by transfinite recursion:

$$\pi(x) \coloneqq \{\pi(y) \mid y \in x\}.$$

This π, by construction, does the trick. □

Remark 92.6.2 (Digression for experts) — Earlier versions of Napkin claimed this was true for general models $\mathscr{X} = (X, E)$ with $\mathscr{X} \vDash$ Foundation + Extensionality. This is false; it does not even imply E is well-founded, because there may be infinite descending chains of subsets of X which do not live in X itself. Another issue is that E may not be set-like.

The picture of this is "collapsing" the elements of M down to the bottom of V, hence the name.



Picture of Mostowski collapse

§92.7 Adding an inaccessible

Prototypical example for this section: V_κ

At this point you might be asking, well, where's my model of ZFC?

I unfortunately have to admit now: ZFC can never prove that there is a model of ZFC (unless ZFC is inconsistent, but that would be even worse). This is a result called Gödel's incompleteness theorem.

Nonetheless, with some very modest assumptions added, we can actually show that a model *does* exist: for example, assuming that there exists a strongly inaccessible cardinal κ would do the trick, V_κ will be such a model (Problem 92D*). Intuitively you can see why: κ is so big that any set of rank lower than it can't escape it even if we take their power sets, use the Replacement axiom, or any other method that ZFC lets us do.

More pessimistically, this shows that it's impossible to prove in ZFC that such a κ exists. Nonetheless, we now proceed under ZFC$^+$ for convenience, which adds the existence of such a κ as a final axiom. So we now have a model V_κ to play with. Joy!

Great. Now we do something *really* crazy.

Theorem 92.7.1 (Countable transitive model)

Assume ZFC$^+$. Then there exists a transitive model X of ZFC such that X is a *countable* set.

Proof. Fasten your seat belts.

First, since we assumed ZFC$^+$, we can take $V_\kappa = (V_\kappa, \in)$ as our model of ZFC. Start with the set $X_0 = \varnothing$. Then for every integer n, we do the following to get X_{n+1}.

- Start with X_{n+1} containing every element of X_n.
- Consider a formula $\phi(x, x_1, \dots, x_n)$ and $b_1, \dots, b_n$ in X_n. Suppose that V_κ thinks there is a $b \in V_\kappa$ for which
$$V_\kappa \vDash \phi[b, b_1, \dots, b_n].$$

We then add in the element b to X_{n+1}.

- We do this for *every possible formula in the language of set theory*. We also have to put in *every possible set of parameters* from the previous set X_n.

At every step X_n is countable. Reason: there are countably many possible finite sets of parameters in X_n, and countably many possible formulas, so in total we only ever add in countably many things at each step. This exhibits an infinite nested sequence of countable sets

$$X_0 \subseteq X_1 \subseteq X_2 \subseteq \dots$$

None of these is an elementary substructure of V_κ, because each X_n relies on witnesses in X_{n+1}. So we instead *take the union*:

$$X = \bigcup_n X_n.$$

This satisfies the Tarski-Vaught test, and is countable.

There is one minor caveat: X might not be transitive. We don't care, because we just take its Mostowski collapse. □

Please take a moment to admire how insane this is. It hinges irrevocably on the fact that there are countably many sentences we can write down.

Remark 92.7.2 — This proof relies heavily on the Axiom of Choice when we add in the element b to X_{n+1}. Without Choice, there is no way of making these decisions all at once.
Usually, the right way to formalize the Axiom of Choice usage is, for every formula $\phi(x, x_1, \dots, x_n)$, to pre-commit (at the very beginning) to a function $f_\phi(x_1, \dots, x_n)$, such that given any $b_1, \dots, b_n$, $f_\phi(b_1, \dots, b_n)$ will spit out the suitable value of b (if one exists). Personally, I think this is hiding the spirit of the proof, but it does make it clear how exactly Choice is being used.
These f_ϕ's have a name: **Skolem functions**.

The trick we used in the proof works in more general settings:

Theorem 92.7.3 (Downward Löwenheim-Skolem theorem)
Let $\mathscr{M} = (M, E)$ be a model, and $A \subseteq M$. Then there exists a set B (called the **Skolem hull** of A) with $A \subseteq B \subseteq M$, such that $(B, E) \prec \mathscr{M}$, and

$$|B| = \max\{\omega, |A|\}.$$

In our case, what we did was simply take A to be the empty set.

Question 92.7.4. Prove this. (Exactly the same proof as before.)

§92.8 FAQ's on countable models

The most common one is "how is this possible?", with runner-up "what just happened?".

Let me do my best to answer the first question. It seems like there are two things running up against each other:

(1) M is a transitive model of ZFC, but its universe is countable.

(2) ZFC tells us there are uncountable sets!

(This has confused so many people it has a name, **Skolem's paradox**.)

The reason this works I actually pointed out earlier: *countability is not absolute, it is a Σ_1 notion.*

Recall that a set x is countable if *there exists* an injective map $x \hookrightarrow \omega$. The first statement just says that *in the universe* V, there is an injective map $F \colon M \hookrightarrow \omega$. In particular, for any $x \in M$ (hence $x \subseteq M$, since M is transitive), x is countable *in* V. This is the content of the first statement.

But for M to be a model of ZFC, M only has to think statements in ZFC are true. More to the point, the fact that ZFC tells us there are uncountable sets means

$$M \vDash \exists x \text{ uncountable.}$$

In other words,

$$M \vDash \exists x \forall f \text{ If } f \colon x \to \omega \text{ then } f \text{ isn't injective.}$$

The key point is the $\forall f$ searches only functions in our tiny model M. It is true that in the "real world" V, there are injective functions $f \colon x \to \omega$.[1] But M has no idea they exist! It is a brain in a vat: M is oblivious to any information outside it.

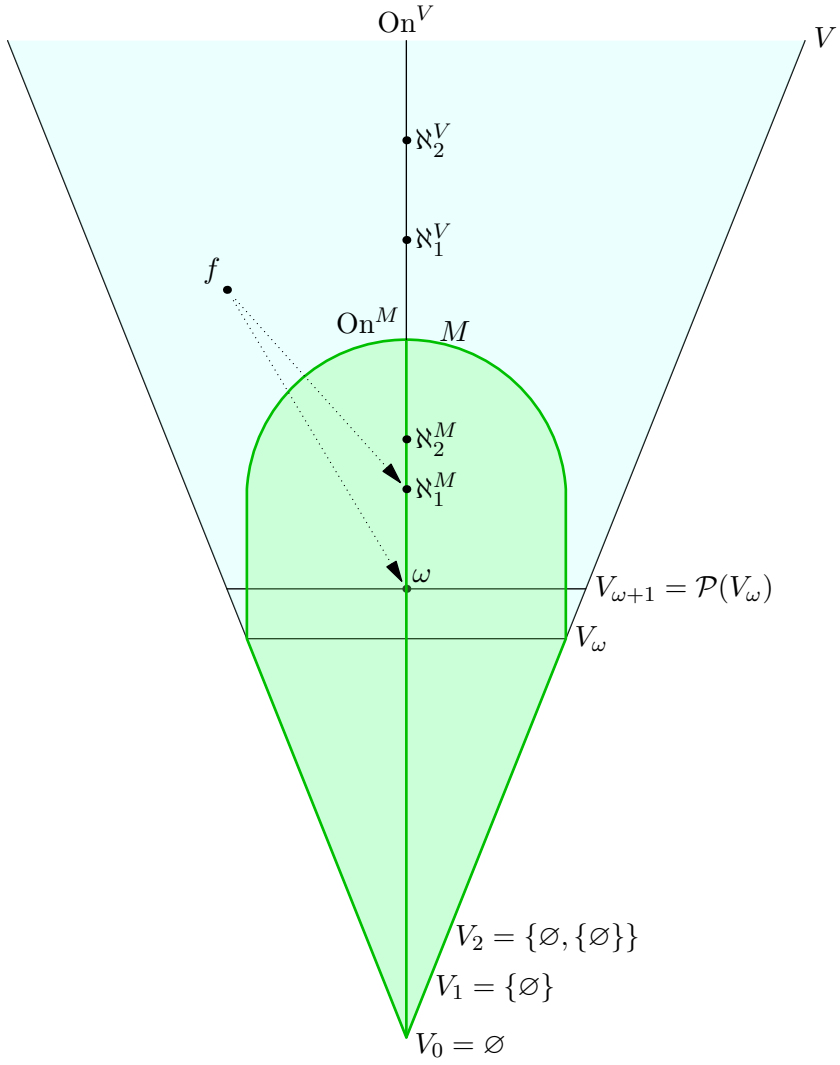
So in fact, every ordinal which appears in M is countable in the real world. It is just not countable in M. Since $M \vDash \text{ZFC}$, M is going to think there is some smallest uncountable cardinal, say $\aleph_1^M$. It will be the smallest (infinite) ordinal in M with the property that there is no bijection *in the model* M between $\aleph_1^M$ and ω. However, we necessarily know that such a bijection is going to exist in the real world V.

Put another way, cardinalities in M can look vastly different from those in the real world, because cardinality is measured by bijections, which I guess is inevitable, but leads to chaos.

§92.9 Picturing inner models

Here is a picture of a countable transitive model M.

[1] Since M is transitive and countable, for any $x \in M$, then x itself can have at most countably many elements. This isn't necessarily true if M is not transitive.



Note that M and V must agree on finite sets, since every finite set has a formula that can express it. However, past V_ω the model and the true universe start to diverge.

The entire model M is countable, so it only occupies a small portion of the universe, below the first uncountable cardinal $\aleph_1^V$ (where the superscript means "of the true universe V"). The ordinals in M are precisely the ordinals of V which happen to live inside the model, because the sentence "α is an ordinal" is absolute. On the other hand, M has only a portion of these ordinals, since it is only a lowly set, and a countable set at that. To denote the ordinals of M, we write On^M, where the superscript means "the ordinals as computed in M". Similarly, On^V will now denote the "set of true ordinals".

Nonetheless, the model M has its own version of the first uncountable cardinal $\aleph_1^M$. In the true universe, $\aleph_1^M$ is countable (below $\aleph_1^V$), but the necessary bijection witnessing this might not be inside M. That's why M can think $\aleph_1^M$ is uncountable, even if it is a countable cardinal in the original universe.

So our model M is a brain in a vat. It happens to believe all the axioms of ZFC, and so every statement that is true in M could conceivably be true in V as well. But M can't see the universe around it; it has no idea that what it believes is the uncountable $\aleph_1^M$ is really just an ordinary countable ordinal.

§92.10 A few harder problems to think about

Problem 92A*. Show that for any transitive model M, the set of ordinals in M is itself some ordinal.

Problem 92B†. Assume $\mathscr{M}_1 \subseteq \mathscr{M}_2$. Show that

(a) If ϕ is Δ_0, then $\mathscr{M}_1 \vDash \phi[b_1, \dots, b_n] \iff \mathscr{M}_2 \vDash \phi[b_1, \dots, b_n]$.

(b) If ϕ is Σ_1, then $\mathscr{M}_1 \vDash \phi[b_1, \dots, b_n] \implies \mathscr{M}_2 \vDash \phi[b_1, \dots, b_n]$.

(c) If ϕ is Π_1, then $\mathscr{M}_2 \vDash \phi[b_1, \dots, b_n] \implies \mathscr{M}_1 \vDash \phi[b_1, \dots, b_n]$.

(This should be easy if you've understood the chapter.)

Problem 92C† (Reflection)**.** Let κ be an inaccessible cardinal such that $|V_\alpha| < \kappa$ for all $\alpha < \kappa$. Prove that for any $\delta < \kappa$ there exists $\delta < \alpha < \kappa$ such that $V_\alpha \prec V_\kappa$; in other words, the set of α such that $V_\alpha \prec V_\kappa$ is *unbounded* in κ. This means that properties of V_κ reflect down to properties of V_α.

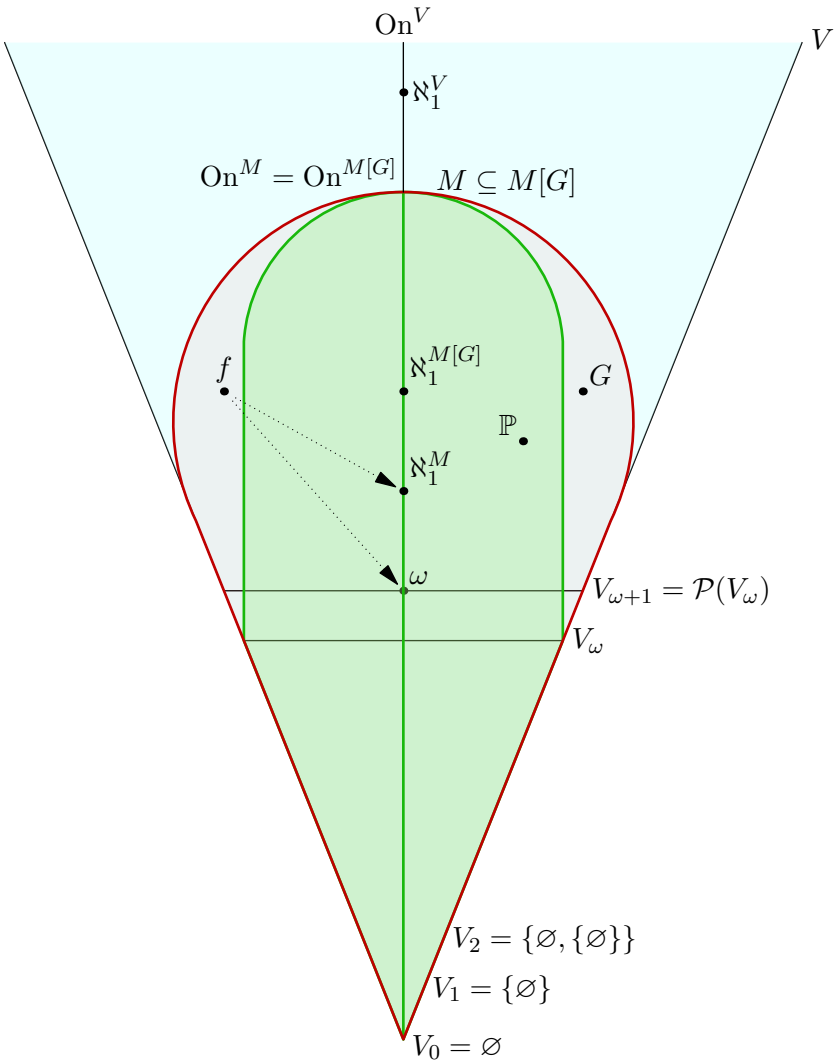
Problem 92D* (Strongly inaccessible cardinals produce models)**.** Let κ be a strongly inaccessible cardinal. Prove that V_κ is a model of ZFC.

93 Forcing

We are now going to introduce Paul Cohen's technique of **forcing**, which we then use to break the Continuum Hypothesis.

Here is how it works. Given a transitive model M and a poset $\mathbb{P}$ inside it, we can consider a "generic" subset $G \subseteq \mathbb{P}$, where G is not in M. Then, we are going to construct a bigger universe $M[G]$ which contains both M and G. (This notation is deliberately the same as $\mathbb{Z}[\sqrt{2}]$, for example – in the algebra case, we are taking $\mathbb{Z}$ and adding in a new element $\sqrt{2}$, plus everything that can be generated from it.) By choosing $\mathbb{P}$ well, we can cause $M[G]$ to have desirable properties.

Picture:



The model M is drawn in green, and its extension $M[G]$ is drawn in red.

The models M and $M[G]$ will share the same ordinals, which is represented here as M being no taller than $M[G]$. But one issue with this is that forcing may introduce some new bijections between cardinals of M that were not there originally; this leads to the phenomenon called **cardinal collapse**: quite literally, cardinals in M will no

longer be cardinals in $M[G]$, and instead just an ordinal. This is because in the process of adjoining G, we may accidentally pick up some bijections which were not in the earlier universe. In the diagram drawn, this is the function f mapping ω to $\aleph_1^M$. Essentially, the difficulty is that "κ is a cardinal" is a Π_1 statement.

In the case of the Continuum Hypothesis, we'll introduce a $\mathbb{P}$ such that any generic subset G will "encode" $\aleph_2^M$ real numbers. We'll then show cardinal collapse does not occur, meaning $\aleph_2^{M[G]} = \aleph_2^M$. Thus $M[G]$ will have $\aleph_2^{M[G]}$ real numbers, as desired.

§93.1 Setting up posets

Prototypical example for this section: Infinite Binary Tree

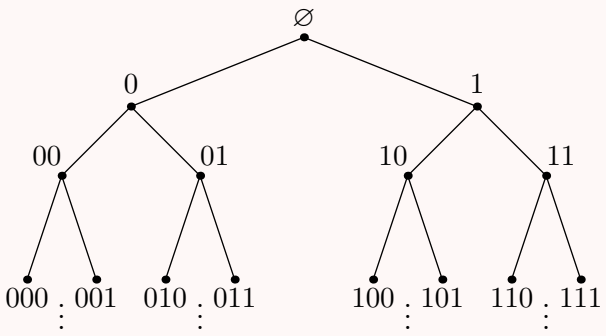
Let M be a transitive model of ZFC. Let $\mathbb{P} = (\mathbb{P}, \le) \in M$ be a poset with a maximum element $1_{\mathbb{P}}$ which lives inside a model M. The elements of $\mathbb{P}$ are called **conditions**; because they will force things to be true in $M[G]$.

Definition 93.1.1. A subset $D \subseteq \mathbb{P}$ is **dense** if for all $p \in \mathbb{P}$, there exists a $q \in D$ such that $q \le p$.

Examples of dense subsets include the entire $\mathbb{P}$ as well as any downwards "slice".

Definition 93.1.2. For $p, q \in \mathbb{P}$ we write $p \parallel q$, saying "p is **compatible** with q", if there exists $r \in \mathbb{P}$ with $r \le p$ and $r \le q$. Otherwise, we say p and q are **incompatible** and write $p \perp q$.

Example 93.1.3 (Infinite binary tree)
Let $\mathbb{P} = 2^{<\omega}$ be the **infinite binary tree** shown below, extended to infinity in the obvious way:



(a) The maximum element $1_{\mathbb{P}}$ is the empty string $\varnothing$.

(b) $D = \{\text{all strings ending in } 001\}$ is an example of a dense set.

(c) No two elements of $\mathbb{P}$ are compatible unless they are comparable.

Example 93.1.4 (Infinite chain)
Let $\mathbb{P} = (\mathbb{N}, \ge)$. This can be considered the "infinite unary tree".

- The maximum element $1_{\mathbb{P}}$ is 1.
- A set is dense if and only if it has infinitely many elements. For example, the

set of all positive even numbers and the set of all primes are dense.

Now, I can specify what it means to be "generic".

Definition 93.1.5. A nonempty set $G \subseteq \mathbb{P}$ is a **filter** if

(a) The set G is upwards-closed: $\forall p \in G(\forall q \geq p)(q \in G)$.

(b) Any pair of elements in G is compatible.

We say G is ***M*-generic** if for all D which are *in the model* M, if D is dense then $G \cap D \neq \varnothing$.

Question 93.1.6. Show that if G is a filter then $1_{\mathbb{P}} \in G$.

Note that the condition that $D \in M$ is important, because:

Question 93.1.7. On the infinite binary tree, show that:

- For every filter G, there's a dense subset D (not necessarily in M) such that $G \cap D = \varnothing$.
- Specifically, if $G \in M$, then such a set D can be chosen such that $D \in M$ — in particular, G is not M-generic.

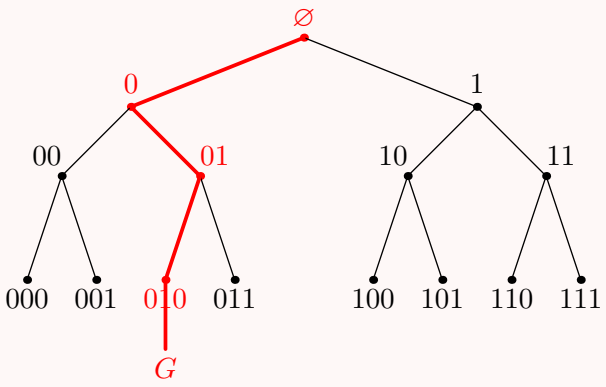
We will formalize this later in Lemma 93.2.5.

Example 93.1.8 (Generic filters on the infinite binary tree)

Let $\mathbb{P} = 2^{<\omega}$. The generic filters on $\mathbb{P}$ are sets of the form

$$\{0,\ b_1,\ b_1b_2,\ b_1b_2b_3,\ \dots\}.$$

So every generic filter on $\mathbb{P}$ corresponds to a binary number $b = 0.b_1b_2b_3\dots$.[a]
It is harder to describe which reals correspond to generic filters, but they should really "look random". For example, the set of strings ending in 011 is dense, so one should expect "011" to appear inside b, and more generally that b should contain every binary string. So one would expect the binary expansion of $\pi - 3$ might correspond to a generic, but not something like $0.010101\dots$. That's why we call them "generic".



[a] Note that it may be the case that two distinct filters correspond to the same real number, such as $0.1000\dots$ and $0.0111\dots$, but such filters are necessarily not generic.

Example 93.1.9 (Generic filters on the infinite chain)
There's only one generic filter on $\mathbb{P} = (\mathbb{N}, \geq)$: $\mathbb{N}$ itself.
This is indeed generic — it hits every dense set. This doesn't "look random" by any measure, you may say — but if you think about it, if you start at the root $1_{\mathbb{P}} = 1$ and move down randomly at each step, there's only one choice where to go!

Exercise 93.1.10. Verify that every generic filter $2^{<\omega}$ has the form above. Show that conversely, a binary number gives a filter, but it need not be generic.

Notice that if $p \geq q$, then the sentence $q \in G$ tells us more information than the sentence $p \in G$. In that sense q is a *stronger* condition. In another sense $1_{\mathbb{P}}$ is the weakest possible condition, because it tells us nothing about G; we always have $1_{\mathbb{P}} \in G$ since G is upwards closed.

§93.2 More properties of posets

We had better make sure that generic filters exist.

Example 93.2.1
If $\mathbb{P}$ is the infinite binary tree and M contains every subset of $\mathbb{P}$, then generic filter does not exist — for every filter $G \subseteq \mathbb{P}$, $\mathbb{P} \setminus G$ is a dense set $D \in M$, and $G \cap M = \varnothing$.

In fact this is kind of tricky, but for countable models it works:

Lemma 93.2.2 (Rasiowa-Sikorski lemma)
Suppose M is a *countable* transitive model of ZFC and $\mathbb{P}$ is a partial order. Then there exists an M-generic filter G.

Proof. Essentially, hit them one by one. Problem 93B. □

Fortunately, for breaking CH we would want M to be countable anyways.

The other thing we want to do to make sure we're on the right track is guarantee that a generic set G is not actually in M. (Analogy: $\mathbb{Z}[3]$ is a really stupid extension.) The condition that guarantees this is:

Definition 93.2.3. A partial order $\mathbb{P}$ is **splitting** if for all $p \in \mathbb{P}$, there exists $q, r \leq p$ such that $q \perp r$.

Example 93.2.4 (Infinite binary tree is (very) splitting)
The infinite binary tree is about as splitting as you can get. Given $p \in 2^{<\omega}$, just consider the two elements right under it.

Lemma 93.2.5 (Splitting posets omit generic sets)
Suppose $\mathbb{P}$ is splitting. Then if $F \subseteq \mathbb{P}$ is a filter such that $F \in M$, then $\mathbb{P} \setminus F$ is dense. In particular, if $G \subseteq \mathbb{P}$ is generic, then $G \notin M$.

Proof. Consider $p \notin \mathbb{P} \setminus F \iff p \in F$. Since $\mathbb{P}$ is splitting, there exist $q, r \le p$ which are not compatible. Since F is a filter it cannot contain both; we must have one of them outside F, say q. Hence every element of $p \in \mathbb{P} \setminus (\mathbb{P} \setminus F)$ has an element $q \le p$ in $\mathbb{P} \setminus F$. That's enough to prove $\mathbb{P} \setminus F$ is dense.

Question 93.2.6. Deduce the last assertion of the lemma about generic G. □

§93.3 Names, and the generic extension

We now define the *names* associated to a poset $\mathbb{P}$.

Definition 93.3.1. Suppose M is a transitive model of ZFC, $\mathbb{P} = (\mathbb{P}, \le) \in M$ is a partial order. We define the hierarchy of **$\mathbb{P}$-names** recursively by

$$\begin{aligned} \text{Name}_0 &= \varnothing \\ \text{Name}_{\alpha+1} &= \mathcal{P}(\text{Name}_\alpha \times \mathbb{P}) \\ \text{Name}_\lambda &= \bigcup_{\alpha<\lambda} \text{Name}_\alpha. \end{aligned}$$

Finally, $\text{Name} = \bigcup_\alpha \text{Name}_\alpha$ denote the class of all $\mathbb{P}$-names.

(These Name_α's are the analog of the V_α's: each Name_α is just the set of all names with rank $\le \alpha$.)

Definition 93.3.2. For a filter G, we define the **interpretation** of τ by G, denoted τ^G, using the transfinite recursion

$$\tau^G = \left\{ \sigma^G \mid \langle \sigma, p \rangle \in \tau \text{ and } p \in G \right\}.$$

We then define the model

$$M[G] = \left\{ \tau^G \mid \tau \in \text{Name}^M \right\}$$

where Name^M are the elements of the class Name that belongs to M. Thus $M[G]$ is a set. In words, $M[G]$ is the interpretation of all the possible $\mathbb{P}$-names (as computed by M).

You should think of a $\mathbb{P}$-name as a "fuzzy set". Here's the idea. Ordinary sets are collections of ordinary sets, so fuzzy sets should be collections of fuzzy sets. These fuzzy sets can be thought of like the Ghosts of Christmases yet to come: they represent things that might be, rather than things that are certain. In other words, they represent the possible futures of $M[G]$ for various choices of G.

Every fuzzy set has an element $p \in \mathbb{P}$ pinned to it. When it comes time to pass judgment, we pick a generic G and filter through the universe of $\mathbb{P}$-names. The fuzzy sets with an element of G attached to it materialize into the real world, while the fuzzy sets with elements outside of G fade from existence. The result is $M[G]$.

Example 93.3.3 (First few levels of the name hierarchy)
Let us compute

$$\begin{aligned} \text{Name}_0 &= \varnothing \\ \text{Name}_1 &= \mathcal{P}(\varnothing \times \mathbb{P}) \\ &= \{\varnothing\} \\ \text{Name}_2 &= \mathcal{P}(\{\varnothing\} \times \mathbb{P}) \\ &= \mathcal{P}\left(\{\langle \varnothing, p\rangle \mid p \in \mathbb{P}\}\right). \end{aligned}$$

Compare the corresponding von Neumann universe.

$$V_0 = \varnothing,\ V_1 = \{\varnothing\},\ V_2 = \{\varnothing, \{\varnothing\}\}.$$

Example 93.3.4 (Example of an interpretation)
As we said earlier, $\text{Name}_1 = \{\varnothing\}$. Now suppose

$$\tau = \{\langle \varnothing, p_1\rangle, \langle \varnothing, p_2\rangle, \dots, \langle \varnothing, p_n\rangle\} \in \text{Name}_2.$$

Then

$$\tau^G = \{\varnothing \mid \langle \varnothing, p\rangle \in \tau \text{ and } p \in G\} = \begin{cases} \{\varnothing\} & \text{if some } p_i \in G \\ \varnothing & \text{otherwise.} \end{cases}$$

In particular, since $1_{\mathbb{P}} \in G$, then:

- when $n = 0$, $\tau = \varnothing$, so $\tau^G = \varnothing$.
- when $n = 1$ and $p_1 = 1_{\mathbb{P}}$, $\tau = \{\langle \varnothing, 1_{\mathbb{P}}\rangle\}$, so $\tau^G = \{\varnothing\}$.

So,

$$\left\{\tau^G \mid \tau \in \text{Name}_2\right\} = V_2.$$

In fact, this holds for any natural number n, not just 2.

So, $M[G]$ and M agree on finite sets.

Now, we want to make sure $M[G]$ contains the elements of M. The proof of $\left\{\tau^G \mid \tau \in \text{Name}_2\right\} = V_2$ above can be easily adapted: Since $1_{\mathbb{P}}$ must be in G, we define for every $x \in M$ the set

$$\check{x} = \{\langle \check{y}, 1_{\mathbb{P}}\rangle \mid y \in x\}$$

by transfinite recursion. Basically, $\check{x}$ is just a copy of x where we tag every element *at every nesting level* with $1_{\mathbb{P}}$.

Example 93.3.5
Compute $\check{0} = 0$ and $\check{1} = \{\langle \check{0}, 1_{\mathbb{P}}\rangle\}$. Thus

$$(\check{0})^G = 0 \quad \text{and} \quad (\check{1})^G = 1.$$

Question 93.3.6. Show that in general, $(\check{x})^G = x$. (Rank induction.)

However, we'd also like to cause G to be in $M[G]$. In fact, we can write down the name exactly: we define

$$\dot{\mathbb{P}} := \{\langle \check{p}, p\rangle \mid p \in \mathbb{P}\}.$$

Question 93.3.7. Show that $(\dot{\mathbb{P}}) \in \text{Name}^M$, and $(\dot{\mathbb{P}})^G = G$.

Question 93.3.8. Verify that $M[G]$ is transitive: that is, if $\sigma^G \in \tau^G \in M[G]$, show that $\sigma^G \in M[G]$. (This is offensively easy.)

In summary,

$M[G]$ is a transitive model extending M (it contains G).

Moreover, it is reasonably well-behaved even if G is just a filter. Let's see what we can get off the bat.

Lemma 93.3.9 (Properties obtained from filters)
Let M be a transitive model of ZFC. If G is a filter, then $M[G]$ is transitive and satisfies Extensionality, Foundation, EmptySet, Infinity, Pairing, and Union.

This leaves PowerSet, Replacement, and Choice.

Proof. We get Extensionality and Foundation for free. Then Infinity and EmptySet follows from $M \subseteq M[G]$.

For Pairing, suppose $\sigma_1^G, \sigma_2^G \in M[G]$. Then

$$\sigma = \{\langle \sigma_1, 1_{\mathbb{P}}\rangle, \langle \sigma_2, 1_{\mathbb{P}}\rangle\}$$

satisfies $\sigma^G = \{\sigma_1^G, \sigma_2^G\}$. (Note that we used $M \vDash$ Pairing.) Union is left as a problem, which you are encouraged to try now. □

Up to here, we don't need to know anything about when a sentence is true in $M[G]$; all we had to do was contrive some names like $\check{x}$ or $\{\langle \sigma_1, 1_{\mathbb{P}}\rangle, \langle \sigma_2, 1_{\mathbb{P}}\rangle\}$ to get the facts we wanted. But for the remaining axioms, we *are* going to need this extra power. For this, we have to introduce the fundamental theorem of forcing.

§93.4 Fundamental theorem of forcing

The model M unfortunately has no idea what G might be, only that it is some generic filter.[1] Nonetheless, we are going to define a relation $\Vdash$, called the *forcing* relation.

[1] You might say this is a good thing; here's why. We're trying to show that ¬CH is consistent with ZFC, and we've started with a model M of the real universe V. But for all we know CH might be true in V (what if $V = L$?), in which case it would also be true of M.

Nonetheless, we boldly construct $M[G]$ an extension of the model M. In order for it to behave differently from M, it has to be out of reach of M. Conversely, if M could compute everything about $M[G]$, then $M[G]$ would have to conform to M's beliefs.

That's why we worked so hard to make sure $G \in M[G]$ but $G \notin M$.

Roughly, we are going to write

$$p \Vdash \varphi(\sigma_1, \dots, \sigma_n)$$

where $p \in \mathbb{P}$, $\sigma_1, \dots, \sigma_n \in M[G]$, if and only if:

> For *any* generic G, if $p \in G$, then $M[G] \vDash \varphi[\sigma_1^G, \dots, \sigma_n^G]$.

Note that $\Vdash$ is defined without reference to G: it is something that M can see. We say p **forces** the sentence $\varphi(\sigma_1, \dots, \sigma_n)$. And miraculously, we can define this relation in such a way that the converse is true: *a sentence holds if and only if some p forces it.*

Theorem 93.4.1 (Fundamental theorem of forcing)

Suppose M is a transitive model of ZF. Let $\mathbb{P} \in M$ be a poset, and $G \subseteq \mathbb{P}$ is an M-generic filter. Then,

(1) Consider $\sigma_1, \dots, \sigma_n \in \text{Name}^M$. Then

$$M[G] \vDash \varphi[\sigma_1^G, \dots, \sigma_n^G]$$

if and only if there exists a condition $p \in G$ such that p *forces* the sentence $\varphi(\sigma_1, \dots, \sigma_n)$. We denote this by $p \Vdash \varphi(\sigma_1, \dots, \sigma_n)$.

(2) This forcing relation is (uniformly) definable in M.

I'll tell you how the definition works in the next section.

§93.5 (Optional) Defining the relation

Here's how we're going to go. We'll define the most generous condition possible such that the forcing works in one direction ($p \Vdash \varphi(\sigma_1, \dots, \sigma_n)$ means $M[G] \vDash \varphi[\sigma_1^G, \dots, \sigma_n^G]$). We will then cross our fingers that the converse also works.

We proceed by induction on the formula complexity. It turns out in this case that the atomic formulas (base cases) are hardest and themselves require induction on ranks.

For some motivation, let's consider how we should define $p \Vdash \tau_1 \in \tau_2$ assuming that we've already defined $p \Vdash \tau_1 = \tau_2$. We need to ensure this holds iff

$$\forall M\text{-generic } G \text{ with } p \in G : \; M[G] \vDash \tau_1^G \in \tau_2^G.$$

So it suffices to ensure that any generic $G \ni p$ hits a condition q which forces τ_1^G to *equal* a member τ^G of τ_2^G. In other words, we want to choose the definition of $p \Vdash \tau_1 \in \tau_2$ to hold if and only if

$$\{q \in \mathbb{P} \mid \exists \langle \tau, r \rangle \in \tau_2 \, (q \le r \wedge q \Vdash (\tau = \tau_1))\}$$

is dense below in p. In other words, if the set is dense, then the generic must hit q, so it must hit r (recall that a filter is upwards-closed), meaning that $\langle \tau_r \rangle \in \tau_2$ will get interpreted such that $\tau^G \in \tau_2^G$, and moreover the $q \in G$ will force $\tau_1 = \tau$.

Now let's write down the definition... In what follows, the $\Vdash$ omits the M and $\mathbb{P}$.

Definition 93.5.1. Let M be a countable transitive model of ZFC. Let $\mathbb{P} \in M$ be a partial order. For $p \in \mathbb{P}$ and $\varphi(\sigma_1, \dots, \sigma_n)$ a formula in the language of set theory, we write $\tau \Vdash \varphi(\sigma_1, \dots, \sigma_n)$ to mean the following, defined by induction on formula complexity plus rank.

(1) $p \Vdash \tau_1 = \tau_2$ means

(i) For all $\langle \sigma_1, q_1 \rangle \in \tau_1$ the set

$$D_{\sigma_1,q_1} := \{r \mid r \leq q_1 \to \exists \langle \sigma_2, q_2 \rangle \in \tau_2 \, (r \leq q_2 \wedge r \Vdash (\sigma_1 = \sigma_2))\} .$$

is dense in p. (This encodes "$\tau_1 \subseteq \tau_2$".)

(ii) For all $\langle \sigma_2, q_2 \rangle \in \tau_2$, the set D_{σ_2,q_2} defined similarly is dense below p.

(2) $p \Vdash \tau_1 \in \tau_2$ means

$$\{q \in \mathbb{P} \mid \exists \langle \tau, r \rangle \in \tau_2 \, (q \leq r \wedge q \Vdash (\tau = \tau_1))\}$$

is dense below p.

(3) $p \Vdash \varphi \wedge \psi$ means $p \Vdash \varphi$ and $p \Vdash \psi$.

(4) $p \Vdash \neg\varphi$ means $\forall q \leq p$, $q \nVdash \varphi$.

(5) $p \Vdash \exists x \varphi(x, \sigma_1, \ldots, \sigma_n)$ means that the set

$$\{q \mid \exists \tau \, (q \Vdash \varphi(\tau, \sigma_1, \ldots, \sigma_n))\}$$

is dense below p.

This is definable in M! All we've referred to is $\mathbb{P}$ and names, which are in M. (Note that being dense is definable.) Actually, in parts (3) through (5) of the definition above, we use induction on formula complexity. But in the atomic cases (1) and (2) we are doing induction on the ranks of the names.

So, the construction above gives us one direction (I've omitted tons of details, but...).

Now, how do we get the converse: that a sentence is true if and only if something forces it? Well, by induction, we can actually show:

Lemma 93.5.2 (Consistency and Persistence)

We have

(1) (Consistency) If $p \Vdash \varphi$ and $q \leq p$ then $q \Vdash \varphi$.

(2) (Persistence) If $\{q \mid q \Vdash \varphi\}$ is dense below p then $p \Vdash \varphi$.

You can prove both of these by induction on formula complexity. From this we get:

Corollary 93.5.3 (Completeness)

The set $\{p \mid p \Vdash \varphi \text{ or } p \Vdash \neg\varphi\}$ is dense.

Proof. We claim that whenever $p \nVdash \varphi$ then for some $\overline{p} \leq p$ we have $\overline{p} \Vdash \neg\varphi$; this will establish the corollary.

By the contrapositive of the previous lemma, $\{q \mid q \Vdash \varphi\}$ is not dense below p, meaning for some $\overline{p} \leq p$, every $q \leq \overline{p}$ gives $q \nVdash \varphi$. By the definition of $p \Vdash \neg\varphi$, we have $\overline{p} \Vdash \neg\varphi$. □

And this gives the converse: the M-generic G has to hit some condition that passes judgment, one way or the other. This completes the proof of the fundamental theorem.

§93.6 The remaining axioms

Theorem 93.6.1 (The generic extension satisfies ZFC)
Suppose M is a transitive model of ZFC. Let $\mathbb{P} \in M$ be a poset, and $G \subseteq \mathbb{P}$ is an M-generic filter. Then
$$M[G] \vDash \text{ZFC}.$$

Proof. We'll just do Comprehension, as the other remaining axioms are similar.

Suppose $\sigma^G, \sigma_1^G, \dots, \sigma_n^G \in M[G]$ are a set and parameters, and $\varphi(x, x_1, \dots, x_n)$ is a formula in the language of set theory. We want to show that the set
$$A = \left\{ x \in \sigma^G \mid M[G] \vDash \varphi[x, \sigma_1^G, \dots, \sigma_n^G] \right\}$$
is in $M[G]$; i.e. it is the interpretation of some name.

Note that every element of σ^G is of the form ρ^G for some $\rho \in \operatorname{dom}(\sigma)$ (a bit of abuse of notation here, σ is a bunch of pairs of names and p's, and the domain $\operatorname{dom}(\sigma)$ is just the set of names). So by the fundamental theorem of forcing, we may write
$$A = \left\{ \rho^G \mid \rho \in \operatorname{dom}(\sigma) \text{ and } \exists p \in G \left(p \Vdash \rho \in \sigma \wedge \varphi(\rho, \sigma_1, \dots, \sigma_n) \right) \right\}.$$
To show $A \in M[G]$ we have to write down a τ such that the name τ^G coincides with A. We claim that
$$\tau = \{ \langle \rho, p \rangle \in \operatorname{dom}(\sigma) \times \mathbb{P} \mid p \Vdash \rho \in \sigma \wedge \varphi(\rho, \sigma_1, \dots, \sigma_n) \}$$
is the correct choice. It's actually clear that $\tau^G = A$ by construction; the "content" is showing that τ is in actually a name of M, which follows from $M \vDash$ Comprehension.

So really, the point of the fundamental theorem of forcing is just to let us write down this τ; it lets us show that τ is in Name^M without actually referencing G. □

§93.7 A few harder problems to think about

Problem 93A. For a filter G and M a transitive model of ZFC, show that $M[G] \vDash$ Union.

Problem 93B (Rasiowa-Sikorski lemma)**.** Show that in a countable transitive model M of ZFC, one can find an M-generic filter on any partial order.

94 Breaking the continuum hypothesis

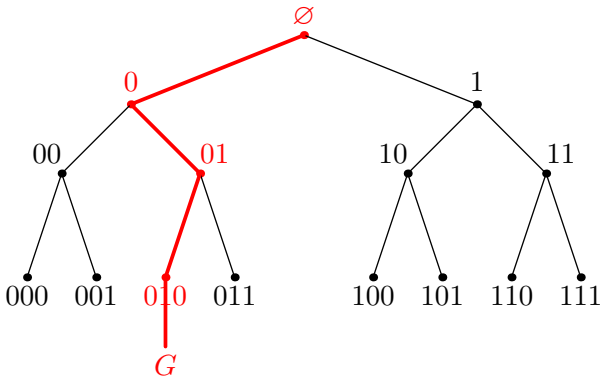
We now use the technique of forcing to break the Continuum Hypothesis by choosing a good poset $\mathbb{P}$. As I mentioned earlier, one can also build a model where the Continuum Hypothesis is true; this is called the *constructible universe*, (this model is often called "$V = L$"). However, I think it's more fun when things break...

§94.1 Adding in reals

Starting with a *countable* transitive model M.

We want to choose $\mathbb{P} \in M$ such that $(\aleph_2)^M$ many real numbers appear, and then worry about cardinal collapse later.

Recall the earlier situation where we set $\mathbb{P}$ to be the infinite complete binary tree; its nodes can be thought of as partial functions $n \to 2$ where $n < \omega$. Then G itself is a path down this tree; i.e. it can be encoded as a total function $G \colon \omega \to 2$, and corresponds to a real number.



We want to do something similar, but with ω_2 many real numbers instead of just one. In light of this, consider in M the poset

$$\mathbb{P} = \operatorname{Add}(\omega, \omega_2) \coloneqq (\{p \colon \omega_2 \times \omega \to 2, \operatorname{dom}(p) \text{ is finite}\}, \supseteq).$$

These elements p (conditions) are "partial functions": we take some finite subset of $\omega_2 \times \omega$ and map it into $2 = \{0, 1\}$. (Here $\operatorname{dom}(p)$ denotes the domain of p, which is the finite subset of $\omega_2 \times \omega$ mentioned.) Moreover, we say $p \le q$ if $\operatorname{dom}(p) \supseteq \operatorname{dom}(q)$ and the two functions agree over $\operatorname{dom}(q)$.

Question 94.1.1. What is the maximum element $1_{\mathbb{P}}$ here?

Exercise 94.1.2. Show that a generic G can be encoded as a function $\omega_2 \times \omega \to 2$.

Lemma 94.1.3 (G encodes distinct real numbers)

For $\alpha \in \omega_2$ define

$$G_\alpha = \{n \mid G(\alpha, n) = 0\} \in \mathcal{P}(\mathbb{N}).$$

Then $G_\alpha \neq G_\beta$ for any $\alpha \neq \beta$.

Proof. We claim that the set

$$D = \{q \mid \exists n \in \omega : q(\alpha, n) \neq q(\beta, n) \text{ are both defined}\}$$

is dense.

> **Question 94.1.4.** Check this. (Use the fact that the domains are all finite.)

Since G is an M-generic it hits this dense set D. Hence $G_\alpha \neq G_\beta$. □

Since $G \in M[G]$ and $M[G] \vDash \mathsf{ZFC}$, it follows that each G_α is in $M[G]$. So there are at least $\aleph_2^M$ real numbers in $M[G]$. We are done once we can show there is no cardinal collapse.

§94.2 The countable chain condition

It remains to show that with $\mathbb{P} = \text{Add}(\omega, \omega_2)$, we have that

$$\aleph_2^{M[G]} = \aleph_2^M.$$

In that case, since $M[G]$ will have $\aleph_2^M = \aleph_2^{M[G]}$ many reals, we will be done.

To do this, we'll rely on a combinatorial property of $\mathbb{P}$:

Definition 94.2.1. We say that $A \subseteq \mathbb{P}$ is a **strong antichain** if for any distinct p and q in A, we have $p \perp q$.

> **Example 94.2.2** (Example of an antichain)
> In the infinite binary tree, the set $A = \{00, 01, 10, 11\}$ is a strong antichain (in fact maximal by inclusion).

This is stronger than the notion of "antichain" than you might be used to![1] We don't merely require that every two elements are incomparable, but that they are in fact *incompatible*.

> **Question 94.2.3.** Draw a finite poset and an antichain of it which is not strong.

> **Question 94.2.4.** Convince yourself that all antichains in the infinite binary tree are strong, but some antichains in the poset $\mathbb{P}$ defined above are not strong.

Definition 94.2.5. A poset $\mathbb{P}$ has the **κ-chain condition** (where κ is a cardinal) if all strong antichains in $\mathbb{P}$ have size less than κ. The special case $\kappa = \aleph_1$ is called the **countable chain condition**, because it implies that every strong antichain is countable.

> **Remark 94.2.6** (Notational digression: Why $<$ instead of $\leq$?) — We could have defined that a poset $\mathbb{P}$ has the κ-chain condition if all strong antichains in $\mathbb{P}$ has size $\leq \kappa$. Nevertheless, this alternative definition is less versatile — for instance, there would be no way to express that all strong antichains in $\mathbb{P}$ are finite!

[1] In the context of forcing, some authors use "antichain" to refer to "strong antichain". I think this is lame.

We are going to show that if the poset has the κ-chain condition then it preserves all cardinals $\geq \kappa$. In particular, the countable chain condition will show that $\mathbb{P}$ preserves all the cardinals. Then, we'll show that $\operatorname{Add}(\omega, \omega_2)$ does indeed have this property. This will complete the proof.

We isolate a useful lemma:

Lemma 94.2.7 (Possible values argument)

Suppose M is a transitive model of ZFC and $\mathbb{P}$ is a partial order such that $\mathbb{P}$ has the κ-chain condition in M. Let $X, Y \in M$ and let $f\colon X \to Y$ be some function in $M[G]$, but $f \notin M$.

Then there exists a function $F \in M$, with $F\colon X \to \mathcal{P}(Y)$ and such that for any $x \in X$,

$$f(x) \in F(x) \quad \text{and} \quad |F(x)|^M < \kappa.$$

What this is saying is that if f is some new function that's generated, M is still able to pin down the values of f to less than κ many values.

Proof. The idea behind the proof is easy: any possible value of f gives us some condition in the poset $\mathbb{P}$ which forces it. Since distinct values must have incompatible conditions, the κ-chain condition guarantees there are at most κ such values.

Here are the details. Let $\dot{f}$, $\check{X}$, $\check{Y}$ be names for f, X, Y. Start with a condition p such that p forces the sentence

$$\text{“}\dot{f} \text{ is a function from } \check{X} \text{ to } \check{Y}\text{”}.$$

We'll work just below here.

For each $x \in X$, we can consider (using the Axiom of Choice) a maximal strong antichain $A(x)$ of incompatible conditions $q \leq p$ which forces $f(x)$ to equal some value $y \in Y$. Then, we let $F(x)$ collect all the resulting y-values. These are all possible values, and there are less than κ of them. □

§94.3 Preserving cardinals

As we saw earlier, cardinal collapse can still occur. For the Continuum Hypothesis we want to avoid this possibility, so we can add in $\aleph_2^M$ many real numbers and have $\aleph_2^{M[G]} = \aleph_2^M$. It turns out that to verify this, one can check a weaker result.

Definition 94.3.1. For M a transitive model of ZFC and $\mathbb{P} \in M$ a poset, we say $\mathbb{P}$ **preserves cardinals** if $\forall G \subseteq \mathbb{P}$ an M-generic, the model M and $M[G]$ agree on the sentence “κ is a cardinal” for every κ. Similarly we say $\mathbb{P}$ **preserves regular cardinals** if M and $M[G]$ agree on the sentence “κ is a regular cardinal” for every κ.

Intuition: In a model M, it's possible that two cardinals which are in bijection in V are no longer in bijection in M. Similarly, it might be the case that some cardinal $\kappa \in M$ is regular, but stops being regular in V because some function $f\colon \overline{\kappa} \to \kappa$ is cofinal but happened to only exist in V. In still other words, “κ is a regular cardinal” turns out to be a Π_1 statement too.

Fortunately, each implies the other. We quote the following without proof.

Proposition 94.3.2 (Preserving cardinals $\iff$ preserving regular cardinals)
Let M be a transitive model of ZFC. Let $\mathbb{P} \in M$ be a poset. Then for any λ, $\mathbb{P}$ preserves cardinalities less than or equal to λ if and only if $\mathbb{P}$ preserves regular cardinals less than or equal to λ. Moreover the same holds if we replace "less than or equal to" by "greater than or equal to".

Thus, to show that $\mathbb{P}$ preserves cardinality and cofinalities it suffices to show that $\mathbb{P}$ preserves regularity. The following theorem lets us do this:

Theorem 94.3.3 (Chain conditions preserve regular cardinals)
Let M be a transitive model of ZFC, and let $\mathbb{P} \in M$ be a poset. Suppose M satisfies the sentence "$\mathbb{P}$ has the κ-chain condition and κ is regular". Then $\mathbb{P}$ preserves regularity greater than or equal to κ.

Proof. Use the Possible Values Argument. Problem 94A. □

In particular, if $\mathbb{P}$ has the countable chain condition then $\mathbb{P}$ preserves *all* the cardinals (and cofinalities). Therefore, it remains to show that $\operatorname{Add}(\omega, \omega_2)$ satisfies the countable chain condition.

§94.4 Infinite combinatorics

We now prove that $\operatorname{Add}(\omega, \omega_2)$ satisfies the countable chain condition. This is purely combinatorial, and so we work briefly.

Definition 94.4.1. Suppose C is an uncountable collection of finite sets. C is a **Δ-system** if there exists a **root** R with the condition that for any distinct X and Y in C, we have $X \cap Y = R$.

Lemma 94.4.2 (Δ-System lemma)
Suppose C is an uncountable collection of finite sets. Then $\exists \overline{C} \subseteq C$ such that $\overline{C}$ is an uncountable Δ-system.

Proof. There exists an integer n such that C has uncountably many guys of length n. So we can throw away all the other sets, and just assume that all sets in C have size n.

We now proceed by induction on n. The base case $n = 1$ is trivial, since we can just take $R = \varnothing$. For the inductive step we consider two cases.

First, assume there exists an $a \in C$ contained in uncountably many $F \in C$. Throw away all the other guys. Then we can just delete a, and apply the inductive hypothesis.

Now assume that for every a, only countably many members of C have a in them. We claim we can even get a $\overline{C}$ with $R = \varnothing$. First, pick $F_0 \in C$. It's straightforward to construct an F_1 such that $F_1 \cap F_0 = \varnothing$. And we can just construct $F_2, F_3, \ldots$ □

Lemma 94.4.3
For all κ, $\operatorname{Add}(\omega, \kappa)$ satisfies the countable chain condition.

Proof. Assume not. Let

$$\{p_\alpha \mid \alpha < \omega_1\}$$

be a strong antichain. Let

$$C = \{\operatorname{dom}(p_\alpha) \mid \alpha < \omega_1\}.$$

Let $\overline{C} \subseteq C$ be such that $\overline{C}$ is uncountable, and $\overline{C}$ is a Δ-system with root R. Then let

$$B = \{p_\alpha \mid \operatorname{dom}(p_\alpha) \in R\}.$$

Each $p_\alpha \in B$ is a function $p_\alpha \colon R \to \{0,1\}$, so there are two that are the same. □

Thus, we have proven that the Continuum Hypothesis cannot be proven in ZFC.

§94.5 A few harder problems to think about

Problem 94A. Let M be a transitive model of ZFC, and let $\mathbb{P} \in M$ be a poset. Suppose M satisfies the sentence "$\mathbb{P}$ has the κ-chain condition and κ is regular". Show that $\mathbb{P}$ preserves regularity greater than or equal to κ.

XXIII

Appendix

Part XXIII: Contents

A Pedagogical comments and references

Here are some higher-level comments on the way specific topics were presented, as well as pointers to further reading.

§A.1 Basic algebra and topology

§A.1.i Linear algebra and multivariable calculus

Following the comments in Section 9.9, I dislike most presentations of linear algebra and multivariable calculus since they miss the two key ideas, namely:

- In linear algebra, we study *linear maps* between spaces.
- In calculus, we *approximate functions at points by linear functions*.

Thus, I believe linear algebra should *always* be taught before multivariable calculus. In particular, I do not recommend most linear algebra or multivariable calculus books.

For linear algebra, I've heard that [Ax97] follows this approach, hence the appropriate name "Linear Algebra Done Right". I followed with heavy modifications the proceedings of Math 55a, see [Ga14].

For multivariable calculus and differential geometry, I found the notes [Sj05] to be unusually well-written. I referred to it frequently while I was enrolled in Math 55b [Ga15].

§A.1.ii General topology

My personal view on spaces is that every space I ever work with is either metrizable or is the Zariski topology.

I adopted the approach of [Pu02], using metric topology first. I find that metric spaces are far more intuitive, and are a much better way to get a picture of what open / closed / compact etc. sets look like. This is the approach history took; general topology grew out of metric topology.

I personally dislike starting any general topology class by defining what a general topological space is, because it doesn't communicate a good picture of open and closed sets to draw pictures of.

§A.1.iii Groups and commutative algebra

I teach groups before commutative rings but might convert later. Rings have better examples, don't have the confusion of multiplicative notation for additive groups, and modding out by ideals is more intuitive.

There's a specific thing I have a qualm with in group theory: the way that the concept of a normal subgroup is introduced. Only [Go11] does something similar to what I do. Most other people simply *define* a normal subgroup N as one with gNg^{-1} and then proceed to define modding out, without taking the time to explain where this definition comes from. I remember distinctly this concept as the first time in learning math where I didn't understand what was going on. Only in hindsight do I see where this definition came from; I tried hard to make sure my own presentation didn't have this issue.

I deliberately don't include a chapter on just commutative algebra; other than the chapter on rings and ideals. The reason is that I always found it easier to learn commutative algebra theorems on the fly, in the context of something like algebraic number theory or algebraic geometry. For example, I finally understand why radicals and the Nullstellensatz were important when I saw how they were used in algebraic geometry. Before then, I never understood why I cared about them.

§A.1.iv Calculus

I do real analysis by using metric and general topology as the main ingredient, since I think it's the most useful later on and the most enlightening. In some senses, I am still following [Pu02].

§A.2 Second-year topics

§A.2.i Measure theory and probability

The main inspiration for these lectures is Vadim Gorin's 18.175 at MIT; [Go18] has really nice lecture notes taken by Tony Zhang. I go into a bit more details of the measure theory, and (for now) less into the probability. But I think probability is a great way to motivate measure theory anyways, and conversely, it's the right setting in to which state things like the central limit theorem.

I also found [Ch08] quite helpful, as another possible reference.

§A.2.ii Complex analysis

I picked the approach of presenting the Cauchy-Goursat theorem as given (rather than proving a weaker version by Stokes' theorem, or whatever), and then deriving the key result that holomorphic functions are analytic from it. I think this most closely mirrors the "real-life" use of complex analysis, i.e. the computation of contour integrals.

The main reference for this chapter was [Ya12], which I recommend.

§A.2.iii Category theory

I enthusiastically recommend [Le14], from which my chapters are based, and which contains much more than I had time to cover.

You might try reading chapters 2-4 in reverse order though: I found that limits were much more intuitive than adjoints. But your mileage may vary.

The category theory will make more sense as you learn more examples of structures: it will help to have read, say, the chapters on groups, rings, and modules.

§A.2.iv Quantum algorithms

The exposition given here is based off a full semester at MIT taught by Seth Lloyd, in 18.435J [Ll15]. It is written in a far more mathematical perspective.

I only deal with finite-dimensional Hilbert spaces, because that is all that is needed for Shor's algorithm, which is the point of this chapter. This is not an exposition intended for someone who wishes to seriously study quantum mechanics (though it might be a reasonable first read): the main purpose is to give students a little appreciation for what this "Shor's algorithm" that everyone keeps talking about is.

§A.2.v Representation theory

I staunchly support teaching the representation of algebras first, and then specializing to the case of groups by looking at $k[G]$. The primary influence for the chapters here is [Et11], and you might think of what I have here as just some selections from the first four chapters of this source.

§A.2.vi Set theory

Set theory is far off the beaten path. The notes I have written are based off the class I took at Harvard College, Math 145a [Ko14].

My general impression is that the way I present set theory (trying to remain intuitive and informal in a logical minefield) is not standard. Possible other reference: [Mi14].

§A.3 Advanced topics

§A.3.i Algebraic topology

I cover the fundamental group π_1 first, because I think the subject is more intuitive this way. A possible reference in this topic is [Mu00]. Only later do I do the much more involved homology groups. The famous standard reference for algebraic topology is [Ha02], which is what almost everyone uses these days. But I also found [Ma13a] to be very helpful, particularly in the part about cohomology rings.

I don't actually do very much algebraic topology. In particular, I think the main reason to learn algebraic topology is to see the construction of the homology and cohomology groups from the chain complex, and watch the long exact sequence in action. The concept of a (co)chain complex comes up often in other contexts as well, like the cohomology of sheaves or Galois cohomology. Algebraic topology is by far the most natural one.

I use category theory extensively, being a category-lover.

§A.3.ii Algebraic number theory

I learned from [Og10], using [Le02] for the part about the Chebotarev density theorem.

When possible I try to keep the algebraic number theory chapter close at heart to an "olympiad spirit". Factoring in rings like $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{-5}]$ is very much an olympiad-flavored topic at heart: one is led naturally to the idea of factoring in general rings of integers, around which the presentation is built. As a reward for the entire buildup, the exposition finishes with the application of the Chebotarev density theorem to IMO 2003, Problem 6.

§A.3.iii Algebraic geometry

My preferred introduction to algebraic geometry is [Ga03] for a first read and [Va17] for the serious version. Both sets of lecture notes are essentially self-contained.

I would like to confess now that I know relatively little algebraic geometry, and in my personal opinion the parts on algebraic geometry are the weakest part of the Napkin. This is reflected in my work here: in the entire set of notes I only barely finish defining a scheme, the first central definition of the subject.

Nonetheless, I will foolishly still make some remarks about my own studies. I think there are three main approaches to beginning the study of schemes:

- Only looking at affine and projective varieties, as part of an "introductory" class, typically an undergraduate course.
- Studying affine and projective varieties closely and using them as the motivating example of a *scheme*, and then developing algebraic geometry from there.
- Jumping straight into the definition of a scheme, as in the well-respected and challenging [Va17].

I have gone with the second approach, I think that if you don't know what a scheme is, then you haven't learned algebraic geometry. But on the other hand I think the definition of a scheme is difficult to digest without having a good handle first on varieties.

These opinions are based on my personal experience of having tried to learn the subject through all three approaches over a period of a year. Your mileage may vary.

I made the decision to, at least for the second part, focus mostly on *affine* schemes. These already generalize varieties in several ways, and I think the jump is too much if one starts then gluing schemes together. I would rather that the student first feel like they really understand how an affine scheme works, before going on into the world where they now have a general scheme X which is locally affine (but probably not itself affine). The entire chapter dedicated to a gazillion examples of affine schemes is a hint of this.

§A.3.iv Riemann surfaces

My friend recommends [Mi95]. The preface of the book reads as follows:

> But I try to stress that the main examples (from the point of view of algebraic geometry) come from projective curves, and slowly but surely the text evolves to the algebraic category, culminating in an algebraic proof of the Riemann-Roch theorem. After returning to the analytic side of things for Abel's theorem, the progression is repeated again when sheaves and cohomology are discussed: first the analytic, then the algebraic category.

Thus you can also use this as a resource to learn algebraic geometry.

Occasionally, a few concepts are not very well-motivated, such as divisors, complex structure induced on plane curves, or line bundles. In these cases, we try to explain the motivation clearly in the Napkin.

§A.4 Topics not in Napkin

§A.4.i Analytic number theory

I never had time to write up notes in Napkin for these. If you're interested though, I recommend [Hi13]. They are highly accessible and delightful to read. The only real prerequisites are a good handle on Cauchy's residue formula.

B Hints to selected problems

1A. Orders.

1B. Copy the proof of Fermat's little theorem, using Lemma 1.2.5.

1C. For the former, decide where the isomorphism should send r and s, and the rest will follow through. For the latter, look at orders.

1D*. Generated groups.

1F†. Use $n = |G|$.

1G. For the lower bound, consider orders (note that 1009 is prime). For the upper bound, consider a 1009-gon.

1H. Draw inspiration from D_6.

1I. Look at the group of 2×2 matrices mod p with determinant ± 1.

2B. No. There is not even a continuous injective map from $\mathbb{Q}$ to $\mathbb{N}$.

2C. You can do this with bare hands. You can also use composition.

2D. $\pm x$ for good choices of $\pm$.

2E. Project gaps onto the y-axis. Use the fact that uncountably many positive reals cannot have finite sum.

2F. First answer the following question: "is $1/x$ a function?".

3A. Write it out: $\phi(ab) = \phi(a)\phi(b)$.

3B. Yes, no.

3C. No.

3D. $\gcd(1000, 999) = 1$.

3F. Find an example of order 8.

3G. Try to show G is the dihedral group of order 18. There is not much group theory content here — just manipulation.

3H. Get yourself a list of English homophones, I guess. Don't try too hard. Letter v is the worst; maybe $felt = veldt$?

4A. $R = \mathbb{R}[i]$.

4B. Show that the map

$$\begin{aligned} \mathbb{C}[x] &\to \mathbb{C} \times \mathbb{C} \\ p &\mapsto (p(0), p(1)) \end{aligned}$$

is surjective and calculate its kernel.

4E. For (b) homomorphism is uniquely determined by the choice of $\psi(x) \in R$

5A. Yes.

5B. The kernel is an ideal of K!

5C*. This is just a definition chase.

5D*. Fermat's little theorem type argument; cancellation holds in integral domains.

5E*. Just keep on adding in elements to get an ascending chain.

5F. Use the fact that both are PID's.

5G†. Show that the quotient $\mathbb{Z}[\sqrt{2017}]/I$ has finitely many elements for any nonzero prime ideal I. Therefore, the quotient is an integral domain, it is also a field, and thus I was a maximal ideal.

6A†. The main task is to show there exists some fixed point. Start at some point x_0 and consider the sequence $x_1 = T(x_0)$, $x_2 = T(x_1)$, $x_3 = T(x_2)$, ..., and so on.

6B. (a): M is complete and bounded but not totally bounded. N is all no. For (b) show that $M \cong \mathbb{R} \cong N$.

6C†. As a set, we let $\overline{M}$ be the set of Cauchy sequences (x_n) in M, modulo the relation that $(x_n) \sim (y_n)$ if $\lim_n d(x_n, y_n) = 0$.

6E. The standard solution seems to be via the so-called "Baire category theorem".

7D. Let p be any point. If there is a real number r such that $d(p,q) \neq r$ for any $q \in M$, then the r-neighborhood of p is clopen.

7E. (a) is yes, and (b) is no even for metric spaces. In fact, a *totally disconnected* space is one for which every connected component consists of only a single point, and there are examples of totally disconnected metric spaces with non-discrete topologies.

7F. Note that $p\mathbb{Z}$ is closed for each p. If there were finitely many primes, then $\bigcup p\mathbb{Z} = \mathbb{Z} \setminus \{-1, 1\}$ would have to be closed; i.e. $\{-1, 1\}$ would be open, but all open sets here are infinite.

7G. The balls at 0 should be of the form $n! \cdot \mathbb{Z}$.

7H. Appeal to $\mathbb{Q}$.

8A. $[0, 1]$ is compact.

8B. If and only if it is finite.

8E. Suppose $p_i = (x_i, y_i)$ is a sequence in $X \times Y$ $(i = 1, 2, \dots)$. Take a sub-sequence such that the x-coordinate converges (throwing out some terms). Then take a sub-sequence of *that* sub-sequence such that y-coordinate converges (throwing out more terms).

8F†. Mimic the proof of Theorem 8.2.2. The totally bounded condition lets you do Pigeonhole.

8H. Assuming M is not compact, construct an unbounded continuous function $F\colon M \to \mathbb{R}$. Once such a function F is defined, the metric

$$d'(x,y) \coloneqq d(x,y) + |F(x) - F(y)|$$

will establish the contrapositive of the problem.

8I. The answer to both parts is no.

For (a) use Problem 8D.

For (b), color each circle in the partition based on whether it contains p but not q, q but not p, or both.

9A†. Use the rank-nullity theorem. Also consider the zero map.

9D. $a + b\sqrt{5} \mapsto \sqrt{5}a + 5b$.

9F. Plug in $y = -1, 0, 1$. Use dimensions of $\mathbb{R}[x]$.

9G. Interpret as $V \oplus V \to W$ for suitable V, W.

9I*. Use the fact that the infinite chain of subspaces

$$\ker T \subseteq \ker T^2 \subseteq \ker T^3 \subseteq \dots$$

and the similar chain for $\operatorname{im} T$ must eventually stabilize (for dimension reasons).

10D. The answer is yes. In fact, the result is true if $\mathbb{C}^{\oplus 2}$ is any finite-dimensional $\mathbb{C}$-vector space.

10F. Only 0 is. Look at degree.

10G. All of them are!

11A. Follows by writing T in an eigenbasis: then the diagonal entries are the eigenvalues.

11B†. Again one can just take a basis.

11C†. One solution is to just take a basis. Otherwise, interpret $T \otimes S \mapsto \operatorname{Tr}(T \circ S)$ as a linear map $(V^\vee \otimes W) \otimes (W^\vee \otimes V) \to k$, and verify that it is commutative.

11D. Look at the trace of T.

12A. The point is that

$$(v_1 + cv_2) \wedge v_2 \cdots \wedge v_n = v_1 \wedge v_2 \cdots \wedge v_n + c(v_2 \wedge v_2 \cdots \wedge v_n)$$

and the latter term is zero.

12B. You can either do this by writing T in matrix form, or you can use the wedge definition of $\det T$ with the basis given by Jordan form.

12C. This is actually immediate by taking any basis in which X is upper-triangular!

12D. You don't need eigenvalues (though they could work also). In one direction, recall that (by Problem 9B†) we can replace "isomorphism" by "injective". In the other, if T is an isomorphism, let S be the inverse map and look at $\det(S \circ T)$.

12E. Consider 1000×1000 matrix M with entries 0 on diagonal and ± 1 off-diagonal. Mod 2.

12F. There is a family of solutions other than just $a = b = c = d$.

One can solve the problem using Cayley-Hamilton. A more "bare-hands" approach is to show the matrix is invertible (unless $a = b = c = d$) and then diagonalize the matrix as $M = \begin{bmatrix} s & -q \\ -r & p \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} p & q \\ r & s \end{bmatrix} = \begin{bmatrix} ps\lambda_1 - qr\lambda_2 & qs(\lambda_1 - \lambda_2) \\ pr(\lambda_2 - \lambda_1) & ps\lambda_2 - qr\lambda_1 \end{bmatrix}$.

12G. Take bases, and do a fairly long calculation.

13B*. Fix an orthonormal basis $e_1, \dots, e_n$. Use the fact that $\mathbb{R}^n$ is complete.

13C. Dot products in $\mathbb{F}_2$.

13D*. Define it on simple tensors then extend linearly.

13E. $k = n^n$. Endow tensor products with an inner form. Note that "zero entry somewhere on its diagonal" is equivalent to the product of those entries being zero.

14A. Use Parseval again, but this time on $f(x) = x^2$.

14B. Define the Boolean function $D \colon \{\pm 1\}^3 \to \mathbb{R}$ by $D(a,b,c) = ab + bc + ca$. Write out the value of $D(a,b,c)$ for each (a,b,c). Then, evaluate its expected value.

15A*. You can *prove* the result just by taking a basis $e_1, \dots, e_n$ of V and showing that it is a linear map sending e_1 to the basis $(e_1^\vee)^\vee$.

15B. Use Theorem 9.7.6 and it will be immediate (the four quantities equal the k in the theorem).

15C†. This actually is just the previous problem in disguise! The row rank is $\dim \operatorname{im} T^\vee$ and the column rank is $\dim \operatorname{im} T$.

15F. If there is a polynomial, check $TT^\dagger = T^\dagger T$ directly. If T is normal, diagonalize it.

16A. Just apply Burnside's lemma directly to get the answer of 198 (the relevant group is D_{14}).

16B. There are multiple ways to see this. One is to just do the algebraic manipulation. Another is to use Cayley's theorem to embed G into a symmetric group.

16C. Double-count pairs (g, x) with $g \cdot x = x$.

16E†. Let H act on the left cosets $\{gH \mid g \in G\}$ by left multiplication: $h \cdot gH = hgH$. Consider any orbit $\mathcal{O}$ By the orbit-stabilizer theorem, $|\mathcal{O}|$ divides $|H|$ which divides n, so $\mathcal{O}$ divides n. But $|\mathcal{O}| \le p$ since there are p cosets. So either $\mathcal{O} = \{H\}$ or $\mathcal{O}$ contains all cosets. Show that the latter is impossible and conclude.

17B. Count Sylow 2 and 7 groups and let them intersect.

17C. Construct a non-abelian group such that all elements have order three.

17D. First, if G abelian it's trivial. Otherwise, let $Z(G)$ be the center of the group, which is always a normal subgroup of G. Do a mod p argument via conjugation (or use the class equation).

18A†. In the structure theorem, $k/(s_i) \in \{0, k\}$.

18B†. By theorem $V \cong \bigoplus_i k[x]/(s_i)$ for some polynomials s_i. Write each block in the form described.

18C†. Copy the previous proof, except using the other form of the structure theorem. Since $k[x]$ is algebraically closed each p_i is a linear factor.

18D. The structure theorem is an anti-result here: it more or less implies that finitely generated abelian groups won't work. So, look for an infinitely generated example.

18E. I think the result is true if you add the assumption A is Noetherian, so look for trouble by picking A not Noetherian.

19B†. For any $a \in A$, the map $v \mapsto a \cdot v$ is intertwining.

19C⋆. For part (b), pick a basis and do $T \mapsto (T(e_1), \dots, T(e_n))$.

19D⋆. Right multiplication.

19E. Apply Problem 9I⋆.

20A. They are all one-dimensional, n of them. What are the homomorphisms $\mathbb{Z}/n\mathbb{Z} \to \mathbb{C}^\times$?

20B. The span of $(1, 0)$ is a subrepresentation.

20C. This is actually easy.

20D. There are only two one-dimensional ones (corresponding to the only two homomorphisms $D_{10} \to \mathbb{C}^\times$). So the remaining ones are two-dimensional.

20E. Let $r, t \in D_{10}$ be rotation and reflection respectively. Then we can sum over all possible bug's moves with

$$\frac{1}{10} \operatorname{Tr}(\rho(r) + \rho(t))^{15}.$$

Then use Problem 20D to compute this trace.

21A†. Obvious. Let $W = \bigoplus V_i^{m_i}$ (possible since $\mathbb{C}[G]$ semisimple) thus $\chi_W = \sum_i m_i \chi_{V_i}$.

21B. Use the previous problem, with $\chi_W = \chi_{\text{refl}_0}^2$.

21C. Characters. Note that $|\chi_W| = 1$ everywhere.

21D. There are five conjugacy classes, 1, -1 and $\pm i$, $\pm j$, $\pm k$. Given four of the representations, orthogonality can give you the fifth one.

21E⋆. Construct two square $r \times r$ matrices A and B such that AB is the identity by the first orthogonality. Then use BA to prove the second orthogonaliy relation.

23A. Rewrite $|\Psi_-\rangle = -\frac{1}{\sqrt{2}} (|{\to}\rangle_A \otimes |{\leftarrow}\rangle_B - |{\leftarrow}\rangle_A |{\to}\rangle_B)$.

23B. 1, 1, 1, -1 respectively. When we multiply them all together, we get that $\text{id}^A \otimes \text{id}^B \otimes \text{id}^C$ has measurement -1, which is the paradox. What this means is that the values of the measurements can't be prepared in advance independently. In other words, this contradicts certain local hidden-variable theories.

This was one of several results for which Zeilinger won a (shared) Nobel Prize in 2022.

24A. One way is to create CCNOT using a few Fredkin gates.

24B. Plug in $|\psi\rangle = |0\rangle$, $|\psi\rangle = |1\rangle$, $|\psi\rangle = |\rightarrow\rangle$ and derive a contradiction.

24C. First show that the box sends $|x_1\rangle \otimes \cdots \otimes |x_m\rangle \otimes |\leftarrow\rangle$ to $(-1)^{f(x_1,\ldots,x_m)}(|x_1\rangle \otimes \cdots \otimes |x_m\rangle \otimes |\leftarrow\rangle)$.

24D†. This is direct computation.

26B. Iff the sequence is convergent!

26D. The nth partial sum is $\frac{1}{1-r}(1 - r^{n+1})$.

26F. This is a very tricky algebraic manipulation. Try setting $a_n = x_1 + \cdots + x_n$ for $x_i \geq 0$.

26G. This is trickier than it looks. We have $x_n = e^{x_n} - e^{x_{n+1}}$ but it requires some care to prove convergences. Helpful hint: $e^t \geq t + 1$ for all real numbers t, therefore all x_n's are nonnegative.

26H. The limit always exists and equals zero. Consequently, f is continuous exactly at irrational points.

28G. First rewrite it as $f(x) = e^{x \log x}$.

29B†. Because you know all derivatives of sin and cos, you can compute their Taylor series, which converge everywhere on $\mathbb{R}$. At the same time, exp was defined as a Taylor series, so you can also compute it. Write them all out and compare.

29C†. Use repeated Rolle's theorem. You don't need any of the theory in this chapter to solve this, so it could have been stated much earlier; but then it would be quite unmotivated.

29D. Use Taylor's theorem.

30A. Contradiction and mean value theorem (again!).

30B*. For every positive integer n, take a partition where every rectangle has width $w = \frac{b-a}{n}$. Use the mean value theorem to construct a tagged partition such that the first rectangle has area $f(a + w) - f(a)$, the second rectangle has area $f(a + 2w) - f(a + w)$, and so on; thus the total area is $f(b) - f(a)$.

30D. Write this as $\frac{1}{n}\sum_{k=1}^{n} \frac{1}{1+\frac{k}{n}}$. Then you can interpret it as a rectangle sum of a certain Riemann integral.

31A*. Look at the Taylor series of f, and use Cauchy's differentiation formula to show that each of the larger coefficients must be zero.

31B*. Proceed by contradiction, meaning there exists a sequence $z_1, z_2, \ldots \to z$ where $0 = f(z_1) = f(z_2) = \ldots$ all distinct. Prove that $f = 0$ on an open neighborhood of z by looking at the Taylor series of f and pulling out factors of z.

31C*. Take the interior of the agreeing points; show that this set is closed, which implies the conclusion.

31E. Liouville. Look at $\frac{1}{f(z)-w}$.

31F. You can adapt part of the proof of Cauchy-Goursat theorem presented above, and apply ML estimation lemma to prove $\oint_\gamma f(z)\,dz = 0$. In this case however, you already know f is holomorphic, so you must have $|\oint_{\gamma_i} f\,dz| \geq |\oint_\gamma f\,dz|$, without the $\frac{1}{4}$ factor.

32C. This is called a "wedge contour". Try to integrate over a wedge shape consisting of a sector of a circle of radius r, with central angle $\frac{2\pi}{n}$. Take the limit as $r \to \infty$ then.

32D. It's $\lim_{a\to\infty} \int_{-a}^{a} \frac{\cos x}{x^2+1}\,dx$. For each a, construct a semicircle.

36B. Show that

$$\mu^*(S) = \begin{cases} 0 & S = \varnothing \\ 1 & S \text{ bounded and nonempty} \\ \infty & S \text{ not bounded.} \end{cases}$$

This lets you solve (b) readily; I think the answer is just unbounded sets, $\varnothing$, and one-point sets.

39A. You can read it off Theorem 39.3.1.

39B. After Pontryagin duality, we need to show G compact implies $\widehat{G}$ discrete and G discrete implies $\widehat{G}$ compact. Both do not need anything fancy: they are topological facts.

41A. This is actually trickier than it appears, you cannot just push quantifiers (contrary to the name), but have to focus on $\varepsilon = 1/m$ for $m = 1, 2, \ldots$.

The problem is saying for each $\varepsilon > 0$, if $n > N_\varepsilon$, we have $\mu(\omega : |X(\omega) - X_n(\omega)| \leq \varepsilon) = 1$. For each m there are some measure zero "bad worlds"; take the union.

42B. There is a cute elementary solution. For the martingale-based solution, show that the fraction of red cards in the deck at time n is a martingale.

42E. Use Problem 42A.

42F. It occurs with probability 1. If X_n is the number on the board at step n, and $\mu = \frac{1}{2.01}\int_0^{2.01} \log t\,dt$, show that $\log(X_n) - n\mu$ is a martingale. (Incidentally, using the law of large numbers could work too.)

43B. Simply induct, with the work having been done on the $k = 2$ case.

44B. This is just a summation. You will need the fact that mixed partials are symmetric.

45A†. Direct application of Stokes' theorem to $\alpha = f\,dx + g\,dy$.

45B. This is just an exercises in sigma notation.

45D. This is a straightforward (but annoying) computation.

45E. We would want $\alpha_p(v) = \|v\|$.

45F. Show that $d^2 = 0$ implies $\int_{\partial c} \alpha = 0$ for exact α. Draw an annulus.

53B. Note that $p(x)$ is a minimal polynomial for r, but so is $q(x) = x^{\deg p} p(1/x)$. So q and p must be multiples of each other.

53C*. $\left|\frac{1}{n}(\varepsilon_1 + \cdots + \varepsilon_n)\right| \le 1$.

53D†. Only the obvious ones. Assume $\cos(q\pi) \in \mathbb{Q}$. Let ζ be a root of unity (algebraic integer as $\zeta^N - 1 = 0$ for some N) and note that $2\cos(q\pi) = \zeta + \zeta^{N-1}$ is both an algebraic integer and a rational number.

53E. View as roots of unity. Note $\frac{1}{2}$ isn't an algebraic integer.

53F. Let $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_n$ be its conjugates. Look at the polynomial $(x-\alpha_1^e)\ldots(x-\alpha_n^e)$ across $e \in \mathbb{N}$. Pigeonhole principle on all possible polynomials.

54A*. The norm is multiplicative and equal to product of Galois conjugates.

54B*. It's isomorphic to K.

54C. Taking the standard norm on $\mathbb{Q}(\sqrt{2})$ will destroy it.

54D. Norm in $\mathbb{Q}(\sqrt[3]{2})$.

54E†. Obviously $\mathbb{Z}[\zeta_p] \subseteq \mathcal{O}_K$, so our goal is to show the reverse inclusion. Show that for any $\alpha \in \mathcal{O}_K$, the trace of $\alpha(1-\zeta_p)$ is divisible by p. Given $x = a_0 + a_1\zeta_p + \cdots + a_{p-2}\zeta^{p-2} \in \mathcal{O}_K$ (where $a_i \in \mathbb{Q}$), consider $(1-\zeta_p)x$.

55C. Copy the proof of the usual Fermat's little theorem.

55D†. Clear denominators!

55E. (a) is straightforward. For (b) work mod p. For (c) use norms.

56A. Repeat the previous procedure.

56B. You should get a group of order three.

56C. Mimic the proof of part (a) of Minkowski's theorem.

56D. Linear algebra.

56E. Factor in $\mathbb{Q}(i)$.

56F. Factor p, show that the class group of $\mathbb{Q}(\sqrt{-5})$ has order two.

57A*. Direct linear algebra computation.

57B*. Let M be the "embedding" matrix. Look at $M^\top M$, where $M^\top$ is the transpose matrix.

57C*. Vandermonde matrices.

57D. $M_K \ge 1$ must hold. Bash.

59A*. Look at the image of ζ_p.

59C. Repeated quadratic extensions have degree 2, so one can only get powers of two.

59E. Hint: $\sigma(x^2) = \sigma(x)^2 \ge 0$ plus Cauchy's Functional Equation.

59F. By induction, suffices to show $\mathbb{Q}(\alpha, \beta) = \mathbb{Q}(\gamma)$ for some γ in terms of α and β. For all but finitely many rational λ, the choice $\gamma = \alpha + \lambda\beta$ will work.

60A†. The Fibonacci sequence is given by $F_n = \frac{\alpha^n - \beta^n}{\alpha - \beta}$ where $\alpha = \frac{1+\sqrt{5}}{2}$ and $\beta = \frac{1-\sqrt{5}}{2}$ are the two roots of $P(X) \stackrel{\text{def}}{=} X^2 - X - 1$. Show the polynomial $P(X)$ is irreducible modulo 127; then work in the splitting field of P, namely $\mathbb{F}_{p^2}$.

Show that $\mathbb{F}_p = -1$, $\mathbb{F}_{p+1} = 0$, $\mathbb{F}_{2p+1} = 1$, $\mathbb{F}_{2p+2} = 0$. (Look at the action of $\text{Gal}(\mathbb{F}_{p^2}/\mathbb{F}_p)$ on the roots of P.)

61A†. Show that no rational prime p can remain inert if $\text{Gal}(K/\mathbb{Q})$ is not cyclic. Indeed, if p is inert then $D_p \cong \text{Gal}(K/\mathbb{Q})$.

62A. Modify the end of the proof of quadratic reciprocity.

62C†. Chebotarev Density on $\mathbb{Q}(\zeta_m)$.

62E. By primitive roots, it's the same as the action of $\times 3$ on $\mathbb{Z}/(p-1)\mathbb{Z}$. Let ζ be a $(p-1)$st root of unity. Take $d = \prod_{i<j}(\zeta^i - \zeta^j)$, think about $\mathbb{Q}(d)$, and figure out how to act on it by $x \mapsto x^3$.

63A†. Pick m so that $\mathfrak{f}(L/\mathbb{Q}) \mid m\infty$.

63B†. Apply the Takagi existence theorem with $\mathfrak{m} = 1$.

63C. The extension $L/\mathbb{Q}$ is not abelian.

64C†. Prove and use the fact that a quotients of compact spaces remain compact.

68A. The category $\mathcal{A} \times \mathbf{2}$ has "redundant arrows".

71A. Take the $n-1$st homology groups.

71B. Build F as follows: draw the ray from x through $f(x)$ and intersect it with the boundary S^{n-1}.

72A. Induction on m, using hemispheres.

72B. One strategy is induction on p, with base case $p = 1$. Another strategy is to let U be the desired space and let V be the union of p non intersecting balls.

72C⋆. Use Theorem 72.2.5. Note that $\mathbb{R}^n \setminus \{0\}$ is homotopy equivalent to S^{n-1}.

72D. $0 \to A_\bullet \to B_\bullet \to C_\bullet \to 0$ is a short exact sequence of chain complexes. Write out the corresponding long exact sequence. Nearly all terms will vanish.

72E⋆. It's possible to use two cylinders with U and V. This time the matrix is $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ or some variant though; in particular, it's injective, so $\widetilde{H}_2(X) = 0$.

72F⋆. Find a new short exact sequence to apply Theorem 72.2.1 to.

73B. Use Theorem 72.2.5.

73E. For any n, prove by induction for $k = 1, \dots, n-1$ that (a) if X is a subset of S^n homeomorphic to D^k then $\widetilde{H}_i(S^n \setminus X) = 0$; (b) if X is a subset of S^n homeomorphic to S^k then $\widetilde{H}_i(S^n \setminus X) = \mathbb{Z}$ for $i = n-k-1$ and 0 otherwise.

74A†. $\mathbb{CP}^n$ has no cells in adjacent dimensions, so all d_k maps must be zero.

74B. The space $S^n - \{x_0\}$ is contractible.

74D. You won't need to refer to any elements. Start with

$$H_2(X) \cong H_2(X^3) \cong H_2(X^2)/\ker\left[H_2(X^2) \twoheadrightarrow H_2(X^3)\right],$$

say. Take note of the marked injective and surjective arrows.

74E†. There is one cell of each dimension. Show that the degree of d_k is $\deg(\mathrm{id})+\deg(-\mathrm{id})$, hence d_k is zero or $\cdot 2$ depending on whether k is even or odd.

76A†. Write $H^k(M;\mathbb{Z})$ in terms of $H_k(M)$ using the UCT, and analyze the ranks.

76B. Use the previous result on Betti numbers.

76C. Use the $\mathbb{Z}/2\mathbb{Z}$ cohomologies, and find the cup product.

76D. Assume that $r\colon S^m \times S^n \to S^m \vee S^n$ is such a map. Show that the induced map $H^\bullet(S^m \vee S^n;\mathbb{Z}) \to H^\bullet(S^m \times S^n;\mathbb{Z})$ between their cohomology rings is monic (since there exists an inverse map i).

77B. Squares are nonnegative.

77C. This is actually an equivalent formulation of the Weak Nullstellensatz.

77D. Use the weak Nullstellensatz on $n+1$ dimensions. Given f vanishing on everything, consider $x_{n+1}f - 1$.

80B. You will need to know about complex numbers in Euclidean geometry to solve this problem.

81B†. Use the standard affine charts.

81C. Examine the global regular functions.

81D. Assume f was an isomorphism. Then it gives an isomorphism $f^\sharp\colon \mathcal{O}_V(V) \to \mathcal{O}_X(X) = \mathbb{C}[x,y]$. Thus we may write $\mathcal{O}_V(V) = \mathbb{C}[a,b]$, where $f^\sharp(a) = x$ and $f^\sharp(b) = y$. Let $f(p) = q$ where $\mathcal{V}(a,b) = \{q\}$. Use the definition of pullback to prove $p \in \mathcal{V}(x,y)$, contradiction.

82C. The stalk is R at points in the closure of $\{p\}$, and 0 elsewhere.

82D. Show that the complement $\{p \mid [s]_p = 0\}$ is open.

83B. Consider zero divisors.

83C*. Only one! A proof will be given a few chapters later.

83D. No. Imagine two axes.

84A. Galois conjugates.

85B. $k[x,y] \times k[z,z^{-1}]$.

85D. It's isomorphic to R!

87A. Use the fact that $\mathsf{AffSch} \simeq \mathsf{CRing}$.

88A. Let $\varepsilon = \pi - 3.141592653 < 10^{-9}$. Find $f(\varepsilon)$.

89E. This is an application of Axiom of Choice.

91A. $\sup_{k\in\omega} |V_k|$.

91B. Rearrange the cofinal maps to be nondecreasing.

92C†. This is very similar to the proof of Löwenheim-Skolem. For a sentence ϕ, let f_ϕ send α to the least $\beta < \kappa$ such that for all $\vec{b} \in V_\alpha$, if there exists $a \in M$ such that $V_\kappa \vDash \phi[a, \vec{b}]$ then $\exists a \in V_\beta$ such that $V_\kappa \vDash \phi[a, \vec{b}]$. (To prove this β exists, use the fact that κ is cofinal.) Then, take the supremum over the countably many sentences for each α.

92D*. Use Lemma 92.5.1. To prove $V_\kappa \vDash$ PowerSet you need κ to be a strong limit cardinal, and to prove $V_\kappa \vDash$ Replacement you need κ to be inaccessible — this is why we cared about cofinality and inaccessibility.

93B. Let D_1, D_2, ... be the dense sets (there are countably many of them).

94A. Assume not, and take $\lambda > \kappa$ regular in M; if $f\colon \overline{\lambda} \to \lambda$, use the Possible Values Argument on f to generate a function in M that breaks cofinality of λ.

C Sketches of selected solutions

1A. The point is that $\heartsuit$ is a group, $G \subsetneq \heartsuit$ a subgroup and $G \cong \heartsuit$. This can only occur if $|\heartsuit| = \infty$; otherwise, a proper subgroup would have strictly smaller size than the original.

1B. Let $\{g_1, g_2, \ldots, g_n\}$ denote the elements of G. For any $g \in G$, this is the same as the set $\{gg_1, \ldots, gg_n\}$. Taking the entire product and exploiting commutativity gives $g^n \cdot g_1 g_2 \ldots g_n = g_1 g_2 \ldots g_n$, hence $g^n = 1$.

1C. One can check manually that $D_6 \cong S_3$, using the map $r \mapsto (1\ 2\ 3)$ and $s \mapsto (1\ 2)$. (The right-hand sides are in "cycle notation", as mentioned in Section 6.iv.) On the other hand D_{24} contains an element of order 12 while S_4 does not.

1D*. Let G be a group of order p, and $1 \neq g \in G$. Look at the group H generated by g and use Lagrange's theorem.

1F†. The idea is that each element $g \in G$ can be thought of as a permutation $G \to G$ by $x \mapsto gx$.

1G. The answer is $n = 1009$. This solution uses the fact that 1009 is prime.

To show that no smaller m is possible, note that D_{2018} has elements of order 1009, a prime. Since S_n has no elements of this order for $n < 1009$, we need $n \geq 1009$.

To give a construction from $n = 1009$, note that D_{2018} can be thought of the symmetries of a 1009-gon. If one labels the vertices of the 1009-gon by $S := \{1, 2, \ldots, 1009\}$, then elements of D_{2018} induces permutations on S, and the set of permutations achieved is the desired subgroup.

1H. We have $www = bb$, $bww = wb$, $wwb = bw$, $bwb = ww$. Interpret these as elements of D_6.

1I. Look at the group G of 2×2 matrices mod p with determinant ± 1 (whose entries are the integers mod p). Let $g = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ and then use $g^{|G|} = 1_G$.

2B. Two possible approaches, one using metric definition and one using open sets.

Metric approach: I claim there is no injective map from $\mathbb{Q}$ to $\mathbb{N}$ that is continuous. Indeed, suppose f was such a map and $f(x) = n$. Then, choose $\varepsilon = 1/2$. There should be a $\delta > 0$ such that everything with δ of x in $\mathbb{Q}$ should land within ε of $n \in \mathbb{N}$ — i.e., is equal to n. This is a blatant contradiction of injectivity.

Open set approach: In $\mathbb{Q}$, no singleton set is open, whereas in $\mathbb{N}$, they all are (in fact $\mathbb{N}$ is discrete). As you'll see at the start of Chapter 7, with the new and improved definition of "homeomorphism", we found out that the structure of open sets on $\mathbb{Q}$ and $\mathbb{N}$ are different, so they are not homeomorphic.

2C. For subtraction, the map $x \mapsto -x$ is continuous so you can view it as a composed map

$$\mathbb{R} \times \mathbb{R} \xrightarrow{(\mathrm{id},-x)} \mathbb{R} \times \mathbb{R} \xrightarrow{+} \mathbb{R}$$

$$(a,b) \longmapsto (a,-b) \longmapsto a-b.$$

Similarly, if you are willing to believe $x \mapsto 1/x$ is a continuous function, then division is composition

$$\mathbb{R} \times \mathbb{R}_{>0} \xrightarrow{(\mathrm{id},1/x)} \mathbb{R} \times \mathbb{R}_{>0} \xrightarrow{\times} \mathbb{R}$$

$$(a,b) \longmapsto (a,1/b) \longmapsto a/b.$$

If for some reason you are suspicious that $x \mapsto 1/x$ is continuous, then here is a proof using sequential continuity. Suppose $x_n \to x$ with $x_n > 0$ and $x > 0$ (since x needs to be in $\mathbb{R}_{>0}$ too). Then

$$\left| \frac{1}{x} - \frac{1}{x_n} \right| = \frac{|x_n - x|}{|x x_n|}.$$

If n is large enough, then $|x_n| > x/2$; so the denominator is at least $x^2/2$, and hence the whole fraction is at most $\frac{2}{x^2}|x_n - x|$, which tends to zero as $n \to \infty$.

2D. Let $f(x) = x$ for $x \in \mathbb{Q}$ and $f(x) = -x$ for irrational x.

2E. Assume for contradiction it is completely discontinuous; by scaling set $f(0) = 0$, $f(1) = 1$ and focus just on $f \colon [0,1] \to [0,1]$. Since it's discontinuous everywhere, for every $x \in [0,1]$ there's an $\varepsilon_x > 0$ such that the continuity condition fails. Since the function is strictly increasing, that can only happen if the function misses all y-values in the interval $(f(x) - \varepsilon_x, f(x))$ or $(f(x), f(x) + \varepsilon_x)$ (or both).

Projecting these missing intervals to the y-axis you find uncountably many intervals (one for each $x \in [0,1]$) all of which are disjoint. In particular, summing the ε_x you get that a sum of uncountably many positive reals is 1.

But in general it isn't possible for an uncountable family $\mathcal{F}$ of positive reals to have finite sum. Indeed, just classify the reals into buckets $\frac{1}{k} \le x < \frac{1}{k-1}$. If the sum is actually finite then each bucket is finite, so the collection $\mathcal{F}$ must be countable, contradiction.

2F. Like most Internet "debates" about math, the question revolves around sloppy definitions. The original posed question (which is ill-formed) is

(1) Is $1/x$ a continuous function?

To make it well-formed, I want to *first* bring up the question:

(2) Is $1/x$ a function?

Technically, this question is *also* ill-formed because it never specifies the domain of the function, which is part of the data needed to specify a function. One reasonable guess what the asker meant would be $\mathbb{R} \setminus \{0\}$, i.e. the set of nonzero real numbers, in which case we get the question

(2') Does $1/x$ define a function from $\mathbb{R} \setminus \{0\}$ to $\mathbb{R}$?

which has the firm answer YES.

On the other hand, it does *not* make sense to try to define $1/x$ as a function on $\mathbb{R}$. The definition a function requires you to specify an output value for every input, so at least if you want a real-valued function[1], there isn't any way to construe $1/x$ as a function on all of $\mathbb{R}$.

Now, returning to (1), we can now ask a well-formed question

(1') Does $1/x$ describe a continuous function from $\mathbb{R} \setminus \{0\} \to \mathbb{R}$?

which again has the firm answer YES.

Of course, you could also consider a question like "does $1/x$ describe a continuous function $\mathbb{R} \to \mathbb{R}$?". However, this feels misleading: it would be like asking "is $\sqrt{2}$ an even integer?". The question doesn't make sense to begin with because $\sqrt{2}$ isn't an integer, and "even" is an adjective used for integers, so trying to ask whether it applies to $\sqrt{2}$ is a type-error. Similarly, "continuous" is an adjective used for functions; it doesn't make sense to ask whether it applies to something that isn't a function.

See https://twitter.com/davidcpvm/status/1481024944830046209 for the Twitter post (in Spanish) and the accompanying Reddit post (one of several) at https://www.reddit.com/r/math/comments/s82vf8.

3A. Abelian groups: $abab = a^2b^2 \iff ab = ba$.

3B. Yes to (a): you can check this directly from the ghg^{-1} definition. For example, for (a) it is enough to compute $(r^a s)r^n(r^a s)^{-1} = r^{-n} \in H$. The quotient group is $\mathbb{Z}/2\mathbb{Z}$.

The answer is no for (b) by following Example 3.5.2.

3C. A subgroup of order 3 must be generated by an element of order 3, since 3 is prime. So we may assume WLOG that $H = \langle (1\ 2\ 3) \rangle$ (by renaming elements appropriately). But then let $g = (3\ 4)$; one can check $gHg^{-1} \neq H$.

3D. $G/\ker G$ is isomorphic to a subgroup of H. The order of the former divides 1000; the order of the latter divides 999. This can only occur if $G/\ker G = \{1\}$ so $\ker G = G$.

3F. Quaternion group.

3G. The answer is $|G| = 18$.

First, observe that by induction we have

$$a^n c = ca^{8n}$$

for all $n \geq 1$. We then note that

$$\begin{aligned} a(bc) &= (ab)c \\ a \cdot ca^6 &= c^2a^4 \cdot c \\ ca^8 \cdot a^6 &= c^2a^4 \cdot c \\ a^{14} &= c(a^4c) = c^2a^{32}. \end{aligned}$$

[1]Those of you that know what $\mathbb{RP}^1$ is could consider it as a function $\mathbb{RP}^1 \to \mathbb{RP}^1$ if you insisted; but it's continuous in that case too.

Hence we conclude $c^2 = a^{-18}$. Then $ab = c^2a^4 \implies b = a^{-15}$.

In that case, since $c^{2018} = b^{2019}$, we conclude $1 = a^{-1009\cdot18+2019\cdot15} = a^{12123}$. Finally,

$$\begin{aligned} bc &= ca^6 \\ a^{-15}c &= ca^6 \\ a^{-15}c^2 &= c(a^6c) = c^2a^{48} \\ a^{-33} &= a^{30} \\ \implies a^{63} &= 1. \end{aligned}$$

Since $\gcd(12123, 63) = 9$, we find $a^9 = 1$, hence finally $c^2 = 1$. So the presentation above simplifies to

$$G = \left\langle a, c \mid a^9 = c^2 = 1,\ ac = ca^{-1} \right\rangle$$

which is the presentation of the dihedral group of order 18. This completes the proof.

3H. You can find many solutions by searching "homophone group"; one is https://math.stackexchange.com/q/843966/229197.

4A. This is just $\mathbb{R}[i] = \mathbb{C}$. The isomorphism is given by $x \mapsto i$, which has kernel (x^2+1).

4B. Note that the map

$$\begin{aligned} \mathbb{C}[x] &\to \mathbb{C} \times \mathbb{C} \\ p &\mapsto (p(0), p(1)) \end{aligned}$$

is indeed a surjective ring homomorphism. Its kernel consists of those polynomials p such that $p(0) = p(1) = 0$; this is the set of polynomials divisible by both x and $x-1$, so it is $x(x-1)$.

5C*. Consider $ab \in \phi^{\text{pre}}(I)$, meaning $\phi(ab) = \phi(a)\phi(b) \in I$. Since I is prime, either $\phi(a) \in I$ or $\phi(b) \in I$. In the former case we get $a \in \phi^{\text{pre}}(I)$ as needed; the latter case we get $b \in \phi^{\text{pre}}(I)$.

5D*. Let $x \in R$ with $x \neq 0$. Look at the powers x, x^2, ... By pigeonhole, eventually two of them coincide. So assume $x^m = x^n$ where $m < n$, or equivalently

$$0 = x \cdot x \cdot \cdots \cdot x \cdot \left(x^{n-m} - 1\right).$$

Since $x \neq 0$, we get $x^{n-m} - 1 = 0$, or $x^{n-m} = 1$. So x^{n-m-1} is an inverse for x.

This means every nonzero element has an inverse, ergo R is a field.

5E*. For part (b), look at the poset of *proper* ideals. Apply Zorn's lemma (again using a union trick to verify the condition; be sure to verify that the union is proper!). In part (a) we are given no ascending infinite chains, so no need to use Zorn's lemma.

5F. The ideal (0) is of course prime in both. Also, both rings are PID's.

For $\mathbb{C}[x]$ we get a prime ideal $(x-z)$ for each $z \in \mathbb{C}$.

For $\mathbb{R}[x]$ a prime ideal $(x-a)$ for each $a \in \mathbb{R}$ and a prime ideal $(x^2 - ax + b)$ for each quadratic with two conjugate non-real roots.

5G†. Only one; the ideal (0) which is not maximal. We contend every other prime ideal is maximal.

Indeed, let I be any ideal (not necessarily prime), and let $a + b\sqrt{2017}$ be a nonzero element of it. Then I also contains $(a^2 - 2017b^2)$. That means when taking modulo I we may take modulo the integer $n \coloneqq |a^2 - 2017b^2| \neq 0$.

So every element in R is equivalent modulo I to an element of the form $x + y\sqrt{2017}$, where $x, y \in \{0, 1, \ldots, n-1\}$. In other words, the quotient R/I has at most finitely many elements.

When I is prime, it follows R/I is an integral domain, too. An integral domain with finitely many elements must be a field. Hence, from R/I being a field, we conclude I is maximal.

5H. The ideals are (0), $(1) = R$, and $(5^n) = 5^n R$ for each $n \geq 1$. The ideal (0) is prime and the ideal (5) is maximal (because the quotient $R/(5) \cong \mathbb{F}_5$ is a field).

6A†. Uniqueness of the fixed point follows from noting that if $T(p) = p$ and $T(q) = q$ and $p \neq q$ then we get a direct contradiction by plugging this into the given statement. Hence the main task is to show there exists some fixed point.

Start with any point x_0. Let $x_1 = T(x_0)$, $x_2 = T(x_1)$, $x_3 = T(x_2)$, ..., and so on. We contend that $(x_0, x_1, x_2, \ldots)$ is a Cauchy sequence. Indeed, if we let $r \coloneqq 0.999 < 1$ and $c \coloneqq d(x_0, x_1)$, then

$$\begin{aligned} d(x_1, x_2) &< r \cdot c \\ d(x_2, x_3) &< r^2 \cdot c \\ d(x_3, x_4) &< r^3 \cdot c \\ &\vdots \end{aligned}$$

and so for large $M < N$ we have

$$d(x_M, x_N) < \left(r^M + r^{M+1} + \cdots + r^N\right) \cdot c < \frac{r^M}{1-r} \cdot c$$

which tends to zero once M is large enough.

Hence, because M is complete, the sequence must converge to some limit x. Because T is continuous, we get

$$T(x) = T\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} T(x_n) == \lim_{n\to\infty} x_{n+1} = x$$

as desired.

6B. Part (a) is essentially by definition. The space M is bounded since no distances exceed 1, but not totally bounded since we can't cover M with finitely many $\frac{1}{2}$-neighborhoods. The space M is complete since a sequence of real numbers converges in M if it converges in the usual sense. As for N, the sequence -1, -2, ... is Cauchy but fails to converge; and it is obviously not bounded.

To show (b), the identity map (!) is an homeomorphism $M \cong \mathbb{R}$ and $\mathbb{R} \cong N$, since it is continuous.

This illustrates that $M \cong N$ despite the fact that M is both complete and bounded but N is neither complete nor bounded. On the other hand, we will later see that

complete and totally bounded implies *compact*, which is a very strong property preserved under homeomorphism.

6D. See https://math.stackexchange.com/q/556150/229197.

7E. Part (a) is straightforward: assume for contradiction that the connected component of p is a disjoint union $U \sqcup V$ of two nonempty sets open in X. WLOG, assume $x \in U$ Let S be one of the subspaces containing X that intersects V. Then $S = (S \cap U) \sqcup (S \cap V)$ rewrites S as the disjoint union of two sets which are open in S, contradicting the connectedness of S.

(Though note that as S is not necessarily open in X, the sets $S \cap U$ and $S \cap V$ are not necessarily open in X either.)

For (b), a counterexample is to take any *totally disconnected* space like the Cantor set or the p-adic numbers.

7G. Let $d(x,y) = 2017^{-n}$, where n is the largest integer such that $n!$ divides $|x-y|$.

7H. You can pick a rational number in each interval and there are only countably many rational numbers. Done!

8A. Compactness is preserved under homeomorphism, but $[0,1]$ is compact while $(0,1)$ is not.

8E. Suppose $p_i = (x_i, y_i)$ is a sequence in $X \times Y$ $(i = 1, 2, \dots)$. Looking on the X side, some subsequence converges: for the sake of illustration say it's $x_1, x_4, x_9, x_{16}, \cdots \to x$. Then look at the corresponding sequence $y_1, y_4, y_9, y_{16}, \dots$. Using compactness of Y, it has a convergent subsequence, say $y_1, y_{16}, y_{81}, y_{256}, \cdots \to y$. Then $p_1, p_{16}, p_{81}, \dots$ will converge to (x,y).

One common mistake is to just conclude that (x_n) has a convergent subsequence and that (y_n) does too. But these sequences could be totally unrelated. For this proof to work, you do need to apply compactness of X first, and then compactness of Y on the resulting *filtered* sequence like we did here.

8H. The following solution is due to Royce Yao. We show the contrapositive: if M is not compact, then there exists a homeomorphic unbounded metric.

The main step is to construction an unbounded continuous function $F \colon M \to \mathbb{R}$. Once such a function F is defined, the metric

$$d'(x,y) \coloneqq d(x,y) + |F(x) - F(y)|$$

will solve the problem.

So, let a_1, a_2, $\dots$ be a sequence in M with no convergent subsequence. For each a_i, there exists a radius r_i such that

$$0 < r_i < \frac{1}{2} \min_j d(a_i, a_j)$$

Define C_i as an open ball at a_i with radius r_i. Note that every ball is disjoint. Then, we define F as follow

$$F(x) = \begin{cases} 0 & x \notin C_i \\ \frac{i}{r_1}(r_i - d(x,a_i)) & x \in C_i \end{cases}$$

which can be seen to be continuous. Then, F is unbounded by considering $F(a_i)$ as i goes to infinity.

8I. Part (a) follows by the Cantor intersection theorem (Problem 8D). Assume for contradiction such a partition existed. Take any of the circles C_0, and let K_0 denote the closed disk with boundary C_0. Now take the circle C_1 passing through the center of C_0, and let K_1 denote the closed disk with boundary C_1. If we repeat in this way, we get a nested sequence $K_0 \supseteq K_1 \supseteq \dots$ and the radii of C_i approach zero (since each is at most half the previous once). Thus some point p lies in $\bigcap_n K_n$ which is impossible.

Now for part (b), again assume for contradiction a partition into circles exists. Color a circle magenta if it contains p but not q and color a circle cyan if it contains q but not p. Color p itself magenta and q itself cyan as well. Finally, color a circle neon yellow if it contains both p and q. (When we refer to coloring a circle, we mean to color all the points on it.)

By repeating the argument in (a) there are no circles enclosing neither p nor q. Hence every point is either magenta, cyan, or neon yellow. Now note that given any magenta circle, its interior is completely magenta. Actually, the magenta circles can be totally ordered by inclusion (since they can't intersect). So we consider two cases:

- If there is a magenta circle which is maximal by inclusion (i.e. a magenta circle not contained in any other magenta circle) then the set of all magenta points is just a closed disk.
- If there is no such magenta circle, then the set of magenta points can also be expressed as the union over all magenta circles of their interiors. This is a union of open sets, so it is itself open.

We conclude the set of magenta points is either a closed disk or an open set. Similarly for the set of cyan points. Moreover, the set of such points is convex.

To finish the problem:

- Suppose there are no neon yellow points. If the magenta points form a closed disk, then the cyan points are $\mathbb{R}^2$ minus a disk which is not convex. Contradiction. So the magenta points must be open. Similarly the cyan points must be open. But $\mathbb{R}^2$ is connected, so it can't be written as the union of two open sets.
- Now suppose there are neon yellow points. We claim there is a neon yellow circle minimal by inclusion. If not, then repeat the argument of (a) to get a contradiction, since any neon yellow circle must have diameter the distance from p to q. So we can find a neon yellow circle $\mathscr{C}$ whose interior is all magenta and cyan. Now repeat the argument of the previous part, replacing $\mathbb{R}^2$ by the interior of $\mathscr{C}$.

9A†.

	T injective	T surjective	T isomorphism
If $\dim V > \dim W \dots$	never	sometimes	never
If $\dim V = \dim W \dots$	sometimes	sometimes	sometimes
If $\dim V < \dim W \dots$	sometimes	never	never

Each "never" is by the rank-nullity theorem. Each counterexample is obtained by the zero map sending every element of V to zero; this map is certainly neither injective or surjective.

9B†. It essentially follows by Theorem 9.7.6.

9D. Since $1 \mapsto \sqrt{5}$ and $\sqrt{5} \mapsto 5$, the matrix is $\begin{bmatrix} 0 & 5 \\ 1 & 0 \end{bmatrix}$.

9G. Let V be the space of real polynomials with degree at most $d/2$ (which has dimension $1 + \lfloor d/2 \rfloor$), and W be the space of real polynomials modulo P (which has dimension d). Then $\dim(V \oplus V) > \dim W$. So the linear map $V \oplus V \to W$ by $(A, B) \mapsto A + Q \cdot B$ has a kernel of positive dimension (by rank-nullity, for example).

9I*. Consider

$$\{0\} \subsetneq \ker S \subseteq \ker S^2 \subseteq \ker S^3 \subseteq \dots \text{ and } V \supsetneq \operatorname{im} S \supseteq \operatorname{im} S^2 \supseteq \operatorname{im} S^3 \supseteq \dots.$$

For dimension reasons, these subspaces must eventually stabilize: for some large integer N, $\ker T^N = \ker T^{N+1} = \dots$ and $\operatorname{im} T^N = \operatorname{im} T^{N+1} = \operatorname{im} T^{N+2} = \dots$. When this happens, $\ker T^N \bigcap \operatorname{im} T^N = \{0\}$, since T^N is an automorphism of $\operatorname{im} T^N$. On the other hand, by Rank-Nullity we also have $\dim \ker T^N + \dim \operatorname{im} T^n = \dim V$. Thus for dimension reasons, $V = \ker T^N \oplus \operatorname{im} T^N$.

10A. It's just $\dim V = 2018$. After all, you are adding the dimensions of the Jordan blocks...

10B. (a): if you express T as a matrix in such a basis, one gets a diagonal matrix. (b): this is just saying each Jordan block has dimension 1, which is what we wanted. (We are implicitly using uniqueness of Jordan form here.)

10C. The $+1$ eigenspace is spanned by $e_1 + e_2$. The -1 eigenspace is spanned by $e_1 - e_2$.

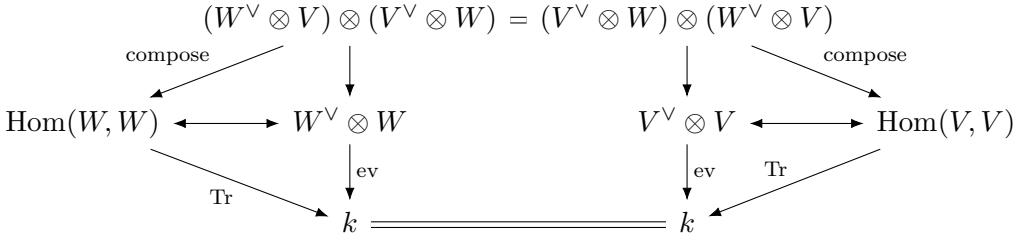
10E. The $+1$ eigenspace is spanned by $1 + x^2$ and x. The -1 eigenspace is spanned by $1 - x^2$.

10F. Constant functions differentiate to zero, and these are the only 0-eigenvectors. There can be no other eigenvectors, since if $\deg p > 0$ then $\deg p' = \deg p - 1$, so if p' is a constant real multiple of p we must have $p' = 0$, ergo p is constant.

10G. e^{cx} is an example of a c-eigenvector for every c. If you know differential equations, these generate all examples!

11A. We saw already the trace is always the sum of the eigenvalues, in *any* basis. In particular, choosing the Jordan form basis from the previous chapter gives the result because the Jordan form has the eigenvalues for its diagonal entries.

11C†. Although we could give a coordinate calculation, we instead opt to give a cleaner proof. This amounts to drawing the diagram



It is easy to check that the center rectangle commutes, by checking it on pure tensors $\xi_W \otimes v \otimes \xi_V \otimes w$. So the outer hexagon commutes and we're done. This is really the same as the proof with bases; what it amounts to is checking the assertion is true for matrices that have a 1 somewhere and 0 elsewhere, then extending by linearity.

11D. See https://mks.mff.cuni.cz/kalva/putnam/psoln/psol886.html.

12D. Recall that (by Problem 9B†) we can replace "isomorphism" by "injective".

If $T(v) = 0$ for any nonzero v, then by taking a basis for which $e_1 = v$, we find $\bigwedge^n(T)$ will map $e_1 \wedge \dots$ to $0 \wedge T(e_2) \wedge \dots = 0$, hence is the zero map, so $\det T = 0$.

Conversely, if T is an isomorphism, we let S denote the inverse map. Then $1 = \det(\mathrm{id}) = \det(S \circ T) = \det S \det T$, so $\det T \neq 0$.

12E. We proceed by contradiction. Let v be a vector of length 1000 whose entries are weight of cows. Assume the existence of a matrix M such that $Mv = 0$, with entries 0 on diagonal and ± 1 off-diagonal. But $\det M \pmod 2$ is equal to the number of derangements of $\{1, \dots, 1000\}$, which is odd. Thus $\det M$ is odd and in particular not zero, so M is invertible. Thus $Mv = 0 \implies v = 0$, contradiction.

12F. The answer is

$$\begin{bmatrix} t & t \\ t & t \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -3t & -t \\ t & 3t \end{bmatrix}$$

for $t \in \mathbb{R}$. These work by taking $k = 3$.

Now to see these are the only ones, consider an arithmetic matrix

$$M = \begin{bmatrix} a & a+e \\ a+2e & a+3e \end{bmatrix}.$$

with $e \neq 0$. Its characteristic polynomial is $t^2 - (2a+3e)t - 2e^2$, with discriminant $(2a+3e)^2 + 8e^2$, so it has two distinct real roots; moreover, since $-2e^2 \leq 0$ either one of the roots is zero or they are of opposite signs. Now we can diagonalize M by writing

$$M = \begin{bmatrix} s & -q \\ -r & p \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} p & q \\ r & s \end{bmatrix} = \begin{bmatrix} ps\lambda_1 - qr\lambda_2 & qs(\lambda_1 - \lambda_2) \\ pr(\lambda_2 - \lambda_1) & ps\lambda_2 - qr\lambda_1 \end{bmatrix}$$

where $ps - qr = 1$. By using the fact the diagonal entries have sum equalling the off-diagonal entries, we obtain that

$$(ps - qr)(\lambda_1 + \lambda_2) = (qs - pr)(\lambda_1 - \lambda_2) \implies qs - pr = \frac{\lambda_1 + \lambda_2}{\lambda_1 - \lambda_2}.$$

Now if $M^k \in S$ too then the same calculation gives

$$qs - pr = \frac{\lambda_1^k + \lambda_2^k}{\lambda_1^k - \lambda_2^k}.$$

Let $x = \lambda_1/\lambda_2 < 0$ (since $-2e^2 < 0$). We appropriately get

$$\frac{x+1}{x-1} = \frac{x^k+1}{x^k-1} \implies \frac{2}{x-1} = \frac{2}{x^k-1} \implies x = x^k \implies x = -1 \text{ or } x = 0$$

and k odd. If $x = 0$ we get $e = 0$ and if $x = -1$ we get $2a + 3e = 0$, which gives the curve of solutions that we claimed.

A slicker approach is by Cayley-Hamilton. Assume that $e \neq 0$, so M has two distinct real eigenvalues as above. We have $M^k = cM + d\mathrm{id}$ for some constants c and d (since M satisfies some quadratic polynomial). Since $M \in S$, $M^k \in S$ we obtain $d = 0$. Thus $M^k = cM$, so it follows the eigenvalues of M are negatives of each other. That means $\operatorname{Tr} M = 0$, and the rest is clear.

12G. Pick a basis $e_1, \dots, e_n$ of V. Let T have matrix (x_{ij}), and let $m = \dim V$. Let δ_{ij} be the Kronecker delta. Also, let $\operatorname{Fix}(\sigma)$ denote the fixed points of a permutation σ and let $\operatorname{NoFix}(\sigma)$ denote the non-fixed points.

Expanding then gives

$$\begin{aligned}
&\det(a \cdot \mathrm{id} - T) \\
&= \sum_{\sigma \in S_m} \left(\operatorname{sign}(\sigma) \cdot \prod_{i=1}^{m} \left(a \cdot \delta_{i\sigma(i)} - x_{i\sigma(i)} \right) \right) \\
&= \sum_{s=0}^{m} \sum_{1 \le i_1 < \dots < i_s \le m} \sum_{\substack{\sigma \in S_m \\ \sigma \text{ fixes } i_k}} \left(\operatorname{sign}(\sigma) \cdot \prod_{i=1}^{m} \left(a \cdot \delta_{i\sigma(i)} - x_{i\sigma(i)} \right) \right) \\
&= \sum_{s=0}^{m} \sum_{1 \le i_1 < \dots < i_s \le m} \sum_{\substack{\sigma \in S_m \\ \sigma \text{ fixes } (i_k)}} \left(\operatorname{sign}(\sigma) \cdot \prod_{i \notin (i_k)} -x_{i\sigma(i)} \prod_{i \in (i_k)}^{n} (a \cdot -x_{ii}) \right) \\
&= \sum_{\sigma \in S_m} \left(\operatorname{sign}(\sigma) \cdot \prod_{i \in \operatorname{NoFix}(\sigma)} -x_{i\sigma(i)} \prod_{i \in \operatorname{Fix} \sigma} (a - x_{ii}) \right) \\
&= \sum_{\sigma \in S_m} \left(\operatorname{sign}(\sigma) \cdot \left(\prod_{i \in \operatorname{NoFix}(\sigma)} -x_{i\sigma(i)} \right) \left(\sum_{t=0}^{|\operatorname{Fix}(\sigma)|} a^{|\operatorname{Fix}(\sigma)| - t} \cdot \sum_{i_1 < \dots < i_t \in \operatorname{Fix}(\sigma)} \prod_{k=1}^{t} -x_{i_k i_k} \right) \right) \\
&= \sum_{\sigma \in S_m} \left(\operatorname{sign}(\sigma) \left(\sum_{t=0}^{|\operatorname{Fix}(\sigma)|} a^{m-t-|\operatorname{NoFix}(\sigma)|} \sum_{\substack{X \subseteq \{1,\dots,m\} \\ \operatorname{NoFix}(\sigma) \subseteq X \\ X \text{ has exactly } t \text{ fixed}}} \prod_{i \in X} -x_{i\sigma(i)} \right) \right) \\
&= \sum_{n=0}^{m} a^{m-n} \left(\sum_{\sigma \in S_m} \operatorname{sign}(\sigma) \sum_{\substack{X \subseteq \{1,\dots,m\} \\ \operatorname{NoFix}(\sigma) \subseteq X \\ |X| = n}} \prod_{i \in X} -x_{i\sigma(i)} \right) \\
&= \sum_{n=0}^{m} a^{m-n} (-1)^n \left(\sum_{\substack{X \subseteq \{1,\dots,m\} \\ |X| = n}} \sum_{\substack{\sigma \in S_m \\ \operatorname{NoFix}(\sigma) \subseteq X}} \operatorname{sign}(\sigma) \prod_{i \in X} x_{i\sigma(i)} \right).
\end{aligned}$$

Hence it's the same to show that

$$\sum_{\substack{X \subseteq \{1,\dots,m\} \\ |X| = n}} \sum_{\substack{\sigma \in S_m \\ \operatorname{NoFix}(\sigma) \subseteq X}} \operatorname{sign}(\sigma) \prod_{i \in X} x_{i\sigma(i)} = \operatorname{Tr}_{\bigwedge^n(V)} \left(\bigwedge^n (T) \right)$$

holds for every n.

We can expand the definition of trace as using basis elements as

$$\begin{aligned}
\operatorname{Tr}\left(\bigwedge^n(T)\right) &= \sum_{1\le i_1<\cdots<i_n\le m} \left(\bigwedge_{k=1}^n e_{i_k}\right)^\vee \left(\bigwedge^n(T)\left(\bigwedge_{k=1}^n e_{i_k}\right)\right) \\
&= \sum_{1\le i_1<\cdots<i_n\le m} \left(\bigwedge_{k=1}^n e_{i_k}\right)^\vee \left(\bigwedge_{k=1}^n T(e_{i_k})\right) \\
&= \sum_{1\le i_1<\cdots<i_n\le m} \left(\bigwedge_{k=1}^n e_{i_k}\right)^\vee \left(\bigwedge_{k=1}^n \left(\sum_{j=1}^m x_{i_k j} e_j\right)\right) \\
&= \sum_{1\le i_1<\cdots<i_n\le m} \sum_{\pi\in S_n} \operatorname{sign}(\pi) \prod_{k=1}^n x_{i_{\pi(k)}k} \\
&= \sum_{\substack{X\subseteq\{1,\ldots,m\} \\ |X|=n}} \sum_{\pi\in S_X} \operatorname{sign}(\pi) \prod_{i\in X} x_{t\pi(t)}
\end{aligned}$$

Hence it remains to show that the permutations over X are in bijection with the permutations over S_m which fix $\{1,\ldots,m\} - X$, which is clear, and moreover, the signs clearly coincide.

13C. Interpret clubs as vectors in the vector space $\mathbb{F}_2^n$. Consider a "dot product" to show that all k vectors are linearly independent: any two different club-vectors have dot product 0, while each club vector has dot product 1 with itself. So these vectors are orthonormal and hence linearly independent. Thus $k \le \dim \mathbb{F}_2^n = n$.

13D*. The inner form given by

$$\langle v_1 \otimes w_1, v_2 \otimes w_2\rangle_{V\otimes W} = \langle v_1, v_2\rangle_V \langle w_1, w_2\rangle_W$$

on pure tensors, then extending linearly. For (b) take $e_i \otimes f_j$ for $1 \le i \le n$, $1 \le j \le m$.

14B. Define the Boolean function $D\colon \{\pm 1\}^3 \to \mathbb{R}$ by

$$D(a,b,c) = ab + bc + ca = \begin{cases} 3 & a,b,c \text{ all equal} \\ -1 & a,b,c \text{ not all equal.} \end{cases}.$$

Thus paradoxical outcomes arise when $D(f(x_\bullet), g(y_\bullet), h(z_\bullet)) = 3$. Now, we compute that for randomly selected $x_\bullet$, $y_\bullet$, $z_\bullet$ that

$$\begin{aligned}
\mathbb{E}D(f(x_\bullet), g(y_\bullet), h(z_\bullet)) &= \mathbb{E}\sum_S \sum_T \left(\widehat{f}(S)\widehat{g}(T) + \widehat{g}(S)\widehat{h}(T) + \widehat{h}(S)\widehat{f}(T)\right)(\chi_S(x_\bullet)\chi_T(y_\bullet)) \\
&= \sum_S \sum_T \left(\widehat{f}(S)\widehat{g}(T) + \widehat{g}(S)\widehat{h}(T) + \widehat{h}(S)\widehat{f}(T)\right)\mathbb{E}\left(\chi_S(x_\bullet)\chi_T(y_\bullet)\right).
\end{aligned}$$

Now we observe that:

- If $S \ne T$, then $\mathbb{E}\chi_S(x_\bullet)\chi_T(y_\bullet) = 0$, since if say $s \in S$, $s \notin T$ then x_s affects the parity of the product with 50% either way, and is independent of any other variables in the product.

- On the other hand, suppose $S = T$. Then
$$\chi_S(x_\bullet)\chi_T(y_\bullet) = \prod_{s\in S} x_s y_s.$$
Note that $x_s y_s$ is equal to 1 with probability $\frac{1}{3}$ and -1 with probability $\frac{2}{3}$ (since (x_s, y_s, z_s) is uniform from $3! = 6$ choices, which we can enumerate). From this an inductive calculation on $|S|$ gives that
$$\prod_{s\in S} x_s y_s = \begin{cases} +1 & \text{with probability } \frac{1}{2}(1 + (-1/3)^{|S|}) \\ -1 & \text{with probability } \frac{1}{2}(1 - (-1/3)^{|S|}). \end{cases}$$
Thus
$$\mathbb{E}\left(\prod_{s\in S} x_s y_s\right) = \left(-\frac{1}{3}\right)^{|S|}.$$

Piecing this altogether, we now have that
$$\mathbb{E}D(f(x_\bullet), g(y_\bullet), h(z_\bullet)) = \left(\widehat{f}(S)\widehat{g}(T) + \widehat{g}(S)\widehat{h}(T) + \widehat{h}(S)\widehat{f}(T)\right)\left(-\frac{1}{3}\right)^{|S|}.$$

Then, we obtain that
$$\begin{aligned} &\mathbb{E}\frac{1}{4}\left(1 + D(f(x_\bullet), g(y_\bullet), h(z_\bullet))\right) \\ =&\frac{1}{4} + \frac{1}{4}\sum_S \left(\widehat{f}(S)\widehat{g}(T) + \widehat{g}(S)\widehat{h}(T) + \widehat{h}(S)\widehat{f}(T)\right)\widehat{f}(S)^2\left(-\frac{1}{3}\right)^{|S|}. \end{aligned}$$

Comparing this with the definition of D gives the desired result.

15B. By Theorem 9.7.6, we may select $e_1, \dots, e_n$ a basis of V and $f_1, \dots, f_m$ a basis of W such that $T(e_i) = f_i$ for $i \le k$ and $T(e_i) = 0$ for $i > k$. Then $T^\vee(f_i^\vee) = e_i^\vee$ for $i \le k$ and $T^\vee(f_i^\vee) = 0$ for $i > k$. All four quantities are above are then equal to k.

15F. First, suppose $T^* = p(T)$. Then $T^*T = p(T)\cdot T = T\cdot p(T) = TT^*$ and we're done.

Conversely, suppose T is diagonalizable in a way compatible with the inner form (OK since V is finite dimensional). Consider the orthonormal basis. Then T consists of eigenvalues on the main diagonals and zeros elsewhere, say
$$T = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}.$$
In that case, we find that for any polynomial q we have
$$q(T) = \begin{pmatrix} q(\lambda_1) & 0 & \dots & 0 \\ 0 & q(\lambda_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & q(\lambda_n) \end{pmatrix}.$$
and
$$T^* = \begin{pmatrix} \overline{\lambda_1} & 0 & \dots & 0 \\ 0 & \overline{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \overline{\lambda_n} \end{pmatrix}.$$

So we simply require a polynomial q such that $q(\lambda_i) = \overline{\lambda_i}$ for every i. Since there are finitely many λ_i, we can construct such a polynomial using Lagrange interpolation.

16E†. https://math.stackexchange.com/a/3012179/229197

17B. Suppose $|G| = 56$ and G is simple. Consider the Sylow 7-subgroups; if there are n_7 of them we assume $n_7 > 1$ (since G is simple) and $n_7 \equiv 1 \pmod 7$, so $n_7 = 8$. That means there are $(7-1) \cdot 8 = 48$ elements of order 7 in G.

But consider the Sylow 2-subgroups. These have 8 elements each, and we conclude therefore that there is at exactly one Sylow 2-subgroup. That subgroup is normal, contradiction.

17C. One example is the group of 3×3 matrices with entries in $\mathbb{F}_3$ that are of the form $\begin{bmatrix} 1 & x & y \\ & 1 & z \\ & & 1 \end{bmatrix}$.

17D. Let G be said group. If G is abelian then all subgroups are normal, and since G is simple, G can't have any subgroups at all. We can clearly find an element of order p, hence G has a subgroup of order p, which can only happen if $n = 1$, $G \cong \mathbb{Z}/p\mathbb{Z}$.

Thus it suffices to show G can't be abelian. For this, we can use the class equation, but let's avoid that and do it directly:

Assume not and let $Z(G) = \{g \in G \mid xg = gx \ \forall x \in G\}$ be the center of the group. Since $Z(G)$ is normal in G, and G is simple, we see $Z(G) = \{1_G\}$. But then let G act on itself by conjugation: $g \cdot x = gxg^{-1}$. This breaks G into a bunch of orbits $\mathcal{O}_0 = \{1_G\}$, $\mathcal{O}_1$, $\mathcal{O}_2$, ..., and since 1_G is the only fixed point by definition, all other orbits have size greater than 1. The Orbit-stabilizer theorem says that each orbit now has size dividing p^n, so they must all have size zero mod p.

But then summing across all orbits (which partition G), we obtain $|G| \equiv 1 \pmod p$, which is a contradiction.

17E. I can't remember what solution I had in mind when I added this problem in. There's a more technical solution outlined at https://github.com/vEnhance/napkin/pull/268#issuecomment-2767676253 but I don't think it's the one I originally had.

18D. Take $G = \mathbb{Z}/3\mathbb{Z} \oplus \mathbb{Z}/9\mathbb{Z} \oplus \mathbb{Z}/9\mathbb{Z} \oplus \mathbb{Z}/9\mathbb{Z} \oplus \dots$ and $H = \mathbb{Z}/9\mathbb{Z} \oplus \mathbb{Z}/9\mathbb{Z} \oplus \mathbb{Z}/9\mathbb{Z} \oplus \mathbb{Z}/9\mathbb{Z} \oplus \dots$. Then there are maps $G \hookrightarrow H$ and $H \hookrightarrow G$, but the groups are not isomorphic since e.g. G has an element $g \in G$ of order 3 for which there's no $g' \in G$ with $g = 3g'$.

18E. Nope! Pick

$$\begin{aligned} A &= \mathbb{Z}[x_1, x_2, \dots] \\ B &= \mathbb{Z}[x_1, x_2, \dots, \varepsilon x_1, \varepsilon x_2, \dots] \\ C &= \mathbb{Z}[x_1, x_2, \dots, \varepsilon]. \end{aligned}$$

where $\varepsilon \neq 0$ but $\varepsilon^2 = 0$. I think the result is true if you add the assumption A is Noetherian.

19D*. The operators are those of the form $T(a) = ab$ for some fixed $b \in A$. One can check these work, since for $c \in A$ we have $T(c \cdot a) = cab = c \cdot T(a)$. To see they are the only ones, note that $T(a) = T(a \cdot 1_A) = a \cdot T(1_A)$ for any $a \in A$.

20C. Pick any $v \in V$, then the subspace spanned by elements $g \cdot v$ for $v \in V$ is G-invariant; this is a finite-dimensional subspace, so it must equal all of V.

21B. $\mathbb{C}_{\text{sign}} \oplus \mathbb{C}^2 \oplus \text{refl}_0 \oplus (\text{refl}_0 \otimes \mathbb{C}_{\text{sign}})$.

21C. First, observe that $|\chi_W(g)| = 1$ for all $g \in G$.

$$\begin{aligned}
\langle \chi_{V \otimes W}, \chi_{V \otimes W} \rangle &= \langle \chi_V \chi_W, \chi_V \chi_W \rangle \\
&= \frac{1}{|G|} \sum_{g \in G} |\chi_V(g)|^2 |\chi_W(g)|^2 \\
&= \frac{1}{|G|} \sum_{g \in G} |\chi_V(g)|^2 \\
&= \langle \chi_V, \chi_V \rangle = 1.
\end{aligned}$$

21D. The table is given by

Q_8	1	-1	$\pm i$	$\pm j$	$\pm k$
$\mathbb{C}_{\text{triv}}$	1	1	1	1	1
$\mathbb{C}_i$	1	1	1	-1	-1
$\mathbb{C}_j$	1	1	-1	1	-1
$\mathbb{C}_k$	1	1	-1	-1	1
$\mathbb{C}^2$	2	-2	0	0	0

The one-dimensional representations (first four rows) follows by considering the homomorphism $Q_8 \to \mathbb{C}^\times$. The last row is two-dimensional and can be recovered by using the orthogonality formula.

23A. By a straightforward computation, we have $|\Psi_-\rangle = -\frac{1}{\sqrt{2}} (|{\rightarrow}\rangle_A \otimes |{\leftarrow}\rangle_B - |{\leftarrow}\rangle_A |{\rightarrow}\rangle_B)$. Now, $|{\rightarrow}\rangle_A \otimes |{\rightarrow}\rangle_B$, $|{\rightarrow}\rangle_A \otimes |{\leftarrow}\rangle_B$ span one eigenspace of $\sigma_x^A \otimes \text{id}_B$, and $|{\leftarrow}\rangle_A \otimes |{\rightarrow}\rangle_B$, $|{\leftarrow}\rangle_A \otimes |{\leftarrow}\rangle_B$ span the other. So this is the same as before: $+1$ gives $|{\leftarrow}\rangle_B$ and -1 gives $|{\leftarrow}\rangle_A$.

24A. To show the Fredkin gate is universal it suffices to reversibly create a CCNOT gate with it. We write the system

$$\begin{aligned}
(z, \neg z, -) &= \text{Fred}(z, 1, 0) \\
(x, a, -) &= \text{Fred}(x, 1, 0) \\
(y, b, -) &= \text{Fred}(y, a, 0) \\
(-, c, -) &= \text{Fred}(b, 0, 1) \\
(-, d, -) &= \text{Fred}(c, z, \neg z).
\end{aligned}$$

Direct computation shows that $d = z + xy \pmod 2$.

24C. Put $|{\leftarrow}\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$. Then we have that U_f sends

$$|x_1\rangle \dots |x_m\rangle |0\rangle - |x_1\rangle \dots |x_m\rangle |1\rangle \xmapsto{U_f} \pm |x_1\rangle \dots |x_m\rangle |0\rangle \mp |x_1\rangle \dots |x_m\rangle |1\rangle$$

the sign being $+$, $-$ exactly when $f(x_1, \dots, x_m) = 1$.

Now, upon inputting $|0\rangle \dots |0\rangle |1\rangle$, we find that $H^{\otimes m+1}$ maps it to

$$2^{-n/2} \sum_{x_1, \dots, x_n} |x_1\rangle \dots |x_n\rangle |{\leftarrow}\rangle .$$

Then the image under U_f is

$$2^{-n/2} \sum_{x_1,\ldots,x_n} (-1)^{f(x_1,\ldots,x_n)} |x_1\rangle \ldots |x_n\rangle |\leftarrow\rangle .$$

We now discard the last qubit, leaving us with

$$2^{-n/2} \sum_{x_1,\ldots,x_n} (-1)^{f(x_1,\ldots,x_n)} |x_1\rangle \ldots |x_n\rangle .$$

Applying $H^{\otimes m}$ to this, we get

$$2^{-n/2} \sum_{x_1,\ldots,x_n} (-1)^{f(x_1,\ldots,x_n)} \cdot \left(2^{-n/2} \sum_{y_1,\ldots,y_n} (-1)^{x_1y_1+\cdots+x_ny_n} |y_1\rangle |y_2\rangle \ldots |y_n\rangle \right)$$

since $H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ while $H|1\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$, so minus signs arise exactly if $x_i = 1$ and $y_i = 1$ simultaneously, hence the term $(-1)^{x_1y_1+\cdots+x_ny_n}$. Swapping the order of summation, we get

$$2^{-n} \sum_{y_1,\ldots,y_n} C(y_1,\ldots,y_n) |y_1\rangle |y_2\rangle \ldots |y_n\rangle$$

where $C_{y_1,\ldots,y_n} = \sum_{x_1,\ldots,x_n} (-1)^{f(x_1,\ldots,x_n)+x_1y_1+\cdots+x_ny_n}$. Now, we finally consider two cases.

- If f is the constant function, then we find that

$$C(y_1,\ldots,y_n) = \begin{cases} \pm 1 & y_1 = \cdots = y_n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

  To see this, note that the result is clear for $y_1 = \cdots = y_n = 0$; otherwise, if WLOG $y_1 = 1$, then the terms for $x_1 = 0$ exactly cancel the terms for $x_1 = 1$, pair by pair. Thus in this state, the measurements all result in $|0\rangle \ldots |0\rangle$.

- On the other hand if f is balanced, we derive that

$$C(0,\ldots,0) = 0.$$

  Thus *no* measurements result in $|0\rangle \ldots |0\rangle$.

In this way, we can tell whether f is balanced or not.

26E. This is an application of Cauchy convergence, since one can show that

$$\left| \sum_{n=M}^{N} (-1)^n a_n \right| \leq a_{\min\{M,N\}}.$$

Indeed, if M and N are even (for simplicity; other cases identical) then

$$\begin{aligned} a_M - a_{M+1} + a_{M+2} - \ldots &= a_M - (a_{M+1} - a_{M+2}) - (a_{M+3} - a_{M+4}) \\ &\quad - \cdots - (a_{N-1} - a_N) \\ &\leq a_M \\ a_M - a_{M+1} + a_{M+2} - \ldots &= a_M - a_{M+1} + (a_{M+2} - a_{M+3}) + (a_{M+4} - a_{M+5}) \\ &\quad + \cdots + (a_{N-2} - a_{N+1}) + a_N \\ &\geq -a_{M+1}. \end{aligned}$$

In this way we see that the sequence of partial sums is Cauchy, hence converges to some limit.

26F. To capture the hypothesis of monotonic and bounded, write $a_n = x_1 + \cdots + x_n$ for some x_i. Then x_2, ... are all the same sign and so $\sum |x_i| = A < \infty$ for some constant A.

We now prove that the partial sums of $\sum a_n b_n$ are a Cauchy sequence. Consider any $\varepsilon > 0$. Let K be such that the tails of b_n starting after K have absolute value less than $\frac{\varepsilon}{A}$. Then for any $N > M \geq K$ we have

$$\begin{aligned}
\left| \sum_{k=M}^{N} a_k b_k \right| &= \left| \sum_{k=M}^{N} \sum_{j=1}^{k} b_k x_j \right| \\
&= \left| \sum_{j=1}^{N} \sum_{k=\max\{j,M\}}^{N} b_k x_j \right| \\
&= \left| \sum_{j=1}^{N} x_j \cdot \sum_{k=\max\{j,M\}}^{N} b_k \right| \\
&\leq \sum_{j=1}^{N} |x_j| \left| \sum_{k=\max\{j,M\}}^{N} b_k \right| \\
&< \sum_{j=1}^{N} |x_j| \cdot \frac{\varepsilon}{A} \\
&< \varepsilon
\end{aligned}$$

as desired.

26G. The answer is $e - 1$.

We begin by noting $x_{n+1} = \log(e^{x_n} - x_n) \geq \log 1 = 0$, owing to $e^t \geq 1 + t$. So $x_n \geq 0$ for all n.

Next notice that

$$x_{n+1} = \log\left(e^{x_n} - x_n\right) < \log e^{x_n} = x_n.$$

So x_1, x_2, ... is strictly decreasing in addition to nonnegative. Thus it must converge to some limit L.

Third, observe that

$$x_n = e^{x_n} - e^{x_{n+1}} \implies x_0 + x_1 + \cdots + x_n = e^{x_0} - e^{x_n} = e - e^{x_n} < e.$$

Since the partial sums are bounded by e, and $x_i \geq 0$, we conclude $L = 0$.

Finally, the limit of the partial sums is then

$$\lim_{n \to \infty} e - e^{x_n} = e - e^0 = e - 1.$$

28G. Write $f(x) = e^{x \log x}$ and then apply the chain rule and product rule:

$$\begin{aligned}
f'(x) &= e^{x \log x} \cdot (x \log x)' \\
&= e^{x \log x} \cdot (1 + \log x) \\
&= x^x \left(1 + \log x\right).
\end{aligned}$$

29E. See https://mathoverflow.net/q/81613 and in particular https://web.archive.org/web/20161009194815/http://mathforum.org/kb/message.jspa?messageID=387148.

31B⋆. Proceed by contradiction, meaning there exists a sequence $z_1, z_2, \ldots \to z$ where $0 = f(z_1) = f(z_2) = \ldots$ all distinct. WLOG set $z = 0$. Look at the Taylor series of f around $z = 0$. Since it isn't uniformly zero by assumption, write it as $a_N z^N + a_{N+1} z^{N+1} + \cdots$, $a_N \neq 0$. But by continuity of $h(z) = a_N + a_{N+1} z + \cdots$ there is some open neighborhood of zero where $h(z) \neq 0$.

31C⋆. Let S be the interior of the points satisfying $f = g$. By definition S is open. By the previous part, S is closed: if $z_i \to z$ and $z_i \in S$, then $f = g$ in some open neighborhood of z, hence $z \in S$. Since S is clopen and nonempty, $S = U$.

31E. Suppose we want to show that there's a point in the image within ε of a given a point $w \in \mathbb{C}$. Look at $\frac{1}{f(z)-w}$ and use Liouville's theorem.

32C. See https://math.stackexchange.com/q/242514/229197, which does it with 2019 replaced by 3.

39A. It is the counting measure.

41A. For each positive integer m, consider what happens when $\varepsilon = 1/m$. Then, by hypothesis, there is a threshold N_m such that the *anomaly set*

$$A_m \coloneqq \left\{ \omega : |X(\omega) - X_n(\omega)| \geq \frac{1}{m} \text{ for some } n > N_m \right\}$$

has measure $\mu(A_m) = 0$. Hence, the countable union $A = \bigcup_{m \geq 1} A_m$ has measure zero too.

So the complement of A has measure 1. For any world $\omega \notin A$, we then have

$$\lim_n |X(\omega) - X_n(\omega)| = 1$$

because when $n > N_m$ that absolute value is always at most $1/m$ (as $\omega \notin A_m$).

41B. https://math.stackexchange.com/a/2201906/229197

55C. If $\alpha \equiv 0 \pmod{\mathfrak{p}}$ it's clear, so assume this isn't the case. Then $\mathcal{O}_K/\mathfrak{p}$ is a finite field with $\mathrm{N}(\mathfrak{p})$ elements. Looking at $(\mathcal{O}_K/\mathfrak{p})^*$, it's a multiplicative group with $\mathrm{N}(\mathfrak{p}) - 1$ elements, so $\alpha^{\mathrm{N}(\mathfrak{p})-1} \equiv 1 \pmod{\mathfrak{p}}$, as desired.

55D†. Suppose it's generated by some elements in K; we can write them as $\frac{\beta_i}{\alpha_i}$ for $\alpha_i, \beta_i \in \mathcal{A}$. Hence

$$J = \left\{ \sum_i \gamma_i \cdot \frac{\beta_i}{\alpha_i} \mid \alpha_i, \beta_i, \gamma_i \in \mathcal{O}_K \right\}.$$

Now "clear denominators". Set $\alpha = \alpha_1 \ldots \alpha_n$, and show that αJ is an integral ideal.

55E. For part (a), note that the $\mathfrak{p}_i$ are prime just because

$$\mathcal{O}_K/\mathfrak{p}_i \cong (\mathbb{Z}[x]/f)/(p, f_i) \cong \mathbb{F}_p[x]/(f_i)$$

is a field, since the f_i are irreducible.

We check (b). Computing the product modulo p yields[2]

$$\prod_{i=1}^{g} (f_i(\theta))^{e_i} \equiv (f(\theta)) \equiv 0 \pmod{p}$$

[2]For example, suppose we want to know that $(3, 1+\sqrt{7})(3, 1-\sqrt{7})$ is contained in (3). We could do the full computation and get $(9, 3+3\sqrt{7}, 3-3\sqrt{7}, 6)$. But if all we care about is that every element is divisible by 3, we could have just taken "mod 3" at the beginning and looked at just $(1+\sqrt{7})(1-\sqrt{7}) = (6)$; all the other products we get will obviously have factors of 3.

so we've shown that $I \subseteq (p)$.

Finally, we prove (c) with a size argument. The idea is that I and (p) really should have the same size; to nail this down we'll use the ideal norm. Since (p) divides I, we can write $(p) = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i'}$ where $e_i' \le e_i$ for each i. Remark $\mathcal{O}_K/(p) \cong \mathbb{Z}/p\mathbb{Z}[x]/(f)$ has size $p^{\deg f}$. Similarly, $\mathcal{O}_K/(\mathfrak{p}_i)$ has degree $p^{\deg f_i}$ for each i. Compute $\mathrm{N}((p))$ using the e_i' now and compare the results.

56F. Let $K = \mathbb{Q}(\sqrt{-5})$. Check that Cl_K has order two using the Minkowski bound; moreover $\Delta_K = 20$. Now note that $\mathcal{O}_K = \mathbb{Z}[\sqrt{-5}]$, and $x^2 + 5$ factors mod p as $(x+k)(x-k)$; hence in $\mathcal{O}_K$ we have $(p) = (p, \sqrt{-5}+k)(p, \sqrt{-5}-k) = \mathfrak{p}_1\mathfrak{p}_2$, say. For $p > 5$ the prime p does not ramify and we have $\mathfrak{p}_1 \neq \mathfrak{p}_2$, since $\Delta_K = 20$.

Then $(p^2) = \mathfrak{p}_1^2 \cdot \mathfrak{p}_2^2$. Because the class group has order two, both $\mathfrak{p}_1^2$ and $\mathfrak{p}_2^2$ are principal, and because $\mathfrak{p}_1 \neq \mathfrak{p}_2$ they are distinct. Thus p^2 is a nontrivial product of two elements of $\mathcal{O}_K$; from this we can extract the desired factorization.

59A*. It's just $\mathbb{Z}/p-1\mathbb{Z}$, since ζ_p needs to get sent to one (any) of the $p-1$ primitive roots of unity.

59D. A similar (but not identical) problem is solved here: https://aops.com/community/c6h149153p842956.

59F. https://www.math.cornell.edu/~kbrown/6310/primitive.pdf

60A†. Recall that the Fibonacci sequence is given by

$$F_n = \frac{\alpha^n - \beta^n}{\alpha - \beta}$$

where $\alpha = \frac{1+\sqrt{5}}{2}$ and $\beta = \frac{1-\sqrt{5}}{2}$ are the two roots of $P(X) \coloneqq X^2 - X - 1$.

Let $p = 127$ and work modulo p. As

$$\left(\frac{5}{p}\right) = \left(\frac{p}{5}\right) = \left(\frac{2}{5}\right) = -1$$

we see 5 is not a quadratic residue mod 127. Thus the polynomial $P(X)$, viewed as a polynomial in $\mathbb{F}_p[X]$, is irreducible (intuitively, α and β are not elements of $\mathbb{F}_p$). Accordingly we will work in the finite field $\mathbb{F}_{p^2}$, which is the $\mathbb{F}_p$-splitting field of $P(X)$. In other words we interpret α and β as elements of $\mathbb{F}_{p^2}$ which do not lie in $\mathbb{F}_p$.

Let $\sigma \colon \mathbb{F}_{p^2} \to \mathbb{F}_{p^2}$ by $t \mapsto t^p$ be the nontrivial element of $\mathrm{Gal}\left(\mathbb{F}_{p^2}/\mathbb{F}_p\right)$; in other words, σ is the non-identity automorphism of $\mathbb{F}_{p^2}$. Since the fixed points of σ are the elements of $\mathbb{F}_p$, this means σ does not fix either root of P; thus we must have

$$\alpha^p = \sigma(\alpha) = \beta$$
$$\beta^p = \sigma(\beta) = \alpha.$$

Now, compute

$$\begin{aligned}
F_p &= \frac{\alpha^p - \beta^p}{\alpha - \beta} = \frac{\beta - \alpha}{\alpha - \beta} = -1. \\
F_{p+1} &= \frac{\alpha^{p+1} - \beta^{p+1}}{\alpha - \beta} = \frac{\alpha\beta - \beta\alpha}{\alpha - \beta} = 0. \\
F_{2p+1} &= \frac{\alpha^{2p+1} - \beta^{2p+1}}{\alpha - \beta} = \frac{\beta^2\alpha - \alpha^2\beta}{\alpha - \beta} = -\alpha\beta = 1. \\
F_{2p+2} &= \frac{\alpha^{2p+2} - \beta^{2p+2}}{\alpha - \beta} = \frac{\beta^2\alpha^2 - \alpha^2\beta^2}{\alpha - \beta} = 0.
\end{aligned}$$

Consequently, the period must divide $2p+2$ but not $p+1$.

We now use for the first time the exact numerical value $p = 127$ to see the period divides $2p + 2 = 256 = 2^8$, but not $p + 1 = 128 = 2^7$. (Previously we only used the fact that $(5/p) = -1$.) Thus the period must be exactly 256.

62A. It is still true that

$$\left(\frac{2}{q}\right) = 1 \iff \sigma_2 \in H \iff 2 \text{ splits in } \mathbb{Z}\left[\tfrac{1}{2}(1+\sqrt{q^*})\right].$$

Now, 2 splits in the ring if and only if $t^2 - t - \frac{1}{4}(1-q^*)$ factors mod 2. This happens if and only if $q^* \equiv 1 \pmod 8$. One can check this is exactly if $q \equiv \pm 1 \pmod 8$, which gives the conclusion.

62C†. Let $K = \text{Gal}(\mathbb{Q}(\zeta_m)/\mathbb{Q})$. One can show that $\text{Gal}(K/\mathbb{Q}) \cong (\mathbb{Z}/m\mathbb{Z})^\times$ exactly as before. In particular, $\text{Gal}(K/\mathbb{Q})$ is abelian and therefore its conjugacy classes are singleton sets; there are $\phi(m)$ of them.

As long as p is sufficiently large, it is unramified and $\sigma_p = \text{Frob}_\mathfrak{p}$ for any $\mathfrak{p}$ above p (as mth roots of unity will be distinct modulo p; differentiate $x^m - 1 \bmod p$ again).

62E. This solution is by David Corwin. By primitive roots, it's the same as the action of $\times 3$ on $\mathbb{Z}/(p-1)\mathbb{Z}$. Let ζ be a $(p-1)$st root of unity.

Consider

$$d = \prod_{0 \le i < j < p-1} (\zeta^i - \zeta^j).$$

This is the square root of the discriminant of the polynomial $X^{p-1} - 1$; in other words $d^2 \in \mathbb{Z}$. In fact, by elementary methods one can compute

$$(-1)^{\binom{p-1}{2}} d^2 = -(p-1)^{p-1}$$

Now take the extension $K = \mathbb{Q}(d)$, noting that

- If $p \equiv 3 \pmod 4$, then $d = (p-1)^{\frac{1}{2}(p-1)}$, so $K = \mathbb{Q}$.
- If $p \equiv 1 \pmod 4$, then $d = i(p-1)^{\frac{1}{2}(p-1)}$, so $K = \mathbb{Q}(i)$.

Either way, in $\mathcal{O}_K$, let $\mathfrak{p}$ be a prime ideal above $(3) \subseteq \mathcal{O}_K$. Let $\sigma = \text{Frob}_\mathfrak{p}$ then be the unique element such that $\sigma(x) = x^3 \pmod{\mathfrak{p}}$ for all x. Then, we observe that

$$\sigma(d) \equiv \prod_{0 \le i < j < p-1} (\zeta^{3i} - \zeta^{3j}) \equiv \begin{cases} +d & \text{if } \pi \text{ is even} \\ -d & \text{if } \pi \text{ is odd} \end{cases} \pmod{\mathfrak{p}}.$$

Now if $K = \mathbb{Q}$, then σ is the identity, thus σ even. Conversely, if $K = \mathbb{Q}(i)$, then 3 does not split, so $\sigma(d) = -d$ (actually σ is complex conjugation) thus π is odd.

Note the condition that $p \equiv 2 \pmod 3$ is used only to guarantee that π is actually a permutation (and thus $d \neq 0$); it does not play any substantial role in the solution.

63A†. Suppose $\mathfrak{f}(L/\mathbb{Q}) \mid m\infty$ for some m. Then by the example from earlier we have the chain

$$P_{\mathbb{Q}}(m\infty) = H(\mathbb{Q}(\zeta)/\mathbb{Q}, m\infty) \subseteq H(L/\mathbb{Q}, m) \subseteq I_{\mathbb{Q}}(m\infty).$$

So by inclusion reversal we're done.

63B†. Apply the Takagi existence theorem with $\mathfrak{m} = 1$ to obtain an unramified extension E/K such that $H(E/K, 1) = P_K(1)$. We claim this works:

- To see it is maximal by inclusion, note that any other extension M/K with this property has conductor 1 (no primes divide the conductor), and then we have $P_K(1) = H(E/K, 1) \subseteq H(M/K, 1) \subseteq I_K(1)$, so inclusion reversal gives $M \subseteq E$.
- We have $\mathrm{Gal}(L/K) \cong I_K(1)/P_K(1) = C_K(1)$ the class group.
- The isomorphism in the previous part is given by the Artin symbol. So $\mathfrak{p}$ splits completely if and only if $\left(\frac{L/K}{\mathfrak{p}}\right) = \mathrm{id}$ if and only if $\mathfrak{p}$ is principal (trivial in $C_K(1)$).

This completes the proof.

68A. The main observation is that in $\mathcal{A} \times \mathbf{2}$, you have the arrows in $\mathcal{A}$ (of the form $(f, \mathrm{id}_{\mathbf{2}})$), and then the arrows crossing the two copies of $\mathcal{A}$ (of the form $(\mathrm{id}_A, 0 \leq 1)$). But there are some more arrows $(f, 0 \leq 1)$: nonetheless, they can be thought of as compositions

$$(f, 0 \leq 1) = (f, \mathrm{id}_{\mathbf{2}}) \circ (\mathrm{id}_A, 0 \leq 1) = (\mathrm{id}_A, 0 \leq 1) \circ (f, \mathrm{id}_{\mathbf{2}}).$$

Now to specify a functor $\alpha \colon \mathcal{A} \times \mathbf{2} \to \mathcal{B}$, we only have to specify where each of these two more basic things goes. The conditions on α already tells us that $(f, \mathrm{id}_{\mathbf{2}})$ should be mapped to $F(f)$ or $G(f)$ (depending on whether the arrow above is in $\mathcal{A} \times \{0\}$ or $\mathcal{A} \times \{1\}$), and specifying the arrow $(\mathrm{id}_A, 0 \leq 1)$ amounts to specifying the Ath component. Where does naturality come in?

The above discussion transfers to products of categories in general: you really only have to think about (f, id) and (id, g) arrows to get the general arrow $(f, g) = (f, \mathrm{id}) \circ (\mathrm{id}, g) = (\mathrm{id}, g) \circ (f, \mathrm{id})$.

70A. Let $c \in C$ with $\gamma(c) = 0$. We show $c = 0$. This proceeds in a diagram chase:

- Note that $0 = r'(\gamma(c)) = \delta(r(c))$, and since δ is injective, it follows that $r(c) = 0$.
- Since the top row is exact, it follows $c = q(b)$ for some $b \in B$.
- Then $q'(\beta(b)) = 0$, so if we let $b' = \beta(b)$, then $b' \in \ker(q')$. As the bottom row is exact, there exists a' with $p'(a') = b'$.
- Since α is injective, there is $a \in A$ with $\alpha(a) = a'$.
- Since β is injective, it follows that $p(a) = b$.
- Since the top row is exact, and b is in the image of p, it follows that $0 = q(b) = c$ as needed.

71A. Applying the functor H_{n-1} we get that the composition $\mathbb{Z} \to 0 \to \mathbb{Z}$ is the identity which is clearly not possible.

72B. The answer is $\widetilde{H}_{n-1}(X) \cong \mathbb{Z}^{\oplus p}$, with all other groups vanishing. For $p = 1$, $\mathbb{R}^n - \{*\} \cong S^{n-1}$ so we're done. For all other p, draw a hyperplane dividing the p points into two halves with a points on one side and b points on the other (so $a + b = p$). Set U and V and use induction.

Alternatively, let U be the desired space and let V be the union of p disjoint balls, one around every point. Then $U \cup V = \mathbb{R}^n$ has all reduced homology groups trivial. From the Mayer-Vietoris sequence we can read $\widetilde{H}_k(U \cap V) \cong \widetilde{H}_k(U) \cap \widetilde{H}_k(V)$. Then $U \cap V$ is p punctured balls, which are each the same as S^{n-1}. One can read the conclusion from here.

72C*. It is $\mathbb{Z}$ for $k = n$ and 0 otherwise.

72F*. Use the short exact sequence

$$0 \to C_\bullet(B, A) \to C_\bullet(X, A) \to C_\bullet(X, B) \to 0$$

of chain complexes.

73B. We have an exact sequence

$$\underbrace{\widetilde{H}_1(\mathbb{R})}_{=0} \to \widetilde{H}_1(\mathbb{R}, \mathbb{Q}) \to \widetilde{H}_0(\mathbb{Q}) \to \underbrace{\widetilde{H}_0(\mathbb{R})}_{=0}.$$

Now, since $\mathbb{Q}$ is path-disconnected (i.e. no two of its points are path-connected) it follows that $\widetilde{H}_0(\mathbb{Q})$ consists of countably infinitely many copies of $\mathbb{Z}$.

73E. This is shown in detail in Section 2.B of Hatcher.

74D. For concreteness, let's just look at the homology at $H_2(X^2, X^1)$ and show it's isomorphic to $H_2(X)$. According to the diagram

$$\begin{aligned}
H_2(X) &\cong H_2(X^3) \\
&\cong H_2(X^2) / \ker \left[H_2(X^2) \twoheadrightarrow H_2(X^3) \right] \\
&\cong H_2(X^2) / \operatorname{im} \partial_3 \\
&\cong \operatorname{im} \left[H_2(X^2) \hookrightarrow H_2(X^2, X^1) \right] / \operatorname{im} \partial_3 \\
&\cong \ker(\partial_2) / \operatorname{im} \partial_3 \\
&\cong \ker d_2 / \operatorname{im} d_3.
\end{aligned}$$

76D. See [Ma13a, Example 3.3.14, pages 68-69].

77B. If $V = \mathcal{V}(I)$ with $I = (f_1, \ldots, f_m)$ (as usual there are finitely many polynomials since $\mathbb{R}[x_1, \ldots, x_n]$ is Noetherian) then we can take $f = f_1^2 + \cdots + f_m^2$.

77C. Let I be an ideal, and let $\mathfrak{m}$ be a maximal ideal contained in it. (If you are worried about the existence of $\mathfrak{m}$, it follows from Krull's Theorem, Problem 5E*). Then $\mathfrak{m} = (x_1 - a_1, \ldots, x_n - a_n)$ by Weak Nullstellensatz. Consequently, $(a_1, \ldots, a_n)$ is the unique point of $\mathcal{V}(\mathfrak{m})$, and hence this point is also in $\mathcal{V}(I)$.

77D. The point is to check that if f vanishes on all of $\mathcal{V}(I)$, then $f \in \sqrt{I}$.

Take a set of generators $f_1, \dots, f_m$, in the original ring $\mathbb{C}[x_1, \dots, x_n]$; we may assume it's finite by the Hilbert basis theorem.

We're going to do a trick now: consider $S = \mathbb{C}[x_1, \dots, x_n, x_{n+1}]$ instead. Consider the ideal $I' \subseteq S$ in the bigger ring generated by $\{f_1, \dots, f_m\}$ and the polynomial $x_{n+1}f - 1$. The point of the last guy is that its zero locus does not touch our copy $x_{n+1} = 0$ of $\mathbb{A}^n$ nor any point in the "projection" of f through $\mathbb{A}^{n+1}$ (one can think of this as $\mathcal{V}(I)$ in the smaller ring direct multiplied with $\mathbb{C}$). Thus $\mathcal{V}(I') = \varnothing$, and by the weak Nullstellensatz we in fact have $I' = \mathbb{C}[x_1, \dots, x_{n+1}]$. So

$$1 = g_1 f_1 + \cdots + g_m f_m + g_{m+1}\left(x_{n+1}f - 1\right).$$

Now the hack: **replace every instance of x_{n+1} by $\frac{1}{f}$**, and then clear all denominators. Thus for some large enough integer N we can get

$$f^N = f^N(g_1 f_1 + \cdots + g_m f_m)$$

which eliminates any fractional powers of f in the right-hand side. It follows that $f^N \in I$.

80A. From the exactness, $h_I(d) = h_I(d-k) + h_{I+(f)}(d)$, and it follows that

$$\chi_{I+(f)}(d) = \chi_I(d) - \chi_I(d-k).$$

Let $m = \dim \mathcal{V}_{\text{pr}}(I) \geq 1$. Now $\dim \mathcal{V}_{\text{pr}}(I + (f)) = m - 1$, so and $c_{\text{new}} = \deg I + (f)$ then we have

$$\frac{\deg(I+(f))d^{m-1} + \dots}{(m-1)!} = \frac{1}{m!}\left(\deg I(d^m - (d-k)^m) + \text{lower order terms}\right)$$

from which we read off

$$\deg(I + (f)) = \frac{(m-1)!}{m!} \cdot k\binom{m}{1} \deg I = k \deg I$$

as needed.

80B. In complex numbers with ABC the unit circle, it is equivalent to solving the two cubic equations

$$\begin{aligned}(p-a)(p-b)(p-c) &= (abc)^2(q-1/a)(q-1/b)(q-1/c) \\ 0 &= \prod_{\text{cyc}}(p+c-b-bcq) + \prod_{\text{cyc}}(p+b-c-bcq)\end{aligned}$$

in p and $q = \overline{p}$. Viewing this as two cubic curves in $(p, q) \in \mathbb{C}^2$, by Bézout's theorem it follows there are at most nine solutions (unless both curves are not irreducible, but one can check the first one cannot be factored). Moreover it is easy to name nine solutions (for ABC scalene): the three vertices, the three excenters, and I, O, H. Hence the answer is just those three triangle centers I, O and H.

81C. If they were isomorphic, we would have $\mathcal{O}_V(V) \cong \mathcal{O}_W(W)$. For irreducible projective varieties, $\mathcal{O}_W(W) \cong \mathbb{C}$, while for affine varieties $\mathcal{O}_V(V) \cong \mathbb{C}[V]$. Thus we conclude V must be a single point.

81D. Assume for contradiction there is an affine variety V and an isomorphism

$$f \colon X \to V.$$

Then taking the pullback we get a ring isomorphism

$$f^\sharp \colon \mathcal{O}_V(V) \to \mathcal{O}_X(X) = \mathbb{C}[x,y].$$

Now let $\mathcal{O}_V(V) = \mathbb{C}[a,b]$ where $f^\sharp(a) = x$, $f^\sharp(b) = y$. In particular, we actually have to have $V \cong \mathbb{A}^2$.

Now in the *affine* variety V we can take $\mathcal{V}(a)$ and $\mathcal{V}(b)$; these have nonempty intersection since (a,b) is a maximal ideal in $\mathcal{O}_V(V)$. Call this point q, and let p be a point with $f(p) = q$.

Then

$$0 = a(q) = (f^\sharp a)(p) = x(p)$$

and so $p \in \mathcal{V}(x) \subseteq X$. Similarly, $p \in \mathcal{V}(y) \subseteq X$, but this is a contradiction since $\mathcal{V}(x,y) = \varnothing$.

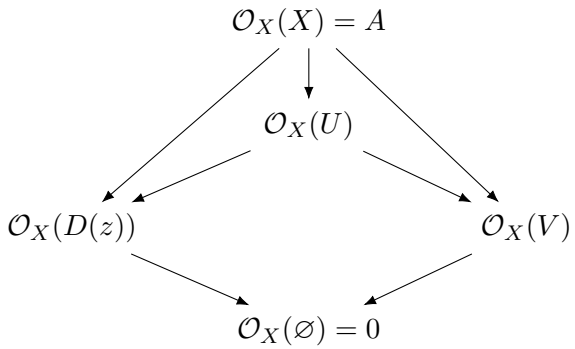
82A. Because the stalks are preserved by sheafification, there is essentially nothing to prove: both sides correspond to sequences of compatible $\mathscr{F}$-germs over U.

83A. One should get $A[1/60] = \mathbb{Z}/7\mathbb{Z}$.

83B. If and only if S has no zero divisors.

83D. Take $A = \mathbb{C}[x,y]/(xy)$.

85B. Let $V = D(x) \cup D(y) \subset U$ denote the punctured plane, so its complement $D(z)$ looks like a punctured line. Then $V \cap D(z) = \varnothing$ and the following diagram of restriction maps commutes



By sheaf axioms we should actually have

$$\mathcal{O}_X(U) = \mathcal{O}_X(D(z)) \times \mathcal{O}_X(V).$$

We have $\mathcal{O}_X(D(z)) = A_z = k[x,y,z,z^{-1}]/(xz,yz) \cong k[z,z^{-1}]$. On the other hand $\mathcal{O}_X(V) = k[x,y]$ as shown in §4.4.1 of Vakil. So

$$\mathcal{O}_X(U) = k[x,y] \times k[z,z^{-1}].$$

87A. Since $\mathbb{Z}$ is the initial object of CRing, it follows $\operatorname{Spec}\mathbb{Z}$ is the final object of AffSch. $\mathfrak{p}$ gets sent to the characteristic of the field $\mathcal{O}_{X,\mathfrak{p}}/\mathfrak{m}_{X,\mathfrak{p}}$.

88A. Let $\varepsilon = \pi - 3.141592653 < 10^{-9}$. Then

$$\frac{22}{7} = f(\pi) = f(3.141592653) + f(\varepsilon) = 3.141592653 + f(\varepsilon).$$

Therefore,

$$f(\varepsilon) = \frac{22}{7} - 3.141592653 = \frac{22 - 21.991148571}{7} > \frac{0.008}{7} > 10^{-3}.$$

So

$$f(10^8\varepsilon) = 10^8 f(\varepsilon) > 10^5 > 9000$$

and $10^8\varepsilon < 1$, as needed.

88B. Every statement is true.

The first statement follows by simply extending f via

$$x \mapsto \begin{cases} f(x) & x > 0 \\ 0 & x = 0 \\ -f(-x) & x < 0. \end{cases}$$

The second statement is true for any additive function $\mathbb{R} \to \mathbb{R}$. Indeed, $f(0) = f(0) + f(0) \implies f(0) = 0$, and odd follows.

The third and fourth statement follow from https://en.wikipedia.org/wiki/Cauchy%27s_functional_equation#Properties_of_nonlinear_solutions_over_the_real_numbers.

The fifth statement is kind of stupid. If f was surjective, there should exist $a > 0$ such that $f(a) = 0$. But then $f(2a) = f(a) + f(a) = 0$, so f is not injective.

For the rest, fix a Hamel basis

$$E = \{e_\alpha \mid \alpha \in S \coloneqq \{0, 1, 2, \dots\} \dots\}.$$

Here S is an uncountable set of ordinals. WLOG, $e_0 = 1$ and $e_\alpha > 0$ for all $\alpha \in S$. Then f is uniquely determined by the value of $f(e_\alpha)$ for each $\alpha \in S$.

- For the sixth statement, let $f(e_0) = e_1$, $f(e_1) = e_0$, and $f(e_\alpha) = e_\alpha$ for all other $\alpha \geq 2$.
- The seventh statement is the most complicated. Since S is infinite, it's possible to construct a 2-to-1 map $\psi \colon S \to S$, meaning every element of the codomain is the image of exactly two elements in the domain. Then if $\psi(\alpha) = \psi(\beta) = \gamma$ for $\alpha \neq \beta$, set $f(e_\alpha) = e_\gamma$, $f(e_\beta) = -e_\gamma$.
- For the eighth statement, let $f(e_\alpha) = 1$ for every $\alpha \in S$.
- For the ninth statement, let $f(e_\alpha) = \sqrt{2}$ for every $\alpha \in S$.

89E. Define an equivalence relation equating two hat configurations if they differ in only finitely many places. Now for each equivalence class, everyone pre-agrees on a particular representative. Finally, note that a person can determine which equiv class the group is in even without their own hat color. Hence they unanimously select the same representative, QED.

92C†. For a sentence ϕ let

$$f_\phi \colon \kappa \to \kappa$$

send α to the least $\beta < \kappa$ such that for all $\vec{b} \in V_\alpha$, if there exists $a \in V_\kappa$ such that $V_\kappa \vDash \phi[a, \vec{b}]$ then $\exists a \in V_\beta$ such that $V_\kappa \vDash \phi[a, \vec{b}]$.

We claim this is well-defined. There are only $|V_\alpha|^n$ many possible choices of $\vec{b}$, and in particular there are fewer than κ of these (since we know that $|V_\alpha| < \kappa$; compare Problem 91C*). Otherwise, we can construct a cofinal map from $|V_\alpha^n|$ into κ by mapping each vector $\vec{b}$ into a β for which the proposition fails. And that's impossible since κ is regular!

In other words, what we've done is fix ϕ and then use Tarski-Vaught on all the $\vec{b} \in V_\alpha^n$. Now let $g \colon \kappa \to \kappa$ be defined by

$$\alpha \mapsto \sup f_\phi(\alpha).$$

Since κ is regular and there are only countably many formulas, $g(\alpha)$ is well-defined.

Check that if α has the property that g maps α into itself (in other words, α is closed under g), then by the Tarski-Vaught test, we have $V_\alpha \prec V_\kappa$.

So it suffices to show there are arbitrarily large $\alpha < \kappa$ which are closed under g. Fix α_0. Let $\alpha_1 = g(\alpha_0)$, et cetera and define

$$\alpha = \sup_{n<\omega} \alpha_n.$$

This α is closed under g, and by making α_0 arbitrarily large we can make α as large as we like.

93B. Since M is countable, there are only countably many dense sets (they live in M!), say

$$D_1, D_2, \ldots, D_n, \ldots \in M.$$

Using Choice, let $p_1 \in D_1$, and then let $p_2 \leq p_1$ such that $p_2 \in D_2$ (this is possible since D_2 is dense), and so on. In this way we can inductively exhibit a chain

$$p_1 \geq p_2 \geq p_3 \geq \ldots$$

with $p_i \in D_i$ for every i.

Hence, we want to generate a filter from the $\{p_i\}$. Just take the upwards closure – let G be the set of $q \in \mathbb{P}$ such that $q \geq p_n$ for some n. By construction, G is a filter (this is actually trivial). Moreover, G intersects all the dense sets by construction.

94A. It suffices to show that $\mathbb{P}$ preserves regularity greater than or equal to κ. Consider $\lambda > \kappa$ which is regular in M, and suppose for contradiction that λ is not regular in $M[G]$. That's the same as saying that there is a function $f \in M[G]$, $f \colon \overline{\lambda} \to \lambda$ cofinal, with $\overline{\lambda} < \lambda$. Then by the Possible Values Argument, there exists a function $F \in M$ from $\overline{\lambda} \to \mathcal{P}(\lambda)$ such that $f(\alpha) \in F(\alpha)$ and $|F(\alpha)|^M < \kappa$ for every α.

Now we work in M again. Note for each $\alpha \in \overline{\lambda}$, $F(\alpha)$ is bounded in λ since λ is regular in M and greater than $|F(\alpha)|$. Now look at the function $\overline{\lambda} \to \lambda$ in M by just

$$\alpha \mapsto \cup F(\alpha) < \lambda.$$

This is cofinal in M, contradiction.

D Glossary of notations

§D.1 General

- $\forall$: for all
- $\exists$: there exists
- $\operatorname{sign}(\sigma)$: sign of permutation σ
- $X \implies Y$: X implies Y

§D.2 Functions and sets

- $f^{\text{img}}(S)$ is the image of $f\colon X \to Y$ for $S \subseteq X$.
- $f^{-1}(y)$ is the inverse for $f\colon X \to Y$ when $y \in Y$.
- $f^{\text{pre}}(T)$ is the pre-image for $f\colon X \to Y$ when $T \subseteq Y$.
- $f\restriction_S$ is the restriction of $f\colon X \to Y$ to $S \subseteq X$.
- f^n is the function f applied n times

Below are some common sets. These may also be thought of as groups, rings, fields etc. in the obvious way.

- $\mathbb{C}$: set of complex numbers
- $\mathbb{R}$: set of real numbers
- $\mathbb{N}$: set of positive integers
- $\mathbb{Q}$: set of rational numbers
- $\mathbb{Z}$: set of integers
- $\varnothing$: empty set

Some common notation with sets:

- $A \subset B$: A is any subset of B
- $A \subseteq B$: A is any subset of B
- $A \subsetneq B$: A is a *proper* subset of B
- $S \times T$: Cartesian product of sets S and T
- $S \setminus T$: difference of sets S and T
- $S \cup T$: set union of S and T
- $S \cap T$: set intersection of S and T

- $S \sqcup T$: disjoint union of S and T
- $|S|$: cardinality of S
- $S/\sim$: if $\sim$ is an equivalence relation on S, this is the set of equivalence classes
- $x + S$: denotes the set $\{x + s \mid s \in S\}$.
- xS: denotes the set $\{xs \mid s \in S\}$.

§D.3 Abstract and linear algebra

Some common groups/rings/fields:

- $\mathbb{Z}/n\mathbb{Z}$: cyclic group of order n
- $(\mathbb{Z}/n\mathbb{Z})^\times$: set of units of $\mathbb{Z}/n\mathbb{Z}$.
- S_n: symmetric group on $\{1, \dots, n\}$
- D_{2n}: dihedral group of order $2n$.
- 0, 1: trivial group (depending on context)
- $\mathbb{F}_p$: integers modulo p

Notation with groups:

- 1_G: identity element of the group G
- $N \trianglelefteq G$: subgroup N is normal in G.
- G/N: quotient group of G by the normal subgroup N
- $Z(G)$: center of group G
- $N_G(H)$: normalizer of the subgroup H of G
- $G \times H$: product group of G and H
- $G \oplus H$: also product group, but often used when G and H are abelian (and hence we can think of them as $\mathbb{Z}$-modules)
- $\text{Stab}_G(x)$: the stabilizer of $x \in X$, if X is acted on by G
- $\text{FixPt}\, g$, the set of fixed points by $g \in G$ (under a group action)

Notation with rings:

- R/I: quotient of ring R by ideal I
- $(a_1, \dots, a_n)$: ideal generated by the a_i
- $R^\times$: the group of units of R
- $R[x_1, \dots, x_n]$: polynomial ring in x_i, or ring obtained by adjoining the x_i to R
- $F(x_1, \dots, x_n)$: field obtained by adjoining x_i to F
- R^d: dth graded part of a graded (pseudo)ring R

Linear algebra:

- id: the identity matrix
- $V \oplus W$: direct sum
- $V^{\oplus n}$: direct sum of V, n times
- $V \otimes W$: tensor product
- $V^{\otimes n}$: tensor product of V, n times
- $V^\vee$: dual space
- $T^\vee$: dual map (for T a vector space)
- $T^\dagger$: conjugate transpose (for T a vector space)
- $\langle -, - \rangle$: a bilinear form
- $\mathrm{Mat}(V)$: endomorphisms of V, i.e. $\mathrm{Hom}_k(V, V)$
- $\mathbf{e}_1, \ldots, \mathbf{e}_n$: the "standard basis" of $k^{\oplus n}$

§D.4 Quantum computation

- $|\psi\rangle$: a vector in some vector space H
- $\langle\psi|$: a vector in some vector space $H^\vee$, dual to $|\psi\rangle$.
- $\langle\phi|\psi\rangle$: evaluation of an element $\langle\phi| \in H^\vee$ at $|\phi\rangle \in H$.
- $|\uparrow\rangle$, $|\downarrow\rangle$: spin z-up, spin z-down
- $|\rightarrow\rangle$, $|\leftarrow\rangle$: spin x-up, spin x-down
- $|\otimes\rangle$, $|\odot\rangle$: spin y-up, spin y-down

§D.5 Topology and real/complex analysis

Common topological spaces:

- S^1: the unit circle
- S^n: surface of an n-sphere (in $\mathbb{R}^{n+1}$)
- D^{n+1}: closed $n+1$ dimensional ball (in $\mathbb{R}^{n+1}$)
- $\mathbb{RP}^n$: real projective n-space
- $\mathbb{CP}^n$: complex projective n-space

Some topological notation:

- ∂Y: boundary of a set Y (in some topological space)
- X/S: quotient topology of X by $S \subseteq X$
- $X \times Y$: product topology of spaces X and Y

- $X \amalg Y$: disjoint union of spaces X and Y
- $X \vee Y$: wedge product of (pointed) spaces X and Y

Real analysis (calculus 101):

- $\liminf$: limit infimum
- $\limsup$: limit supremum
- $\inf$: infimum
- $\sup$: supremum
- $\mathbb{Z}_p$: p-adic integers
- $\mathbb{Q}_p$: p-adic numbers
- f': derivative of f
- $\int_a^b f(x)\,dx$: Riemann integral of f on $[a,b]$

Complex analysis:

- $\int_\alpha f\,dz$: contour integral of f along path α
- $\operatorname{Res}(f;p)$: the residue of a meromorphic function f at point p
- $\mathbf{I}(\gamma,p)$: winding number of γ around p.

§D.6 Measure theory and probability

- $\mathscr{A}^{\text{cm}}$: the σ-algebra of Caratheory-measurable sets
- $\mathscr{B}(X)$: the Borel space for X
- μ^{cm}: the induced measure on $\mathscr{A}^{\text{cm}}$.
- λ: Lebesgue measure
- $\mathbf{1}_A$: the indicator function for A
- $\int_\Omega f\,d\mu$: the Lebesgue integral of f
- $\lim_{n\to\infty} f_n$: pointwise limit of f_n
- $\widehat{G}$: Pontryagin dual for G

§D.7 Algebraic topology

- $\alpha \simeq \beta$: for paths, this indicates path homotopy
- $*$: path concatenation
- $\pi_1(X) = \pi_1(X, x_0)$: the fundamental group of (pointed) space X
- $\pi_n(X) = \pi_n(X, x_0)$: the nth homotopy group of (pointed) space X
- $f_\sharp$: the induced map $\pi_1(X) \to \pi_1(Y)$ of $f\colon X \to Y$

- Δ^n: the standard n-simplex
- $\partial\sigma$: the boundary of a singular n-simplex σ
- $H_n(A_\bullet)$: the nth homology group of the chain complex $A_\bullet$
- $H_n(X)$: the nth homology group of a space X
- $\widetilde{H}_n(X)$: the nth reduced homology group of X
- $H_n(X, A)$: the nth relative homology group of X and $A \subseteq X$
- f_*: the induced map on $H_n(A_\bullet) \to H_n(B_\bullet)$ of $f\colon A_\bullet \to B_\bullet$, or $H_n(X) \to H_n(Y)$ for $f\colon X \to Y$
- $\chi(X)$: Euler characteristic of a space X
- $H^n(A^\bullet)$: the nth cohomology group of a cochain complex $A^\bullet$
- $H^n(A_\bullet; G)$: the nth cohomology group of the cochain complex obtained by applying $\mathrm{Hom}(-, G)$ to $A_\bullet$
- $H^n(X; G)$: the nth cohomology group/ring of X with G-coefficients
- $\widetilde{H}^n(X; G)$: the nth reduced cohomology group/ring of X with G-coefficients
- $H^n(X, A; G)$: the nth relative cohomology group/ring of X and $A \subset X$ with G-coefficients
- $f^\sharp$: the induced map on $H^n(A^\bullet) \to H^n(B^\bullet)$ of $f\colon A^\bullet \to B^\bullet$, or $H^n(X) \to H^n(Y)$ for $f\colon X \to Y$
- $\mathrm{Ext}(-,-)$: the Ext functor
- $\phi \smile \psi$: cup product of cochains ϕ and ψ

§D.8 Category theory

Some common categories (in alphabetical order):

- Grp: category of groups
- CRing: category of commutative rings
- Top: category of topological spaces
- Top_*: category of pointed topological spaces
- Vect_k: category of k-vector spaces
- FDVect_k: category of finite-dimensional vector spaces
- Set: category of sets
- hTop: category of topological spaces, whose morphisms are homotopy classes of maps
- hTop_*: pointed version of hTop

- hPairTop: category of pairs (X, A) with morphisms being pair-homotopy equivalence classes
- OpenSets(X): the category of open sets of X, as a poset

Operations with categories:

- obj $\mathcal{A}$: objects of the category $\mathcal{A}$
- $\mathcal{A}^{\text{op}}$: opposite category
- $\mathcal{A} \times \mathcal{B}$: product category
- $[\mathcal{A}, \mathcal{B}]$: category of functors from $\mathcal{A}$ to $\mathcal{B}$
- $\ker f$: $\operatorname{Ker} f \to B$: for $f\colon A \to B$, categorical kernel
- $\operatorname{coker} f\colon A \to \operatorname{Coker} f$: for $f\colon A \to B$, categorical cokernel
- $\operatorname{im} f\colon A \to \operatorname{Im} f$: for $f\colon A \to B$, categorical image

§D.9 Differential geometry

- Df: total derivative of f
- $(Df)_p$: total derivate of f at point p
- $\frac{\partial f}{\partial e_i}$: i^{th} partial derivative
- α_p: evaluating a k-form α at p
- $\int_c \alpha$: integration of the differential form α over a cell c
- $d\alpha$: exterior derivative of a k-form α
- $\phi^*\alpha$: pullback of k-form α by ϕ

§D.10 Algebraic number theory

- $\overline{\mathbb{Q}}$: ring of algebraic numbers
- $\overline{\mathbb{Z}}$: ring of algebraic integers
- $\overline{F}$: algebraic closure of a field F
- $\mathrm{N}_{K/\mathbb{Q}}(\alpha)$: the norm of α in extension $K/\mathbb{Q}$
- $\mathrm{Tr}_{K/\mathbb{Q}}(\alpha)$: the trace of α in extension $K/\mathbb{Q}$
- $\mathcal{O}_K$: ring of integers in K
- $\mathfrak{a} + \mathfrak{b}$: sum of two ideals $\mathfrak{a}$ and $\mathfrak{b}$
- $\mathfrak{ab}$: ideal generated by products of elements in ideals $\mathfrak{a}$ and $\mathfrak{b}$
- $\mathfrak{a} \mid \mathfrak{b}$: ideal $\mathfrak{a}$ divides ideal $\mathfrak{b}$
- $\mathfrak{a}^{-1}$: the inverse of $\mathfrak{a}$ in the ideal group

- $\mathrm{N}(I)$: ideal norm
- Cl_K: class group of K
- Δ_K: discriminant of number field K
- $\mu(\mathcal{O}_K)$: set of roots of unity contained in $\mathcal{O}_K$
- $[K : F]$: degree of a field extension
- $\mathrm{Aut}(K/F)$: set of field automorphisms of K fixing F
- $\mathrm{Gal}(K/F)$: Galois group of K/F
- $D_{\mathfrak{p}}$: decomposition group of prime ideal $\mathfrak{p}$
- $I_{\mathfrak{p}}$: inertia group of prime ideal $\mathfrak{p}$
- $\mathrm{Frob}_{\mathfrak{p}}$: Frobenius element of $\mathfrak{p}$ (element of $\mathrm{Gal}(K/\mathbb{Q})$)
- $P_K(\mathfrak{m})$: ray of principal ideals of a modulus $\mathfrak{m}$
- $I_K(\mathfrak{m})$: fractional ideals of a modulus $\mathfrak{m}$
- $C_K(\mathfrak{m})$: ray class group of a modulus $\mathfrak{m}$
- $\left(\frac{L/K}{\bullet}\right)$: the Artin symbol
- $\mathrm{Ram}(L/K)$: primes of K ramifying in L
- $\mathfrak{f}(L/K)$: the conductor of L/K

§D.11 Representation theory

- $k[G]$: group algebra
- $V \oplus W$: direct sum of representations $V = (V, \rho_V)$ and $W = (W, \rho_W)$ of an algebra A
- $V^\vee$: dual representation of a representation $V = (V, \rho_V)$
- $\mathrm{Reg}(A)$: regular representation of an algebra A
- $\mathrm{Hom}_{\mathrm{rep}}(V, W)$: algebra of morphisms $V \to W$ of representations
- χ_V: the character $A \to k$ attached to an A-representation V
- $\mathrm{Classes}(G)$: set of conjugacy classes of G
- $\mathrm{Fun}_{\mathrm{class}}(G)$: the complex vector space of functions $\mathrm{Classes}(G) \to \mathbb{C}$
- $V \otimes W$: tensor product of representations $V = (V, \rho_V)$ and $W = (W, \rho_W)$ of a *group* G (rather than an algebra)
- $\mathbb{C}_{\mathrm{triv}}$: the trivial representation
- $\mathbb{C}_{\mathrm{sign}}$: the sign representation

§D.12 Algebraic geometry

- $\mathcal{V}(-)$: vanishing locus of a set or ideal
- $\mathbb{A}^n$: n-dimensional (complex) affine space
- $\sqrt{I}$: radical of an ideal I
- $\mathbb{C}[V]$: coordinate ring of an affine variety V
- $\mathcal{O}_V(U)$: ring of rational functions on U
- $D(f)$: distinguished open set
- $\mathbb{CP}^n$: complex projective n-space (ambient space for projective varieties)
- $(x_0 : \cdots : x_n)$: coordinates of projective space
- U_i: standard affine charts
- $\mathcal{V}_{\text{pr}}(-)$: projective vanishing locus.
- h_I, h_V: Hilbert function of an ideal I or projective variety V
- $\pi^\sharp$ or $\pi^\sharp_U$: the pullback $\mathcal{O}_Y \to \mathcal{O}_X(\pi^{\text{pre}}(U))$ obtained from $\pi\colon X \to Y$
- $\mathscr{F}_p$: the stalk of a (pre-)sheaf $\mathscr{F}$ at a point p
- $[s]_p$: the germ of $s \in \mathscr{F}(U)$ at the point p
- $\mathcal{O}_{X,p}$: shorthand for $(\mathcal{O}_X)_p$.
- $\mathscr{F}^{\text{sh}}$: sheafification of pre-sheaf $\mathscr{F}$
- $\alpha_p\colon \mathscr{F}_p \to \mathscr{G}_p$: morphism of stalks obtained from $\alpha\colon \mathscr{F} \to \mathscr{G}$
- $\mathfrak{m}_{X,p}$: the maximal ideal of $\mathcal{O}_{X,p}$
- $\operatorname{Spec} A$: the spectrum of a ring A
- $S^{-1}A$: localization of ring A at a set S
- $A[1/f]$: localization of ring A away from element f
- $A_{\mathfrak{p}}$: localization of ring A at prime ideal $\mathfrak{p}$
- $f(\mathfrak{p})$: the value of f at $\mathfrak{p}$, i.e. $f \pmod{\mathfrak{p}}$
- $\kappa(\mathfrak{p})$: the residue field of $\operatorname{Spec} A$ at the element $\mathfrak{p}$.
- $\pi^\sharp_{\mathfrak{p}}$: the induced map of stalks in $\pi^\sharp$.

§D.13 Set theory

- ZFC: standard theory of ZFC
- ZFC^+: standard theory of ZFC, plus the sentence "there exists a strongly inaccessible cardinal"
- 2^S or $\mathcal{P}(S)$: power set of S
- $A \wedge B$: A and B
- $A \vee B$: A or B
- $\neg A$: not A
- V: class of all sets (von Neumann universe)
- ω: the first infinite ordinal, also the set of nonnegative integers
- V_α: level of the von Neumann universe
- On: class of ordinals
- $\bigcup A$: the union of elements inside A
- $A \approx B$: sets A and B are equinumerous
- $\aleph_\alpha$: the aleph numbers
- $\operatorname{cof} \lambda$: the cofinality of λ
- $\mathscr{M} \vDash \phi[b_1, \ldots, b_n]$: model $\mathscr{M}$ satisfies sentence ϕ with parameters $b_1, \ldots, b_n$
- Δ_n, Σ_n, Π_n: levels of the Levy hierarchy
- $\mathscr{M}_1 \subseteq \mathscr{M}_2$: $\mathscr{M}_1$ is a substructure of $\mathscr{M}_2$
- $\mathscr{M}_1 \prec \mathscr{M}_2$: $\mathscr{M}_1$ is an elementary substructure of $\mathscr{M}_2$
- $p \parallel q$: elements p and q of a poset $\mathbb{P}$ are compatible
- $p \perp q$: elements p and q of a poset $\mathbb{P}$ are incompatible
- $\operatorname{Name}_\alpha$: the hierarchy of $\mathbb{P}$-names
- τ^G: interpretation of a name τ by filter G
- $M[G]$: the model obtained from a forcing poset $G \subseteq \mathbb{P}$
- $p \Vdash \varphi(\sigma_1, \ldots, \sigma_n)$: $p \in \mathbb{P}$ forces the sentence φ
- $\check{x}$: the name giving an $x \in M$ when interpreted
- $\dot{G}$: the name giving G when interpreted

E Terminology on sets and functions

This appendix will cover some notions on sets and functions such as "bijections", "equivalence classes", and so on.

Remark for experts: I am not dealing with foundational issues in this chapter. See Chapter 89 (and onwards) if that's what you're interested in. Consequently I will not prove most assertions.

§E.1 Sets

A **set** for us will just be a collection of elements (whatever they may be). For example, the set $\mathbb{N} = \{1, 2, 3, 4, \dots\}$ is the positive integers, and $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ is the set of all integers. As another example, we have a set of humans:

$$H = \{x \mid x \text{ is a featherless biped}\}.$$

(Here the "|" means "such that".)

There's also a set with no elements, which we call the **empty set**. It's denoted by $\varnothing$.

It's conventional to use capital letters for sets (like H), and lowercase letters for elements of sets (like x).

Definition E.1.1. We write $x \in S$ to mean "x is in S", for example $3 \in \mathbb{N}$.

Definition E.1.2. If every element of a set A is also in a set B, then we say A is a **subset** of B, and denote this by $A \subseteq B$. If moreover $A \neq B$, we say A is a **proper subset** and write $A \subsetneq B$. (This is analogous to $\leq$ and $<$.)

Given a set A, the set of all subsets is denoted 2^A or $\mathcal{P}(A)$ and called the **power set** of A.

Example E.1.3 (Examples of subsets)

(a) $\{1, 2, 3\} \subseteq \mathbb{N} \subseteq \mathbb{Z}$.

(b) $\varnothing \subseteq A$ for any set A. (Why?)

(c) $A \subseteq A$ for any set A.

(d) If $A = \{1, 2\}$ then $2^A = \{\varnothing, \{1\}, \{2\}, \{1, 2\}\}$.

Definition E.1.4. We write

- $A \cup B$ for the set of elements in *either* A or B (possibly both), called the **union** of A and B.
- $A \cap B$ for the set of elements in *both* A and B, and called the **intersection** of A and B.
- $A \setminus B$ for the set of elements in A but *not* in B.

Example E.1.5 (Examples of set operations)
Let $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$. Then

$$A \cup B = \{1, 2, 3, 4, 5\}$$
$$A \cap B = \{3\}$$
$$A \setminus B = \{1, 2\}.$$

Exercise E.1.6. Convince yourself: for any sets A and B, we have $A \cap B \subseteq A \subseteq A \cup B$.

Here are some commonly recurring sets:

- $\mathbb{C}$ is the set of complex numbers, like $3.2 + \sqrt{2}i$.
- $\mathbb{R}$ is the set of real numbers, like $\sqrt{2}$ or π.
- $\mathbb{N}$ is the set of positive integers, like 5 or 9.
- $\mathbb{Q}$ is the set of rational numbers, like $7/3$.
- $\mathbb{Z}$ is the set of integers, like -2 or 8.

(These are pronounced in the way you would expect: "see", "are", "en", "cue", "zed".)

§E.2 Functions

Given two sets A and B, a **function** f from A to B is a mapping of every element of A to some element of B.

We call A the **domain** of f, and B the **codomain**. We write this as $f \colon A \to B$ or $A \xrightarrow{f} B$.

Abuse of Notation E.2.1. If the name f is not important, we will often just write $A \to B$.

We write $f(a) = b$ or $a \mapsto b$ to signal that f takes a to b.

If B has 0 as an element and $f(a) = 0$, we often say a is a **root** or **zero** of f, and that f **vanishes** at a.

§E.2.i Injective / surjective / bijective functions

Definition E.2.2. A function $f \colon A \to B$ is **injective** if it is "one-to-one" in the following sense: if $f(a) = f(a')$ then $a = a'$. In other words, for any $b \in B$, there is *at most* one $a \in A$ such that $f(a) = b$.

Often, we will write $f \colon A \hookrightarrow B$ to emphasize this.

Definition E.2.3. A function $f \colon A \to B$ is **surjective** if it is "onto" in the following sense: for any $b \in B$ there is *at least* one $a \in A$ such that $f(a) = b$.

Often, we will write $f \colon A \twoheadrightarrow B$ to emphasize this.

Definition E.2.4. A function $f \colon A \to B$ is **bijective** if it is both injective and surjective. In other words, for each $b \in B$, there is *exactly* one $a \in A$ such that $f(a) = b$.

Example E.2.5 (Examples of functions)

By "human" I mean "living featherless biped".

(a) There's a function taking every human to their age in years (rounded to the nearest integer). This function is **not injective**, because for example there are many people with age 20. This function is also **not surjective**: no one has age 10000.

(b) There's a function taking every USA citizen to their social security number. This is also **not surjective** (no one has SSN equal to 3), but at least it **is injective** (no two people have the same SSN).

Example E.2.6 (Examples of bijections)

(a) Let $A = \{1, 2, 3, 4, 5\}$ and $B = \{6, 7, 8, 9, 10\}$. Then the function $f : A \to B$ by $a \mapsto a + 5$ is a bijection.

(b) In a classroom with 30 seats, there is exactly one student in every seat. Thus the function taking each student to the seat they're in is a bijection; in particular, there are exactly 30 students.

Remark E.2.7 — Assume for convenience that A and B are finite sets. Then:

- If $f : A \hookrightarrow B$ is injective, then the size of A is at most the size of B.
- If $f : A \twoheadrightarrow B$ is surjective, then the size of A is at least the size of B.
- If $f : A \to B$ is a bijection, then the size of A equals the size of B.

Now, notice that if $f : A \to B$ is a bijection, then we can "apply f backwards": (for example, rather than mapping each student to the seat they're in, we map each seat to the student sitting in it). This is called an **inverse function**; we denote it $f^{-1} : B \to A$.

§E.2.ii Images and pre-images

Let $X \xrightarrow{f} Y$ be a function.

Definition E.2.8. Suppose $T \subseteq Y$. The **pre-image** $f^{\text{pre}}(T)$ is the set of all $x \in X$ such that $f(x) \in T$. Thus, $f^{\text{pre}}(T)$ is a subset of X.

Example E.2.9 (Examples of pre-image)

Let $f : H \to \mathbb{Z}$ be the age function from earlier. Then

(a) $f^{\text{pre}}(\{13, 14, 15, 16, 17, 18, 19\})$ is the set of teenagers.

(b) $f^{\text{pre}}(\{0\})$ is the set of newborns.

(c) $f^{\text{pre}}(\{1000, 1001, 1002, \dots\}) = \varnothing$, as I don't think anyone is that old.

Abuse of Notation E.2.10. By abuse of notation, we may abbreviate $f^{\text{pre}}(\{y\})$ to $f^{\text{pre}}(y)$. So for example, $f^{\text{pre}}(\{0\})$ above becomes shortened to $f^{\text{pre}}(0)$.

The dual notion is:

Definition E.2.11. Suppose $S \subseteq X$. The **image** $f^{\text{img}}(S)$ is the set of all things of the form $f(s)$.

Example E.2.12 (Examples of images)
Let $A = \{1, 2, 3, 4, 5\}$ and $B = \mathbb{Z}$. Consider a function $f \colon A \to B$ given by

$$f(1) = 17 \quad f(2) = 17 \quad f(3) = 19 \quad f(4) = 30 \quad f(5) = 234.$$

(a) The image $f^{\text{img}}(\{1, 2, 3\})$ is the set $\{17, 19\}$.

(b) The image $f^{\text{img}}(A)$ is the set $\{17, 19, 30, 234\}$.

Question E.2.13. Suppose $f \colon A \twoheadrightarrow B$ is surjective. What is $f^{\text{img}}(A)$?

§E.3 Equivalence relations

Let X be a fixed set now. A binary relation $\sim$ on X assigns a truth value "true" or "false" to $x \sim y$ for each x or y. Now an **equivalence relation** $\sim$ on X is a binary relation which satisfies the following axioms:

- Reflexive: we have $x \sim x$.
- Symmetric: if $x \sim y$ then $y \sim x$
- Transitive: if $x \sim y$ and $y \sim z$ then $x \sim z$.

An **equivalence class** is then a set of all things equivalent to each other. One can show that X becomes partitioned by these equivalence classes:

Example E.3.1 (Example of an equivalence relation)
Let $\mathbb{N}$ denote the set of positive integers. Then suppose we declare $a \sim b$ if a and b have the same last digit, for example $131 \sim 211$, $45 \sim 125$, and so on.
Then $\sim$ is an equivalence relation. It partitions $\mathbb{N}$ into ten equivalence classes, one for each trailing digit.

Often, the set of equivalence classes will be denoted $X/\!\sim$ (pronounced "X mod sim").

Image Attributions

[1207] 127"RECT". *Cantor set in seven iterations*. Public domain. 2007. URL: https://en.wikipedia.org/wiki/File:Cantor_set_in_seven_iterations.svg (cited p. 398)

[ca] POP-UP CASKET. *Omega exp*. Public domain. URL: https://commons.wikimedia.org/wiki/File:Omega-exp-omega-labeled.svg (cited p. 921)

[Ee] EEYORE22. *Weierstrass function*. Public domain. URL: https://commons.wikimedia.org/wiki/File:WeierstrassFunction.svg (cited p. 348)

[Fr] FROPUFF. *Klein bottle*. Public domain. URL: https://en.wikipedia.org/wiki/File:KleinBottle-01.png (cited p. 658)

[Ge] TOPOLOGICAL GIRL'S GENERATION. *Topological Girl's Generation*. (dead link). URL: https://topologicalgirlsgeneration.tumblr.com/ (cited p. 53)

[gk] G.KOV. *Normal surface vector*. URL: https://tex.stackexchange.com/a/235142/76888 (cited p. 452)

[Go08] ABSTRUSE GOOSE. *Math Text*. CC 3.0. 2008. URL: https://abstrusegoose.com/12 (cited p. x)

[Go09] ABSTRUSE GOOSE. *Zornaholic*. CC 3.0. 2009. URL: https://abstrusegoose.com/133 (cited p. 924)

[Ho] GEORGE HODAN. *Apple*. Public domain. URL: https://www.publicdomainpictures.net/view-image.php?image=20117 (cited p. 935)

[In] INDUCTIVELOAD. *Klein Bottle Folding*. Public domain. URL: https://commons.wikimedia.org/wiki/File:Klein_Bottle_Folding_1.svg (cited p. 658)

[Kr] PETR KRATOCHVIL. *Velociraptor*. Public domain. URL: https://www.publicdomainpictures.net/view-image.php?image=93881 (cited p. 935)

[Ma12] MATHMATHSMATHEMATICS. *How to Multiply Matrices - A 2x2 Matrix by various sizes*. May 2012. URL: https://youtu.be/pJwslaulUMU (cited p. 149)

[Mu] RANDALL MUNROE. *Rolle's theorem*. CC 2.5. URL: https://xkcd.com/2042/ (cited p. 324)

[Na] KRISH NAVEDALA. *Stokes patch*. Public domain. URL: https://en.wikipedia.org/wiki/File:Stokes_patch.svg (cited p. 462)

[Or] BEN ORLIN. *The Math Major Who Never Reads Math*. URL: https://mathwithbaddrawings.com/2015/03/17/the-math-major-who-never-reads-math/ (cited p. xi)

[To] TOAHIGHERLEVEL. *Projection color torus*. Public domain. URL: https://en.wikipedia.org/wiki/File:Projection_color_torus.jpg (cited p. 656)

[Wa] BILL WATTERSON. *Calvin and Hobbes*. I think this is fair use. (cited p. 873)

[Wo] WORDSLAUGH. *Covering space diagram*. CC 3.0. URL: https://commons.wikimedia.org/wiki/File:Covering_space_diagram.svg (cited p. 675)

Bibliography

[**Ax97**] Sheldon Axler. *Linear algebra done right*. New York: Springer, 1997. ISBN: 978-0-387-98258-8 (cited pp. 159, 987)

[**Ba10**] Joseph Bak. *Complex analysis*. New York: Springer Science+Business Media, LLC, 2010. ISBN: 978-1-4419-7287-3 (cited p. 347)

[**Ch08**] Steve Cheng. "A Crash Course on the Lebesgue Integral and Measure Theory". Apr. 2008. URL: https://www.gold-saucer.org/math/lebesgue/lebesgue-new.pdf (cited p. 988)

[**Et11**] Pavel Etingof. "Introduction to Representation Theory". 2011. URL: https://math.mit.edu/~etingof/replect.pdf (cited pp. vii, 989)

[**Ga03**] Andreas Gathmann. "Algebraic Geometry". 2003. URL: https://www.mathematik.uni-kl.de/~gathmann/de/alggeom.php (cited pp. vii, 916, 989)

[**Ga14**] Dennis Gaitsgory. "Math 55a: Honors Abstract and Linear Algebra". 2014. URL: https://web.evanchen.cc/coursework.html (cited pp. vii, 987)

[**Ga15**] Dennis Gaitsgory. "Math 55b: Honors Real and Complex Analysis". 2015. URL: https://web.evanchen.cc/coursework.html (cited pp. 337, 987)

[**Go11**] Timothy Gowers. "Normal subgroups and quotient groups". 2011. URL: https://gowers.wordpress.com/2011/11/20/normal-subgroups-and-quotient-groups/ (cited pp. 72, 987)

[**Go18**] Vadim Gorin. "18.175: Theory of Probability". 2018. URL: https://web.archive.org/web/20190617235844/http://web.mit.edu/txz/www/links.html (cited pp. vii, 988)

[**Ha02**] Allen Hatcher. *Algebraic topology*. Cambridge, New York: Cambridge University Press, 2002. ISBN: 0-521-79160-X. URL: https://opac.inria.fr/record=b1122188 (cited pp. 655, 741, 752, 762, 769, 781, 786, 792, 989)

[**Hi13**] A. J. Hildebrand. "Introduction to Analytic Number Theory". 2013. URL: https://web.archive.org/web/20230326025121/https://faculty.math.illinois.edu/~hildebr/ant/main.pdf (cited p. 990)

[**Ko14**] Peter Koellner. "Math 145a: Set Theory I". 2014. URL: https://web.evanchen.cc/coursework.html (cited pp. vii, 989)

[**Le**] Holden Lee. "Number Theory". URL: https://github.com/holdenlee/number-theory (cited p. 628)

[**Le02**] Hendrik Lenstra. "The Chebotarev Density Theorem". 2002. URL: https://websites.math.leidenuniv.nl/algebra/ (cited pp. 633, 989)

[**Le14**] Tom Leinster. *Basic category theory*. Cambridge: Cambridge University Press, 2014. ISBN: 978-1-107-04424-1. URL: https://arxiv.org/abs/1612.09375 (cited pp. vii, 687, 690, 708, 711, 927, 988)

[**Ll15**] Seth Lloyd. "18.435J: Quantum Computation". 2015. URL: https://web.evanchen.cc/coursework.html (cited pp. vii, 988)

[**Ma13a**] Laurentiu Maxim. "Math 752 Topology Lecture Notes". 2013. URL: https://www.math.wisc.edu/~maxim/752notes.pdf (cited pp. 781, 989, 1023)

[**Ma13**b] Maxima. "Burnside's Lemma, post 6". 2013. URL: https://www.aops.com/Forum/viewtopic.php?p=3089768#p3089768 (cited p. 214)

[**Mi14**] Alexandre Miquel. "An Axiomatic Presentation of the Method of Forcing". 2014. URL: https://www.fing.edu.uy/~amiquel/forcing/ (cited p. 989)

[**Mi95**] R. Miranda. *Algebraic Curves and Riemann Surfaces*. Dimacs Series in Discrete Mathematics and Theoretical Comput. American Mathematical Society, 1995. ISBN: 9780-8218-0268-7 (cited pp. 498, 530, 990)

[**Mu00**] James Munkres. *Topology*. 2nd. Prentice-Hall, Inc., Jan. 2000. ISBN: 97881-203-2046-8. URL: https://amazon.com/o/ASIN/8120320468/ (cited p. 989)

[**Og10**] Frederique Oggier. "Algebraic Number Theory". 2010. URL: https://feog.github.io/ANT10.pdf (cited pp. 539, 989)

[**Pu02**] C. C. Pugh. *Real mathematical analysis*. New York: Springer, 2002. ISBN: 978-0-387-95297-0 (cited pp. vii, 109, 123, 296, 304, 348, 452, 457, 987, 988)

[**Sc07**] W. H. Schikhof. *Ultrametric Calculus: An Introduction to P-Adic Analysis*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2007. ISBN: 9780-521-03287-2. URL: https://books.google.com/books?id=cBT05R7TH1QC (cited pp. 313, 314)

[**Sj05**] Reyer Sjamaar. "Manifolds and Differential Forms". 2005. URL: https://pi.math.cornell.edu/~sjamaar/manifolds/manifold.pdf (cited pp. vii, 486, 488, 987)

[**Ul08**] Brooke Ullery. "Minkowski Theory and the Class Number". 2008. URL: https://www.math.uchicago.edu/~may/VIGRE/VIGRE2008/REUPapers/Ullery.pdf (cited p. 557)

[**Va17**] Ravi Vakil. "The Rising Sea: Foundations of Algebraic Geometry". Nov. 2017. URL: https://math.stanford.edu/~vakil/216blog/ (cited pp. iii, vii, 807, 808, 814, 815, 881, 893, 916, 989, 990)

[**Ya12**] Andrew Yang. "Math 43: Complex Analysis". 2012. URL: https://math.dartmouth.edu/~m43s12/syllabus.html (cited pp. 347, 353, 988)