Contents

List of Definitions

Preface

The idea of these notes is to provide a gentle and informal introduction to Interesting Math, which for me generally means number theory and associated fields. The basic idea is that I don't think this sort of mathematics is nearly as inaccessible to a general audience as it is generally perceived to be, but there don't exist many sources that don't assume more experience. I don't intend for this to function as a textbook, and the reader should not expect to "fully" understand any subject covered herein; my goal is for the reader to be able to have an understanding of some of the problems and ideas of the fields discussed, and be able to have some intuition for why an idea might be interesting or useful.

My intention is that these notes should be accessible to an interested reader with essentially no post-high school math background. Some computations involve calculus, but generally speaking they can be taken on faith or verified in WolframAlpha or similar; at some point I intend to have calculus notes up written in a similar style as an additional resource. That said I have the fatal flaw of most mathematicians which is a tendency to write like other mathematicians, so especially later on I've regressed towards a more formal style, and there is almost certainly a good amount of unexplained notation. Theoretically I'll try to improve these issues at some point.

One piece of advice for the reader: if you encounter a symbol you are not familiar with, there is a very good chance that it is a Greek letter. If you, like me, find it easier to read symbols if you know how to mentally pronounce them, I recommend having a table of the Greek alphabet on hand (e.g. Wikipedia's).

I'll update this with new material periodically (and unpredictably); my goal, for now, is to cover classical analytic and algebraic number theory, and say something about modern algebraic number theory and related areas. For the purpose of doing so we'll have to say something about complex analysis, topology, and abstract algebra, among others.

None of the material in these notes is original. I learned it from many places, which I am unable to list in full, but particularly noteworthy are the lecture notes of A.J. Hildebrand [2] for chapter 1 and Mike Artin's algebra classes 18.701-2 at MIT, with textbook Artin's *Algebra* [1]; the section on (classical) algebraic number theory will be largely based on Andrew Sutherland's number theory course 18.785 and its accompanying lecture notes [3].

Credit to my mom for being the primary reader of these notes and catching a number of errors, and to Wojtek Wawrów for catching some more. (Nevertheless there are undoubtedly many more remaining, for which I and not them are responsible.)

Notation

For quick reference, let's list some common notation that will be used throughout.

Notation	Meaning
$s \in S$	The element s is in the set S
$S \subseteq T$	The set S is a subset of the set T, possibly equal to all of T
$S \subset T$	The set S is a proper subset of the set T, i.e. a subset which is not equal to all of T
$S \times T$	The direct product of sets S and T, i.e. the set of pairs (s, t) for $s \in S$ and $t \in T$
$\mathbb{N}$	The natural numbers $\{1, 2, 3, \ldots\}$
$\mathbb{Z}$	The integers $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$
$\mathbb{Q}$	The rationals $\{\frac{a}{b}\}$, where $a, b \in \mathbb{Z}$
$\mathbb{R}$	The real numbers, such as π, e, $\sqrt{2}$, $-\frac{3}{2}$, 0, $\ldots$
$\mathbb{C}$	The complex numbers $\{a + bi\}$, where $a, b \in \mathbb{R}$
$f : S \to T$	f is a function from S to T, assigning to each element $s \in S$ an element $f(s) \in T$
$s \mapsto t$	The element s in a given set S maps to the element t in T under some function $S \to T$
$f : S \twoheadrightarrow T$	f is a surjection from S to T
$f : S \hookrightarrow T$	f is an injection from S to T
T^S	The set of functions $S \to T$
$\mathbf{1}_S(x)$	The indicator function of S, which returns 1 if $x \in S$ and 0 otherwise
$c_S(x)$	The counting function of S, which returns the number of elements of S which are less than or equal to x
$\pi(x)$	The counting function of the primes, i.e. the number of primes less than or equal to x
$\Lambda(x)$	The von Mangoldt function, which returns $\log p$ if $x = p^k$ for some $k \geq 1$ and prime p, i.e. if x is a prime power, and returns 0 otherwise
$\zeta(s)$	The Riemann zeta function
$\varphi(x)$	Euler's totient function
$G \circlearrowright S$	A group G acts on a set S
S_n	The symmetric group on n elements
$R^\times$	The group of units, or multiplicatively invertible elements, of a ring R
$(x_1, x_2, \ldots)$	The ideal of a ring generated by $x_1, x_2, \ldots$
μ_n	The set of nth roots of unity
$\mathrm{GL}_n(k)$	The group of invertible $n \times n$ matrices over a field k
$\oplus$	Direct sum
$\otimes$	Tensor product

Chapter 1

Analytic number theory: distribution of primes

1.1 Introduction

Note: I'll try to explain all notation as I go in footnotes, but I'll probably forget some; email me if any is unclear. It should be noted that small numbers floating around equations that otherwise don't make any sense are likely to be footnotes.

For our purposes for the moment, number theory is concerned with properties of the set of natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$ or the integers $\mathbb{Z} = \{\ldots -2, -1, 0, 1, 2, \ldots\}$.[1] In particular, $\mathbb{N}$ is equipped with two natural operations: addition and multiplication.

The additive structure of $\mathbb{N}$ is fairly simple: it is *generated* by 1, i.e. we can construct every element of $\mathbb{N}$ by adding 1 to itself some number of times. The multiplicative structure is more complicated. Given an arbitrary positive integer $n > 1$, we obviously cannot construct it by multiplying 1 with itself, since this would just give 1. If we add in the next-smallest possibility 2 we can generate some subset of $\mathbb{N}$, specifically those n of the form 2^k for some positive integer k; but this is a small subset of all positive integers.

Of course, there exists some subset S of $\mathbb{N}$ such that any $n \in \mathbb{N}$ [2] can be written as a product of elements of S. To see this, note that we can simply take $S = \mathbb{N}$; then $n \in S$, so just writing n as the product of one element[3] satisfies the claim.

Let's make this more precise. For any subset $S \subseteq \mathbb{N}$ of the natural numbers,[4] write $\overline{S}$ for the *multiplicative closure of* S, meaning the set of all natural numbers that can be written as a product of elements of S.[5] For example, if $S = \{3\}$, then $\overline{S} = \{1, 3, 9, 27, \ldots\}$; if $S = \{2, 5\}$ then $\overline{S} = \{1, 2, 4, 5, 8, 10, 16, 20, 25, \ldots\}$. We then have the following.

[1] Sets will in general be written using these curly brackets; the font of $\mathbb{N}$ and $\mathbb{Z}$ will frequently be used for important sets.

[2] the symbol $\in$ denotes membership in a set; this reads n in $\mathbb{N}$.

[3] This may be counterintuitive; we will often be taking sums or products of sets, which simply means adding or multiplying the elements together, but when there is only one element the sum or product of the set is just that element. The empty sum, the sum of the empty set, is 0, while the empty product is 1.

[4] the notation $S \subset T$ for two sets S and T means that S is a subset of T; $S \subseteq T$ means that S is either a subset of T or equal to T, analogous to $<$ and $\leq$ notation. I will often say that one set is a subset of another when I actually mean $\subseteq$; when precision is needed the symbolic notation will be used.

[5] This includes 1, since as in a previous note it can be written as the empty product.

Proposition 1.1.1. *There exists some $S \subseteq \mathbb{N}$ such that $\overline{S} = \mathbb{N}$.*

Proof. $S = \mathbb{N}$ satisfies the claim, as above. □

Of course, we can do much better than this. Let $S = \mathbb{N}$; what numbers can we remove from S and still have it satisfy the claim of Proposition 1.1.1 above?

Well, we can certainly remove 1, since 1 is the empty product and so can certainly be written as the product of elements in S, specifically as the product of none of them. There's no particular reason to think we don't need 2 or 3, but we definitely don't need 4: we can write $4 = 2 \cdot 2$,[6] so if $n = 4 \cdot x \cdot y \cdots$ is the representation of some number n as a product of elements of S then $n = 2 \cdot 2 \cdot x \cdot y \cdots$ represents it as a product of elements of $S \setminus \{2\}$.[7] Using the same logic, we see that we can remove any natural number which can be written as the product of two smaller numbers greater than 1 (a *composite* number). This gives us the following algorithm.

Algorithm 1.1.2.

(1) Let $S = \mathbb{N} \setminus \{1\} = \{2, 3, 4, \ldots\}$.

(2) Let $m = 2$.

(3) For each $k \in S$ greater than m, if k is divisible by m remove it from S.

(4) Set m to be the next-smallest remaining element of S and go back to step 3.

If we were to run this algorithm, it would in fact run forever. However, if in step 1 we instead set S to be the set of integers between 1 and N for some large integer N and run the algorithm from there on, it would terminate after at most $\sqrt{N}$ loops, and the remaining elements of S at the end would be exactly the prime numbers less than or equal to N. This is because we remove every integer divisible by a smaller number greater than 1, i.e. all of the composite numbers, and the prime numbers are by definition the natural numbers greater than 1 which are not composite. This is called the sieve of Eratosthenes, and it is a pretty good way of generating all prime numbers up to a certain point. A good exercise is to try it with $N = 100$ to see how it works.

Proposition 1.1.3. *If S satisfies the conditions of Proposition 1.1.1, then after doing each of steps 3 and 4 once it still satisfies these conditions.*

Proof. The assumption that we are making is that S is a subset of $\mathbb{N}$ such that $\overline{S} = \mathbb{N}$. If this is the case and m is the current element, for any $k > m$ such that k is divisible by m we can write k as $m \cdot \frac{k}{m}$ with both m and $\frac{k}{m}$ positive integers greater than 1 (and less than k), so k is composite and so as we have seen above we can replace it in any product formula by m and $\frac{k}{m}$. Therefore we can remove any such k from S without changing $\overline{S}$. □

[6] The dot $\cdot$ denotes multiplication.

[7] For two sets S and T with $T \subseteq S$, the notation $S \setminus T$ means the set difference of S and T, i.e. the elements of S which are not in T.

In combination with our discussion above, this suggests that the smallest possible S would be the set of prime numbers. Is this true?

Well, what do we mean by "smallest possible"? Let $\mathcal{P}$ be the set of prime numbers. First, we claim that $\mathcal{P}$ does in fact generate $\mathbb{N}$.

Proposition 1.1.4. $\overline{\mathcal{P}} = \mathbb{N}$.

Proof. We want to show that every $n \in \mathbb{N}$ can be written as the product of some number of primes. Suppose that this were false; we will try to derive a contradiction, thus proving it true.

Suppose that not every $n \in \mathbb{N}$ is in $\overline{\mathcal{P}}$, and let n be the minimal such integer. If n is prime or 1, as above it must be in $\overline{\mathcal{P}}$, so n must be composite. Then it can be written as $a \cdot b$ for some a, b less than n and greater than 1. If a and b are prime, then $n \in \overline{\mathcal{P}}$, so at least one of them must be composite, say a. Then we can repeat the argument above: a can be written as $a_1 \cdot b_1$, and b similarly. We could then repeat this argument with $a_1 = a_2 \cdot b_2$ and so on.

This gives us a sequence $a, a_1, a_2, \ldots$, with $1 < a_k < a_{k-1}$. Since this is a sequence of decreasing integers, all of which are greater than 1, it must terminate eventually, i.e. eventually a_k must be prime for some k sufficiently large. Since a_k divides a_{k-1}, which in turn divides a_{k-2} and so on, this shows that n has some prime factor a_k. Since n cannot be written as a product of primes by assumption, neither can n/a_k: if it could, then $a_k \cdot \frac{n}{a_k}$ would give a prime factorization of n by inserting the prime factorization of n/a_k and multiplying by the prime a_k, so since this would contradict our assumption we must conclude that n/a_k is not in $\overline{\mathcal{P}}$. But $n/a_k < n$ since $a_k > 1$, and we assumed that n was the smallest positive integer not in $\overline{\mathcal{P}}$. This is a contradiction, so our assumption must have been wrong and we must have $\overline{\mathcal{P}} = \mathbb{N}$. □

This shows that $\mathcal{P}$ is a valid choice for S. Now we want to see that it is minimal.

Proposition 1.1.5. *Suppose that S is a subset of $\mathbb{N}$ such that $\overline{S} = \mathbb{N}$. Then S contains $\mathcal{P}$.*

Proof. Let p be a prime number. Since $p \in \mathbb{N}$, we have $p \in \overline{S}$, so there exist some elements of S whose product is p; taking a to be the first of these that is not equal to 1 (if they are all 1 then their product is $1 \neq p$), we see that a divides p and is greater than 1, so since p is prime $a = p$. But $a \in S$, so $p \in S$. Since this holds true for every prime p, every prime must be in S, so $\mathcal{P} \subseteq S$. □

Thus since every valid S must contain $\mathcal{P}$ and $\mathcal{P}$ itself satisfies $\overline{\mathcal{P}} = \mathbb{N}$, we really are justified in saying that $\mathcal{P}$ is the minimal generating set for $\mathbb{N}$ under multiplication, just as $\{1\}$ was the minimal generating set under addition. $\mathcal{P}$, however, is a much more mysterious set than $\{1\}$, and most of our time will be spent trying to understand it, and thus the multiplicative structure of $\mathbb{N}$.

1.2 Elementary prime-counting

The most obvious question about any set, and $\mathcal{P}$ in particular, is: how large is it? Though in one sense this has a very simple answer, understanding it better will turn out to be very difficult.

Proposition 1.2.1. *There are infinitely many prime numbers.*

Proof. Suppose that $\mathcal{P}$ were finite. Let P be the product of every element of $\mathcal{P}$. Then for every $p \in \mathcal{P}$, P is divisible by p, i.e. $P = pN$ for some integer N. If $P+1$ were also divisible by p, i.e. $P + 1 = pM$ for some integer M, then $pM = pN + 1$, so $p(M - N) = 1$, so 1 would be divisible by p, which is clearly false. Thus $P + 1$ is not divisible by any $p \in \mathcal{P}$. But by Proposition 1.1.4 we can write $P + 1$ as the product of elements of $\mathcal{P}$, so since no prime divides $P + 1$ we conclude that $P + 1$ is the empty product 1. But then $P = 0$, and since every element of $\mathcal{P}$ is greater than 1 so is P, a contradiction. Therefore $\mathcal{P}$ cannot be finite. □

Strictly speaking, this answers our question. Nevertheless this is somewhat unsatisfying. After all, we would like to be able to say that, as a subset of $\mathbb{N}$, $\mathbb{N}$ itself is "large" in some sense—after all, it contains everything!—while $T_{10} = \{1, 10, 100, 1000, \ldots\}$ is "small" in that it seems to contain relatively few numbers, even though both are infinite. The natural measure for this is *density*.

Definition 1.2.2. Let S be a subset of the natural numbers. We define its *counting function* $c_S(x)$ to be the number of elements of S less than or equal to x.

Thus for example $c_{\mathbb{N}}(x) = x$ for every natural number x, while for example $c_{T_{10}}(100) = c_{T_{10}}(500) = 3$.

Definition 1.2.3. The *density* $\delta(S)$ of a set $S \subseteq \mathbb{N}$ is defined as the limit $\lim_{x\to\infty} \frac{c_S(x)}{x}$.

The density of any subset S of $\mathbb{N}$, when it exists (it need not be well-defined in general!), will be a real number between 0 and 1.

We can compute $\delta(\mathbb{N}) = \lim_{x\to\infty} \frac{x}{x} = 1$; on the other hand, $c_{T_{10}}(x)$ is at most $\log_{10}(x)+1$, so $\delta(T_{10}) \leq \lim_{x\to\infty} \frac{\log_{10}(x)+1}{x} = 0$, so $\delta(T_{10})$ must be 0.[8] This gives us the kind of measure of infinite sets we were looking for; indeed, it's pretty easy to see that if S is a finite set then $\delta(S) = 0$ (good exercise!).

The natural question, then, is: what is $\delta(\mathcal{P})$? Unfortunately, this doesn't actually tell us very much:

Proposition 1.2.4. $\delta(\mathcal{P}) = 0$.

We could actually prove this now, but it wouldn't be very enlightening; we'll come back to this. Nevertheless, we see that density, while a more sensitive measure than the size of a set, is still too coarse to tell us very much about $\mathcal{P}$.

Instead, let's look at the counting function of $\mathcal{P}$. This has a canonical and somewhat unfortunate name: we write $\pi(x)$ for $c_{\mathcal{P}}(x)$. This has nothing to do with $\pi = 3.14159\ldots$, but instead is 'pi' for 'p' for 'prime'. Henceforth when we talk about prime counting what we really mean is estimating $\pi(x)$.

What do we know so far? Well, from Proposition 1.2.1, we know that $\lim_{x\to\infty} \pi(x) = \infty$, and Proposition 1.2.4 claims that $\lim_{x\to\infty} \frac{\pi(x)}{x} = 0$ (though we have not yet proven this). This leaves a *huge* range of possibilities. Before we can attack it directly, though, we need to

[8] It's a pretty good exercise to work out why $\log_{10}(x)/x$ tends to 0 as x goes to infinity; the underlying principle, which is generally quite useful, is that anything "logarithmic" (such as $\log_{10}(x)$) is much smaller than anything "polynomial" (such as x).

complete our analysis of the primes as the generators of the positive integers. In particular, though we have proved that every $n \in \mathbb{N}$ can be written as a product of primes, we have not yet proven that such a decomposition is (essentially) unique. This statement will be the source of most of what we can say about the distribution of the prime numbers; due to its deepness it gets a shmancy title.

Theorem 1.2.5 (Fundamental Theorem of Arithmetic)**.** *Every natural number n can be written as a product of prime numbers in exactly one way, up to the ordering of the factors.*

Proof. The existence of such a product representation for any $n \in \mathbb{N}$ is Proposition 1.1.4, so it remains only to prove uniqueness. Fix some $n \in \mathbb{N}$. We can gather duplicate copies of each prime factor into prime powers. Suppose then that n has two different prime product representations

$$n = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r} = q_1^{b_1} q_2^{a_2} \cdots q_s^{b_s}.$$

First, if any of the p_i and any of the q_i are the same, we can divide out by those factors until one side or the other has no remaining factors of these primes; thus we can convert any such formula into one where all of the p_i are different from all of the q_i.

Since p_1 divides $n = p_1^{a_1} \cdots p_r^{a_r}$, it also divides $n = q_1^{b_1} \cdots q_s^{b_s}$. We now use the following lemma, which we will prove afterward.

Lemma 1.2.6. *If a prime p divides ab where a and b are positive integers, then p divides at least one of a and b.*

By Lemma 1.2.6, p_1 divides at least one of $q_1^{b_1} \cdots q_{s-1}^{b_{s-1}}$ and $q_s^{b_s}$. In the former case, we can repeat the process s times to see that p_1 divides at least one of $q_1^{b_1}$, $q_2^{b_2}$, etc. up through $q_s^{b_s}$. Let $q_i^{b_i}$ be one of these that is divisible by p_1. If $b_i = 1$ then this is impossible since $p_1 \neq q_i$, so assume $b_i > 1$. Then by the same argument p_1 divides either $q_i^{b_i - 1}$ or q_i. Since q_i is prime, the latter is impossible as above, so p_1 divides $q_i^{b_i - 1}$. Repeating the above argument, we conclude that p_1 divides $q_i^{b_i - 2}$, and so repeating the same argument we will eventually reduce the exponent down to 1, so p_1 divides q_i. This is impossible as above, so n cannot have two distinct prime factorizations. □

It remains only to prove the lemma.

Proof of Lemma 1.2.6. If the result were not true, then there would be some subset $\mathcal{P}_0$ of $\mathcal{P}$ of primes p for which there exist some integers a, b such that p divides ab but neither of a and b. Let p be the smallest such prime. If $a = pm + k$, then $ab = pmb + bk$, so $bk = ab - pmb$. The right-hand side is divisible by p, so so is the left-hand side. By choosing m such that k is the remainder of a when divided by p, we can choose k such that $0 \leq k < p$; so we are left with a case like the original one, where now we can assume $0 \leq a < p$. We can do the same thing for b, so we can assume without loss of generality that a and b are both less than p. Since ab is divisible by p, we can write $ab = pt$ for some integer t. Since a and b are less than p, $ab < p^2$, so $t < p$.

If either of a or b is equal to 1, then ab is just equal to the other one, so if p divides ab it must divide one of a and b. Therefore assume that a and b are greater than 1. Then ab is composite, so t cannot be equal to 1, since that would imply that $ab = p$ is prime. Therefore, by Proposition 1.1.4, there exists at least one prime q that divides t. Since q divides t, it also

divides $pt = ab$; since $t < p$ and q divides t, $q < p$ and so since we assumed that p was the smallest element of $\mathcal{P}_0$ we conclude that q must satisfy the claim of the lemma, i.e. q must divide either a or b. Let's say that it divides a; if it divides b we can do the same argument after switching the two. Then dividing by q gives $p\frac{t}{q} = \frac{a}{q}b$. If $a = q$, then since p divides the left-hand side it must divide the right-hand side, which is then just b, contradicting our original claim. If not, we can repeat the argument above until, after dividing through by enough prime factors (all of which, as above, will be less than p), we have divided out by all of the factors of either a, b, or t. In either of the first two cases, we are left with $p\frac{t}{a} = \frac{a}{a}\frac{b}{a}$ or $p\frac{t}{b} = \frac{a}{b}\frac{b}{b}$, in which case p divides $\frac{b}{a}$ or $\frac{a}{b}$ respectively and therefore either b or a, and in the third case we get $p = \frac{a}{t}\frac{b}{t}$. Since both factors on the right-hand side are integers since we have only divided by their prime factors, either one of them is equal to 1, in which case we can apply the argument above, or p is composite, which is impossible. Thus p cannot be an element of $\mathcal{P}_0$, a contradiction. □

There is one more notion of prime-counting we will explore here. Namely, we have already found an algorithm to produce all of the primes less than or equal to a given integer x, the sieve of Eratosthenes (Algorithm 1.1.2); surely we can also use this to count these primes.

This is in fact true, but it is more difficult than it might appear. What we might naively think that the algorithm is doing is starting with the integers from 1 to x, of which of course there are x, and then for each prime p less than x removing all of the integers between p and x which are divisible by p, of which there are roughly $\frac{x}{p} - 1$ (we subtract one because we do not want to remove p itself!). The most obvious problem with this is that it already requires us to know all of the primes up to x, which is not terribly helpful for counting; instead, we will fix some smaller number y, and work as follows.

Algorithm 1.2.7.

(1) Determine all of the primes less than or equal to y.

(2) Set $S = \{2, 3, 4, \ldots, x-1, x\}$.

(3) For each prime $p \leq y$, remove *all* integers divisible by p from S (including p itself).

This is a modified version of the sieve of Eratosthenes. The remaining elements of S at its completion will be the integers less than or equal to x not divisible by any prime less than or equal to y.

Proposition 1.2.8. *The elements of S resulting from running Algorithm 1.2.7 with $y = \sqrt{x}$ are exactly the primes between $\sqrt{x}$ and x.*

Proof. By Lemma 1.1.4, every integer less than $y = \sqrt{x}$ has some prime divisor, which will necessarily be less than y. Therefore the algorithm removes every integer less than or equal to $\sqrt{x}$. If $n > \sqrt{x}$, there are two possibilities. If n is prime, it will not be removed, since it has no smaller prime divisors. If n is not prime, it has at least two prime divisors, say p and q. Since $x \geq n \geq pq$, p and q cannot both be larger than $\sqrt{x}$, so n will be removed by the algorithm. Therefore the remaining elements will be exactly the primes between $\sqrt{x}$ and x. □

This does not let us directly count $\pi(x)$; instead it counts $\pi(x) - \pi(\sqrt{x})$. However, that will be good enough for our purposes.

However, we still have another problem with the scheme outlined above. Namely, based on the above discussion we expect that we will get

$$\pi(x) - \pi(\sqrt{x}) \approx x - \sum_{p \leq \sqrt{x}} \frac{x}{p}$$

(note that here we do not subtract 1, since we are not interested in counting primes less than or equal to $\sqrt{x}$). Unfortunately, it turns out that this sum actually diverges, so this would imply that $\pi(x) - \pi(\sqrt{x})$ tends to negative infinity, which, since $\pi(x)$ can only increase, is blatantly impossible.

Where have we gone wrong? Working out the sieve for as little as $x = 10$ will quickly reveal the answer: double counting. At first, we remove everything divisible by 2: there are 5 of them, well and good. Next, we remove everything divisible by 3. There are 3 of them, roughly as we would expect. But wait! We have already removed 6; we cannot do it again. Therefore we need to add back in those numbers which are divisible by two different primes less than or equal to $\sqrt{x}$.

If we continue to a larger sieve, we will find that this is still not right: we will also need to subtract back out those numbers that are divisible by *three* such primes, and then add back those divisible by four, and so on. This is called the principle of inclusion-exclusion, though we won't go into it further here. This leads us naturally to define another function which will be useful later.

Definition 1.2.9. We say that a positive integer is *squarefree* if it is not divisible by any perfect square greater than 1.

Definition 1.2.10. The *Möbius function* $\mu(x)$ is defined to be 1 if x is squarefree and has an even number of prime factors, -1 if x is squarefree and has an odd number of prime factors, and 0 if x is not squarefree.

It is reasonably intuitive why we care about the parity of the number of prime factors: we just saw that whether we add or subtract each number back in depends on whether it has an even or odd number of prime factors. The squarefreeness condition is a little more confusing. It may help to think of it this way: if we write $n = p_1^{a_1} \cdots p_r^{a_r}$, the property of being squarefree is that all of the a_i are equal to 1. The reason that this is relevant is that we are in no danger of multiply counting any number that is not squarefree.

This function is important for several reasons, but its main power is encapsulated in the following proposition, which will be proved in the next section. Henceforth we will use the notation $a|b$ for "a divides b", and summing over $d|n$ means the sum over all divisors of n (including 1 and n).

Proposition 1.2.11. *For any integer $n > 1$, we have*

$$\sum_{d|n} \mu(d) = 0.$$

This may appear obscure now, but we will see that in fact this means that μ is in some sense an inverse of the constant function 1, and as such will be very useful to us.

1.3 Dirichlet series and the zeta function

In 1740, Leonhard Euler introduced the zeta function

$$\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}$$

where s is an integer greater than 1, so that for example $\zeta(2) = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots$. (It turns out that this sum is equal to $\frac{\pi^2}{6}$, but we won't prove this.) More than a century later, Pafnuty Chebyshev considered the same function but allowing s to be any complex number with real part greater than 1, and Bernhard Riemann showed that it could be defined for any complex number. It is often called the Riemann zeta function, both in his honor and to distinguish it from other zeta functions (of which there are many, all analogous in some way to this prime example).

Proposition 1.3.1. *The infinite sum*

$$\sum_{k=1}^{\infty} \frac{1}{k^s}$$

converges if the real part of s, denoted $\mathrm{Re}(s)$*, is greater than* 1.

Proof. Recall from calculus that for any monotonically decreasing function[9] f the sum

$$\sum_{k=a}^{\infty} f(k)$$

converges if and only if the integral

$$\int_a^{\infty} f(x)\,dx$$

converges. Here, $a = 1$ and $f(k) = \frac{1}{k^s}$, so the integral is

$$\int_1^{\infty} \frac{1}{x^s}\,dx = \frac{x^{1-s}}{1-s}\Big|_1^{\infty} = \frac{1}{s-1}$$

and converges so long as the real part of s is at least 1.[10] □

[9] That is, decreasing across the whole domain with which we are considered, here over all positive integers.

[10] To see why we only care about the real part, note that the convergence is actually about the absolute value of $f(x)$. Writing $x = \sigma + it$, we have $f(x) = \frac{1}{x^\sigma x^{it}}$. Recalling Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$ for any real θ, we see that $x^{it} = (e^{\log x})^{it} = e^{it\log x} = \cos(t\log x) + i\sin(t\log x)$. Since the absolute value of a complex number is $|a+bi| = \sqrt{a^2+b^2}$, we get $|x^{it}| = \cos(t\log x)^2 + \sin(t\log x)^2 = 1$ by the trigonometric formula, so $|f(x)| = \frac{1}{|x^\sigma|\cdot|x^{it}|} = \frac{1}{|x^\sigma|} = \frac{1}{x^\sigma}$ since x^σ is a positive real number for x and σ both positive and real.

The above bound is in a sense tight: in particular, the sum diverges at $s = 1$.

Why should we care about the Riemann zeta function? There are two reasons. The first requires the theory of Dirichlet series.

Let $f : \mathbb{N} \to \mathbb{C}$ be an arithmetic function, i.e. a function from the natural numbers $\mathbb{N}$ to the complex numbers $\mathbb{C}$. We define its Dirichlet series $\mathcal{D}[f](s)$ by the sum

$$\mathcal{D}[f](s) = \sum_{k=1}^{\infty} \frac{f(k)}{k^s}.$$

Thus we see that, writing 1 for the constant function $f(x) = 1$, the corresponding Dirichlet series is $\mathcal{D}[1](s) = \zeta(s)$. (We will often simply write $\mathcal{D}[f]$ for the function taking s to $\mathcal{D}[f](s)$, so that $\mathcal{D}[1] = \zeta$.)

There is a natural notion of addition on arithmetic functions: given two function f and g, define $(f + g)(n)$ to be $f(n) + g(n)$. Dirichlet series are well-behaved with respect to this operation:

$$\mathcal{D}[f + g] = \mathcal{D}[f] + \mathcal{D}[g].$$

There is similarly an obvious notion of multiplication of functions: $(f \cdot g)(n) = f(n)g(n)$. However, Dirichlet series do *not* behave well in this case: $\mathcal{D}[f \cdot g]$ need not be equal to $\mathcal{D}[f] \cdot D[g]$. To see this, consider the zeta function $\zeta = \mathcal{D}[1]$. We have $1 \cdot 1 = 1$, so this would predict $\zeta^2 = \mathcal{D}[1] \cdot \mathcal{D}[1] = \mathcal{D}[1] = \zeta$. At any point s, this yields $\zeta(s)^2 = \zeta(s)$ and so $\zeta(s)^2 - \zeta(s) = \zeta(s)(\zeta(s) - 1) = 0$, which implies that either $\zeta(s) = 0$ or $\zeta(s) = 1$. But certainly this is not true for every s: $\zeta(2) = 1 + \frac{1}{4} + \frac{1}{9} + \cdots > 1 + \frac{1}{4}$, so neither of these is possible.

Instead, we will try to find another notion of multiplication of arithmetic functions, which we will write $f * g$, such that $\mathcal{D}[f * g] = \mathcal{D}[f] \cdot \mathcal{D}[g]$. The correct notion turns out to be as follows.

Definition 1.3.2. The *Dirichlet convolution* of two arithmetic functions f and g is given by

$$(f * g)(n) = \sum_{d|n} f(d)g(n/d),$$

where the summation is over all divisors d of n; the notation $d|n$ will frequently be used for "d divides n." The divisors of n include 1 and n.

It is not immediately obvious that this makes sense as a form of multiplication. We'll verify some nice properties in a moment, but let's first compute an example to get a better sense for this sort of object.

Example 1.3.3. Consider the constant function 1. We can compute

$$(1 * 1)(n) = \sum_{d|n} 1$$

is the number of divisors of n, written $\tau(n)$. It might be surprising that 1 is not the multiplicative identity[11] under Dirichlet convolution, as it is with "normal" multiplication; instead,

[11] That is, the element e such that for any f we have $e \cdot f = f \cdot e = f$ for some notion $\cdot$ of multiplication.

here the multiplicative identity is the function ϵ defined by $\epsilon(1) = 1$ and $\epsilon(n) = 0$ for every $n > 1$. To see this, for any arithmetic function f we can compute

$$(\epsilon * f)(n) = \sum_{d|n} \epsilon(d) f(n/d) = f(n)$$

and

$$(f * \epsilon)(n) = \sum_{d|n} f(d)\epsilon(n/d) = f(n).$$

We say that any operation $a \star b$ (for example, $\star$ could be $+$, $\times$, or $*$) is commutative if $a \star b = b \star a$ for every a, b, and that it is associative if for any a, b, c we have $a \star (b \star c) = (a \star b) \star c$. We want our multiplication to satisfy both of these properties.[12]

Proposition 1.3.4. *Dirichlet convolution satisfies the following properties:*

a) It is commutative.

b) It is associative.

c) It distributes over addition: for any arithmetic functions f, g, h, we have $f * (g + h) = f * g + f * h$.

d) It distributes over scalar multiplication: for any complex numbers a, b and arithmetic functions f, g, we have $(a \cdot f) * (b \cdot g) = (a \cdot b) \cdot (f * g)$.

Proof. The proof of a) follows immediately from noting that in the sum for $(f * g)(n)$ each of f and g loop over all of the divisors of f and g but in opposite directions. Thus reversing both of them does not change the sum. More formally:

$$(f * g)(n) = \sum_{d|n} f(d)g(n/d) = \sum_{(n/d)|n} f(n/d)g(d) = \sum_{d|n} f(n/d)g(d) = (g * f)(n)$$

since summing over those integers n/d dividing n is equivalent to summing over divisors d of n.

The proof of b) is a tedious expansion and change of variables, so we will skip it; the reader is welcome to verify it for themselves should they so desire.

Distribution over addition is immediate:

$$(f*(g+h))(n) = \sum_{d|n} f(d)(g+h)(n/d) = \sum_{d|n} f(d)g(n/d) + \sum_{d|n} f(d)h(n/d) = (f*g)(n) + (f*h)(n).$$

The same argument, collecting scalar multiples, shows d). □

The most fundamental property of Dirichlet convolution is the following.

[12] Usually we want anything that we are thinking of as a form of addition to be both associative and commutative. Things we think of as multiplication, such as Dirichlet convolution, can in some cases be noncommutative, but usually it's easier to work with them when they are commutative; and with a few exceptions any self-respecting multiplication should be associative.

Proposition 1.3.5. *For any two arithmetic functions f and g, we have*

$$\mathcal{D}[f * g] = \mathcal{D}[f] \cdot \mathcal{D}[g].$$

Proof. This is pure algebraic manipulation:

$$\mathcal{D}[f * g](s) = \sum_{k=1}^{\infty} \frac{1}{k^s} \sum_{d|k} f(d)g(k/d) = \sum_{d=1}^{\infty} \sum_{k:d|k} \frac{f(d)g(k/d)}{k^s}$$

by switching the order of the summation,[13] where the second sum is now over all positive integers k such that $d|k$. Writing $k = dm$, this is equivalent to

$$\sum_{d=1}^{\infty} \sum_{m=1}^{\infty} \frac{f(d)g(m)}{(dm)^s} = \sum_{d=1}^{\infty} \frac{f(d)}{d^s} \sum_{m=1}^{\infty} \frac{g(m)}{m^s} = \sum_{d=1}^{\infty} \frac{f(d)}{d^s} \mathcal{D}[g](s) = \mathcal{D}[f](s) \cdot \mathcal{D}[g](s).$$

This yields the result. □

Remark. This may not make sense on a first reading (or at all) and is a bit out of our way, but it is interesting to note that in fact more is true. In combination with our observation above that Dirichlet series respect addition, this shows that there is a ring homomorphism from arithmetic functions, with addition as usual and multiplication given by Dirichlet convolution, and the ring of Dirichlet series with standard addition and multiplication.[14] It is not difficult to show that in fact this is an isomorphism,[15] which gives another proof of Proposition 1.3.4, since all of the corresponding assertions are easy in the ring of Dirichlet series.

Remark. Note that in everything so far we have not been worrying about the convergence of the Dirichlet series but treating them "formally," i.e. as a sum of symbols whose meaning we are not concerned with. By the same methods as in the proof of Proposition 1.3.1 it can be shown that either a Dirichlet series converges nowhere or there exists some real number σ such that the series converges whenever the real part of s is greater than σ. This number σ is called the abscissa of convergence of the series; in general for series we are concerned with it will usually be at most 1, so in the region $\mathrm{Re}(s) > 0$ the above results hold analytically over the complex numbers $\mathbb{C}$ as well as formally.

[13] We will frequently do this sort of thing: given a double sum, or a sum contained within an integral or similar, we will often want to switch the order of the two. Strictly speaking, doing so is *not* justified in general, and in some cases can give explicitly false results. However, it is valid whenever the inner sum (or integral) converges *absolutely*, i.e. converges even when we take the sum of the absolute values of the terms. It is not difficult (and a decent exercise) to check that the proof of Proposition 1.3.1 also shows that in fact the sum converges absolutely. We will usually gloss over this sort of concern, and the reader is left to justify to their satisfaction that this is in fact justified whenever used; however when the justification is more subtle we will mention it explicitly.

[14] A ring is a structure with some form of addition and multiplication; a ring homomorphism is a function respecting these structures, i.e. $f(a + b) = f(a) + f(b)$ and $f(ab) = f(a)f(b)$, with the homomorphism f here given by the function taking an arithmetic function g to $\mathcal{D}[g]$.

[15] An isomorphism is an invertible homomorphism, and acts as an equivalence for many algebraic structures, including rings. Thus if a suitable statement is true in one ring it is true in any isomorphic ring, i.e. a ring with an isomorphism between it and the original.

We introduced this section by claiming that there were two reasons that we care about the Riemann zeta function. The first, as above, is that it is an important example of a Dirichlet series. The second is that it has deep connections to the prime numbers.

Proposition 1.3.8. *For any complex s with $\mathrm{Re}(s) > 1$ we have*

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1},$$

where the product is taken over all prime numbers p.

This is called the product formula, and it is essential to our understanding of the distribution of the prime numbers. Indeed it is the primary way, at least for our purposes, in which the prime numbers are related to any analytic structure, and we will derive most of our results about the primes from it.

Proof of Proposition 1.3.8. We have

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$$

for any x with absolute value less than 1 from calculus. Choosing $x = \frac{1}{p^s}$, which as above has absolute value less than 1 for any s with real part greater than 0 and so certainly in our case, gives

$$\prod_p \left(1 - \frac{1}{p^s}\right)^{-1} = \prod_p \left(1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots\right).$$

Let's try expanding this product.

If it was a finite product, say $(1 + p^{-s})(1 + q^{-s})$ for two primes p and q, we could expand this out by adding together the product of the first term from each, the product of the first term of the first factor with the second term of the second factor, and so on: $(1 + p^{-s})(1 + q^{-s}) = 1 + q^{-s} + p^{-s} + p^{-s}q^{-s}$. Let's try the same thing here: taking the first term of every factor gives 1. Taking the second term of one factor and the first factor of every other term gives p^{-s} for some prime p, so we sum over all the primes p to get every such term. Continuing on, each term corresponds to some finite collection S of primes p, where to each prime p we take the k_pth term for some integer $k_p > 0$, and we take the 0th term 1 of every other prime. If we call 1 the 0th term, p^{-s} the first term, and so on, then the corresponding term is

$$\prod_{p \in S} p^{-k_p s}.$$

Now, by the fundamental theorem of arithmetic, every finite set S of primes p with integers $k_p > 0$ for each prime p correspond exactly to the natural numbers: each natural number n has exactly one representation as

$$\prod_{p \in S} p^{k_p},$$

and every such representation is clearly a natural number. Therefore each term is n^{-s} for a unique integer n, and so expanding the product we get

$$\prod_p \left(1 - \frac{1}{p^s}\right)^{-1} = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

wherever the sum converges. In combination with Proposition 1.3.1 this gives the result. □

The proof shows that we can view Proposition 1.3.8 as a formulaic encapsulation of the fundamental theorem of arithmetic, and indeed this is exactly how we transform it into an analytic statement.

We should expect from this result that analytic information about $\zeta(s)$ should give us arithmetic information about the distribution of primes. The remainder of our work is figuring out how.

1.4 The von Mangoldt function

We can now introduce a "prime-detecting" function that we will use to indirectly count primes.[16]

Definition 1.4.1. The *von Mangoldt function* $\Lambda(n)$ is an arithmetic function defined to be $\log p$ if n is a prime power $n = p^k$ and 0 otherwise, i.e. $\Lambda(n)$ detects prime powers and counts all powers of a given prime equally.

Proposition 1.4.2. *We have* $\Lambda * 1 = \log$.

Proof. For any $n \in N$, we have

$$(\Lambda * 1)(n) = \sum_{d|n} \Lambda(d).$$

If d is not a prime power, then $\Lambda(d) = 0$, so we can restrict the sum to the divisors of n which are prime powers. For each prime p dividing n, examine all the powers p^k of p dividing n. For each, we add one term of $\Lambda(p^k) = \log p$, so if p divides n m times, i.e. p^k divides n and p^{k+1} does not divide n, then the terms of the sum coming from powers of p is $m \log p = \log(p^m)$. Writing $n = p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r}$ as the unique prime factorization of n (up to order), we see that the total sum is then $\log(p_1^{m_1}) + \cdots + \log(p_r^{m_r}) = \log(p_1^{m_1} \cdots p_r^{m_r}) = \log n$. Thus $(\Lambda * 1)(n) = \log n$ for every n, and so $\Lambda * 1 = \log$. □

[16] In this section we will freely use the properties of logarithms, as well as some calculus. The relevant properties of logarithms are as follows: $\log(xy) = \log x + \log y$, $\log(x^y) = y \log x$, and the Taylor series of $-\log(1-x)$ is

$$-\log(1-x) = \sum_{k=1}^{\infty} \frac{x^k}{k}$$

for $|1 - x| < 1$. The relevant calculus is largely in the form of some explicit computations and as such can be verified with the help of a computer (e.g. WolframAlpha), though it may also be useful to review some concepts such as Taylor series.

This also encodes the fundamental theorem of arithmetic in an equation, here in the setting of arithmetic functions rather than the zeta function. Recall that the zeta function is a Dirichlet series, and arithmetic functions are naturally related to Dirichlet series: in fact it will turn out that this formulation is essentially equivalent to that of Proposition 1.3.8.

Proposition 1.4.3. *The Dirichlet series of Λ is*

$$\mathcal{D}[\Lambda] = -\frac{\zeta'}{\zeta}$$

where $\zeta'(s)$ is the derivative of $\zeta(s)$ with respect to s.

Proof. This is equivalent to the claim that $\zeta \cdot \mathcal{D}[\Lambda] = \mathcal{D}[1] \cdot \mathcal{D}[\Lambda] = -\zeta'$, which by Propositions 1.3.5 and 1.4.2 is equivalent to $\mathcal{D}[1 * \Lambda] = \mathcal{D}[\log] = -\zeta'$. Now

$$-\zeta'(s) = -\frac{d}{ds}\zeta(s) = -\frac{d}{ds}\sum_{k=1}^{\infty}\frac{1}{k^s} = -\sum_{k=1}^{\infty}\frac{d}{ds}\frac{1}{k^s} = \sum_{k=1}^{\infty}\frac{\log k}{k^s} = \mathcal{D}[\log],$$

which by the above gives the result. □

Remark. With the machinery we've developed, this is the easiest way to prove this result: a direct application of our previous results plus some calculus. However, we could also have worked directly with the product formula. First, we can verify the calculus identity

$$\frac{d}{ds}\log f(s) = \frac{f'(s)}{f(s)}$$

using the chain rule for any differentiable function $f(s)$ wherever $f(s) \neq 0$, where $f'(s) = \frac{d}{ds}f(s)$. Thus

$$-\frac{\zeta'(s)}{\zeta(s)} = -\frac{d}{ds}\log\zeta(s) = \frac{d}{ds}\sum_{p}\log\left(1-\frac{1}{p^s}\right) = \sum_{p}\frac{d}{ds}\log\left(1-\frac{1}{p^s}\right) = \sum_{p}\frac{\log p}{p^s-1}.$$

We have $\frac{1}{p^s-1} = \frac{p^{-s}}{1-p^{-s}} = p^{-s}\sum_{k=0}^{\infty}p^{-ks}$, so

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_{p}\frac{\log p}{p^s}\sum_{k=0}^{\infty}p^{-ks} = \sum_{p}\sum_{k=0}^{\infty}\frac{\log p}{p^{(k+1)s}} = \sum_{p}\sum_{k=1}^{\infty}\frac{\log p}{p^{ks}}.$$

For prime powers $n = p^k$, we have

$$\frac{\log p}{p^{ks}} = \frac{\Lambda(n)}{n^s}$$

and so we conclude

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_{n=1}^{\infty}\frac{\Lambda(n)}{n^s} = \mathcal{D}[\Lambda].$$

Like the prime indicator function $1_{\mathcal{P}}(n)$, which is 1 for prime n and 0 otherwise, Λ is a prime-detecting function. Thus Proposition 1.4.3 gives us an equation with an arithmetic, prime-detecting object $\mathcal{D}[\Lambda]$ on one side and a purely analytic object $-\frac{\zeta'}{\zeta}$ on the other. Thus we can hope to be able to use analysis of the Riemann zeta function ζ to extract information about the primes, and this is what we will do in the next section.

1.5 Prime-counting functions

In the previous sections, we saw how arithmetic information is encoded by the Riemann zeta function. In particular, Proposition 1.4.3 directly relates the prime-detecting function Λ to an analytic transform of the zeta function. In this section, we will see how to use this connection to turn information about the zeta function into information about the primes, and in particular an estimation for $\pi(x)$.

Let $f : \mathbb{N} \to \mathbb{C}$ be any arithmetic function, and define

$$F(x) = \sum_{k \le x} f(n),$$

where the sum is over all positive integers k less than or equal to x. (Note that we do not require that x be an integer for this definition.)

Proposition 1.5.1. *Whenever both sides converge, we have*

$$\mathcal{D}[f](s) = s \int_1^\infty \frac{F(x)}{x^{s+1}}\,dx.$$

Proof. Expanding the definition of $F(x)$, the right-hand side is

$$s \int_1^\infty \sum_{k \le x} \frac{f(k)}{x^{s+1}}\,dx.$$

Since integration is linear[17], it distributes over finite sums. We can think of the sum over $k \le x$ as the sum over all positive integers k with the condition that if $k > x$ then we multiply by 0: defining $[k \le x]$ to be the function of k and x defined by $[k \le x] = 1$ if $k \le x$ and 0 otherwise[18] the above is

$$s \int_1^\infty \sum_{k=1}^\infty \frac{f(k)}{x^{s+1}}[k \le x]\,dx = s \sum_{k=1}^\infty f(k) \int_1^\infty \frac{[k \le x]}{x^{s+1}}\,dx = s \sum_{k=1}^\infty f(k) \int_k^\infty \frac{1}{x^{s+1}}\,dx$$

since if $x < k$ then the integrand vanishes and if $x \ge k$ then it is just $\frac{1}{x^{s+1}}$. We can compute this integral by standard calculus methods[19] to get that this is

$$s \sum_{k=1}^\infty f(k) \frac{1}{sk^s} = \sum_{k=1}^\infty \frac{f(k)}{k^s} = \mathcal{D}[f](s)$$

as claimed. □

[17] That is,

$$\int_a^b f(x) + g(x)\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$$

for any bounds a, b and any integrable functions f, g.

[18] This is a more general notation called the Iverson bracket: for any statement, such as $k \le x$ or $yz = 3$, encasing it in brackets denotes 1 if the statement is true and 0 if it is false.

[19] Or by plugging into WolframAlpha.

This isn't obviously helpful, since this integral expression doesn't necessarily seem any easier to deal with than the sum. The key lies in the fact that although $f(x)$ may behave very strangely—for example, $f = \Lambda$ is pretty unpredictable without doing some relatively hard computations—in many cases of interest, the summatory function $F(x)$ is often much better-behaved. Here's an example, which will be quite useful to us in the future. For any real number x, write $\lfloor x \rfloor$ for the integer part of x, i.e. the greatest integer less than or equal to x, and $\{x\}$ for the fractional part of x, i.e. $x - \lfloor x \rfloor$. For example, $\lfloor \pi \rfloor = 3$, $\lfloor -\pi \rfloor = -4$, and $\{-\pi\} = 4 - \pi \approx 0.858$. Note that for any real x we have $0 \leq \{x\} < 1$.

Proposition 1.5.2. *For all s with $\operatorname{Re} s > 1$, we have*

$$\zeta(s) = \frac{s}{s-1} - s \int_1^\infty \frac{\{x\}}{x^{s+1}}\, dx.$$

Proof. This is a straightforward application of Proposition 1.5.1. Since $\zeta = \mathcal{D}[1]$ and

$$\sum_{n \leq x} 1 = \lfloor x \rfloor$$

for positive real x, we have

$$\zeta(s) = s \int_1^\infty \frac{\lfloor x \rfloor}{x^{s+1}}\, dx.$$

Now $\lfloor x \rfloor = x - \{x\}$, so

$$\zeta(s) = s \int_1^\infty \frac{x}{x^{s+1}}\, dx - s \int_1^\infty \frac{\{x\}}{x^{s+1}}\, dx.$$

The first term is

$$s \int_1^\infty \frac{1}{x^s}\, dx = \frac{s}{s-1},$$

and combining the two terms the result follows immediately. □

This result may or may not immediately look exciting, but it should. Why? Consider the integral

$$\int_1^\infty \frac{\{x\}}{x^{s+1}}\, dx.$$

Since $\{x\}$ is between 0 and 1, this integral is bounded below by 0 and above by

$$\int_1^\infty \frac{1}{x^{s+1}}\, dx = \frac{1}{s},$$

which converges for every s with $\operatorname{Re} s > 0$. Since every term of this integral is positive, it follows that the integral converges for every s with $\operatorname{Re} s > 0$.

Why is this exciting? Because now for any such s (other than $s = 1$) we can plug it into this formula and obtain a value for $\zeta(s)$. We have thus defined an *analytic continuation* of $\zeta(s)$ to the region $\operatorname{Re} s > 0$ except at $s = 1$; the theory of analytic continuations tells us that any such differentiable continuation must be unique, i.e. all such continuations are the same. Thus we can really say that we have extended the definition of $\zeta(s)$ to this region.

Here's another reason that Proposition 1.5.1 is useful. Suppose that we have some interesting arithmetic function $f(x)$, and we want to estimate its summatory function $F(x)$. If we can show that $\mathcal{D}[f](s)$ is close to

$$s\int_1^\infty \frac{G(x)}{x^{s+1}}\,dx$$

for some well-behaved function $G(x)$, then we can conclude that $F(x)$ is close to $G(x)$. To make sense of this, we'll need some notation to explain what we mean by "close."

Let $f(x)$ and $g(x)$ be two real- or complex-valued functions, and fix a point x_0, which may be (and often is) ∞. We say that $f(x)$ is in $O(g(x))$, which may also be written $f(x) = O(g(x))$,[20] if there exists a positive real constant C such that as x approaches x_0 we have $|f(x)| \leq C|g(x)|$. This should be thought of as roughly "f is of order at most that of g," i.e. f increases asymptotically no faster than g.

Example 1.5.3. For any such function f, we have $f(x)$ in $O(cf(x))$ for any constant c for any limit x_0. In particular, $f(x) = O(-f(x))$, and generally$-O(f(x))$ and $O(f(x))$ denote the same thing.

Example 1.5.4. For any two integers a, b, we have x^a in $O(x^b)$ if and only if $a \leq b$ as $x \to \infty$.

Example 1.5.5. For any integer x, we have x^a in $O(e^x)$, and e^x is not in $O(x^a)$ as $x \to \infty$.

Example 1.5.6. We have $\sin(x)$ in $O(1)$ over the real numbers, for any limit x_0. In fact the same is true for any bounded function.

In general when we use this notation, known as "big-O notation," we will take the limit x_0 to be positive infinity unless stated otherwise.

A similar notation is "little-o notation," defined as above except that $f(x)$ is in $O(g(x))$ if for any positive real constant C for all x sufficiently close to x_0 we have $|f(x)| \leq C|g(x)|$. This should be thought of as "the order of f is strictly smaller than the order of g." For example, $f(x) = o(1)$ as $x \to x_0$ simply means that $f(x)$ goes to 0 as $x \to x_0$. We write $f(x) \sim g(x)$ as $x \to x_0$ if $f(x) = g(x) + o(g(x))$ as $x \to x_0$, and again assume that x_0 is positive infinity unless stated otherwise. (There are a number of analogous notations, but these are the most useful ones.)

We can now give an example of the philosophy mentioned above.

Example 1.5.7. Recall that the divisor function

$$\tau(n) = \sum_{d|n} 1$$

is equal to the Dirichlet convolution of 1 with itself. Therefore by Proposition 1.3.5 we have $\mathcal{D}[\tau] = \zeta^2$. Let

$$D(x) = \sum_{n \leq x} \tau(n)$$

[20] This is somewhat unfortunate notation, since one might then think that $O(g(x))$ and $f(x)$ are equal in some sense, which is very much false; $O(g(x))$ is really a class of functions. It is however very convenient notation.

be the summatory function of τ. Then it follows from Proposition 1.5.1 that

$$\zeta(s)^2 = s\int_1^\infty \frac{D(x)}{x^{s+1}}\,dx.$$

On the other hand by Proposition 1.5.2 we have

$$s\int_1^\infty \frac{D(x)}{x^{s+1}}\,dx = \zeta(s)^2 = \left(\frac{s}{s-1} - \int_1^\infty \frac{\{x\}}{x^{s+1}}\,dx\right)^2.$$

Since the integral part of the latter expression converges for $\operatorname{Re} s > 0$, as noted above, as $s \to 1$ the integral approaches some constant c_0; we can therefore say that as $s \to 1$ we have

$$\zeta(s) = \frac{s}{s-1} + c_0 + o(1),$$

and indeed since $\frac{s}{s-1} = \frac{1}{s-1} + 1$ we can define $c = c_0 + 1$ to get

$$\zeta(s)^2 = \left(\frac{1}{s-1} + c + o(1)\right)^2 = \frac{1}{(s-1)^2} + \frac{2c}{s-1} + o\left(\frac{1}{s-1}\right)$$

as $s \to 1$. One can check[21] that

$$s\int_1^\infty \frac{x\log x}{x^{s+1}}\,dx = \frac{s}{(s-1)^2} = \frac{1}{(s-1)^2} + \frac{1}{s-1}$$

as $s \to 1$. Therefore

$$s\int_1^\infty \frac{D(x) - x\log x}{x^{s+1}}\,dx = \frac{2c-1}{s-1} + o\left(\frac{1}{s-1}\right)$$

as $s \to 1$. We know that

$$s\int_1^\infty \frac{x}{x^{s+1}}\,dx = \frac{1}{s-1} + 1 = \frac{1}{s-1} + o\left(\frac{1}{s-1}\right)$$

since a constant function is certainly bounded by $\frac{1}{s-1}$ as $s \to 1$, since the latter goes to infinity, so we conclude that

$$s\int_1^\infty \frac{D(x) - x\log x - (2c-1)x}{x^{s+1}}\,dx = o\left(\frac{1}{s-1}\right).$$

We would like to conclude that $D(x) - x\log x - (2c-1)x = o(x)$, and therefore $D(x) = x\log x + (2c-1)x + o(x)$, and this is in fact true. Later, we will introduce a tool that will permit this conclusion; for now, we cannot prove this with the tools we have at our disposal. Nevertheless this gives us a way to at least guess at formulas it would otherwise be hard to develop intuition for: here we have suggested the formula

$$\frac{1}{N}\sum_{n\le N} \tau(n) = \log x + 2c - 1 + o(1),$$

which predicts the "average value" of $\tau(n)$. (The constant c turns out to be the Euler-Mascheroni constant $\gamma \approx 0.5772$, and is important throughout analytic number theory.)

[21] Again, either by calculus (here integration by parts) or WolframAlpha.

Recall that $\pi(x)$ is the prime-counting function

$$\pi(x) \sum_{p \leq x} 1$$

where the summation over p denotes the sum over all primes less than or equal to p. Writing $1_{\mathcal{P}}(x)$ for the indicator function on the primes, given by $1_{\mathcal{P}}(x) = 1$ if x is prime and 0 otherwise, we could connect $\mathcal{D}[1_{\mathcal{P}}]$ with $\pi(x)$ using Proposition 1.5.1. However, $\mathcal{D}[1_{\mathcal{P}}]$ turns out not to be very tractable, though it has some interesting properties. Instead, we want to use the prime-detecting function we have defined which we know has nice properties for our purposes: the von Mangoldt function $\Lambda(x)$. We denote by $\psi(x)$ its summatory function

$$\psi(x) = \sum_{n \leq x} \Lambda(n).$$

We asserted earlier that Λ is a prime-detecting function, and in a sense this is obviously true. For the purpose of prime-counting, we would like a way of determining the behavior of $\pi(x)$ based on that of $\psi(x)$, and indeed we have one.

Proposition 1.5.8. *As $x \to \infty$ we have*

$$\psi(x) = \pi(x) \log x + o(x).$$

Proof. Fix a prime $p \leq x$. We want to know what the contribution of powers of p is to the sum defining $\psi(x)$. Suppose that $p^k \leq x$ but $p^{k+1} > x$. Then there are k powers of p less than or equal to x, each of which contributes $\log p$ to the sum, so the total contribution is $k \log p$. On the other hand $k = \lfloor \log_p x \rfloor = \left\lfloor \frac{\log x}{\log p} \right\rfloor$, so

$$\psi(x) = \sum_{p \leq x} \left\lfloor \frac{\log x}{\log p} \right\rfloor \log p$$

and so

$$\pi(x) \log x - \psi(x) = \sum_{p \leq x} \left(\frac{\log x}{\log p} - \left\lfloor \frac{\log x}{\log p} \right\rfloor \right) \log p.$$

Fix some small constant $\epsilon > 0$. We can split this sum into two halves:

$$\pi(x) \log x - \psi(x) = \sum_{p \leq x^{1-\epsilon}} \left(\frac{\log x}{\log p} - \left\lfloor \frac{\log x}{\log p} \right\rfloor \right) \log p + \sum_{x^{1-\epsilon} < p \leq x} \left(\frac{\log x}{\log p} - \left\lfloor \frac{\log x}{\log p} \right\rfloor \right) \log p.$$

For the first sum, note that $\left\lfloor \frac{\log x}{\log p} \right\rfloor = \frac{\log x}{\log p} + O(1)$, so the sum is

$$O\left(\sum_{p \leq x^{1-\epsilon}} \log p \right) = O(\pi(x^{1-\epsilon}) \log(x^{1-\epsilon})) = o(x),$$

where we use the trivial bound $\pi(x) \leq x$ and the fact that $\log x = o(x^\epsilon)$ for any $\epsilon > 0$.[22] For the second sum, note that $(1-\epsilon)\log x < \log p \leq \log x$ and so $1 \leq \frac{\log x}{\log p} < \frac{1}{1-\epsilon}$, while for $\epsilon < \frac{1}{2}$ we have $\left\lfloor \frac{\log x}{\log p} \right\rfloor = 1$. Thus the second sum is bounded between 0 and

$$\frac{\epsilon}{1-\epsilon}\pi(x)\log x.$$

We can choose ϵ to depend on x, so if we let it be any decreasing function of x we conclude that this sum is also $o(x)$, so $\pi(x)\log x - \psi(x) = o(x)$ and the result follows. □

From Propositions 1.4.3 and 1.5.1, we know that

$$-\frac{\zeta'(s)}{\zeta(s)} = s\int_1^\infty \frac{\psi(x)}{x^{s+1}}\,dx.$$

From the discussion above, we know that near $s = 1$ we have

$$\zeta(s) = \frac{1}{s-1} + O(1),$$

so we might expect that near $s = 1$ the derivative of $\zeta(s)$ is close to that of $\frac{1}{s-1}$, i.e. $\zeta'(s) \approx \frac{d}{ds}\frac{1}{s-1} = -\frac{1}{(s-1)^2}$, which would then imply that

$$-\frac{\zeta'(s)}{\zeta(s)} \approx \frac{1}{s-1}.$$

From our analysis of the zeta function itself, we know that this is close to $\zeta(s)$, so this suggests that $\psi(x)$ should be close to $\lfloor x \rfloor = x + O(1)$. This motivates the following theorem.

Theorem 1.5.9 (Prime number theorem)**.** *As $x \to \infty$ we have*

$$\psi(x) \sim x,$$

or equivalently

$$\pi(x) \sim \frac{x}{\log x}.$$

The equivalence of these two statements follows from Proposition 1.5.8.

The proof of this theorem will be our first major goal, and in fact we are already fairly well-positioned to prove it. What remains is a more formal analysis of $-\frac{\zeta'}{\zeta}$ and the missing ingredient mentioned in Example 1.5.7, which will allow us to turn the prime number theorem into a complex-analytic statement; and finally we'll develop some analytic tools to prove that statement directly.

[22] This is equivalent to seeing that $x^a = o(e^x)$ for any real a, which is a good exercise to do elementarily; another way to see this is via the power series

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots,$$

which shows that for any x^a a higher polynomial x^{a+1} also has nontrivial contribution to the series.

1.6 An introduction to complex analysis

Before we can prove Theorem 1.5.9, we will need first to better understand the Riemann zeta function and second to be able to apply the tools of complex analysis to transfer this understanding to information about the primes. As it turns out, the first of these goals also requires substantial complex analysis, so we'll begin there. As we are not so much concerned with complex analysis for its own sake, we will skip the proofs of most results and focus on conceptual understanding.

It should also be noted that there exists a significantly easier version of the proof of the prime number theorem using much less analysis, due to Newman. However, it does not yield the explicit formula (Theorem 1.8.3) and I think is less conducive to understanding what is really happening.

In calculus, we are typically concerned with function $f : \mathbb{R} \to \mathbb{R}$ which take in a real number and spit out another real number. We can then attempt to differentiate and integrate these functions.

We will mostly be concerned with integration, and so will skip a bunch of material that otherwise is important for complex analysis. However, one very important distinction when working over the complex numbers as opposed to the reals is as follows: if we have a real function $f : \mathbb{R} \to \mathbb{R}$, we can attempt to differentiate it, and we say that this is well-behaved, or that f is differentiable, whenever the defining limit

$$f'(x) = \lim_{y \to x} \frac{f(x) - f(y)}{x - y}$$

converges, i.e. we get the same real number taking y to x either from above or below. For example, the absolute value function $f(x) = |x|$ is differentiable everywhere except at $x = 0$: for $x > 0$, the function is given by $f(x) = x$ and so its derivative is well-defined, and for $x < 0$ it is given by $f(x) = -x$ and so is similarly well-behaved; but at $x = 0$, if we take the limit from above we get

$$f'(0) = \lim_{y \to 0^+} \frac{f(y) - f(0)}{y - 0} = \frac{y}{y} = 1,$$

while if we take the limit from below we get

$$\lim_{y \to 0^-} \frac{f(y) - f(0)}{y - 1} = \frac{-y}{y} = -1,$$

since in the first case $f(y) = y$ and in the second $f(y) = -y$. In the region where $f(x)$ is differentiable (for any function f) we can then look at $f'(x)$ and ask where it is differentiable, and so on, so on a given region we can meaningfully ask how many times a function f is differentiable—possibly infinitely many! In particular, most function that we typically work with—polynomials, exponential functions, trigonometric functions—are infinitely differentiable everywhere. This is because they are analytic functions: functions which can be defined (on some region, say near x_0) by a power series

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n.$$

Taylor's theorem tells us that $a_n = \frac{f^{(n)}(x_0)}{n!}$ where $f^{(n)}$ is the nth derivative of f, so any function which is analytic on a certain region must be infinitely differentiable on that region. For example, the nth derivative of $f(x) = e^x$ is e^x, and so taking the Taylor series about $x_0 = 0$ gives $f^{(n)}(0) = e^0 = 1$ and so

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

This converges everywhere, so the exponential function is analytic everywhere. Similarly polynomials are analytic everywhere, since their higher derivatives are all 0; in particular, a polynomial is its own Taylor series.

In the complex world, much of this still holds true: we still have differentiable and analytic functions, we still have to worry about where they are differentiable and analytic, and the same formulas hold. However, the major difference is that on a given region, if a complex function $f : \mathbb{C} \to \mathbb{C}$ is differentiable (once!), then it is analytic (and therefore infinitely differentiable). (In the complex world, we will refer to analytic functions as "holomorphic."

We won't prove this, but here's some intuition about why this should be true: recall that we said that a function f is differentiable at x if

$$\lim_{y \to x} \frac{f(x) - f(y)}{x - y}$$

converges to the same value for every way of taking $y \to x$. When working over the reals, there are essentially only two ways of taking this limit: from above or from below. But over the complex numbers, given a point x in the complex plane there are infinitely many ways of approaching it: we can come from any angle! Thus we see that this is a much stronger condition in the complex-analytic world, so it should not be too surprising that it has much stronger consequences.

Now that we've talked a bit about what differentiation means, let's think about what integration is. In the real world, an integral is defined by a function $f(x)$ and two real numbers a and b, and then we can write down the integral

$$\int_a^b f(x)\, dx.$$

If we want to be really fancy, we can take one or both of a and b to be positive or negative infinity by taking the limit of the integral as whichever term goes to infinity.

In the complex plane, things are similar: an integral requires two points a and b in the complex plane and a function $f : \mathbb{C} \to \mathbb{C}$, and then we can write down the integral

$$\int_a^b f(x)\, dx.$$

This works well enough, so long as f is holomorphic everywhere. But if not, we can have some problems.

When we're just given two points in the complex plane and asked to take the integral between them, it's natural to draw a straight line between them and integrate over that.

But there's no real reason we had to pick that particular path other than simplicity: what if we took a different path? If this was a real integral, this would be a somewhat odd idea: we're going from a to b, and there's really only one path—we could double back and then go forward again, but the terms from this would cancel and we'd be left with the same thing. It would be nice if the same thing was true for complex integrals: no matter what path you take, you get the same result. This turns out to be true if f is holomorphic everywhere, but not otherwise.

To see why, consider the function $f(x) = \frac{1}{x}$. This is holomorphic for $x \neq 0$, with derivative $-\frac{1}{x^2}$, but has a pole at $x = 0$ and so cannot be holomorphic there.

Let's consider two complex integrals, both from 1 to 1. The first of these takes the obvious path: it stays at 1, and so the integral is 0, much as a real integral $\int_1^1 \frac{1}{x}\,dx$ would be 0.

For our second integral, we'll move in a counterclockwise circle around the origin, along the unit circle. We can parametrize this by integrating with respect to a parameter t, and defining the path that we'll integrate along by $x(t) = e^{it}$; by Euler's formula as t goes from 0 to 2π this traces a counterclockwise circle as desired, beginning and ending at $e^0 = e^{2\pi i} = 1$. If $x = e^{it}$, then $dx = d(e^{it}) = ie^{it}dt$, so our integral is

$$\int_0^{2\pi} \frac{1}{e^{it}}\,dx = \int_0^{2\pi} e^{-it} i e^{it}\,dt = \int_0^{2\pi} i\,dt = 2\pi i.$$

Since this is nonzero, we see that our choice of path matters.

Thus instead of integrating between two points, we'll usually think of complex integrals as along some path. Given a path γ, i.e. a curve in the complex plane, we'll write the integral as

$$\int_\gamma f(x)\,dx.$$

If γ is a closed curve[23], as will often be the case, then we'll use the symbol $\oint$ to make this clear:

$$\oint_\gamma f(x)\,dx.$$

If γ is a closed path, such as for example the counterclockwise circle mentioned above, then it encloses some subset of the complex plane.[24]

To properly state the result above—that the integral over a closed curve of a holomorphic function is 0—we also need a small amount of topology. (This is not really all that necessary here, but it's good stuff to know.)

Definition 1.6.1. A *topological space* is a set X equipped with a collection τ of subsets of X satisfying the following properties:

(1) The empty set $\{\}$ is in τ;

[23]Beginning and ending at the same point.

[24]This is not quite so simple in practice, as it might enclose multiple disconnected regions, or one region "several times." To understand this, consider the path making a counterclockwise circle around the origin n times, or spiraling out from the origin, circling it n times, and then connecting back to the origin; in this case the path is said to have "winding number" n, and the enclosed region must be counted n times.

(2) The whole set X, viewed as a subset of itself, is in τ;

(3) The intersection of any finite number of subsets of τ is also in τ;

(4) The union of sets in τ (potentially infinitely or even uncountably[25] many) is also in τ.

Subsets of X in τ are called *open*. The complement of an open set U, i.e. the set of points in X which are not in U, is called *closed*. Note that it is possible for a set to be both open and closed: in particular, since the empty set $\{\}$ and the entire space X are always open, since they are each other's complement they are also both closed. Sets which are both open and closed are called *clopen* (yes, this is the actual technical terminology).

Often, topologies can be defined by a set X and a collection T of "basic open sets" which, via taking finite intersections and arbitrary unions, generate a larger collection τ of open sets for X.

Definition 1.6.2. A *metric space* is a set X together with a real-valued function $d : X \times X \to \mathbb{R}$, i.e. a function taking in a pair $(x, y) \in X \times X$ of elements of X,[26] satisfying the following properties:

(1) For every $x \in X$, we have $d(x, x) = 0$;

(2) For every $x, y \in X$, we have $d(x, y) \geq 0$, and if $x \neq y$ then $d(x, y) > 0$;

(3) For every $x, y \in X$, we have $d(x, y) = d(y, x)$;

(4) For every $x, y, z \in X$, we have $d(x, y) \leq d(x, z) + d(z, y)$.

Example 1.6.3. Consider the set of real numbers $\mathbb{R}$, and let $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be defined by $d(x, y) = |x - y|$. We can check that this satisfies all of the requirements of the definition: $|x - y|$ is always nonnegative, it is equal to 0 if and only if $x = y$, and $d(y, x) = |y - x| = |-(x - y)| = |x - y| = d(x, y)$. The only tricky property is the last one. But this is not so hard either: the distance from x to y cannot be greater than the distance from x to a third point z plus the distance from that third point z to y, since any path other than a straight line would be less efficient.

This last property is known as the triangle inequality because of this geometric interpretation: if x, y, and z are three points in the plane, then they define a triangle with corners at these points. The triangle inequality says that the length of any side is no more than the sum of the lengths of the two other sides, which from drawing some triangles should be clear (of course, this can also be proven algebraically).

Proposition 1.6.4. *There is a natural topology on a metric space, i.e. any metric space is naturally a topological space.*

[25] Countably infinite refers to a set which has a one-to-one mapping with the natural numbers, such as the prime numbers (via assigning to each natural number n the nth prime) or the rational numbers (by a more complicated mapping). Countably infinite sets are the smallest type of infinite sets; any larger set, such as the real or complex numbers, is called uncountably infinite, or simply uncountable. Proving a given set to be countable or uncountable can be highly nontrivial.

[26] Generally the notation $X \times Y$ means the set of pairs (x, y) with $x \in X$ and $y \in Y$.

Proof. We'll show this by constructing a topology. Let X be our metric space. For any point $x \in X$ and positive real number $\epsilon > 0$ we can define the open ball $B_\epsilon(X)$ to be the set of points $y \in X$ such that $d(x, y) < \epsilon$. Let T be the set of all such balls, for any x and ϵ. Then T forms a collection of basic open subsets.

More explicitly, we can define a set U to be open in X if for every $x \in U$ we can choose a positive ϵ sufficiently small that the open ball $B_\epsilon(x)$ is a subset of U. It is not too hard to see that this satisfies the definition of a topological space; one perhaps non-obvious thing is that the empty set is open, since although no open ball will be contained in it the condition "for every $x \in U$" can never be satisfied for $U = \{\}$, and so the condition for it to be open is vacuously true. □

Example 1.6.5. As in Example 1.6.3, for $\mathbb{R}$ our metric space the basic open sets are then intervals of the form (a, b), i.e. the set of real numbers x such that $a < x < b$; for any x in this interval, by zooming in sufficiently we can find a small open ball around x contained in (a, b). We can combine these together and take intersections to get all open sets; taking complements, the closed sets will be things like $[a, b]$, the interval of reals x such that $a \leq x \leq b$, as well as all finite subsets of $\mathbb{R}$ (in particular, any subset consisting of only one element is closed). Note that the only clopen subsets are $\{\}$ and $\mathbb{R}$, while there are many sets (e.g. $(a, b]$, the set of reals x with $a < x \leq b$) which are neither open nor closed.

Primarily, though, we are concerned not with the reals but with the complex numbers. Here again we have a natural metric $d(x, y) = |x - y|$, where $|x$ is the absolute value defined by $|x| = |a + bi| = \sqrt{a^2 + b^2}$ where a and b are real numbers such that $x = a + bi$. One can check that this is indeed a metric, and so by Proposition 1.6.4 we get a natural topology on the complex plane. In general one should think of open sets as regions of the complex plane with "fuzzy" boundaries, and of closed sets as ones with "hard" boundaries.

Theorem 1.6.6 (Cauchy's integral theorem)**.** *Let U be an open subset of the complex plane, and let f be a complex function holomorphic on U. Then for any*[27] *closed path γ on U we have*

$$\oint_\gamma f(x)\,dx = 0.$$

This is just a formalization of our result above, and follows from the statement that the fundamental theorem of calculus holds for holomorphic functions. However it has powerful consequences.

Theorem 1.6.7 (Cauchy's integral formula)**.** *Let U be an open subset of the complex plane, and let f be a complex function holomorphic on U. Then for any*[28] *closed path γ on U, for any a in the region bounded by γ we have*

$$f(a) = \frac{1}{2\pi i} \oint_\gamma \frac{f(x)}{x - a}\,dx.$$

[27] This is not strictly true: we must also impose the condition that γ is "rectifiable," i.e. as a curve its length is well-defined (for certain pathogenic curves this will fail). However for our purposes this will always hold.

[28] Again, we additionally require γ to be rectifiable; here we also require that it have winding number 1 so that it really does bound some region (see Footnote 24).

Proof. Since f is holomorphic on U, the only possible place within U for $\frac{f(x)}{x-a}$ to fail to be holomorphic is at $x = a$. By Cauchy's integral theorem, the integral over γ is equal to an integral over a counterclockwise circle of arbitrary radius r around a: the difference between these two integrals is an integral along a path (consisting of γ and a clockwise circle around a) whose "interior" is the interior of γ minus the interior of the circle, and on this region $\frac{f(x)}{x-a}$ is holomorphic. Therefore by Cauchy's integral theorem the difference is 0 and so the integrals are equal.

Let C be the counterclockwise circle around a with radius r. It remains only to show that

$$\int_C \frac{f(x)}{x-a}\,dx - f(a)$$

is 0. Parametrize C by $x(t) = a + re^{it}$. Write C_0 for the curve with $a = 0$ and $r = 1$; we saw above that

$$\int_{C_0} \frac{1}{x}\,dx = 2\pi i,$$

and by examining that argument it is not difficult to see that it is independent of r. If we have some arbitrary a, then

$$\int_C \frac{1}{x-a}\,dx = \int_{C_0} \frac{1}{x}\,dx = 2\pi i,$$

so

$$f(a) = \frac{1}{2\pi i}\int_C \frac{f(a)}{x-a}\,dx.$$

Therefore

$$\int_C \frac{f(x)}{x-a}\,dx - f(a) = \int_C \frac{f(x)-f(a)}{x-a}\,dx.$$

Since this is true for any r, we can take r to 0; then the limit

$$\lim_{r\to 0} \frac{f(x)-f(a)}{x-a}$$

converges for any x on the circle C of radius r about a, since, this is one of the ways of taking the limit $x \to a$ and we know that

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = \lim_{x\to a} \frac{f(a)-f(x)}{a-x} = f'(a)$$

converges since f is holomorphic. Therefore at every x as we take r to 0 the integrand converges to $f'(a)$. Thus since the integrand converges to a constant and the length of the path over which we are integrating goes to 0, the limit of

$$\int_C \frac{f(x)}{x-a}\,dx - f(a)$$

as $r \to 0$ is 0. But since this value is independent of r, it must always be 0, and so

$$\int_C \frac{f(x)}{x-a}\,dx = \int_\gamma \frac{f(x)}{x-a}\,dx = f(a).$$

□

Corollary 1.6.8. *With the hypotheses of Theorem 1.6.7, the function f is infinitely differentiable, with nth derivative at any a in the region bounded by γ given by*

$$f^{(n)}(a) = \frac{n!}{2\pi i} \oint_\gamma \frac{f(x)}{(x-a)^{n+1}}\, dx.$$

Proof sketch. This can be computed explicitly from Cauchy's integral formula for the first few n by differentiating under the integral sign; after that it's a proof by induction using the same technique. □

This is somewhat remarkable: not only have we proven our earlier assertion that every holomorphic (which, at first glance, just meant once-differentiable!) function is infinitely differentiable (and indeed the same argument shows analyticity) but also gives a formula for not only the value of f at every point in the region bounded by our curve γ, but also every derivative of f at every such point, given only the values of f along the boundary γ. The main takeaway here is that holomorphic functions are very highly constrained, and as a result very well-behaved.

We can also move away from the holomorphic case to the situation in which there are countably many singularities or poles. This is not an uncommon case, because many of the functions with which we will be dealing are meromorphic.

Definition 1.6.9. A *meromorphic function* is one which can be written as $\frac{f(x)}{g(x)}$ where f and g are holomorphic functions.

As mentioned above, holomorphic functions are generally well-behaved; meromorphic functions are the next-best one can do. In particular, holomorphic functions have the property that although they may have infinitely many zeros their zeros are discrete, i.e. the set of zeros (under the usual complex topology) has no limit points:

Definition 1.6.10. Given a topological space X containing a set of points S, we say that $x \in X$ (not necessarily in S!) is a *limit point* of S if for every open set U containing X there exists some $y \in S$ not equal to x which is also in U. More concretely, if X is a metric space with metric d then x is a limit point for S if for every $\epsilon > 0$ there exists some $y \in S$ not equal to x with $d(x, y) < \epsilon$.

Although meromorphic functions may have poles, they satisfy the same property.

Just as holomorphic functions are analytic and so can (at least locally) be written as a Taylor series, meromorphic functions can locally be written as a Laurent series, which is the same thing as a Taylor series except that it can have finitely many negative-degree terms. That is, if $f(x)$ is a meromorphic function near x_0 then (in some open set containing x_0) there exists an integer m and complex coefficients c_n such that

$$f(x) = \sum_{n=m}^{\infty} c_n (x - x_0)^n.$$

For all $n < m$, say that $c_n = 0$. Then Taylor's theorem tells us that if $n \geq 0$ then $c_n = \frac{f^{(n)}(x_0)}{n!}$; the coefficient in degree -1, c_{-1}, is called the residue of f at x_0, denoted $\mathrm{res}_f(x_0)$. For example, if $f(x) = x^n$, then the Laurent series of f about 0 is simply x^n, and so $\mathrm{res}_f(0)$ is 1 if $n = -1$ and 0 otherwise.

If f is holomorphic (or equivalently has a well-defined value) at x_0, then $\text{res}_f(x_0) = 0$, since near x_0 our function f is locally analytic, i.e. it has a Taylor series with no negative terms (we can take m to be 0) and so $c_{-1} = 0$.

Recall the example we computed with $f(x) = \frac{1}{x}$ to show that the integral along a closed path of a non-holomorphic function need not be 0. If we restrict our focus to meromorphic functions, then at every point where f fails to be holomorphic we can compute its residue, and it turns out that this information lets us complete our picture of closed path integrals of holomorphic functions.

Theorem 1.6.11 (The residue theorem)**.** *Let U be an open subset[29] of the complex plane, γ be a closed path[30] on U, and f be a complex function meromorphic on U and holomorphic on the boundary γ. Let $\{x_i\}$ be the set of points in the region enclosed by γ at which f is not holomorphic. Then*

$$\oint_\gamma f(x)\,dx = 2\pi i \sum_i \text{res}_f(x_i).$$

This is proved where there is only one x_i by writing f as a Laurent series and applying Cauchy's integral theorem to the holomorphic portion (the part coming from nonnegative degrees) and applying Cauchy's integral formula to $x - x_i$ times the non-holomorphic part. The total result is obtained by summing the result for a single singularity over all singularities.

As a check, observe that if $f(x) = \frac{1}{x}$ and γ is a counterclockwise circle about the origin then since we have seen that $\text{res}_f(0) = 1$ we have

$$\oint_\gamma \frac{1}{x}\,dx = 2\pi i$$

as previously computed.

This will be our main technique in evaluating path integrals. Note that as suggested by our previous results the residue theorem shows that closed path integrals are independent of the path taken except insofar as which singularities they include, and so we can push contours (paths) around freely so long as we do not push them through any singularities.

Integrals that we care about will often not be over closed paths; instead, our method will be to complete the curve by adding a segment connecting the endpoints on which the integral is easy to evaluate. Thus the original integral will be the difference between the formula given by Theorem 1.6.11 and the "easy" integral over the added segment.

1.7 The functional equation

With the techniques of complex analysis in hand, our next goal is to say something more substantial about the Riemann zeta function. Originally, we defined $\zeta(s)$ for s with real part

[29] In fact we also require that it be "simply connected," a term which we will not define until some theoretical future unit on topology but should be understood as a simplicity condition, which includes the idea that any two points in U should be connected by a path lying entirely in U.

[30] Again, rectifiable and with winding number 1. Note that this is different from a winding number of -1, which would again be one loop but in the opposite direction; this is for the *positive orientation*, corresponding to e.g. the counterclockwise circle.

greater than 1; then from Proposition 1.5.2 we saw that we can extend our definition to all s with real part greater than 0 except for $s = 1$. Our goal in this section is to show that $\zeta(s)$ extends to the entire complex plane, except for the pole at $s = 1$, by giving a formula relating $\zeta(1-s)$ to $\zeta(s)$, so that for s with real part less than or equal to 0 we can find $\zeta(s)$ in terms of $\zeta(1-s)$, which since $1-s$ has real part greater than or equal to 1 in this case we can find $\zeta(1-s)$ using our previous formulae.

This formula is known as the functional equation for the zeta function, and can be written in many ways; one of the most common forms is

$$\zeta(1-s) = 2^s \pi^{s-1} \sin\left(\frac{\pi}{2}s\right) \Gamma(1-s)\zeta(s),$$

which even aside from the mysterious $\Gamma(1-s)$ factor (which will be explained momentarily) is somewhat inexplicable, though it does have the virtue of giving a formula for $\zeta(1-s)$ given $\zeta(s)$. We will prove a slightly different (though equivalent) form.

First, let's define this function $\Gamma(s)$:

$$\Gamma(s) := \int_0^\infty e^{-x} x^{s-1}\, dx.$$

Why should we care about this function? Recall the factorial function $n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$, giving $0! = 1$ (by convention, since an empty product is 1), $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$, and so on. One might wonder whether it is possible to interpolate a smooth function through these points, so that we could give meaning to an expression like 3.5!. As it turns out, the answer is yes, and in essentially only one way: via the gamma function.

Proposition 1.7.1. *For any nonnegative integer n, we have*

$$n! = \Gamma(n+1) = \int_0^\infty e^{-x} x^n\, dx.$$

Proof. First, how is the factorial function defined? Its essential properties are first that $n! = n \cdot (n-1)!$ and that $0! = 1$; given these two rules, we can compute any value of $n!$ by

$$\begin{aligned} n! = n \cdot (n-1)! = n \cdot (n-1) \cdot (n-2)! = \cdots &= n \cdot (n-1) \cdots 3 \cdot 2 \cdot 1 \cdot 0! \\ &= n \cdot (n-1) \cdots 3 \cdot 2 \cdot 1. \end{aligned}$$

Thus if we can verify that $\Gamma(n+1)$ satisfies these two properties, i.e. $\Gamma(n+1) = n\Gamma(n)$ and $\Gamma(0+1) = \Gamma(1) = 1$, then we're done.

The first of these follows from integration by parts:

$$\Gamma(n+1) = \int_0^\infty e^{-x} x^n\, dx = -e^{-x} x^n|_0^\infty - \int_0^\infty -n e^{-x} x^{n-1}\, dx = n \int_0^\infty e^{-x} x^{n-1}\, dx = n\Gamma(n).$$

The second is just directly evaluating an integral:

$$\Gamma(1) = \int_0^\infty e^{-x} x^0\, dx = \int_0^\infty e^{-x}\, dx = 1.$$

□

Thus the gamma function extends the factorial to a larger domain.[31] What exactly is this domain? Certainly we can evaluate $\Gamma(s)$ when s is a positive integer, as then as we have seen it is just $(s-1)!$. More generally if s is a positive real number, or even a complex number with positive real part, then the integral defining $\Gamma(s)$ converges. If the real part of s is less than or equal to 0 then the integral diverges; but we have the functional equation

$$\Gamma(s+1) = s\Gamma(s),$$

as proven above, and so we can define for example $\Gamma(-\frac{1}{2})$ by plugging in $s = -\frac{1}{2}$ to this formula:

$$\Gamma\left(\frac{1}{2}\right) = -\frac{1}{2}\Gamma\left(-\frac{1}{2}\right).$$

(Explicitly, it turns out that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, and so $\Gamma(-\frac{1}{2}) = -2\sqrt{\pi}$.) Applying this formula repeatedly allows us to compute $\Gamma(s)$ in terms of $\Gamma(s+n)$ for some integer n by dividing by $s+k$ for integers $0 \leq k < n$; this yields values of $\Gamma(s)$ unless one of these $s+k$ is equal to 0. Therefore $\Gamma(s)$ is defined for any complex number s except nonpositive integers $0, -1, -2, \ldots$.

Okay, we've defined and somewhat motivated the gamma function. Why should we care? Abstractly: it turns out that the zeta function is not quite the right object to be looking at to have a good functional equation. The zeta function can be defined by an Euler product

$$\zeta(s) = \prod_p \frac{1}{1-p^{-s}}$$

with one factor for every prime, but it turns out that there is a "missing factor": from a certain point of view there is a "prime at infinity," and we need to account for it as well. Whereas for the "finite" primes the factors are all of the form

$$\frac{1}{1-p^{-s}},$$

for the "infinite prime" the factor is somewhat more complicated: it turns out to be

$$\pi^{-s/2}\Gamma\left(\frac{s}{2}\right).$$

This motivates us to construct the *completed zeta function*

$$\Lambda(s) = \pi^{-s/2}\Gamma\left(\frac{s}{2}\right)\zeta(s),$$

and indeed this is the function for which we will prove a functional equation (which will then imply one for $\zeta(s)$).[32]

[31] One might ask why $\Gamma(s)$ has that off-by-one feature, with the exponent of $s-1$ in the integral such that $n! = \Gamma(n+1)$ instead of $\Gamma(n)$. The answer is that Legendre defined it that way for some reason and no one's ever changed it. It does turn out though that this definition is more appropriate for certain uses, especially in analytic or algebraic number theory; it would be inconvenient to keep having to write the $s-1$ term all the time.

[32] All this will, hopefully, be better explained and generalized later in algebraic number theory when we talk about Tate's thesis.

A more grounded motivation is simply that we want to find a functional equation for $\zeta(s)$ and already know of one for $\Gamma(s)$, so it is perhaps natural to try to go from one to the other. Indeed there's a rather slick way of doing so, though the functional equation we reduce to is a different one. Start with the formula

$$\Gamma(s/2) = \int_0^\infty e^{-x} x^{s/2-1}\, dx.$$

Multiplying both sides by n^{-s}, where n is some positive integer, and making the substitution $x = \pi n^2 t$,[33] so that $dx = \pi n^2\, dt$, we get

$$n^{-s}\Gamma(s/2) = n^{-s} \int_0^\infty e^{-\pi n^2 t} (\pi n^2 t)^{s/2-1} \pi n^2\, dt = \pi^{s/2} \int_0^\infty e^{-\pi n^2 t} t^{s/2-1}\, dt.$$

Dividing both sides by $\pi^{s/2}$ and summing over n from 1 to infinity gives

$$\pi^{-s/2}\Gamma(s/2) \sum_{n=1}^\infty n^{-s} = \pi^{-s/2}\Gamma(s/2)\zeta(s) = \sum_{n=1}^\infty \int_0^\infty e^{-\pi n^2 t} t^{s/2-1}\, dt = \int_0^\infty t^{s/2-1} \sum_{n=1}^\infty e^{-\pi n^2 t}\, dt.$$

Note that the left-hand side is now precisely the completed zeta function $\Lambda(s)$. Now, if you happen to be a late nineteenth-century analytic number theorist, the sum in the rightmost integral would look very familiar to you. In particular, it would recall the theta function

$$\theta(x) = \sum_{n=-\infty}^\infty e^{-\pi n^2 x},$$

where the sum is over all integers. Since the summand is the same for n and $-n$ and for $n = 0$ it is just 1, we have

$$\theta(x) = 1 + 2\sum_{n=1}^\infty e^{-\pi n^2 x},$$

and so from the above we have

$$\Lambda(s) = \int_0^\infty t^{s/2-1} \frac{1}{2}(\theta(t) - 1)\, dt.$$

Again, as a nineteenth-century analytic number theorist you are also aware of the following lemma due to Jacobi, which requires too much Fourier analysis for us to prove just now but which we may come back to in the future.

Lemma 1.7.2. *For any complex number x with positive real part, the theta function $\theta(x)$ converges and satisfies*

$$\theta\left(\frac{1}{x}\right) = \theta(x)\sqrt{x}.$$

[33] This part is the only one that's really unmotivated at this point in the proof; the presence of the factor of π is to simplify the constant factors later, while the factor of n^2 is to combine with the exponentiation to a power of $s/2$ to cancel the n^{-s} term.

If an integral from 0 to ∞ diverges and its integrand is well-defined for positive real numbers, then its divergence is due to its behavior near one of the endpoints, i.e. the integrand grows too large or ill-behaved either as t goes to infinity or as it goes to 0. Thus it makes sense to split such an integral into the portion from 0 to 1 and the portion from 1 to ∞ and work with each separately.

Carrying this out for our expression for $\Lambda(s)$, we have

$$\Lambda(s) = \frac{1}{2}\int_0^1 t^{s/2-1}(\theta(t)-1)\,dt + \frac{1}{2}\int_1^\infty t^{s/2-1}(\theta(t)-1)\,dt.$$

Now we know that our integral converges for s with large positive real part, since it comes from $\Gamma(s/2)$ and $\zeta(s)$, so in the second integral although for such s we might be concerned that $t^{s/2-1}$ might get very large as $t \to \infty$ we should be confident that this nevertheless converges. Indeed it is not hard to see that as $t \to \infty$ correspondingly $\theta(t) - 1$ becomes exponentially small, so in this case the second integral converges, and if s has small or negative real part then the second integral converges all the better. Thus the second integral is universally convergent, and all our problems are coming from the first integral where t is small.

But Lemma 1.7.2 handles this perfectly for us: it transforms the analysis of $\theta(t)$ for small t into the analysis of $\theta(1/t)$, and where t is small $1/t$ must be large. Explicitly, applying Lemma 1.7.2 the first integral becomes

$$\frac{1}{2}\int_0^1 t^{s/2-1}\left(\frac{1}{\sqrt{t}}\theta\left(\frac{1}{t}\right)-1\right)\,dt.$$

Let $u = \frac{1}{t}$, so that $dt = d(1/u) = -\frac{1}{u^2}du$. Then this is

$$\frac{1}{2}\int_1^\infty u^{-s/2-1}\left(\sqrt{u}\theta(u)-1\right)\,du$$

and so, renaming u back to t for convenience, we get that

$$\begin{aligned}
\Lambda(s) &= \frac{1}{2}\int_1^\infty t^{-s/2-1}(\sqrt{t}\theta(t)-1)\,dt + \frac{1}{2}\int_1^\infty t^{s/2-1}(\theta(t)-1)\,dt \\
&= \frac{1}{2}\int_1^\infty \theta(t)\left(t^{s/2-1}+t^{-s/2-1/2}\right) + t^{s/2-1} - t^{-s/2-1}\,dt \\
&= \frac{1}{2}\int_1^\infty (\theta(t)-1)(t^{s/2-1}+t^{(1-s)/2-1})\,dt + \frac{1}{2}\int_1^\infty t^{(1-s)/2-1}\,dt - \frac{1}{2}\int_1^\infty t^{-s/2-1}\,dt \\
&= \frac{1}{2}\int_1^\infty (\theta(t)-1)(t^{s/2-1}+t^{(1-s)/2-1})\,dt - \frac{1}{1-s} - \frac{1}{s}.
\end{aligned}$$

There are two key properties to notice here. First, although initially we have no reason to believe that $\Lambda(s)$ should converge other than for $\mathrm{Re}(s) > 1$, the formula we have derived for it is convergent in the entire complex plane except at $s = 0$ and $s = 1$, since the integral is everywhere convergent by the same argument as above. Second, by examining this formula it is clear that replacing s with $1 - s$ does not change the value of the formula, since $1 - s$ is then replaced with $1 - (1 - s) = s$. We have proven the following.

Theorem 1.7.3 (The functional equation)**.** *The completed zeta function $\Lambda(s)$ extends to a meromorphic function on the entire complex plane, except for simple poles*[34] *at $s = 0$ and $s = 1$. Further, for any complex s not equal to 0 or 1, we have*

$$\Lambda(s) = \Lambda(1 - s).$$

Expanding the definition of $\Lambda(s)$, we get the immediate corollary that

$$\pi^{-s/2}\Gamma(s/2)\zeta(s) = \pi^{(s-1)/2}\Gamma((1 - s)/2)\zeta(1 - s).$$

From here, using certain properties of the gamma function, one can derive the form originally stated. More interestingly, we get the following corollary.

Corollary 1.7.4. *For every positive integer n, we have $\zeta(-2n) = 0$. Any other zeros of the zeta function must satisfy $0 \leq \mathrm{Re}(s) \leq 1$.*

Proof. Suppose that $\mathrm{Re}(s) > 1$. Then the product formula

$$\zeta(s) = \prod_p \frac{1}{1 - p^{-s}}$$

holds and the product converges. Since no factor of the product is 0 and the product converges, it cannot be 0, so there are no zeros of the zeta function with real part greater than 1.

Next, suppose that $\mathrm{Re}(s) < 0$, so that $\mathrm{Re}(1 - s) > 1$. By the above, $\zeta(1 - s) \neq 0$, so if $\zeta(s) = 0$ then by the expanded form of the functional equation we must have either $\pi^{(s-1)/2} = 0$ or $\Gamma((1 - s)/2) = 0$, or the zero of $\zeta(s)$ must be cancelled by $\Gamma(s/2)$ having a pole. The first is clearly impossible. For the second, it turns out that the gamma function has no zeros in the entire complex plane; proving this is a decent exercise. Thus if $\zeta(s) = 0$ then $\Gamma(s/2)$ must have a pole, which we know occurs only at negative integers (the case $s = 0$ is irrelevant here, since we assume that $\mathrm{Re}(s) < 0$). Therefore if $\mathrm{Re}(s) < 0$ then $\zeta(s) = 0$ if and only if $s = -2n$ for some positive integer n. □

Remark. This, by the way, is also where the heinous idea that

$$1 + 2 + 3 + 4 + \cdots = -\frac{1}{12}$$

comes from: naively, one might think that $\zeta(-1)$ is given by the original sum definition of $\zeta(s)$ evaluated at $s = -1$, which would give

$$\zeta(-1) = \sum_{n=1}^{\infty} \frac{1}{n^{-1}} = \sum_{n=1}^{\infty} n = 1 + 2 + 3 + 4 + \cdots .$$

The functional equation at $s = -1$ shows that

$$\pi^{1/2}\Gamma(-1/2)\zeta(-1) = \pi^{-1}\Gamma(1)\zeta(2);$$

[34] A simple pole is one that locally looks like $\frac{1}{x-x_0}$ up to a constant factor.

we saw before that $\Gamma(-1/2) = -2\sqrt{\pi}$ and $\Gamma(1) = 1$, and it turns out that $\zeta(2) = \frac{\pi^2}{6}$, so we conclude that

$$\zeta(-1) = \frac{\zeta(2)}{-2\pi^2} = -\frac{1}{12}.$$

It does *not*, however, follow that the sum of positive integers is $-\frac{1}{12}$, since this sum diverges and so is not the definition of the zeta function in this region.

The reader might at this point feel, not unreasonably, that we have cheated: rather than proving the functional equation outright, we reduced it to the functional equation for $\theta(x)$, which we then refused to prove. It is in fact possible to prove Theorem 1.7.3 via the path integral methods of Section 1.6, and indeed this was Riemann's first proof (his second was essentially that given above). We won't give it here due to its highly technical and not especially enlightening nature, but we will eventually come back to this issue of the functional equation for $\theta(x)$; though not terribly difficult to prove, it turns out to be a pretty deep fact, coming from the "automorphicity" of the theta function.

1.8 The prime number theorem

After an analytic detour, we are finally at a point where we can prove Theorem 1.5.9, the prime number theorem. The form which we will attack is that which states that

$$\psi(x) = \sum_{n \leq x} \Lambda(n) \sim x,$$

where Λ is the von Mangoldt function.

From Propositions 1.4.3 and 1.5.1, we have

$$-\frac{\zeta'(s)}{\zeta(s)} = s \int_1^\infty \frac{\psi(x)}{x^{s+1}}\, dx.$$

This is a special case of the transform taking a function $f(x)$ to the function

$$s \int_1^\infty \frac{f(x)}{x^{s+1}}\, dx,$$

known as the Mellin transform; thus since we want to estimate $\psi(x)$ and the analytic function

$$-\frac{\zeta'(s)}{\zeta(s)}$$

is something we can more or less get our hands on and compute directly, it would suffice to be able to *invert* the Mellin transform—that is, given a function $F(s)$ which we know is the Mellin transform of some $f(x)$, find $f(x)$ from $F(s)$. This inversion is given by Perron's formula.

Lemma 1.8.1 (Perron's formula)**.** *Suppose that*

$$F(s) = s \int_1^\infty \frac{f(x)}{x^{s+1}}\, dx$$

converges for every s with real part greater than σ. Then

$$f(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} F(s)\frac{x^s}{s}\,ds$$

where c is any positive real number greater than σ.

This is a line integral over the line in the complex plane which is parallel to the imaginary axis and intersects the real axis at c.

Technically, this formula is only true for non-integer x: if x is an integer, then this formula will give

$$\lim_{\epsilon\to 0}\frac{f(x)+f(x-\epsilon)}{2}.$$

However, since we only care about the asymptotic result we'll neglect this.

Perron's formula is not hard to prove using the contour integral methods of Section 1.6, but still annoying so we'll skip it. The interesting question is how to apply it in our case. On the one hand this is exactly what we're looking for: set $F(s) = -\frac{\zeta'(s)}{\zeta(s)}$, with $\sigma = 1$, and $f(x) = \psi(x)$. Then this lemma gives us a way of computing $\psi(x)$ from information about $-\frac{\zeta'(s)}{\zeta(s)}$. On the other hand, it requires us to evaluate this line integral, which seems no easier!

Here, we use on of the central tricks of computing integrals via complex analysis. We have an integral that we want to compute, along a certain line. We know how to compute integrals along curves that bound a given region: use the residue theorem (Theorem 1.6.11). Thus we want to transform our line into a closed curve, i.e. one that bounds some region, without changing the value of the integral. How can we do this? Well, by adding to our line some curve on which the integral is 0.

Let's restrict to a finite line, which we'll then take the limit of to infinity, so what we want to evaluate is

$$\int_{c-iL}^{c+iL} F(s)\frac{x^s}{s}\,ds$$

where L is some large positive real number. Let's call this line from $c-iL$ to $c+iL$ the line segment λ_L (λ short for "line"). Next, we complete the curve by adding a semicircle of radius L connecting the endpoints; see Figure 1.1, which shows this modification with $L = 4$ (and $c = 1.2$). We'll call this semicircle C_L.

What we want to show is that as L goes to infinity, the integral over C_L goes to 0, so that we can evaluate the integral over the line via the residue theorem on the region bounded by λ_L and C_L:

$$\int_{c-i\infty}^{c+i\infty} F(s)\frac{x^s}{s}\,ds = \lim_{L\to\infty}\int_{\lambda_L} F(s)\frac{x^s}{s}\,ds = \lim_{L\to\infty}\left(\int_{\lambda_L} F(s)\frac{x^s}{s}\,ds + \int_{C_L} F(s)\frac{x^s}{s}\,ds\right),$$

and since the last is an integral over a closed curve we can compute it using the residue theorem.

Proposition 1.8.2. *For $F(s) = -\frac{\zeta'(s)}{\zeta(s)}$ and notation as above, we have*

$$\lim_{L\to\infty}\int_{C_L}\frac{\zeta'(s)}{\zeta(s)}\frac{x^s}{s}\,ds = 0.$$

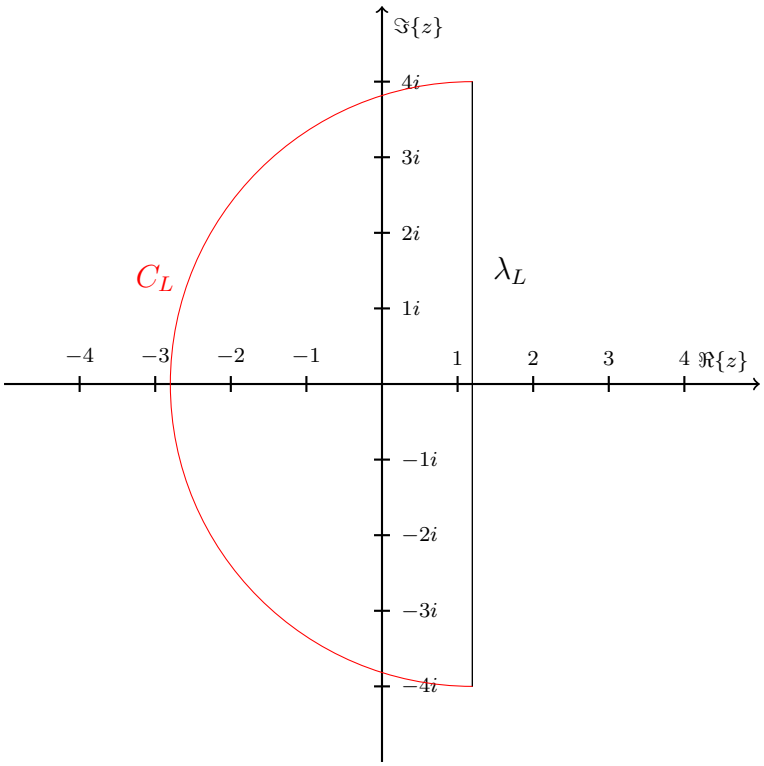


Figure 1.1: Modified line integral.

This can be worked out from the functional equation, with some care; there are two basic cases, one where s is near the endpoints of C_L, and so has small real part but large imaginary part, and one where s is far from them, and so has large negative real part but potentially small imaginary part. The latter case is easy to handle with the functional equation, and the former can be bounded by some analytic trickery, some of which we'll see later.

With this in hand, let's apply the residue theorem. Our integrand

$$-\frac{\zeta'(s)}{\zeta(s)}\frac{x^s}{s}$$

has poles at 0, the poles of $\zeta'(s)$, and the zeros of $\zeta(s)$, all of which are within our region of integration (upon taking the limit). The pole at 0 is the easiest: its residue is just

$$-\lim_{s\to 0}\frac{\zeta'(s)}{\zeta(s)}x^s = -\frac{\zeta'(0)}{\zeta(0)}.$$

Computing this number is not entirely trivial, but differentiating the functional equation[35] shows that it is $-\log(2\pi)$. Next, the only pole of $\zeta(s)$ is at $s=1$, and since it is analytic everywhere else similarly the only pole of $\zeta'(s)$ is at $s=1$; since near $s=1$ we know that

$$\zeta(s) \sim \frac{1}{s-1},$$

[35] We'll hide several computations in this section under the label of "you can work this out by differentiating the functional equation"; this is true, but it also requires knowing something about the behavior of the gamma function, which is not hard to work out but is too much of a pain to include and not really all that relevant to us anyway.

we have

$$\zeta'(s) \sim \frac{d}{ds}\frac{1}{s-1} = -\frac{1}{(s-1)^2},$$

so

$$\frac{\zeta'(s)}{\zeta(s)}\frac{x^s}{s} \sim \frac{x^s}{s(s-1)}$$

has residue x at $s = 1$. This leaves only the zeros of $\zeta(s)$.

By Corollary 1.7.4, all of these zeros are either of the form $-2n$ for some positive integer n or have real part between 0 and 1. The contribution of each zero ρ is then the residue $-\frac{x^\rho}{\rho}$, so for the first type the total contribution is

$$-\sum_{n=1}^{\infty}\frac{x^{-2n}}{-2n} = -\frac{1}{2}\log(1-x^{-2}),$$

and for the second simply

$$-\sum_{\rho}\frac{x^\rho}{\rho}$$

where the sum is taken over all nontrivial zeros, i.e. with real parts between 0 and 1. Thus we have proven the following.

Theorem 1.8.3 (Explicit formula)**.** *With the modification at integers as before, we have*

$$\psi(x) = x - \sum_{\rho}\frac{x^\rho}{\rho} - \frac{1}{2}\log(1-x^{-2}) - \log(2\pi),$$

where the sum is taken over nontrivial zeros ρ of $\zeta(s)$.

Proof. This follows from applying the discussion above to Perron's formula for $-\frac{\zeta'(s)}{\zeta(s)}$. □

This is quite remarkable, and better than we had hoped for: we have not only an analytic approximation to $\psi(x)$, but in fact an exact formula! Indeed, if we compute the first few nontrivial zeros of $\zeta(s)$, we can see that our approximation to $\psi(x)$ starts looking like a "staircase" or step function, as it should since it increments only at integers (and indeed only at prime powers).

We're very close to proving the prime number theorem. The latter two terms in Theorem 1.8.3 are clearly much smaller order than the main term; what we want to show is that $\psi(x) \sim x$, and so it remains only to show that

$$\sum_{\rho}\frac{x^\rho}{\rho}$$

is smaller order than x.

Technically speaking, this requires us to prove some bounds on the zeta function in the "critical region" $0 \le \mathrm{Re}(s) \le 1$, since there are infinitely many zeros ρ and we need to make sure their contributions are not adding up in inconvenient ways. Morally speaking, though,

it's enough to show that each term $\frac{x^\rho}{\rho}$ has order smaller than x, i.e. $\mathrm{Re}(\rho) < 1$, and indeed this is the heart of the proof after some nastiness (which we'll skip).[36]

Therefore the following proposition completes the proof of the prime number theorem.

Proposition 1.8.4. *There is no zero of the Riemann zeta function with real part equal to* 1*.*

This ensures (up to some more analytic shenanigans) that the sum must be of order strictly less than x, so that Theorem 1.8.3 immediately implies the prime number theorem. In fact, in a certain sense this proposition *is* the prime number theorem, and everything else we have been doing is just demonstrating the equivalence between them.

Proof of Proposition 1.8.4. The first proof of this proposition was a slick trick due to Mertens, which we duplicate here; it's basically completely unmotivated, and it's something of a mystery how Mertens came up with it in the first place. Its basis is the following trigonometric identity: for any angle θ, we have

$$3 + 4\cos\theta + \cos(2\theta) \geq 0.$$

This is easy to prove:

$$\begin{aligned} 3 + 4\cos\theta + \cos(2\theta) &= 2(1 + 2\cos\theta) + 1 + \cos(2\theta) \\ &= 2(1 + 2\cos\theta + (\cos\theta)^2) \\ &= 2(1 + \cos\theta)^2 \end{aligned}$$

is a square of a real number and therefore nonnegative.[37] As before, $\log|z| = \mathrm{Re}(\log z)$, so we have

$$\begin{aligned} \log|\zeta(s)| &= \log\left|\prod_p \left(\frac{1}{1-p^{-s}}\right)\right| \\ &= \mathrm{Re}\left(-\sum_p \log(1-p^{-s})\right) \\ &= \mathrm{Re}\left(\sum_p \sum_{k\geq 1} \frac{1}{kp^{ks}}\right) \\ &= \sum_p \sum_{k\geq 1} \frac{1}{k}\mathrm{Re}(p^{-ks}). \end{aligned}$$

[36]The fact that we only care about the real part for the absolute value comes from Euler's formula: $x^{\sigma+it}$ for real σ and t is equal to $x^\sigma e^{it\log x}$, and $e^{it\log x}$ is a point on the complex unit circle, with absolute value 1, so only x^σ contributes to the absolute value.

[37]The second line is from the trigonometric identity $(\cos\theta)^2 - 1 = \cos(2\theta)$, which can be derived from Euler's formula: $\mathrm{Re}(e^{i\theta}) = \cos\theta$, so $\cos(2\theta) = \mathrm{Re}(e^{2i\theta}) = \mathrm{Re}((e^{i\theta})^2) = \mathrm{Re}((\cos\theta + i\sin\theta)^2) = \mathrm{Re}((\cos\theta)^2 - (\sin\theta)^2 + 2i\sin\theta\cos\theta) = (\cos\theta)^2 - (\sin\theta)^2$ and then applying the Pythagorean formula $(\sin\theta)^2 + (\cos\theta)^2 = 1$ and doing some manipulations.

Write $s = \sigma + it$. Then $p^{-ks} = e^{-ks\log p} = e^{-k\sigma\log p}e^{-kt\log p\cdot i} = p^{-k\sigma}(\cos(-kt\log p) + i\sin(-kt\log p))$, and so taking real parts[38] we have $\mathrm{Re}(p^{-ks}) = p^{-k\sigma}\cos(kt\log p)$. Now,

$$\sum_p \sum_{k\geq 1} \frac{1}{kp^{k\sigma}}(3 + 4\cos(kt\log p) + \cos(2kt\log p))$$

is nonnegative since every term is nonnegative by the identity above, applied to $\theta = kt\log p$. On the other hand,

$$\sum_p \sum_{k\geq 1} \frac{1}{kp^{k\sigma}}(3+4\cos(kt\log p)+\cos(2kt\log p)) = 3\log|\zeta(\sigma)|+4\log|\zeta(\sigma+it)|+\log|\zeta(\sigma+2it)|$$

by the construction above, so exponentiating we have

$$|\zeta(\sigma)|^3|\zeta(\sigma+it)|^4|\zeta(\sigma+2it)| \geq 1.$$

Since $s = \sigma + it$ was arbitrary, suppose that there is a zero of ζ at $1 + iT$; T must be nonzero, since $\zeta(s)$ has a pole at $s = 1$. Then as $\sigma + it \to 1 + iT$, the above quantity must remain greater than 1; fix $t = T$ and let $\sigma + iT$ approach $1 + iT$ by letting σ approach 1. First, note that as $\sigma \to 1$ we have $(\sigma - 1)\zeta(s)$ bounded by some constant b, since the pole of $\zeta(s)$ at 1 is simple. Next, since $\zeta(s)$ is analytic, near $1 + iT$ we must have $\zeta(s) = a_1(s - (1 + iT)) + O((s - (1 + iT))^2)$ for some constant a_1, so in particular

$$|\zeta(\sigma + iT)|^4 = a_1(\sigma - 1)^4$$

and so

$$1 \leq |\zeta(\sigma)|^3|\zeta(\sigma+it)|^4|\zeta(\sigma+2it)| \leq |\zeta(\sigma)|^3 a_1^3(\sigma-1)^4|\zeta(\sigma+2iT)| \leq a_1 b(\sigma-1)|\zeta(\sigma+2iT)|.$$

Since $\zeta(s)$ has no poles other than at $s = 1$, and $T \neq 0$ so $1 + 2iT \neq 1$, the right-hand side goes to 0 as $\sigma \to 1$; but it must be greater than or equal to 1, by the above. Therefore we have a contradiction and no such zero can exist. □

This concludes the proof of Theorem 1.5.9, our primary goal in this chapter. However, we have learned more, especially in the form of Theorem 1.8.3, and it is natural to wonder if we can do better: we have shown that no zero can have nontrivial part as large as 1, but how large can they be?

This was a question that concerned Riemann, and in fact he computed several hundred zeros of $\zeta(s)$ in an attempt to answer it. He found that in fact *every* nontrivial zero of $\zeta(s)$ has real part precisely $\frac{1}{2}$.[39] This led him to conjecture the Riemann hypothesis, which is that this is in fact true for all nontrivial zeros. As might be guessed from the above, this gives a much better error bound for $\psi(x)$: in fact, the Riemann hypothesis is equivalent to the conjecture that

$$\psi(x) = x + O(x^{\frac{1}{2}+\epsilon})$$

[38] And recalling that $\cos(-x) = \cos(x)$ for all x.

[39] This is equivalent to the bound that they must all have real part *at most* $\frac{1}{2}$: the functional equation then implies that by symmetry they cannot be less than $\frac{1}{2}$ either.

for every $\epsilon > 0$, i.e. an error bound arbitrarily close (but not equal) to $x^{\frac{1}{2}} = \sqrt{x}$. It is known that this is the best possible bound, in the sense that

$$\psi(x) = x + O(x^{\frac{1}{2}-\epsilon})$$

does not hold for any $\epsilon > 0$.

1.9 Primes in arithmetic progressions

So now we know how to estimate the density of primes, via the prime number theorem. But really given any subset S of the natural numbers this is the least we can say about it, and certainly the prime numbers have a great deal more structure. Let's say we're looking at an integer, say 2308, and we want to know whether it's prime. Well, that's pretty easy: it's even (and not equal to 2), so it can't be prime. Similarly, 9335 is divisible by 5, so it can't be prime. On the other hand, 4089 is a little harder to tell (in fact it's divisible by 3, so not too much harder).

What's the difference here? With the first two examples, we can immediately tell whether they're prime by looking only at the last digit. In fact, working in base 10, we can always be sure that if our number has last digit 0, 2, 4, 6, or 8 then it can't be prime (unless our number is actually 2) since it must be even; and if it's 0 or 5 then it's divisible by 5, and so not prime (unless equal to 5). Thus other than 2 or 5 every prime number must have last digit 1, 3, 5, or 7. Can we do any better?

Well, in one sense clearly not: there exist "sufficiently large" prime numbers, such as 11, 13, 17, and 19, ending with all of these, so we can't eliminate any of them. Could we do any better—for example, can we say that there are only finitely many prime numbers with last digit 3? The answer will turn out to be no, and in fact primes are *equidistributed* across these possibilities: if we pick a "random prime" number p, the probabilities of it having last digit 1, 3, 5, or 7 are all equal to $\frac{1}{4}$.

Let's generalize a little. In order to do so, we need to think about modular arithmetic.

Let a and b be two integers, and n be a positive integer. We say that a and b are equivalent modulo n, written $a \equiv b \pmod{n}$, if the remainders of a and b upon being divided by n are equal, i.e. if $a - b$ is divisible by n. For example, let $n = 5$. Then 2 and 7 are equivalent modulo 5, since $2 - 7 = -5$ is divisible by 5. On the other hand, 3 and 4 are not equivalent modulo 5: $3 - 4 = -1$ is not divisible by 5.

What we really care about, for fixed n, is equivalence classes modulo n, of which there are n. For example, if $n = 2$, there are two equivalence classes, corresponding to even and odd integers: we have $\cdots \equiv -4 \equiv -2 \equiv 0 \equiv 2 \equiv 4 \equiv \cdots$ modulo 2, and similarly $\cdots \equiv -3 \equiv -1 \equiv 1 \equiv 3 \equiv \cdots$ modulo 2, so we have the two equivalence classes $\{\ldots, -4, -2, 0, 2, 4, \ldots\}$ and $\{\ldots, -3, -1, 1, 3, \ldots\}$, which are disjoint[40] and together cover all integers. These are often written just as 0 and 1 respectively. The set of equivalence classes of integers modulo n is written $\mathbb{Z}/n\mathbb{Z}$.

Of course, we have not defined what we mean by "equivalence."

[40] i.e. have no elements in common.

Definition 1.9.1. An *equivalence relation* $\sim$ is a relation[41] such that for any two elements a and b we have $a \sim b$ if and only if $b \sim a$; $a \sim a$ for every a; and if $a \sim b$ and $b \sim c$ then $a \sim c$.

The canonical equivalence relation is just equality: $a = b$ if and only if $b = a$, every element is equal to itself, and if $a = b$ and $b = c$ then $a = c$. It's not hard, and a good exercise, to verify that equivalence modulo n as defined above is an equivalence relation for any n.

In fact, this equivalence is even better than an equivalence relation: it's one that respects addition and multiplication. What do we mean by this? Fix some n, and suppose that $a \equiv b$ and $c \equiv d$ modulo n. Then we can add these two relations, to get $a + c \equiv b + d \pmod{n}$, and multiply them, to get $a \cdot c \equiv b \cdot d \pmod{n}$, both of which still hold for this equivalence. (Exchanging c and d, by the reversibility required of an equivalence relation, shows that the same thing holds for subtraction; division is a little more complicated.) There's a strong sense in which equality and equivalence modulo n, for some n, are the *only* equivalence relations on integers which respect addition and multiplication, though on more general sets there are sometimes more.

Reduction modulo n can introduce some strange phenomena. For example, suppose that $a \equiv 0 \pmod{n}$. Then $a + b \equiv 0 + b \equiv b \pmod{n}$, so any such a is an additive identity; and similarly if $a \equiv 1 \pmod{n}$ then $a \cdot b \equiv b \pmod{n}$. This isn't too surprising; this is just saying that we can treat anything equivalent to 0 or 1 as *being* 0 or 1 respectively. On the other hand, for any nonzero integer a we can find a rational number, though not necessarily an integer, a^{-1} such that $a \cdot a^{-1} = 1$. Suppose that $a \equiv 0 \pmod{n}$, i.e. a is divisible by n. Then if there exists an equivalence class a^{-1} such that $a \cdot a^{-1} \equiv 1 \pmod{n}$, then the left-hand side is equivalent to $0 \cdot a^{-1} = 0 \pmod{n}$, and since $0 \not\equiv 1 \pmod{n}$ for any $n \neq 1$ this is impossible. Thus not every element of $\mathbb{Z}/n\mathbb{Z}$ is invertible.

But still, this isn't too surprising: really since $a \equiv 0 \pmod{n}$, what we've shown is really that the equivalence class $0 \in \mathbb{Z}/n\mathbb{Z}$ isn't invertible (for $n \neq 1$), and of course 0 is not normally invertible. Let's consider another example, with $n = 6$. Then $2 \cdot 3 = 6 \equiv 0 \pmod{6}$. Suppose that 2 has an inverse 2^{-1} modulo 6. Then $2^{-1} \cdot 2 \cdot 3 \equiv 1 \cdot 3 \equiv 3$, but on the other hand $2^{-1} \cdot 2 \cdot 3 \equiv 2^{-1} \cdot 0 \equiv 0$; but 3 and 0 are not equivalent modulo 6, so 2 cannot be invertible. (The same argument shows that 3 cannot be invertible.)

Still, this isn't all that surprising: reducing modulo n is an operation on integers, and so it's not shocking that the rational number (but not an integer) $2^{-1} = \frac{1}{2}$ isn't in $\mathbb{Z}/6\mathbb{Z}$. Perhaps more surprising, then, is that some integers *are* invertible: for example, $5 \equiv -1 \pmod{6}$, so $5^{-1} \equiv (-1)^{-1} \equiv -1 \equiv 5 \pmod{6}$, i.e. 5 is its own inverse modulo 6. Indeed, $5 \cdot 5 = 25 \equiv 1 \pmod{6}$, so $\frac{1}{5} \equiv 5 \pmod{6}$ does exist.

More generally, for any n, some of the equivalence classes in $\mathbb{Z}/n\mathbb{Z}$ will be invertible, and some will not; the only guarantee is that the equivalence class 0 of numbers divisible by n will never be invertible. The set of invertible classes is written $(\mathbb{Z}/n\mathbb{Z})^\times$. (All this notation will be explained eventually, by the way, probably when we get to our unit on algebra.)

Let's fix the convention of referring to equivalence classes in $\mathbb{Z}/n\mathbb{Z}$ by integers, i.e. (in context) an integer k would refer to the class of integers congruent[42] to k modulo n.

[41] Which really just means a pairing of two elements, for our purposes integers.

[42] Here, a synonym for "equivalent."

Definition 1.9.2. The *greatest common divisor* of two integers, written $\gcd(a, b)$, is the largest positive integer d such that both a and b are divisible by d. If $\gcd(a, b) = 1$, we say that a and b are relatively prime.[43]

Note that if d divides both a and b, then it also divides $\gcd(a, b)$: write a and b as products of primes $a = p_1^{e_1} \cdots p_s^{e_s}$, $b = q_1^{f_1} \cdots q_t^{e_t}$. Then if $r_1, \ldots, r_u$ are the primes dividing both a and b, with r_i dividing a g_i times and dividing b h_i times, then $\gcd(a, b) = r_1^{\min(g_1,h_1)} \cdots r_u^{\min_{g_u,h_u}}$, while divisors of both a and b must be products of the prime powers dividing them and therefore must also divide $\gcd(a, b)$.

Proposition 1.9.3. *Suppose that* $\gcd(a, b) = d$*, for positive integers* a *and* b*. Then there exist integers* x, y *such that* $ax + by = d$*.*

This is for some reason known as Bézout's identity, though it was certainly known long before him; he proved a generalization of it to polynomial rings.

Proof. Consider the set S of integers of the form $ax+by$ for some integers x and y. We want to show that S contains d. Since S is a set of integers, it has a minimal positive element, which we'll call m. We certainly have $m \leq a$, since $a \cdot 1 + b \cdot 0 = a$ is a positive element of S, and by definition m is the smallest such element; so we can consider the division of a by m, with remainder r, i.e. $a = mt + r$ for some integer t, with $r < m$ a nonnegative integer. But then $r = a - mt = a - (ax + by)t = a(1 - x) - by$ for some integers x and y, so r is in S; since it is a nonnegative integer strictly smaller than m, by the definition of m we must have $r = 0$, since otherwise r would be a smaller positive element of S than m. Therefore $a = mt$, so m divides a.

Now, we chose a in the above, but exactly the same proof works for b, so we see that m divides both a and b, and therefore divides $d = \gcd(a, b)$ by the remark above. Since $m = ax + by$, writing $d = m \cdot m'$ we have $d = m'(ax + by) = am'x + bm'y$, so d is in S. □

In fact, d will itself always be the smallest positive element of S, and indeed S is just multiples of d; deducing this from the proof is not too hard.

Proposition 1.9.4. *An equivalence class in* $\mathbb{Z}/n\mathbb{Z}$ *represented by an integer* k *is invertible if and only if* n *and* k *are relatively prime.*

Proof. Suppose that k and n are relatively prime, i.e. $\gcd(k, n) = 1$. Then by Proposition 1.9.3 there exist integers x, y such that $kx + ny = 1$, so $kx \equiv 1 \pmod{n}$. Therefore x is an inverse of k modulo n, so the equivalence class of k is invertible.

Conversely, suppose that k and n are not relatively prime, i.e. there is some integer $d \geq 2$ dividing both k and n. Then $d \cdot \frac{n}{d} \equiv 0 \pmod{n}$ and $d \cdot \frac{k}{d} \equiv k \pmod{n}$, and by the multiplicative property of equivalence modulo n we have $d \cdot \frac{k}{d} \cdot \frac{n}{d} \equiv k \cdot \frac{n}{d} \pmod{n}$; by the above the left-hand side is $d \cdot \frac{n}{d} \cdot \frac{k}{d} \equiv 0 \cdot \frac{k}{d} \equiv 0 \pmod{n}$, so we have $k \cdot \frac{n}{d} \equiv 0 \pmod{n}$. But then the same proof that showed that 2 could not be invertible in $\mathbb{Z}/6\mathbb{Z}$ above shows that k cannot be invertible: if so, then $k^{-1}k\frac{n}{d} \equiv \frac{n}{d} \pmod{n}$, but on the other hand this is equivalent to $k^{-1} \cdot 0 \equiv 0 \pmod{n}$, so we have shown that $\frac{n}{d} \equiv 0 \pmod{n}$, which is impossible since $d \geq 2$. Therefore k cannot be invertible modulo n. □

[43] This is equivalent to the statement that the fraction $\frac{a}{b}$ is in reduced form. For example 3 and 8 are relatively prime, but 4 and 6 are not, since both are divisible by 2.

Note that if n is prime, then every integer $0 \leq k < n$ is relatively prime to n, since n has no divisors other than itself and 1. Therefore $(\mathbb{Z}/n\mathbb{Z})^\times = \mathbb{Z}/n\mathbb{Z} - \{0\}$ if n is prime; and if n is not prime, then it has nontrivial divisors, which will not be invertible modulo n, so this is not true. Thus we have shown that every nonzero element of $\mathbb{Z}/n\mathbb{Z}$ is invertible if and only if n is prime, an observation which will be useful later (again, in the algebra section).

Now that we've learned how to work with congruence classes, let's take a moment to see why they're useful before delving into their relationships with primes (beyond the above). Suppose that we have some question about the integers, involving only addition and multiplication (and their inverses). Then since these are well-behaved with respect to reduction modulo n, we can ask the same question modulo n; and in this case, since there are only finitely many equivalence classes, it can be much easier, and sometimes the information can be lifted back to the setting of the integers.

For example, here's a problem: can (say) 99 be written as the sum of two squares? That is: do there exist integers a and b such that $a^2 + b^2 = 99$?[44] If we think of this in terms of integers, there's no obvious way of deciding, other than checking all sufficiently small integers a and b, which gets hard if we were to look at e.g. the number 9999999999999. Instead, let's use the power of modular arithmetic! In particular, we're using the secret of *quadratic residues*, which is essentially the observation that, modulo n, only half the classes can be written as squares. For example, for $n = 4$, our classes are (say) $0, 1, 2, 3$. We have $0^2 = 0$, $1^2 = 1$, $2^2 = 4 \equiv 0 \pmod 4$, and $3^2 = 9 \equiv 1 \pmod 4$, so every square is congruent to 0 or 1 modulo 4 (since if we were to go larger we could reduce first, e.g. $7^2 \equiv 3^2 \pmod 4$ since $7 \equiv 3$, and so we only need to make these four computations). These squares, here 0 and 1, are called the quadratic residues modulo 4; the other classes, 2 and 3, are called quadratic nonresidues.[45]

The reason for this phenomenon is that $x^2 = (-x)^2$. When we only have finitely many classes, this means that the n classes get sent by squaring to (approximately) $n/2$ classes (in fact potentially slightly more: if $x \equiv -x$ then x^2 has a unique square root, i.e. x, corresponding to an extra square. This happens when $x \equiv 0$ or, when n is even, when $x \equiv n/2$).

Thus, working modulo 4, $a^2 + b^2$ is the sum of two squares, so it is congruent to either $0 + 0 = 0$, $0 + 1 = 1 + 0 = 1$, or $1 + 1 = 2$. Since $99 \equiv 3 \pmod 4$, we see that it cannot be written as the sum of two squares.

Let's go back to our discussion of the distribution of primes. For $n = 10$, the classes relatively prime to 10 are 1, 3, 7, and 9, i.e. $(\mathbb{Z}/10\mathbb{Z})^\times = \{1, 3, 7, 9\}$. This suggests the following proposition.

Proposition 1.9.5. *For every integer $n \geq 2$, if p is a prime not dividing n then the congruence class of p is in $(\mathbb{Z}/n\mathbb{Z})^\times$.*

Proof. By Proposition 1.9.4, this is just the assertion that either p is relatively prime to n

[44] This can also be thought of geometrically: if we draw a circle of radius $\sqrt{99}$ with circle at the origin, are there any points on the circle whose coordinates are integers? In this formulation it might sound like the answer should trivially be no, but for example for a circle of radius $\sqrt{97}$ we have the points $(4, 9)$, since $4^2 + 9^2 = 97$.

[45] Sometimes 0 is not considered a quadratic residue, and is treated differently; then we would classify the equivalence classes as either a quadratic residue, a quadratic nonresidue, or 0.

or p divides n. But since p is prime, $\gcd(p, n)$ is either p, if n is divisible by p, or 1; the first case is impossible by assumption, and the second case is the desired result. □

The equidistribution result we really want is this: let $\varphi(n)$ be the number of integers $1 \leq k < n$ relatively prime to n; this is called Euler's totient function. Write $\pi(x; a, n)$ for the number of primes less than or equal to x which are congruent to a modulo n. Then we have the following theorem.

Theorem 1.9.6 (Prime number theorem in arithmetic progressions[46])**.** *For every integer $n > 1$ and $a \in (\mathbb{Z}/n\mathbb{Z})^\times$ we have*

$$\pi(x; a, n) \sim \frac{1}{\varphi(n)} \frac{x}{\log x}$$

as $x \to \infty$.

For $n = 10$, this yields the desired equidistribution in the last digit. More generally, it should be understood as saying that the prime numbers are evenly distributed over all residues modulo n where there is not a particular reason for primes not to be in the progression, i.e. an integer greater than 1 dividing both a and n.

We probably won't prove this theorem in full strength, mostly due to analytic difficulties. The interesting part of the proof is in proving the following weaker statement.

Theorem 1.9.7 (Dirichlet's theorem)**.** *For any integer $n > 1$ and $a \in (\mathbb{Z}/n\mathbb{Z})^\times$ there are infinitely many primes congruent to a modulo n.*

Theorem 1.9.6 is strongly similar to the prime number theorem, and our techniques will mostly be quite similar. In particular, the central formula underlying our entire method to prove the prime number theorem was the product formula, Proposition 1.3.8, which in fact can be generalized.

Let $f : \mathbb{N} \to \mathbb{C}$ be an arithmetic function. We say that it is multiplicative if for any two relatively prime integers m and n we have $f(mn) = f(m)f(n)$, and completely multiplicative if the same formula holds for *any* m and n, relatively prime or not. In the first case, since any integer can be written as a product of prime powers and all the prime powers are relatively prime for different prime numbers, f is totally determined by its values at prime powers; in the second case, since $f(p^k) = f(p)^k$, f is totally determined by its values at primes.

We have the following generalized product formula.

Proposition 1.9.8. *Let $f : \mathbb{N} \to \mathbb{C}$ be a completely multiplicative function. Then*

$$\mathcal{D}[f](s) = \sum_{k \geq 1} \frac{f(k)}{k^s} = \prod_p \left(1 - \frac{f(p)}{p^s}\right)^{-1}.$$

Such formulas involving a product over primes are known as Euler products, and can be vastly generalized. Proposition 1.3.8 is the special case of this formula where f is the constant function 1.

[46] An arithmetic progression is just a sequence of the form $a, a + n, a + 2n, a + 3n, \ldots$ for some a and n.

Proof. We can mimic the proof of Proposition 1.3.8: we have

$$\frac{1}{1-f(p)p^s} = \sum_{k=0}^{\infty} \frac{f(p)^k}{p^{ks}} = \sum_{k=0}^{\infty} \frac{f(p^k)}{p^{ks}},$$

and so

$$\prod_p \left(1 - \frac{f(p)}{p^{-s}}\right)^{-1} = \prod_p \left(1 + \frac{f(p)}{p^s} + \frac{f(p^2)}{p^{2s}} + \cdots\right),$$

and expanding the product as before gives

$$\sum_{k\geq 1} \frac{f(k)}{k^s}$$

since f is completely multiplicative. □

Remark. The same proof holds for multiplicative rather than completely multiplicative functions, except for the step that $f(p)^k = f(p^k)$; thus the correct formula in this case is

$$\mathcal{D}[f](s) = \prod_p \sum_{k=0}^{\infty} \frac{f(p^k)}{p^{ks}}.$$

Now, what we'd like to do is analyze the expression

$$\prod_p \left(1 - \frac{\mathbf{1}_{a,n}(p)}{p^s}\right)$$

where $\mathbf{1}_{a,n}(x)$ is the function given by 1 if $x \equiv a \pmod{n}$ and 0 otherwise, since this is essentially the zeta function, restricted to primes congruent to a modulo n as desired; then we can apply the same techniques as used for the zeta function to this expression to derive an estimate for $\pi(x; a, n)$. Unfortunately, $\mathbf{1}_{a,n}$ is neither completely multiplicative nor multiplicative: for example, if $n = 5$ and $a = 2$, then $6 \cdot 7 = 42 \equiv 2 \pmod{5}$, so $\mathbf{1}_{2,5}(6 \cdot 7) = 1$. But $\mathbf{1}_{2,5}(6) = 0$, while $\mathbf{1}_{2,5}(7) = 1$, so $\mathbf{1}_{2,5}(6) \cdot \mathbf{1}_{2,5}(7) = 0 \cdot 1 = 0$. Thus we cannot apply Proposition 1.9.8, and in particular our desired product is not the Euler product of $\mathcal{D}[\mathbf{1}_{a,n}]$, or indeed any obvious Dirichlet series.

How can we fix this? Well, it would be a good start if we could decompose $\mathbf{1}_{a,n}$ into completely multiplicative functions, so that we could apply Proposition 1.9.8 to them and go from there. How can we do this? Dirichlet characters!

Definition 1.9.10. A *Dirichlet character* modulo n is an arithmetic function $\chi : \mathbb{N} \to \mathbb{C}$ satisfying the following properties:

(1) χ is completely multiplicative;

(2) χ is periodic with period n, i.e. for any integer m we have $\chi(m) = \chi(m + n)$;

(3) if the equivalence class of m is not invertible, then $\chi(m) = 0$.

From the definition, it's not obvious that there should be *any* such functions, let alone why they're interesting. First, note that there's always at least one character, known as the trivial character: $\chi(m)$ is 0 if $\gcd(m, n) > 1$, as it must be by the third property, and otherwise $\chi(m) = 1$. This character is often written χ_0, or sometimes simply 1 (though it is not, in fact, constant). There is also the zero character, with constant value 0, but we will usually ignore this one and require our characters to not be everywhere zero.

From the second property, we see that $\chi(m)$ depends only on the equivalence class of m modulo n, so we can really think of χ as a function on $\mathbb{Z}/n\mathbb{Z}$, and by the third property the only nonzero part of χ is a function on $(\mathbb{Z}/n\mathbb{Z})^\times$. We can also note some other useful properties: since χ is completely multiplicative, we have $\chi(m) = \chi(1 \cdot m) = \chi(1) \cdot \chi(m)$ for every m. Since we are assuming that χ is not everywhere zero, there exists some m such that $\chi(m) \neq 0$, so $\chi(1)$ must be equal to 1.

We have the following group-theoretic lemma.

Lemma 1.9.11 (Euler's totient theorem)**.** *For any integer $n \geq 2$ and $a \in (\mathbb{Z}/n\mathbb{Z})^\times$, we have $a^{\varphi(n)} \equiv 1 \pmod{n}$.*

Proof. Consider a set of representatives $x_1, \ldots, x_{\varphi(n)}$ of all the invertible classes modulo n, including a. If we take these and multiply them all by a, since a is an invertible class for each x_i we get another invertible class ax_i. Further, these are all distinct: if $ax_i = ax_j$, then since a is invertible we can multiply by a^{-1} to get $x_i = x_j$. Therefore $ax_1, \ldots, ax_{\varphi(n)}$ is just a rearrangement of $x_1, \ldots, x_{\varphi(n)}$ modulo n. Therefore the product of the elements in each of these lists is congruent modulo n:

$$\prod_{i=1}^{\varphi(n)} x_i \equiv \prod_{i=1}^{\varphi(n)} ax_i \pmod{n}.$$

But the right-hand side, rearranging, is just $\varphi(n)$ copies of a times $x_1 x_2 \cdots x_{\varphi(n)}$, i.e. the right-hand side is equal to $a^{\varphi(n)}$ times the left-hand side. Since the two are congruent modulo n and $x_1 \cdots x_{\varphi(n)}$ is invertible since all of its factors are, with inverse $x_1^{-1} \cdots x_{\varphi(n)}^{-1}$, we conclude that $a^{\varphi(n)} \equiv 1 \pmod{n}$. □

Therefore for any Dirichlet character χ modulo n and any integer m relatively prime to n, we have $\chi(m^{\varphi(n)}) = \chi(1)$, since χ only cares about the value of its input modulo n and $m^{\varphi(n)} \equiv 1 \pmod{n}$ by the above lemma. On the other hand $\chi(m^{\varphi(n)}) = \chi(m)^{\varphi(n)}$ and $\chi(1) = 1$, so $\chi(m)^{\varphi(n)} = 1$. Therefore $\chi(m)$ is a $\varphi(n)$th root of unity for all m relatively prime to n.[47]

Since there are only finitely many $\varphi(n)$th roots of unity and χ is determined by $\varphi(n)$ values modulo n, this shows that there are only finitely many Dirichlet characters of any given modulus.[48] With that in hand, let's try to actually construct some.

[47] Recalling that χ is complex-valued, the kth roots of unity are the k complex numbers z such that $z^k = 1$. These all have absolute value 1, and so lie on the unit circle in the complex plane. For example, the fourth roots of unity are 1, i, -1, and $-i$, since clearly $1^4 = 1$, $(-1)^4 = ((-1)^2)^2 = 1^2 = 1$, and $i^4 = (i^2)^2 = (-1)^2 = 1$, and similarly for $-i$.

[48] A Dirichlet character modulo n is said to be of modulus n.

Let $n = 4$. By periodicity, it's enough to describe χ at 0, 1, 2, and 3. Since 0 and 2 are not invertible, $\chi(0) = \chi(2) = 0$, and we know that $\chi(1) = 1$, so it remains only to determine χ at 3. If $\chi(3) = 1$, then χ is the trivial character χ_0. On the other hand, $\varphi(4) = 2$, since the only integers between 1 and 4 relatively prime to 4 are 1 and 3, so $\chi(3)$ must be a square root of 1, i.e. 1 or -1. The only remaining possibility is -1, so there are two characters modulo 4 total: the trivial character and the one defined by $\chi(3) = -1$ (it's easy to check that both of these are multiplicative). Let's make a table, discarding the parts that don't give us any information.

	1	3
χ_0	1	1
χ_1	1	-1

Well, that wasn't too hard, we didn't even have to use multiplicativity to restrict the possibilities. Let's try $n = 5$, which has a lot more nontrivial terms: we have $\chi(0) = 0$ and $\chi(1) = 1$, but we need to determine all the others. This is still not too hard, though: we have $2^2 = 4$ and $2^3 = 8 \equiv 3 \pmod 5$, so $\chi(4) = \chi(2)^2$ and $\chi(3) = \chi(2)^3$. Therefore it suffices to determine $\chi(2)$. Since $\varphi(5) = 4$ (in general $\varphi(p) = p - 1$ for any prime p) $\chi(2)$ is a fourth root of unity, i.e. 1, -1, i, or $-i$, corresponding to four possible characters:

	1	2	3	4
χ_0	1	1	1	1
χ_1	1	i	$-i$	-1
χ_2	1	$-i$	i	-1
χ_3	1	-1	-1	1

Note that in both of these the columns as well as the rows form Dirichlet characters, and indeed they are symmetric under exchanging rows and columns (for a particular choice of ordering of the characters!).

In particular, it turns out that there will always be exactly $\varphi(n)$ (nonzero) Dirichlet characters modulo n. Also observe that the product of two Dirichlet characters is also a Dirichlet character, which is easy to verify from the definition, and can make it faster to generate them.

The major result for Dirichlet characters, and the reason that we care about them, is the following.

Proposition 1.9.12 (Orthogonality relations)**.** *Let χ_1 and χ_2 be two Dirichlet characters modulo n, and a and b be two integers which are invertible modulo n. Then we have*

$$\sum_{a=0}^{n-1} \chi_1(a)\overline{\chi_2(a)} = \begin{cases} \varphi(n) & \chi_1 = \chi_2 \\ 0 & \text{else} \end{cases}$$

and

$$\sum_{\chi} \chi(a)\overline{\chi(b)} = \begin{cases} \varphi(n) & a = b \\ 0 & \text{else} \end{cases}$$

where $\overline{z}$ denotes complex conjugation and the second sum is taken over all Dirichlet characters χ modulo n.[49]

[49]Recall that if $z = x + iy$ then $\overline{z} = x - iy$.

Proof. First, observe that $\chi_1\overline{\chi_2}$ is itself a Dirichlet character, since the complex conjugate of a character is a character and the product of two characters is a character, so it suffices for the first claim to show that

$$\sum_{a=0}^{n-1} \chi(a)$$

is equal to $\varphi(n)$ if $\chi = \chi_0$ and is 0 otherwise. The first claim is clear, since χ_0 is nonzero at exactly $\varphi(n)$ points and has value 1 at all of those, so the sum of its values is $\varphi(n)$. Suppose that χ is nontrivial, i.e. it takes on some value not equal to 1 at some $b \in (\mathbb{Z}/n\mathbb{Z})^\times$. Let

$$S = \sum_{a=0}^{n-1} \chi(a),$$

and fix such a $b \in (\mathbb{Z}/n\mathbb{Z})^\times$. Then

$$\chi(b)S = \sum_{a=0}^{n-1} \chi(a)\chi(b) = \sum_{a=0}^{n-1} \chi(ab).$$

Write $c = ab$; then since b is invertible we have $a \equiv b^{-1}c \pmod{n}$, so summing over a is equivalent to summing over c up to order. Therefore

$$\chi(b)S = \sum_{c=0}^{n-1} \chi(c) = S.$$

Since $\chi(b) \neq 1$ by assumption, we must have $S = 0$.

Similarly, for the second claim since $\overline{z} = z^{-1}$ for z a root of unity[50] we have $\chi(a)\overline{\chi(b)} = \chi(a)\chi(b)^{-1} = \chi(ab^{-1})$ since b is invertible, and $\chi(b)\chi(b^{-1}) = \chi(bb^{-1}) = \chi(1) = 1$. Therefore it suffices to show that

$$T := \sum_{\chi} \chi(a)$$

is $\varphi(n)$ if $a = 1$ and 0 otherwise, since ab^{-1} is 1 if and only if $a = b$. Again the first part is easy: $\chi(1) = 1$ for every nonzero χ and there are $\varphi(n)$ nonzero Dirichlet characters modulo n, so the sum of their values at n is $\varphi(n)$. Assume therefore that $a \not\equiv 1 \pmod{n}$. Fix some nontrivial character χ_1. Then

$$\chi_1(a)T = \sum_{\chi} \chi(a)\chi_1(a).$$

Since the product of any two characters is a character and, at the invertible classes, χ_1 is invertible (like all Dirichlet characters), multiplication by χ_1 just permutes the sum over the characters: letting $\chi' = \chi_1\chi$, we can just as well sum over χ' to get

$$\chi_1(a)T = \sum_{\chi'} \chi'(a) = T.$$

[50]This is not hard to see: by Euler's formula if z is a kth root of unity then $z = e^{2\pi im/k}$ for some integer m, so $\overline{z} = e^{-2\pi im/k} = z^{-1}$.

Since $a \not\equiv 1 \pmod{n}$, there exists some character χ_1 such that $\chi_1(a) \neq 1$. To see this, consider the subset of $(\mathbb{Z}/n\mathbb{Z})^\times$ consisting of the equivalence classes of $a, a^2, a^3, \ldots$; since $(\mathbb{Z}/n\mathbb{Z})^\times$ is finite, this sequence must start repeating eventually. It may take up all of $(\mathbb{Z}/n\mathbb{Z})^\times$, or may be just a subset; either way denote its size by k. Fix χ_1 to be 1 outside of this subset. On it, the only restriction on $\chi_1(a)$ is that $\chi_1(a)^k = 1$, since $a^k \equiv 1 \pmod{n}$, so for $k > 1$ we can always find some χ_1 with $\chi_1(a)$ not equal to 1; and $k = 1$ occurs only if $a = 1$, so we're safe. Therefore $\chi_1(a)T = T$ and $\chi_1(a) \neq 1$, so T must be 0 as above. □

In particular, note that the function

$$\sum_\chi \overline{\chi(a)}\chi$$

is precisely $\varphi(n) \cdot \mathbf{1}_{a,n}$ (where we've put the complex conjugation on the other factor for convenience, just by taking the complex conjugate of the whole thing), and so by multiplying by $\frac{1}{\varphi(n)}$ we've found our desired decomposition of $\mathbf{1}_{a,n}$ into completely multiplicative functions.

Now we can apply our Dirichlet series and Euler product methods. In particular, the Dirichlet series of Dirichlet characters, like of the constant function, each get a special name:

$$L(\chi, s) = \mathcal{D}[\chi](s) = \sum_{k \geq 1} \frac{\chi(k)}{k^s} = \prod_p \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}$$

by Proposition 1.9.8. These are called Dirichlet L-functions.

But how, after all, does this help? We don't really know how to say anything about the right-hand side, and it's not clear how we can put it together to get information about primes congruent to a modulo n. But actually we can say something about the sum formula. Let

$$F_\chi(x) = \sum_{k \leq x} \chi(k).$$

We have

$$F_\chi(x) = \sum_{k=0}^{n-1} \chi(k) + \sum_{k=n}^{2n-1} \chi(k) + \cdots + \sum_{k=(\lfloor x/n \rfloor - 1)n}^{\lfloor x/n \rfloor n - 1} \chi(k) + \sum_{k=\lfloor x/n \rfloor n}^{x} \chi(k),$$

and by the orthogonality relations all of these sums except the last one are 0 when χ is nontrivial. When χ is trivial, each of these sums is precisely $\varphi(n)$ and there are $\lfloor x/n \rfloor$ of them, so

$$F_{\chi_0}(x) = \frac{\varphi(n)}{n} x + O(1).$$

This is enough to let us prove the following proposition.

Proposition 1.9.13. *If χ is nontrivial, $L(\chi, s)$ converges for all $\mathrm{Re}(s) > 0$, and for χ_0 the trivial character $L(\chi_0, s)$ converges for $\mathrm{Re}(s) > 1$, and extends to the region $\mathrm{Re}(s) > 0$ except for a simple pole at $s = 1$.*

Proof. By Proposition 1.5.1, we have

$$L(\chi, s) = s \int_1^\infty \frac{F_\chi(x)}{x^{s+1}}\, dx.$$

By the discussion above, we have $F_\chi(x) = O(1)$ when χ is nontrivial, and $\frac{\varphi(n)}{n}x + O(1)$ when χ is trivial. In the first case this means that we can bound the integrand by

$$\frac{C}{x^{s+1}}$$

for some constant C, so the integral converges for all s with real part greater than 0; in the second case we have

$$L(\chi_0, s) - s\int_1^\infty \frac{\frac{\varphi(n)}{n}x}{x^{s+1}}\, dx = L(\chi_0, s) - s\frac{\varphi(n)}{n}\int_1^\infty \frac{1}{x^s}\, dx = L(\chi_0, s) - \frac{\varphi(n)}{n}\frac{s}{s-1}$$

equal to

$$s\int_1^\infty \frac{O(1)}{x^{s+1}}\, dx,$$

which as above converges for $\mathrm{Re}(s) > 0$. Therefore $L(\chi_0, s)$ is equal to $\frac{\varphi(n)}{n}\frac{s}{s-1}$ plus something convergent for $\mathrm{Re}(s) > 0$, and so has a simple pole at $s = 1$ and extends analytically to the region with $\mathrm{Re}(s) > 0$. □

This is most of the analytic information we'll need (though there's a key point still to come). Let's look on the other side of our formula: consider

$$\log L(\chi, s) = -\sum_p \log(1 - \chi(p)p^{-s}) = \sum_p \sum_{k\geq 1} \frac{\chi(p)^k}{kp^{ks}}.$$

Now let's sum over all χ. Then we have

$$\sum_\chi \log L(\chi, s) = \sum_\chi \sum_p \sum_{k\geq 1} \frac{\chi(p^k)}{kp^{ks}} = \sum_p \sum_{k\geq 1} \frac{1}{kp^{ks}} \sum_\chi \chi(p^k).$$

By the orthogonality relations, the innermost sum is 0 unless $p^k \equiv 1 \pmod n$ and $\varphi(n)$ otherwise. This only works for $a = 1$, so we need to make a twist:

$$\sum_\chi \overline{\chi(a)} \log L(\chi, s) = \sum_p \sum_{k\geq 1} \frac{\mathbf{1}_{a,n}(p^k)}{kp^{ks}}.$$

While not quite equal to our original desired product formula, this still looks like the right sort of object by comparison with the prime number theorem, where we look at

$$\log \zeta(s) = \sum_p \sum_{k\geq 1} \frac{1}{kp^{ks}}.$$

In fact, we can make the same simplification: let $\Lambda_{a,n}(x)$ be $\log p$ if $x = p^k$ for some prime p and integer k, and 0 otherwise. Then differentiating the above formula gives

$$-\sum_{\chi} \overline{\chi(a)} \frac{L'(\chi, s)}{L(\chi, s)} = \sum_{k \geq 1} \frac{\Lambda_{a,n}(k)}{k^s},$$

which, defining $\psi_{a,n}$ appropriately, eventually leads to a proof of Theorem 1.9.6.

However, we're trying to prove a weaker result, so we don't need to go there. In fact, to prove that there are infinitely many primes congruent to a modulo n, it suffices to show that this has a pole at $s = 1$: observe that the sum over $k \geq 2$ converges for $\mathrm{Re}(s) > \frac{1}{2}$, since

$$\sum_{p} \sum_{k \geq 2} \frac{1}{k p^{ks}} = \sum_{k \geq 2} \frac{1}{k} \sum_{p} p^{-ks} \leq \sum_{k \geq 2} \frac{1}{k} \sum_{j \geq 1} j^{-ks} = \sum_{k \geq 2} \frac{\zeta(ks) - 1}{k},$$

and since $\zeta(ks) \sim \frac{1}{ks-1}$ in this region this is approximately

$$\sum_{k \geq 2} \frac{1}{k^2 s - k}.$$

Since $k^2 s - k$ increases quadratically, this sum converges unless $k^2 s - k = 0$ for some k, i.e. $s = \frac{1}{k}$, so as long as $\mathrm{Re}(s) > \frac{1}{2}$ this sum will always converge.[51] Therefore any divergence must come from the $k = 1$ term, which is just

$$\sum_{p} \frac{\mathbf{1}_{a,n}(p)}{p^s},$$

which of course can only diverge if there are infinitely many primes congruent to a modulo n.

Thus it suffices to show that

$$\sum_{\chi} \overline{\chi(a)} \log L(\chi, s)$$

has a pole at $s = 1$. Around $s = 1$, we know that $L(\chi, s)$ is bounded for χ nontrivial and has a simple pole if χ is trivial, so if $L(\chi, s) \neq 0$ for every nontrivial χ then

$$\sum_{\chi \neq \chi_0} \overline{\chi(a)} \log L(\chi, s)$$

is a bounded term approaching a constant as $s \to 1$, so that the pole of $L(\chi_0, s)$ dominates. Therefore it suffices to show that $L(\chi, s) \neq 0$ for every nontrivial χ.

Proving this for real χ (i.e. taking only real values) turns out to be much harder than for complex χ (i.e. not everywhere real). It's possible to do both directly, in part with similar ideas as used in Proposition 1.8.4, but there's a more elegant proof following from methods in algebraic number theory. Rather than work out the details here let's use this as motivation to push forward into algebra.

[51] This is not a rigorous proof as is, but can be made into one.

Chapter 2

Introduction to algebra

2.1 Sets

In number theory, broadly speaking, we're interested in the properties of the natural numbers $\mathbb{N}$, and related objects such as the integers $\mathbb{Z}$ or the rationals $\mathbb{Q}$. These are very special objects in various ways, but we can take some of their properties and ask what other objects have these properties. Enlarging the set of objects we're interested allows us to develop much more powerful tools to attack problems in number theory,[1] and also gives rise to interesting mathematics in its own right. Algebra deals with these more general objects, and with the relationships between them.

The most basic algebraic structure is a set, which is a term we've already been using freely. What is a set? Well, it's a collection of elements, where the elements are...something; we don't actually care what, only that they exist. They could be numbers, functions, themselves sets, or anything else.

What properties do sets have? Given a set S, we can think about its size or order, i.e. how many elements it contains, written $|S|$; we can also think about its subsets. For example, the set $\{0, 1\}$ has exactly four subsets: the empty set, $\{\}$, which contains no elements; the set $\{0\}$; the set $\{1\}$; and the set $\{0, 1\}$ itself.[2] Write $\mathcal{P}(S)$ for the *power set* of S: the set of all subsets of S. Thus for example

$$\mathcal{P}(\{0, 1\}) = \{\{\}, \{0\}, \{1\}, \{0, 1\}\}.$$

Proposition 2.1.1. *For any finite set S, its power set $\mathcal{P}(S)$ is finite, with order $|\mathcal{P}(S)| = 2^{|S|}$.*

Proof. Suppose that S has n elements, written $x_1, \ldots, x_n$. We want to know how many ways there are of choosing a subset $T \subseteq S$. For each x_i, we can choose it to either be in T or not, so there are two possibilities for each x_i; since these are all independent, the total number of ways of choosing T is $2^n = 2^{|S|}$. □

[1] Not unlike how developing methods in complex analysis lets us prove the prime number theorem, despite complex numbers not being obviously related to prime numbers.

[2] Sets are denoted by curly braces like these, where the elements are whatever's inside. Also note that any set S has both the empty set (generally a very important set) and S itself as subsets, written $\{\} \subseteq S$ and $S \subseteq S$ (if we are trying to denote subsets not equal to all of S, these are called proper subsets, and are written $\subset S$. These two operators should be thought of as analogous to $\leq$ and $<$.

For finite sets, we can talk meaningfully about their size, but for infinite sets (such as the natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$ all we can say is that they are infinite. Can we do better? How can we compare them? For example, we might be tempted to say that $\mathbb{N}$ is strictly smaller than $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, since $\mathbb{N}$ is a proper subset of $\mathbb{Z}$; similarly, we might be tempted to say that $\mathbb{Z}$ is strictly smaller than $\mathbb{Q}$, which is strictly smaller than the real numbers $\mathbb{R}$. But after all, these are all infinite sets: how can we compare their size?

As will frequently be a theme in algebra, the answer is that we should look not at sets themselves, but at functions between them. A function $f : S \to T$ for two sets S and T is just something that assigns to every element $s \in S$ a corresponding element $f(s) \in T$.

Definition 2.1.2. A function $f : S \to T$ is called *injective* or an *injection* if it does not take any two $s \in S$ to the same value of T, i.e. if $f(s) = f(s')$ then $s = s'$.

Example 2.1.3. If $S \subseteq T$, then sending $s \in S$ to the corresponding element $s \in T$ is an injection, since distinct elements of S are still distinct when viewed as an element of T.

Definition 2.1.4. A function $f : S \to T$ is called *surjective* or a *surjection* if for every $t \in T$ there is some $s \in S$ such that $f(s) = T$.

Example 2.1.5. If T is a set with one object, say $T = \{0\}$, then every function $f : S \to T$ from a nonempty set S is a surjection: the only $t \in T$ is 0, and since $f(s) \in T$ for every s we must have $f(s) = 0$, so as long as there exists some $s \in S$ any map to T must be surjective.

These two terms will be useful throughout algebra: after all, much of algebra is about sets with additional structure, and so maps between these objects are just functions (of sets) satisfying certain properties.

Definition 2.1.6. A function $f : S \to T$ is called *bijective* or a *bijection* if it is both injective and surjective, or equivalently if for every $t \in T$ there is *exactly one* $s \in S$ such that $f(s) = t$. Such functions are also called one-to-one.

For example, there is a bijection between the sets $\{0, 1\}$ and $\{2, 3\}$, given by sending 0 to 2 and 1 to 3 (or vice versa). We say that two sets are in bijection if there is a bijection between them—notice that if there is a bijection $f : S \to T$, then there is an inverse bijection $f^{-1} : T \to S$ sending $t \in T$ to the unique $s \in S$ such that $f(s) = t$. These are inverses in that first applying f and then f^{-1} gives the identity function on S, sending each $s \in S$ to itself, and first applying f^{-1} and then f gives the identity function on T. (In fact, this is an equivalent definition of a bijection: a function with an inverse, i.e. some f^{-1} satisfying these properties.)

Thus in fact being in bijection is an equivalence relation: the above shows that it is symmetric, S and T are in bijection if and only if T and S are; any set is in bijection with itself, with bijection given by the identity function; and if S and T are in bijection and so are T and U, then composing the bijections $f : S \to T$ and $g : T \to U$ gives a bijection $g \circ f : S \to U$, so S and U are in bijection (you can check that $g \circ f$, the function given by first applying f and then g, is still a bijection if f and g are; indeed, the same thing is true for injections and bijections).

For finite sets, it's not hard to see that two sets are in bijection if and only if they have the same size: a bijection is just a one-to-one pairing of elements of the two sets, which is possible if and only if they have the same size. We will use this rule to talk about the sizes

of infinite sets as well: we say that two sets S and T, finite or infinite, have the same size if and only they are in bijection.

Thus when we talk about sizes of sets we really mean equivalence classes of sets under the equivalence relation of being in bijection. Indeed, this is one of the ways of defining natural numbers: a natural number is an equivalence class of *finite* sets.[3] If $S \subseteq T$, then we certainly don't want to say that S is strictly bigger than T, so we'll define an order on these equivalence classes by saying that $|S| \leq |T|$ if $S \subseteq T$, and therefore $|S'| \leq |T|$ if S' is in bijection with S. Equivalently, $|S| \leq |T|$ if there is an injection $S \to T$, since this can be separated into a bijection between S and some subset of T.[4]

Does this say anything useful? We can distinguish finite sets of different size, by noting that e.g. $\{0, 1\}$ and $\{2, 3, 4\}$ can never be in bijection, and we can say that a finite set and an infinite set can never be in bijection and therefore have different sizes, but we knew all this before. Can we use this to distinguish larger sets?

Let's look at some examples. Consider the set $E = \{2, 4, 6, 8, \ldots\}$ of even positive integers. Since this is a proper subset of the natural numbers $\mathbb{N}$ we might expect that $|E|$ is strictly smaller than $|\mathbb{N}|$. But in fact there is a natural bijection $\mathbb{N} \to E$, given by $f(n) = 2n$. This is a bijection: every integer n will be sent to a different even number $2n$, so it is injective, and every even number $2n$ is twice some integer n, so it is surjective. Therefore $\mathbb{N}$ and E have the same size.

This is characteristic of infinite sets, and in fact a set is in bijection with some proper subset of itself if and only if it is infinite. This leads to thought experiments such as Hilbert's hotel: imagine a hotel with infinitely many rooms, numbered 1, 2, 3, and so on. Suppose that they are all full. A new guest comes to the hotel. Despite the fact that the rooms are all full, the hotel can still accommodate them: move the guest in room 1 to room 2, the guest in room 2 to room 3, and so on. Then room 1 will be empty, and the new guest can move into it. The reason for this counterintuitive behavior is that the sets $\{1, 2, 3, \ldots\}$ and $\{2, 3, 4, \ldots\}$ are in bijection, with the map $f(n) = n + 1$ going from the first to the second. Generally speaking we should expect our intuition to fail quite a lot around infinite sets, especially very large ones.[5]

We asked before about the relative sizes of $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. A similar argument to the above shows that $|\mathbb{N}| = |\mathbb{Z}|$ via the following bijection: $f(1) = 0$, $f(2) = 1$, $f(3) = -1$, $f(4) = 2$, $f(5) = -2$, and so on: $f(n) = (-1)^n \cdot \lfloor n/2 \rfloor$.[6] You can check that this is a bijection.

The situation for $\mathbb{Q}$ is complicated, but essentially similar. A rational number can be written in the form $\frac{a}{b}$, for a and b relatively prime; therefore there is a natural injection $\mathbb{Q} \to \mathbb{Z}^2$, i.e. pairs of two integers, taking $\frac{a}{b}$ to the pair (a, b). Therefore $|\mathbb{Q}| \leq |\mathbb{Z}^2|$. Now imagine the following "spiral" bijection: start at $(0, 0)$, then $(0, 1)$, then $(1, 1)$, then $(1, 0)$, then $(1, -1)$, then $(0, -1)$, and so on, spiraling outwards from the origin. Sending n to the nth element of this sequence gives the desired bijection between $\mathbb{N}$ and $\mathbb{Z}^2$, so $|\mathbb{Q}| \leq |\mathbb{N}|$. On

[3] This includes 0, since the empty set has size 0, which is why some people consider $\mathbb{N}$ to contain 0, though that's not our convention.

[4] It can be shown that this is equivalent to the statement that $|S| \geq |T|$ if there is a surjection $S \to T$. Generally injections and surjections are "dual" in this sort of way.

[5] We'll see what this means shortly!

[6] Recall that $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

the other hand, $\mathbb{N}$ is the smallest possible infinite set, i.e. there is an injection from $\mathbb{N}$ to any infinite set: if S is an infinite set, we can pick an element x_1 in S, then pick another distinct element x_2, and so on, so that the map $n \mapsto x_n$[7] is an injection.[8]

When we're talking about infinite sets, the term "size" no longer makes as much intuitive sense, since what we really mean is an equivalence class of sets; we'll also start using the term "cardinality." Every nonnegative integer is a cardinality, since there exists a set with that size; the only other cardinality we've seen so far is the cardinality of sets in bijection with $\mathbb{N}$. Since these are the smallest infinite sets, in the sense explained above, we'll refer to this cardinality as $\aleph_0$. It has the property that $n < \aleph_0$ for every integer n.

Do there exist any larger sets, i.e. infinite sets which are not in bijection with $\mathbb{N}$ and therefore have cardinality greater than $\aleph_0$? The following proposition tells us that the answer must be yes.

Proposition 2.1.7. *For any set S, its power set $\mathcal{P}(S)$ has cardinality strictly larger than S.*

Proof. Suppose that $|\mathcal{P}(S)| \leq |S|$. As mentioned above, this is equivalent to the statement that there is some surjection $f : S \to \mathcal{P}(S)$, i.e. for every $s \in S$ the function f assigns to it a subset $T \subseteq S$, since the elements of $\mathcal{P}(S)$ are subsets of S. Consider the subset $U \subseteq S$ of elements $s \in S$ such that s is not in the subset $f(s)$. Since f is surjective, there exists some element $t \in S$ such that $f(t) = U$. Now: is t itself an element of U?

If it is, then, by the definition of U, t is not in $f(t)$. But $f(t)$ is just U, a contradiction. On the other hand, if it is not in U, then by definition t is in $f(t)$; but again $f(t) = U$, so $t \in U$, another contradiction. Therefore no such surjection can exist, and so $|\mathcal{P}(S)| > |S|$. □

Therefore for any set of cardinality $\aleph_0$, e.g. $\mathbb{N}$, its power set has strictly larger cardinality.

Consider the real numbers $\mathbb{R}$, written as binary expansions. For example, 13 is written 1101, and $\frac{1}{2}$ is written 0.1. We can think of both of these as extending infinitely in both directions, with all but finitely many of the digits being 0; for some real numbers, the digits to the right will not all be 0 but will continue. For example, $\frac{1}{3}$ in binary is $0.01010101\ldots$. Regardless, in this way we can think of $\mathbb{R}$ as the set of infinite[9] strings of zeros and ones.[10] These can be though of as functions from $\mathbb{N}$ to the finite set $\{0, 1\}$, sending n to the binary digit in the nth place. The set of functions between two sets $S \to T$ is written T^S, and its cardinality can be thought of as the exponentiation $|T|^{|S|}$: for each element of S, we are choosing an element of T, and if T is finite then each element of S contributes $|T|$ possibilities, which are then multiplied together. Therefore in this language we have shown that $|\mathbb{R}| = 2^{|\mathbb{N}|} = 2^{\aleph_0}$. But in fact this is just the cardinality of $\mathcal{P}(\mathbb{N})$, as Proposition **??** suggests: choosing 0 or 1 for each integer can be thought of as choosing whether that integer

[7] I don't remember if we've had this notation before; it just means that n maps to x_n.

[8] This also means that any set in bijection with $\mathbb{N}$ has this property.

[9] Specifically, extending to a finite but arbitrarily large number of digits to the left and infinitely to the right. For simplicity, we can think of these as extending only to the right, by restricting to real numbers between 0 and 1: indeed, the interval $[0, 1]$ of such real numbers is in bijection with all of $\mathbb{R}$, as can be seen for example by graphing the function $\tan(\pi x)$.

[10] There is a subtlety we're eliding, which is that real numbers can have more than one binary representation; for example, 1 can be written as both 1 and as $0.11111\ldots$, in much the same way as in decimal digits it can be written as either 1 or $0.99999\ldots$. However, this turns out not to matter: there are "few enough" representations that the same idea works.

is in our subset, i.e. we can identify each real number with (say) the set of 1's in its binary expansion. Thus $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$, which is strictly larger than $|\mathbb{N}| = \aleph_0$.

This process can be continued to reach increasingly large sets, and there are more complicated processes that generate much larger ones that cannot be reached just via power sets in any obvious way; but it's rare that we'll interact with sets much larger than $\mathbb{R}$. (Note that, via a process similar to that used to show that $\mathbb{Q}$ is in bijection with $\mathbb{N}$ though somewhat more complicated, it's possible to show that $\mathbb{R}$ is in bijection with $\mathbb{R}^n$ for any natural number n; in particular, $|\mathbb{R}| = |\mathbb{R}^2|$, and there's a natural bijection between $\mathbb{C}$ and $\mathbb{R}^2$ by sending $x+iy$ to the pair (x, y), so $|\mathbb{C}| = |\mathbb{R}| = 2^{\aleph_0}$.) Before we move on, though, there are two notes that should be made.

First, we have successfully generated a set with cardinality larger than $\aleph_0$, by using the "exponentiation" operator $\mathcal{P}$. Can we do any better? That is, can we find a set with cardinality strictly larger than $\mathbb{N}$, but smaller than $\mathbb{R}$? Is there a cardinality between $\aleph_0$ and $2^{\aleph_0}$?

In the early days of set theory, this was a hotly debated question. Since $\mathbb{R}$ is sometimes called the continuum, as the infinite, continuous number line, the conjecture that in fact there exists no such cardinality is known as the continuum hypothesis.

Before discussing the truth of the continuum hypothesis, though, we need to address a more threatening problem: set theory, as described above, is not consistent![11]

For example, let S be the set of all sets. Now consider its power set $\mathcal{P}(S)$. On the one hand, all the elements of $\mathcal{P}(S)$ are sets, so they are all elements of S and so $\mathcal{P}(S) \subseteq S$. Therefore $|\mathcal{P}(S)| \leq |S|$. On the other hand, by Proposition 2.1.7 the opposite is true: $|\mathcal{P}(S)| > |S|$! This is an unavoidable contradiction: we have not made any assumptions to reach this point, yet the contradiction remains. Have we broken math?

We might be suspicious that the problem lies in the power set operation, and that for some reason Proposition 2.1.7 should not apply here (though in fact we've proven it for arbitrary sets). Consider therefore the following example, known as Russell's paradox, after Bertrand Russell: perhaps surprisingly, this rather than the above was the original paradox that led to the downfall of naive set theory. Let S be the set of all sets which do not contain themselves. Is S an element of itself? Take a moment to figure out why both answers are impossible.

This helps show that the problem is not in any of our operations, but in the constructions of our sets themselves. The hidden assumption that we were making, which led to these contradictions, is that it is always possible to consider the set of objects of some kind which satisfy some property. In fact this is not true, and sets can only be formed according to specific rules. The question of which rules these should be occupied early set theorists for some time, and ultimately led to a set of eight axioms defining sets due to Zermelo and Fraenkel, as well as one slightly controversial one known as the Axiom of Choice (which we won't get into now, but has various counterintuitive results). These nine axioms together are known as ZFC, for Zermelo-Fraenkel plus Choice, and together can be used to construct a version of set theory without these contradictions; indeed, in principle all of mathematics can be derived from these axioms, since set theory provides a basis for mathematics.

The resolution to the paradoxes above is then that e.g. the set of all sets *is not a set*!

[11] For this reason it is known as "naive set theory."

Instead, it's something called a proper class: a class is something that can be defined in the way we wanted to in naive set theory, i.e. by some property of its elements, such as being sets or being sets which do not contain themselves, and a proper class is a class which is not a set. But classes have no notion of cardinality or of containing each other, and so the paradoxes above do not arise.

Another way of getting around these issues is to fix a very large set V, called a universe, and work only with objects which are contained within that universe. In this setting, it is true that all classes are sets, and therefore we can ignore most issues of size, except that for sufficiently large sets we cannot do operations like taking the power set.

With the development of ZFC (as well as many alternate logic systems, with which some set theorists and logicians are still concerned today), many people became hopeful that it was possible to develop a complete basis for mathematics, such that all possible coherent mathematical statements could (at least in principle, with sufficient work) be either proven or disproven from the axioms. Gödel dashed these hopes with his incompleteness theorems (which we will not get into at the moment, but approximately state that for any sufficiently powerful[12] consistent mathematical framework there must exist statements which can be neither proven nor disproven. This was a major blow, but quite abstract for some time, until in 1963 Paul Cohen proved that the continuum hypothesis was such a statement for ZFC: the continuum hypothesis is *independent* of ZFC, meaning that adding in either the continuum hypothesis or its negation as a tenth axiom still leads to a consistent theory (if ZFC is consistent, which cannot be proved in ZFC).

2.2 Groups

Much of the content of algebra is in analyzing objects which consist of a set with some additional structure. The first examples of such objects that we'll look at are groups.

Definition 2.2.1. A *group* is a pair $(G, *)$, where G is a set and $*$ is an operation[13], satisfying the following conditions:

(1) The set G is *closed* under the operation $*$, i.e. for any two elements a, b of G their product[14] is indeed an element of G;[15]

(2) The operation $*$ is associative;[16]

[12]This in practice is quite a weak requirement; the only framework that ever arises in practice which is not Gödel incomplete is Euclidean geometry, which is actually complete and for which there exists an algorithm which can automatically prove or disprove any coherent statement (though this is not done in practice).

[13]An operation on a set G can be thought of as a function $G \times G \to G$ taking a pair of elements $a, b \in G$ to another element $a * b$ of G. It is written in this way, rather than in functional notation $*(a, b)$, by analogy to addition and multiplication, $a + b$ or $a \times b$, which are two basic examples of operations. Sometimes these are referred to as *binary* operations, since they act on a pair of elements, but (at least if associative) they can be extended to act on arbitrarily many elements.

[14]We will often use the terminology of multiplication for the group operation, such as "product" or "factor," or the multiplicative notion x^{-1} for the group inverse; this should not be taken to mean that the operation is necessarily actually multiplication, it's just a convention for convenience.

[15]This property can also be taken to be part of the definition of an operator, rather than of a group.

[16]Recall from the discussion preceding Proposition 1.3.4 that an operation is *associative* if for every $a, b, c \in$

(3) There exists an *identity element* $e \in G$, i.e. an element such that for any $x \in G$, we have $e * x = x * e = x$;

(4) For every element x of G, there exists some element x^{-1}, the *inverse* of x, such that $x^{-1} * x = x * x^{-1} = e$.

We'll often refer to a group $(G, *)$ just as G, with the operation implicit.

Proposition 2.2.2. *For a fixed group G, its identity element is unique; and for any $x \in G$, the inverse of x is unique.*

Proof. What this is saying is that the properties which e and x^{-1} must satisfy uniquely define them. To see this, suppose that there exist two identity elements e_1 and e_2. Then $e_1 e_2 = e_2$, since e_1 is an identity element; but on the other hand $e_1 e_2 = e_1$, since e_2 is an identity. Therefore $e_1 = e_2$, so the identity is unique. Similarly, suppose that x has two inverses, x_1^{-1} and x_2^{-1}. Then $x_1^{-1} x x_2^{-1} = x_2^{-1}$, since x_1^{-1} is an inverse of x, but on the other hand $x_1^{-1} x x_2^{-1} = x_1^{-1}$, since x_2^{-1} is an inverse of x, so $x_1^{-1} = x_2^{-1}$. In other words, if we assume that these elements exist, we also get for free that they are unique. □

Example 2.2.3. Let $G = \mathbb{Z}$, with the operation $+$. This is a group. Let's verify each requirement: for any two integers a and b, their sum $a + b$ is also an integer; addition is associative, since $a + (b + c) = (a + b) + c$ for any integers a, b, c; there exists an identity element, specifically 0: $0 + x = x + 0 = x$ for any integer x; and for any integer x there exists an integer $-x$ such that $-x + x = x + (-x) = 0$, so $-x$ is the inverse of x.

The same argument shows that the set of real numbers under addition $(\mathbb{R}, +)$ is also a group. What about the same set with a different operation, specifically multiplication?

Example 2.2.4. The pair $(\mathbb{R}, \times)$ is *not* a group: the product of any two real numbers is real, you can check that multiplication is associative, and it has an identity, specifically 1, since $1 \times x = x \times 1 = x$ for every real number x. But there is no real number x such that $0 \times x = x \times 0 = 1$, since $x \times 0 = 0 \times x = 0$, so 0 has no inverse in $\mathbb{R}$ and so $(\mathbb{R}, \times)$ cannot be a group.

However, this is relatively easy to fix: 0 is the only problematic element, and so by removing 0 from $\mathbb{R}$ we get a group, whose underlying set is the set of nonzero reals, written $\mathbb{R}^\times$, and whose operation is multiplication, with inverse $x^{-1} = \frac{1}{x}$.

What *is* $(\mathbb{R}, \times)$, if not a group? Well, it's a set with an operation, and there's no guarantee that this should be a useful or interesting enough structure to be worth thinking about any more. But in fact this particular type of structure can be interesting and has a name.

Definition 2.2.5. A *monoid* is a pair $(G, *)$ of a set G and an operation $*$, satisfying all of the properties of groups except that not every element must have an inverse.[17]

Let $G^\times \subseteq G$ be the set of invertible elements of G, i.e. those with inverses.[18] Then $(G^\times, *)$ is a group, called its group of units.

G we have $(a * b) * c = a * (b * c)$, i.e. the order in which we put the elements together doesn't matter, and so $a * b * c$ is well-defined. (If the operation was not associative, $a * b * c$ would be ambiguous: does it mean $(a * b) * c$ or $a * (b * c)$?)

[17] Thus all groups are monoids, and some but not all monoids are groups.

[18] Recall the use of this notation for invertible classes modulo n in Section 1.9.

Thus $(\mathbb{R}, \times)$ is a monoid, and $(\mathbb{R}^\times, \times)$ is its group of units.

Example 2.2.6. The pair $(\mathbb{Z}, \times)$ is a monoid, with identity 1; its group of units has underlying set just $\{1, -1\}$. This also explains the terminology "group of units": the "units" of $\mathbb{Z}$ are ± 1, and the term is extended to other monoids even when it is less intuitive.

All of our examples so far, both of groups and monoids, have had a certain nice property: they're all abelian.

Definition 2.2.7. An operation $*$ on a set G is *commutative* or *abelian*[19] if for every $a, b \in G$ we have $a * b = b * a$. If $(G, *)$ is a monoid or a group, it is called a commutative or abelian monoid or group if $*$ is commutative.

Abelian groups are very nice, and make our lives simple. Unfortunately not all groups are abelian.

Example 2.2.8. Let T be a set of n elements, say

$$T = \{1, 2, 3, \ldots, n\}.$$

Consider the set of *permutations* of T, i.e. bijections $\sigma : T \to T$. For a fixed permutation σ, write

$$\{\sigma(1), \sigma(2), \ldots, \sigma(n)\}$$

for its image, even though as sets these are equal; this is a rearrangement of the elements of T. For example, for $n = 2$ the only possible permutations of T are $\{1, 2\}$, the identity permutation which takes each element to itself, and $\{2, 1\}$, which exchanges the two elements.

We'll also use the following notation for permutations, called cycle notation: our permutation σ sends 1 to $\sigma(1)$, and then $\sigma(1)$ to some other element $\sigma(\sigma(1))$, and so on. Since T is finite, eventually this pattern starts to repeat. For example, let $n = 4$ and consider the permutation taking $\{1, 2, 3, 4\}$ to $\{2, 3, 1, 4\}$, which we can think of as a chain

$$1 \mapsto 2 \mapsto 3 \mapsto 1,$$

together with 4 which maps to itself. This is written

$$(1, 2, 3)(4);$$

here 4 is a fixed point of the permutation, i.e. $\sigma(4) = 4$, and we'll often leave fixed points out of the notation, so that if we know that we're talking about permutations on 4 elements then $(1, 2, 3)$ means the one with image $\{2, 3, 1, 4\}$.[20] We could also have multiple nontrivial cycles: for example, the permutation with image $\{4, 3, 2, 1\}$ sends $1 \mapsto 4 \mapsto 1$ and $2 \mapsto 3 \mapsto 2$, i.e. it's just exchanging two pairs of elements, so this is written $(1, 4)(2, 3)$.[21]

[19] Named after the mathematician Niels Henrik Abel, a short-lived but phenomenally productive mathematician from the early 19th century whose work had virtually nothing to do with abelian groups.

[20] Thus the identity permutation, which fixes every element, is written as the "empty cycle" ().

[21] If you want to gain more familiarity with these, spend some time writing down permutations of sets of various size, and find the corresponding cycle notation. Permutations turn out to be pretty important objects, so understanding them is useful.

Write S_n for the set of permutations of $T = \{1, 2, 3, \ldots, n\}$. We can define an operation on S_n: given two permutations σ_1 and σ_2, which recall are just bijections $T \to T$, we can form their composition $\sigma_1 \circ \sigma_2$, taking some $k \in T$ to $\sigma_1(\sigma_2(k))$.[22] I claim that this operation makes S_n a group.

Let's check: it's not hard to see that the composition of two permutations is also a permutation, since the composition of bijections is bijective. Composition of functions is generally associative, essentially because it's defined to be: $(\sigma_1 \circ \sigma_2) \circ \sigma_3$ takes k to $\sigma_3(k)$ and then to $(\sigma_1 \circ \sigma_2)(\sigma_3(k))$, which by definition is $\sigma_1(\sigma_2(\sigma_3(k)))$, which itself is $\sigma_1 \circ (\sigma_1 \circ \sigma_2)$ applied to k. The identity element is the identity permutation: if we call it e, then $e(\sigma(k)) = \sigma(k)$ by definition, and $\sigma(e(k)) = \sigma(k)$, so

$$e \circ \sigma = \sigma \circ e = \sigma.$$

To see that every permutation has an inverse, it's best to think of them as reordering the elements of T: if σ reorders the set in one way, then we can put the elements back where they originally were, and this operation is the inverse of σ. (Alternatively, we could note that σ is a bijection and so has an inverse, which is itself a bijection and therefore also a permutation of T.)[23]

So S_n is a group, called the *symmetric group* on n elements. Is it abelian? Well, let's examine its structure for some small values of n. If $n = 1$, then $T = \{1\}$ has only one map $T \to T$, i.e. the identity taking $1 \mapsto 1$, so S_1 is the group with one element, which we'll call the trivial group: it has an identity and no other elements. Next, S_2 is the group with two elements, the identity e and one other, which we'll call σ: the only possible action which isn't determined by the fact that e is the identity is $\sigma \circ \sigma$, but since σ just switches the two elements of T if we apply it twice it switches them back to the starting configuration, i.e. $\sigma \circ \sigma = e$.[24] Thus both of S_1 and S_2 are commutative: in fact, we can write down their multiplication tables[25] and observe that they are symmetric, and therefore commutative.

	e
e	e

	e	σ
e	e	σ
σ	σ	e

What about for $n = 3$? Well, the possible permutations reorder $\{1, 2, 3\}$ to $\{1, 2, 3\}$, $\{1, 3, 2\}$, $\{2, 3, 1\}$, $\{2, 1, 3\}$, $\{3, 1, 2\}$, or $\{3, 2, 1\}$. Let's label the first of these as e, since

[22] This is a more general notation for composition of functions: if we have two functions $f : X \to Y$ and $g : Y \to Z$, we can think of their compositions $g \circ f : X \to Z$ taking x to $f(x)$ and then to $g(f(x))$. It's important to note that when evaluating a composition $g \circ f$, the function on the right acts first: first we apply f, then g, so that the result is $g(f(x))$ rather than $f(g(x))$. This point can be confusing; a good check is that the composition has to go $X \to Y \to Z$, so we need to apply whichever one acts on X first. (Here where $X = Y = Z$ that doesn't help, though.)

[23] This perspective of a group as a set of invertible functions $T \to T$ for some object T turns out to be a very powerful idea, which we'll come back to when we talk about category theory.

[24] Continuing with the multiplicative notation, we will start to drop the group operation from our notation and treat it as if it was multiplication, so that $\sigma_1 \circ \sigma_2$ might be written $\sigma_1 \sigma_2$, and $\sigma \circ \sigma$ will usually be written σ^2.

[25] In these, the entry with row corresponding to some group element g and column corresponding to a group element h will always be $g * h$ (rather than say $h * g$, which will be important for the noncommutative case).

it's the identity, and call the next one, say, x. What x does is fix 1 and switch 2 and 3, so if we apply it twice we get back the identity, i.e. $x^2 = e$. The next one, call it y, sends $1 \mapsto 3 \mapsto 2 \mapsto 1$, i.e. it rotates the elements forward by one place; since there are three total, if we do this operation three times we get back to the identity, i.e. $y^3 = e$.

What about the remaining three permutations? I claim we can generate all of them from x and y. The next one, $\{2, 1, 3\}$, we can get by first applying y and then switching the last two elements, i.e. applying x; thus this is $x \circ y$, or just xy. Next, $\{3, 1, 2\}$ sends $1 \mapsto 3 \mapsto 2 \mapsto 1$, i.e. it is the inverse of y, written y^{-1}. Since $y^3 = e$, we have

$$y^{-1} = ey^{-1} = y^3y^{-1} = y^2(yy^{-1}) = y^2e = y^2,$$

so this is just y^2. Finally, we can get $\{3, 2, 1\}$ by applying y^2 and then x, so it is xy^2.

Now that we've identified all of the permutations in this way, let's examine the multiplication table for S_3, which we can compute directly from calculating what happens when we compose the permutations:[26]

	e	x	y	xy	y^2	xy^2
e	e	x	y	xy	y^2	xy^2
x	x	e	xy	y	xy^2	y^2
y	y	xy^2	y^2	x	e	xy
xy	xy	y^2	xy^2	e	x	y
y^2	y^2	xy	e	xy^2	y	x
xy^2	xy^2	y	x	y^2	xy	e

We can observe from this table that S_3 is nonabelian: for example, $yx = xy^2 \neq xy$.

In fact, the computation that $yx = xy^2$, together with the knowledge that $x^2 = y^3 = e$, is enough to allow us to compute the entire multiplication table, without even knowing what x and y "really" are, and in fact is enough to show that this group has exactly these six elements. A priori, an element of the group generated by x and y could look like

$$x^{a_1}y^{b_1}x^{a_2}y^{b_2} \cdots x^{a_n}y^{b_n}$$

for some sets of exponents a_i and b_i, and since we don't know that these commute (and in fact they don't) it's not initially clear that this group is even finite. However, knowing that $yx = xy^2$ allows us to rearrange this: think of our product as something like $xxxyyyxxyyyyx$, expanding out the exponentials; then this rule allows us to push each y past a following x at the cost of replacing y with y^2, and we can continue to do this until we have something of the form x^ay^b. For example, consider the group element

$$g = x^2yxy^3x.$$

Then y^3x is just

$$y^2yx = y^2xy^2 = yyxy^2 = yxy^2y^2 = xy^2y^2y^2 = xy^6,$$

[26]This is also a good exercise to get more familiar with permutations. For extra credit, do it in terms of cycle notation. (If you find something that disagrees with my table, ask me, it's very possible that you're right and I'm wrong.)

so our formula is now

$$g = x^2yxy^3x = x^2yx^2y^6.$$

Next,

$$yx^2 = yxx = xy^2x = xyyx = xyxy^2 = xxy^2y^2 = x^2y^4,$$

so we have

$$g = x^2yx^2y^6 = x^2x^2y^4y^6 = x^4y^{10}.$$

So we can put any group element in the form x^ay^b for some integers a and b. Next, the requirement that $x^2 = e$ and $y^3 = e$ ensures that we can set $0 \leq a \leq 1$ and $0 \leq b \leq 2$, so in fact $e = x^0y^0$, $x = x^1y^0$, $y = x^0y^1$, $xy = x^1y^1$, $y^2 = x^0y^2$, and $xy^2 = x^1y^2$ are the only elements of our group. For example, $x^4 = x^2x^2 = ee = e = x^0$, and $y^{10} = y^3y^3y^3y = e^3y = y = y^1$, so above we actually have $g = y$.

The more general version of the problem above is called the group word problem: given an arbitrary group G and two strings of elements of G (also known as words), for example x^2yxy^3x and y, we want to know whether they are equal in G. In this case the answer is yes, and actually we can see this more easily than we did above:

$$x^2 = y^3 = e,$$

so the first string is actually

$$eyxex = yx^2 = ye = y.$$

In general, though, the word problem is quite hard, and for sufficiently complicated groups is actually insolvable.

Determining a group from a set of generators, here x and y, and relations, here $x^2 = y^3 = e$ and $yx = xy^2$, is a common method of defining a group. Groups which can be defined by a *finite* set of generators and relations are called finitely generated; this is one of the most important classes of groups, and most (but not all) groups which we will look at will be finitely generated.

The simplest type of finitely generated groups are *free groups*: those with no relations whatsoever. For example, consider the free group on three generators x, y, and z (written F_3). An element of this group is a string

$$x^{a_1}y^{b_1}z^{c_1}x^{a_2}y^{b_2}z^{c_2}\cdots x^{a_n}y^{b_n}z^{c_n}$$

for integers a_i, b_i, c_i (note that these can also be negative, since by requiring F_3 to be a group we automatically get the inverses x^{-1}, y^{-1}, and z^{-1}.) In a free group, the only "relation" is that

$$xx^{-1} = x^{-1}x = e$$

and similarly for the other generators, so two distinct strings can represent different elements of the group only if they can be transformed into each other by expanding or contracting inverses. For example,

$$xy^{-1}zz^{-1}y = xy^{-1}y = x.$$

The simplest free group is the free group on one generator x, F_1. An element of this group is x^a for some integer a; in fact, in a very strong sense we can think of F_1 as *being* the

integers $\mathbb{Z}$. Specifically, there is a clear bijection $\mathbb{Z} \to F_1$ sending an integer a to x^a; and it turns out that this respects the group structure on each of $\mathbb{Z}$ (under addition) and F_1. We'll come back to this idea.

The simplest finitely generated groups, then, are those with one generator x and either no relations, in which case we get F_1, or one relation. A relation is an equality between two words in the group, e.g. $yx = xy^2$ in S_3 as above, and can in fact always be phrased as setting some element equal to the identity e; for example, in that case it is is

$$yx(xy^2)^{-1} = yxyx = e.^{27}$$

So in our case this is a relation of the form $x^a = e$ for some integer a, and therefore $x^{ma} = e^m = e$ for every integer e. Thus $e, x, x^2, \ldots, x^{a-1}$ are all of the elements of the group, with $x^a = e$, $x^{a+1} = x$, and so on, and $x^{-1} = x^{a-1}$, $x^{-2} = x^{a-2}$, and so on. This is a finite group of order a, generated by x. Generally any group generated by a single element is called *cyclic*; the cyclic group of infinite order is F_1, and there is a cyclic group C_n of order n for every positive integer n.

Proposition 2.2.9. *Every cyclic group is abelian.*

Proof. Let x be a generator of our cyclic group G, so that any element $g \in G$ can be written as $g = x^a$ for some integer a. Then if $h = x^b$ is another element, we have

$$gh = x^a x^b = x^{a+b} = x^{b+a} = x^b x^a = hg,$$

so G is abelian. □

As above, C_1 is the trivial group $\{e\}$, and S_2 is cyclic of order 2 and so essentially equivalent to C_2.

What does this mean? We'd like to have some notion of equivalence of groups, since we can have two definitions for objects that should be essentially the same thing, such as S_2 and C_2. For a sillier example, consider the free group on one generator x and the free group on one generator y. Clearly these are essentially the same thing: the only difference is that we've changed the name of the generator!

As with sets, to find an appropriate equivalence relation we should look not just at groups themselves but at maps between them. But if we look at just functions between groups, we lose the group structure: if $(G, *)$ and $(H, \star)$ are groups, then the collection of maps $f : G \to H$ doesn't know about their group structure, but only about their structure as sets. Instead, we'll define a better notion.

Definition 2.2.10. A (group) *homomorphism* $f : G \to H$ for groups $(G, *)$ and $(H, \star)$ is a function $f : G \to H$ respecting the group structure, i.e. for every two elements $x, y \in G$, we have $f(x * y) = f(x) \star f(y)$.

Since a homomorphism is also a function, we can talk about injective, surjective, or bijective homomorphisms.

[27] Exercise: verify that for any two elements x, y of any group G we have $(xy)^{-1} = y^{-1}x^{-1}$. In this case, we have $(y^2)^{-1} = y^{-2} = y$, since $y^3 = e$, and $x^{-1} = x$, since $x^2 = e$.

Definition 2.2.11. An *isomorphism* is a bijective homomorphism, or equivalently a homomorphism with an inverse which is also a homomorphism.

This turns out to be the correct notion of equivalence for groups, just as bijection did for sets, and henceforth when we talk about groups we will usually really mean isomorphism classes of groups. Check for yourself that isomorphism is an equivalence relation, and that the properties of groups we have mentioned so far, such as order or whether or not they are abelian, are preserved under isomorphism.

Example 2.2.12. Consider the bijection $\mathbb{Z} \to F_1$ mentioned before takin $a \in \mathbb{Z}$ to $x^a \in F_1$, where F_1 is generated by x. We already know that it is a bijection, so to show that it is an isomorphism we need only show that it respects the group operation. On $\mathbb{Z}$, the group operation is $+$, so what we need to show is that $x^{a+b} = x^a x^b$. But this is true almost by definition: x^{a+b} is the product of $a + b$ copies of x, and $x^a x^b$ is the product of a copies of x with b copies of x for a total of $a + b$ copies. Therefore $\mathbb{Z}$ and F_1 are isomorphic, written $\mathbb{Z} \cong F_1$ or $\mathbb{Z} \simeq F_1$,[28] and so we can say that the cyclic groups are C_n for every integer n and $\mathbb{Z}$.

One of the advantages of talking about isomorphism classes rather than groups directly is that, as with sets, it makes the set of objects much smaller. There are infinitely many groups, which we can see by noting that there are already infinitely many finite cyclic groups, one for each positive integer. However, just as for sets there is only one isomorphism class of each cardinality, for any positive integer n there are finitely many groups of order n. How do we know? Well, the structure of a group is encoded, up to isomorphism, in its multiplication table, in which there are n^2 slots, each of which is an element of the group, so there are n possibilities for each. Therefore there are at most n^3 isomorphism classes of groups of order n.[29]

Sometimes, however, the situation is much simpler.

Proposition 2.2.13. *Suppose that G is a group of prime order p. Then G is cyclic: $G \simeq C_p$.*

To prove this, we'll need to introduce some more machinery.

Definition 2.2.14. The *order* of an element g of a group G with identity e is the smallest positive integer n such that $g^n = e$, if it exists.[30]

Definition 2.2.15. A *subgroup* of a group $(G, *)$ is a subset $H \subseteq G$ such that $(H, *)$ is also a group.

Given a group G and some element g, we will often talk about the set $\langle g \rangle$ of elements generated by g, i.e. $\langle g \rangle = \{\ldots, g^{-2}, g^{-1}, e, g, g^2, \ldots\}$. If g has finite order n, then $\langle g \rangle$ will be a finite cyclic group of order n, generated by g. In particular, it is a subgroup of G.

[28] There is a vague and only occasionally-followed convention that $\cong$ is used to denote general isomorphism, and $\simeq$ is used to denote a *canonical* isomorphism, which we will not define but which means something like that the two objects are isomorphic for a good reason in a particular way. In the context of groups, this does not make much difference, and especially for finitely generated groups there is a sense in which all isomorphisms are canonical; but for more general objects this difference can be important, and we may come back to it.

[29] In fact the true bound is much smaller, but is in general still quite large: for example, there are 267 groups of order 64. Exercise: how many groups are there of order 4? What are they?

[30] Not to be confused with the order of a group, which is the number of elements in that group.

Definition 2.2.16. Let G be a group, $H \subseteq G$ be a subgroup, and $x \in G$ be an element. The *left coset* of H in G at x is the set of elements xh for every $h \in H$, written xH; similarly the *right coset* of H in G at x is the set of elements hx for every $h \in H$, written Hx.

Cosets in G are *not* necessarily subgroups.

Suppose that $y = xh$, for some $h \in H$. Then yH gives the same left coset as xH: the elements of yH are xhj for every $j \in H$, and multiplication by h just rearranges the elements of H, so that yH is the same as xH in a different order.

For example, let $G = S_3$, generated by x and y, and let $H = \langle y \rangle = \{e, y, y^2\}$. Then $xH = \{x, xy, xy^2\}$, and $xyH = \{xy, xy^2, x\}$. These are the same sets reordered, and since we don't care about order we conclude that $xH = yH$.

Given a group G and a subgroup H, define an equivalence relation on G by $x \sim_H y$ if $xH = yH$ (check for yourself that this is in fact an equivalence relation). If x and y are not equivalent under this relation, then $xH \neq yH$, so there exists some element $xh \in xH$ that is not in yH. Suppose that xH and yH have some common element $j = xh_1 = yh_2$. Then $jh_1^{-1}h = xh_1h_1^{-1}h = xh$, which we know is in xH but not in yH. But on the other hand $jh_1^{-1}h$ is in jH, since $h_1^{-1}h$ is an element of H; and since $j \in yH$ we have $j \sim_H y$ and so $jh_1^{-1}h = xh$ is also in yH, a contradiction. Therefore we have proven that no such j can exist, i.e. if xH and yH are at all different then they are completely disjoint: left cosets of H in G split G into some number of disjoint sets. Let $[G : H]$ be the number of left cosets of H in G, called the index of H in G.[31]

Proposition 2.2.17 (Lagrange's theorem)**.** *Let G be a finite group, and $H \subseteq G$ be a subgroup. Then*

$$|G| = [G : H] \cdot |H|.$$

In particular, the order of G is divisible by the order of H.

Proof. From the discussion above, we are almost done: we know that G is split into $[G : H]$ distinct left cosets of H. If we can prove that each of these has size $|H|$, then we're done.

Each coset xH is, first of all, a set, and we know that to show that two sets are the same size it suffices to give a bijection between them. Let xH and yH be two cosets. Then the map $xH \to yH$ given by multiplication by yx^{-1} takes xh to $yx^{-1}xh = yh$, and since it does this for every h this is a bijection. Thus all of the cosets are the same size. But the coset eH for the identity e is just H, so all of the cosets have size $|H|$. □

Proof of Proposition 2.2.13. Let G be a finite group of prime order p. Let $x \in G$ be any element other than the identity. Then $\langle x \rangle$ is a subgroup of G; since x is not the identity, $\langle x \rangle$ contains at least e and x, so its size is at least 2. By Proposition 2.2.17, $|\langle x \rangle|$ divides p; but since p is prime and $|\langle x \rangle| \geq 2$, we must have $|\langle x \rangle| = p$, i.e. x generates all of G. Therefore G is cyclic, with generator x.[32] □

[31] It is not hard to see that we could have done all of the above for right cosets instead; prove for yourself that there are the same number of left cosets as right cosets, so that we are not picking one perspective or the other in defining $[G : H]$.

[32] Note that x was arbitrary; this is a good example of the fact that the choice of generator is *not* unique, and indeed in this case any element of the group other than the identity generates the group equally well (though this will not always be true).

Thus there is only one isomorphism class of groups of prime order. It is also possible to work out similar classification theorems for groups of more complicated order, such as for the squares of primes; but we won't bother for the moment.

Instead, we'll state two more classification results: one for finite abelian groups, and one for finitely generated abelian groups. The proofs are somewhat long and not terribly enlightening so we'll skip them for now

Given two groups $(G, *)$ and $(H, \star)$, we can always produce a third from them via the *direct product* $G \times H$, whose elements are pairs (x, y), with $x \in G$ and $y \in H$. The group operation on these is defined by $(x_1, y_1) \circ (x_2, y_2) = (x_1 * x_2, y_1 \star y_2)$.

Theorem 2.2.18. *Every finite abelian group is isomorphic to the direct product of cyclic groups C_{p^r} for some prime number p and positive integer r.*

For example, consider the cyclic group of order 72. I claim that this is isomorphic to the product of the cyclic groups corresponding to its largest prime power factors, i.e. $C_{72} \simeq C_8 \times C_9$. To see this, think of C_{72} as the integers $0, 1, 2, \ldots, 70, 71$ under the operation of addition modulo 72, and of C_8 as the numbers $0, 1, \ldots, 7$ modulo 8 and C_9 as $0, 1, \ldots, 8$ modulo 9. Consider the map $C_8 \times C_9 \to C_{72}$ taking (x, y) to $9x + y$. You can check that this map is both injective and surjective; and since it is linear in both factors, it is also a homomorphism.[33] Therefore $C_{72} \simeq C_8 \times C_9$.

Theorem 2.2.19. *Every finitely generated abelian group is isomorphic to the direct product $\mathbb{Z}^r \times T$ for some nonnegative integer r and some finite abelian group T.*

Given a finitely generated group G, writing $G \simeq \mathbb{Z}^r \times T$ for appropriate r and T the integer r is called the *rank* of G, and T is its *torsion part*: T corresponds to elements in G of finite order, while $\mathbb{Z}^r$ corresponds to elements of infinite (or undefined) order. The rank of a finite group is 0.

Let's pull back to the greater generality of all groups, rather than just abelian or finitely generated ones. Suppose we have two groups, G and H, and a homomorphism $f : G \to H$. Then this defines two new sets. First is the *kernel* of f, written $\ker(f)$. This is the set of elements $x \in G$ such that $f(x)$ is the identity e_H of H. The second is the image of f in H, i.e. the set of $y \in H$ such that there exists $x \in G$ such that $f(x) = y$, written $\mathrm{im}(f)$.

Proposition 2.2.20. *For any homomorphism $f : G \to H$, both $\ker(f)$ and $\mathrm{im}(f)$ are groups, with $\ker(f)$ a subgroup of G and $\mathrm{im}(f)$ a subgroup of H.*

Proof. First, let's focus on $\ker(f)$; it's a subset of G, so to show that it's a subgroup of G it suffices to prove that it's a group. Suppose that $x, y \in \ker(f)$. Then $f(xy) = f(x)f(y) = e_H e_H = e_H$, so $xy \in \ker(f)$, i.e. $\ker(f)$ is closed under the group operation. The group operation is associative on G, so it is also associative on $\ker(f)$; and the identity e_G on G is also the identity on $\ker(f)$,[34] so it remains only to prove that if $x \in \ker(f)$ then so is

[33] Take a moment to write down the expression for applying this map to a product $(x_1, y_1) * (x_2, y_2)$ to verify that this is a homomorphism.

[34] Any homomorphism must take the identity to the identity, $f(e_G) = e_H$, since $f(x) = f(e_G x) = f(e_G)f(x)$ and $f(x) = f(xe_G) = f(x)f(e_G)$, so $f(e_G)$ is an identity element for H and since the identity element is unique, as discussed, it follows that $f(e_G) = f(e_H)$.

x^{-1}. But $f(x^{-1})f(x) = f(x^{-1}x) = f(e_G) = e_H$, so $f(x^{-1}) = f(x)^{-1} = e_H^{-1} = e_H$. Therefore $x^{-1} \in \ker(f)$ for any $x \in \ker(f)$ and so $\ker(f)$ is a group, and thus a subgroup of G.

Similarly, if x and y are in $\mathrm{im}(f)$, say with preimages $\tilde{x}$ and $\tilde{y}$ such that $f(\tilde{x}) = x$ and $f(\tilde{y}) = y$, then $f(\tilde{x}\tilde{y}) = f(\tilde{x})f(\tilde{y}) = xy$, so $xy \in \mathrm{im}(f)$. Again, since H is a group the associativity of its group operation on $\mathrm{im}(f)$ follows, and since $f(e_G) = e_H$ it contains the identity, so it remains only to show that $\mathrm{im}(f)$ contains all inverses. But as above, if $x \in \mathrm{im}(f)$, with preimage $\tilde{x}$, then $f(\tilde{x}^{-1})x = f(\tilde{x}^{-1})f(\tilde{x}) = f(\tilde{x}^{-1}\tilde{x}) = f(e_G) = e_H$ and similarly for $xf(\tilde{x}^{-1})$, so $f(\tilde{x}^{-1})$ is an inverse for x. Therefore $\mathrm{im}(f)$ has all inverses, and so is a group and therefore, as above, a subgroup of H. □

Corollary 2.2.21. *A homomorphism $f : G \to H$ is injective if and only if its kernel is trivial.*[35]

Proof. First, assume that $f : G \to H$ is injective. The kernel of f is the set of elements $x \in G$ such that $f(x) = e_H$. Since f is injective, there can be at most one such element; but since $\ker(f)$ is a subgroup of G, it must contain the identity e_G, so $\ker(f) = \{e_G\}$ is the trivial group.

On the other hand, suppose that $\ker(f)$ is trivial, i.e. the only element of G whose image is e_H under f is e_G. Let x and y be two elements of G such that $f(x) = f(y)$. Then since f is a homomorphism, we have $f(xy^{-1}) = f(x)f(y^{-1}) = f(x)f(y)^{-1}$; but since $f(x) = f(y)$ we have $f(x)f(y)^{-1} = f(x)f(x)^{-1} = e_H$, so $f(xy^{-1}) = e_H$, i.e. xy^{-1} is in the kernel of f. Since $\ker(f)$ is trivial, we must then have $xy^{-1} = e_H$ and so $x = y$, so f is in fact injective. □

Thus where previously we had to check for every element $x \in H$ that the preimage $f^{-1}(x)$ had at most one element in order to know that f is injective, now we can just check $f^{-1}(e_H) = \ker(f)$.[36]

This seems strangely asymmetric: we think of injectivity and surjectivity as a pair, but we can check whether a homomorphism $f : G \to H$ is injective just by whether $\ker(f)$ is trivial, whereas to check if it's surjective we have to verify that every element of H is in $\mathrm{im}(f)$.[37] This asymmetry can be repaired, but to do so we need to introduce another tool for constructing groups: quotients.

Definition 2.2.22. Let G be a group, and $N \subseteq G$ a subgroup. We say that N is a *normal subgroup* of G if the left cosets of N in G are the same as the right cosets of N in G, i.e. for any $g \in G$ we have $gN = Ng$. Equivalently, for any $n \in N$ and $g \in G$, the product gng^{-1} is always in N.[38]

[35] Saying that a group is "trivial" means that it has only one element, the identity. The trivial group will often be written just as 1, with the multiplicative notation, or if we're using additive notation then it can be written just as 0.

[36] I'm not sure if we've used the notation $f^{-1}(x)$ before: what it means is, given a map $f : Y \to X$, the set of $y \in Y$ such that $f(y) = x$. Thus if $f : G \to H$ is a homomorphism then $f^{-1}(e_H)$ is by definition the kernel of f. More broadly, if $S \subseteq X$ is a subset of X, we will write $f^{-1}(S)$ for the subset of Y consisting of elements $y \in Y$ such that $f(y) \in S$. (Thus really the proper notation for $f^{-1}(x)$ would be $f^{-1}(\{x\})$, the preimage of the set containing only x, but this becomes excessively cumbersome.)

[37] Of course, writing down the equation $\mathrm{im}(f) = H$ is just as easy as $\ker(f) = 1$, but elementwise it's much easier to check if a group is trivial than to check if it contains every desired element.

[38] This operation of taking some element n and applying g to make it into gng^{-1} is called *conjugation* by g, and is an important way in which groups can act, especially on each other.

This definition might read as annoyingly technical, and indeed it should be thought of as a technical condition on subgroups, allowing us to define quotient groups. For abelian groups, the case is much simpler: all subgroups are normal.

Even in full generality, though, we already know one class of normal subgroups.

Proposition 2.2.23. *Let $f : G \to H$ be a homomorphism. Then $\ker(f)$ is a normal subgroup of G.*

Proof. Suppose that $h \in \ker(f)$, and $g \in G$. Then it suffices to show that $ghg^{-1} \in \ker(f)$, i.e. $f(ghg^{-1}) = e_H$. Since f is a homomorphism, we have $f(ghg^{-1}) = f(g)f(h)f(g^{-1})$. Since $h \in \ker(f)$, we have $f(h) = e_H$, so $f(ghg^{-1}) = f(g)e_H f(g)^{-1} = f(g)f(g)^{-1} = e_H$, so $ghg^{-1} \in \ker(f)$ as desired and so $\ker(f)$ is normal. □

Thus given a normal subgroup $N \subseteq G$, one might wonder whether N is the kernel of some homomorphism $G \to H$. In fact, we can always construct such a homomorphism, i.e. every normal subgroup is the kernel of some homomorphism. The desired construction is that of a quotient group.

Definition 2.2.24. Let G be a group, and N a normal subgroup. Then the *quotient group* G/N is the set of left cosets[39] of N in G, with group operation $(g_1N)(g_2N) = (g_1g_2)N$.

It's not hard to verify that this is a group; consider this an exercise. Instead, the hard part is verifying that this is well-defined at all. We've seen that being in the same left N-coset is an equivalence relation, $\sim_N$; the claim is that the set of these equivalence classes form a group. Suppose that $g_1 \sim_N g_1'$ and $g_2 \sim_N g_2'$, i.e. $g_1N = g_1'N$ and $g_2N = g_2'N$. Then for our group action to be well defined, we need to have $(g_1g_2)N = (g_1'g_2')N$: that is, the equivalence class that we get from the group action should not depend on the representative of the classes that we choose.

In fact, this well-definedness turns out to be exactly equivalent to requiring that N be normal! If N is normal, we have

$$g_1g_2N = g_1(g_2N) = g_1(g_2'N) = g_1(Ng_2') = (g_1N)g_2' = (g_1'N)g_2' = g_1'(Ng_2') = g_1'(g_2'N) = g_1'g_2'N$$

and so all is well; but if not, then we can choose g_1', g_2' such that at least one of these steps fail and so equality does not hold, so that the group action is not well-defined.

A quotient G/N comes equipped with a homomorphism $G \to G/N$, taking g to gN. Convince yourself that this is surjective and a homomorphism, with kernel N. It is sometimes called the quotient or structure homomorphism.

In fact, slightly more can be said.

Theorem 2.2.25 (First isomorphism theorem)**.** *Let $f : G \to H$ be a homomorphism. Then $G/\ker(f) \simeq \mathrm{im}(f)$. In fact, more is true: there is an isomorphism $\bar{f} : G/\ker(f) \to \mathrm{im}(f)$ such that the composition*

$$G \to G/\ker(f) \xrightarrow{\bar{f}} \mathrm{im}(f) \hookrightarrow H$$

[39] We can also do this with right cosets, in which case this is sometimes written $N\backslash G$; this doesn't usually matter, since the two groups are equivalent, but sometimes if we're going to quotient out G on the right (i.e. take left cosets) and then on the left (right cosets) by two different subgroups N_1 and N_2, this can make a difference and will be written $N_2\backslash G/N_1$.

is equal to $f : G \to H$, where the first map is the quotient map and the last one is the inclusion of $\operatorname{im}(f)$ *into* H.[40]

Proof. We'll prove this by constructing $\bar{f}$. Recall that elements of $G/\ker(f)$ are left cosets $g\ker(f)$ for $g \in G$, up to equivalence. Then we'd like to send $g\ker(f)$ to $f(g)$. Of course, this is only well-defined if $f(g_1) = f(g_2)$ for $g_1 \sim_{\ker(f)} g_2$, since otherwise $\bar{f}$ could give different values for different representations of the same element. But if $g_1\ker(f) = g_2\ker(f)$, then there exists some $x \in \ker(f)$ such that $g_1 = g_2x$, since $\ker(f)$ contains the identity and therefore $g_1\ker(f)$ contains $g_1e_G = g_1$. Therefore $f(g_1) = f(g_2x) = f(g_2)f(x) = f(g_2)e_H = f(g_2)$, so this map is well-defined.

Since its target is $\operatorname{im}(f)$, it's not hard to see that it is surjective: for any $f(g) \in \operatorname{im}(f)$, we have $\bar{f}(g\ker(f)) = f(g)$. Next, it is also injective: the kernel of $\bar{f}$ is the set of cosets $g\ker(f)$ such that $f(g) = e_H$, i.e. $g \in \ker(f)$. But any such g is $\ker(f)$-equivalent to the identity, and so $\ker(\bar{f}) = 1$. By Corollary 2.2.21, it follows that $\bar{f}$ is injective.[41]

Thus we've proven the first half of the claim, that there's an isomorphism $G/\ker(f) \to \operatorname{im}(f)$. To get the rest, we want to show that the composition sending $g \mapsto g\ker(f) \mapsto f(g)$ is the same as that taking $g \mapsto f(g)$; but since they both start and end in the same place, they are the same map. Thus $\bar{f}$ is as desired. □

There are more isomorphism theorems, but for the most part we won't need to deal with them.

Another way of stating Theorem 2.2.25 would be to say that *f factors through* the quotient $G/\ker(f)$, i.e. when we apply f to G we're really applying it to $G/\ker(f)$, and getting from G to $G/\ker(f)$ by some canonical route that doesn't depend on f, here the quotient homomorphism. This will be a useful idea going forward.

The last concept we need before progressing beyond groups is that of a group action.

Definition 2.2.26. Let G be a group, and S be a set. A *group action*[42] of G on S defines a function $S \to S$ for each element g of G satisfying

(1) The group identity e acts as the identity function, i.e. for any $s \in S$ we have $e \cdot s = s$;

(2) and for any two elements g and h of G, their action is *compatible*, i.e. for any $s \in S$ we have $g \cdot (h \cdot s) = (gh) \cdot s$.[43]

We will sometimes write $G \circlearrowright S$ for "G acts on S". A G-set is a set equipped with a particular G-action.

Example 2.2.27. Let G be the symmetric group on n elements S_n, and let S be any set with n elements. Then G acts on S by permutations, i.e. an element of G, interpreted as a permutation of $\{1, 2, \ldots, n\}$, permutes the elements of S the same way. (Verify that this is a group action.)

[40] The hooked arrow $\hookrightarrow$ is often used to denote injections; similarly the two-headed arrow $\twoheadrightarrow$ is sometimes used to denote surjections.

[41] See, I told you it would be useful.

[42] Strictly speaking, this defines a *left group action*, meaning that G acts on S by $s \mapsto gs$; we could also define a *right group action*, where $s \mapsto sg$, with similar requirements.

[43] That is, if we first act on s by h, yielding another element of S, and then act on this new element by g, the result is the same as from acting on s by gh.

Example 2.2.28. Let $G = \mathbb{R}^\times$, and $S = \mathbb{R}$. Then G acts on S by sending a real number x to rx, for any nonzero real number r. The identity of $G = \mathbb{R}^\times$ is 1, which takes $x \mapsto 1 \cdot x = x$, so it is the identity; and for two nonzero real numbers r and s, we have $r(xs) = rxs = (rx)s$, so this is indeed a group action.

Note, however, that this is not the *only* action of $\mathbb{R}^\times$ on $\mathbb{R}$. For example, $r \in \mathbb{R}^\times$ could take $x \in \mathbb{R}$ to r^2x; you can verify that this is also a group action. Thus $\mathbb{R}$ with the first action of $\mathbb{R}^\times$ and $\mathbb{R}$ with this second action, though the same set, are different $\mathbb{R}^\times$-sets, since they carry different $\mathbb{R}^\times$-actions.

Group actions come in many different flavors, such as free, faithful, transitive, etc., but we'll ignore them for now and define them if we need them. We would however like to define some related notions.

Definition 2.2.29. Let G be a group acting on a set S. The *orbit* of an element s of S, written $G \cdot s$, is the subset of S consisting of elements $g \cdot s$ for some $g \in G$.

Definition 2.2.30. Let G be a group acting on a set S. The *stabilizer* of an element s of S, written $\mathrm{Stab}(s)$, is the set of $g \in G$ such that $g \cdot s = s$.

For the first example above for S_n, convince yourself that the orbit of any element of our set S is all of S, and that every stabilizer is equal to all of S_n. For $\mathbb{R}^\times$ acting on $\mathbb{R}$, with either action there are two orbits: one contains every nonzero element, i.e. for any $x \in \mathbb{R}^\times$ its orbit is $R^\times \cdot x = \mathbb{R}^\times$, and one contains only 0, i.e. $\mathbb{R}^\times \cdot 0 = \{0\}$. For the first action, the stabilizer of any nonzero element is trivial, i.e. is only the trivial subgroup $\{1\}$, while the stabilizer of 0 is all of $\mathbb{R}^\times$. For the second action, though, there is one more element r of $\mathbb{R}^\times$ such that $r^2x = x$ for *any* x, namely $r = -1$. Therefore the stabilizer of any nonzero real number is $\{1, -1\} \simeq C_2$ (under multiplication), while the stabilizer of 0 is again all of $\mathbb{R}^\times$.

Lemma 2.2.31. *Let G be a group acting on a set S. For any $s \in S$, the stabilizer* $\mathrm{Stab}(s)$ *is a subgroup of G.*

Proof. If g and h both stabilize s, i.e. $g \cdot s = h \cdot s = s$, then $(gh) \cdot s = g \cdot (hs) = g \cdot s = s$, so gh also stabilizes s. The stabilizer $\mathrm{Stab}(s)$ always contains the identity, since $e \cdot s = s$ for any s, and if g stabilizes s then $g^{-1} \cdot s = g^{-1} \cdot (gs) = (g^{-1}g) \cdot s = e \cdot s = s$, so so does g^{-1}; thus $\mathrm{Stab}(s)$ is a group, and so a subgroup of G. □

Theorem 2.2.32 (Orbit-stabilizer theorem)**.** *Let G be a finite group acting on any set S. For any $s \in S$, we have*

$$|G| = |G \cdot s| \cdot |\,\mathrm{Stab}(s)|,$$

i.e. the size of the group G is equal to the product of the sizes of the orbit and the stabilizer of any element $s \in S$.

Proof. Fix some s, and let x be an element of $G \cdot x$. Suppose that $g, h \in G$ are such that $g \cdot s = h \cdot s = x$. Then $(g^{-1}h) \cdot s = g^{-1} \cdot (h \cdot s) = g^{-1} \cdot (g \cdot s) = e \cdot s = s$, so $g^{-1}h \in \mathrm{Stab}(s)$. Therefore the left coset $g\,\mathrm{Stab}(s)$ contains both g and h; in particular this shows that if $h \cdot s = x$ then $h \in g\,\mathrm{Stab}(s)$. On the other hand for any $j \in g\,\mathrm{Stab}(s)$ we can write $j = gj'$ for some $j' \in \mathrm{Stab}(s)$, so that $j \cdot s = gj' \cdot s = g \cdot s = x$, so $g\,\mathrm{Stab}(s)$ is exactly the subset of G which takes s to x.

Therefore every $x \in G \cdot x$ gives a distinct left coset of $\mathrm{Stab}(s)$, i.e. there is a bijection between $G \cdot x$ and the set of left cosets of $\mathrm{Stab}(s)$ in G. Since there are by definition $[G : \mathrm{Stab}(s)]$ of these, we have

$$|G \cdot s| = [G : \mathrm{Stab}(s)].$$

By Proposition 2.2.17, it follows that

$$|G| = |G \cdot s| \cdot |\mathrm{Stab}(s)|.$$

□

Thus for example for the permutation action of S_n on S, one of the computations above was redundant: once we know that e.g. the orbit of any element is all of S, for S of order n, the orbit-stabilizer theorem tells us that the stabilizer of any element must then be trivial, without need for any further computation.

2.3 Rings

From the point of view that algebra is concerned with generalizing properties of arithmetic, groups are only a good first step. We've successfully defined $\mathbb{Z}$ as a group, with the operation of addition; but in fact $\mathbb{Z}$ has another operation on it, that of multiplication. Under multiplication, $\mathbb{Z}$ is not a group but only a monoid. But studying the theories of groups and monoids can only tell us so much about the integers: the key property of $\mathbb{Z}$ is that it has these two structures simultaneously, and they are compatible, in the sense that multiplication is in some sense repeated addition. This can be formalized by the property that $a \cdot (b + c)$ should be something like $b + c$ copies of a, which should be equal to b copies of a plus c copies of a: that is, $a \cdot (b + c) = a \cdot b + a \cdot c$, the distributive law.

Definition 2.3.1. A *ring* is a triple $(R, +, \times)$, where R is a set and $+$ and $\times$ are binary operations, satisfying the following conditions:

(1) The pair $(R, +)$ is an *abelian* group;

(2) The pair $(R, \times)$ is a monoid (not necessarily abelian);[44]

(3) For any $a, b, c \in R$, we have $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(a + b) \cdot c = a \cdot c + b \cdot c$.[45]

As with groups, we will usually refer to a ring $(R, +, \times)$ just as R, and multiplication will usually be written with a dot $a \cdot b$ or just by juxtaposition ab rather than $a \times b$. The additive identity is written 0, and the multiplicative identity is 1.

[44] Some sources do not require that R have a multiplicative identity, and call a ring with multiplicative identity *unital*. For our purposes, all rings will be unital, so we won't use the term.

[45] These requirements are left and right distributivity, and are distinct requirements if multiplication is not commutative.

Remark. In fact, the assumption that the addition operator $+$ is commutative is not required, in the sense that it follows from the other axioms: consider the product $(1+1)\cdot(a+b)$. By left distributivity, this is $(1+1)a+(1+1)b$, which then by right distributivity is $a+a+b+b$. On the other hand, by right distributivity $(1+1)\cdot(a+b) = 1\cdot(a+b)+1\cdot(a+b)$, which since 1 is the multiplicative identity is just $a+b+a+b$. Therefore $a+a+b+b = a+b+a+b$. Since our ring is a group under addition, a and b have additive inverses $-a$ and $-b$, so we can add $-a$ on the left and $-b$ on the right of both sides to get $-a+(a+a+b+b)+(-b) = -a+(a+b+a+b)+(-b)$. Canceling inverses, this is just $a+b = b+a$, so addition is commutative.

Example 2.3.3. The most obvious example is the integers $\mathbb{Z}$, under normal addition and multiplication. Indeed the ring axioms are set up to generalize the properties of the integers.

Example 2.3.4. The real numbers $\mathbb{R}$ form a ring, again under normal addition and multiplication.

Example 2.3.5. The set of arithmetic functions $\mathbb{N} \to \mathbb{C}$ has (at least) two natural ring structures. For both, addition is normal ("pointwise") addition, with $f+g$ the function taking $n \mapsto f(n)+g(n)$ for any two arithmetic functions f and g, but for one of the ring structures multiplication is the more obvious pointwise multiplication $f \times g : n \mapsto f(n)\cdot g(n)$, while the other is an operation we've already explored, that of Dirichlet convolution: $f \times g = f * g$ takes

$$n \mapsto \sum_{d|n} f(d)g(n/d).$$

Example 2.3.6. More generally, let S be any set, and R be any ring. Then the set of functions $S \to R$ is a ring, with addition given by $f+g : s \mapsto f(s)+f(g)$ and $f \times g : s \mapsto f(s)\cdot g(s)$. In this case we say that the ring structure on the set of functions $S \to R$ is *induced by* the ring structure on R.

There are many ways of forming new rings from old ones. A particularly important one is the following.

Definition 2.3.7. Let R be a ring, and fix a formal variable x.[46] The *polynomial ring* in x over R, written $R[x]$, is the set of polynomials in the variable x with coefficients in R. It is a ring under the usual polynomial operations of addition and multiplication.[47]

This notation is called *adjoining* x to R, and might be said as "R adjoin x." This is to indicate that what we are really doing is taking R, as a set, and adding the single element x to get a new set, $R \cup \{x\}$.[48] This new set is no longer closed under the ring operations of addition and multiplication, so we take the closure

$$\overline{R \cup \{x\}}^{49}$$

[46]This just means that x should not be thought of as a "variable" in the sense of having some unknown value, but as being something of a different type than the elements of R.

[47]e.g. $(x+1)\cdot(x^2-2x+3) = x\cdot(x^2-2x+3)+1\cdot(x^2-2x+3) = x^3-2x^2+3x+x^2-2x+3 = x^3-x^2+x+3$.

[48]Oops, forgot to cover this in the section on sets; $\cup$ is an operator on sets meaning take the union, and $\cap$ is an operator meaning intersection.

[49]Closures of various kinds are often written with an overline $\overline{S}$ like this. In general, this means that we should add to our set precisely those elements which make it closed in whatever sense that might mean.

to get the set of all sums and products of elements of R with x.

Thus for example for $R = \mathbb{Z}$, the polynomial ring $\mathbb{Z}[x]$ contains $x + 2$, 3, and x, and therefore also contains $x(x + 2) + 3 = x^2 + 2x + 3$. Thus this really is the set of polynomials over our ring.[50]

It is also possible to adjoin other elements besides variables. For example, if R is a ring and $r \in R$, then we could consider $R[r]$. Since this is formed by taking the union of R with $\{r\}$ and then taking the closure, since r is already in R the union $R \cup \{r\}$ is already just R, so it closed under the ring operations. Thus $R[r]$ is just R.

We could also do something in between. Consider the ring $\mathbb{Z}$, and the *rational* number $\frac{1}{2}$. This is not an integer, but we could consider the ring $\mathbb{Z}[\frac{1}{2}]$. What does this mean? Well, it's the set of all numbers we can get by adding or multiplying integers and $\frac{1}{2}$. Therefore this contains first all the integers; $\frac{1}{2}$; $\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$; $\frac{29}{16} = 1 + 13 \cdot \left(\frac{1}{2}\right)^4$; etc. In other words, $\mathbb{Z}[\frac{1}{2}]$ is the ring of rational numbers whose denominator is a power of 2.

Since each ring R has an associated multiplicative monoid, we can think about its group of units $R^\times$, the subset of R which has a multiplicative inverse. For $\mathbb{Z}$, the group of units $\mathbb{Z}^\times$ is just $\{1, -1\}$. But for $\mathbb{Z}[\frac{1}{2}]$, we've enlarged it quite a bit: 1 and -1 are invertible, but now so is 2, and therefore so is 4, with inverse $\frac{1}{4}$, and so is -2, with inverse $-\frac{1}{2}$, and so on. Convince yourself that in fact $\mathbb{Z}[\frac{1}{2}]^\times = \{\ldots, -16, -8, -4, -2, -1, 1, 2, 4, 8, 16, \ldots\}$.

There are many different types of rings, even once we've restricted to looking only at commutative rings. One of the simplest and most important types is fields.

Definition 2.3.8. A *field* is a commutative[51] ring R such that $R^\times = R - \{0\}$, i.e. the only element of R which does not have a multiplicative inverse is 0.[52]

Why is this a natural type of ring to consider? Well, we require that $(R, +)$ be a group, and that $(R, \times)$ be a monoid. Since all groups are monoids (though not vice versa), we might wonder what happens if we require $(R, \times)$ to be not just a monoid but also a group. Unfortunately, this is impossible: the additive inverse 0 can never have a multiplicative inverse 0^{-1}, since $0 \cdot 0^{-1} = 0 \neq 1$ (assuming that R is not the zero ring). Thus, we look at the next-best case: that where $(R, \times)$ becomes a group after removing 0.

Example 2.3.9. We've encountered a number of fields so far: the rationals $\mathbb{Q}$, real numbers $\mathbb{R}$, and complex numbers $\mathbb{C}$. Verify first that these are all commutative rings, and second that any nonzero element of them has a multiplicative inverse (in that ring), so that they are all fields.

Example 2.3.10. On the other hand, most rings are not fields. For example, $\mathbb{Z}^\times = \{1, -1\}$; but e.g. 2 does not have a multiplicative inverse in $\mathbb{Z}$, since $2^{-1} = \frac{1}{2}$ is not an integer, so $\mathbb{Z}$ cannot be a field. Similarly, polynomial rings are not in general fields.

[50] Keep in mind that this closure adds additive inverses, so for example $-x$ is in $\mathbb{Z}[x]$, since those are required for the set of polynomials to be a ring, but it does not add multiplicative inverses since those are not required, so $\frac{1}{x}$ is *not* in $\mathbb{Z}[x]$.

[51] There also exist noncommutative versions, called *skew fields*, but they are less fundamental and important and more annoying to study.

[52] The simplest ring of all, which contains only one element 0, is called the "zero ring," often written just as 0, and is usually neglected in definitions of this type for simplicity because it does not have a distinct multiplicative and additive identity, i.e. $0 = 1$ in $\{0\}$. In this case, $\{0\}^\times = \{0\}$, because 0 *does* have a multiplicative inverse, namely 0, since $0 \cdot 0 = 0$ and 0 is also the multiplicative identity; so $\{0\}$ is *not* a field.

We can also define ring homomorphisms by analogy to group homomorphisms.

Definition 2.3.11. A (ring) *homomorphism* f from R to S, for two rings R and S, is a function $f : R \to S$ respecting the ring structure, i.e. for any two elements $x, y \in R$ we have $f(x+y) = f(x)+f(y)$ and $f(xy) = f(x)f(y)$.[53] In other words, f is a group homomorphism $(R, +) \to (S, +)$ and a monoid homomorphism[54] $(R, \times) \to (S, \times)$.

Just as for groups, a ring homomorphism can be injective, surjective, bijective, or none of the above. As for groups, a bijective ring homomorphism is called an *isomorphism* (or a ring isomorphism, or an isomorphism of rings). As with group homomorphisms, ring homomorphisms take identities to identities: for any ring homomorphism f, we have $f(0) = 0$ and $f(1) = 1$.

If fields are the "nicest"[55] type of ring, since our primary motivating example is $\mathbb{Z}$ and $\mathbb{Z}$ is not a field we should find some weaker notion that still includes the good properties of $\mathbb{Z}$. In particular, though most nonzero elements of $\mathbb{Z}$ are not invertible in $\mathbb{Z}$, they do still have inverses in some larger set: for example, 2 does not have an inverse in $\mathbb{Z}$, but it does have an inverse, $\frac{1}{2}$, in the rationals $\mathbb{Q}$. In other words $\mathbb{Z}$ can be embedded[56] into some field (here $\mathbb{Q}$).

The essence of this property turns out to be the following: for any two integers x and y, if their product xy is equal to 0 then at least one of x and y must be 0. This might seem like a simple property of multiplication that we might expect to be true in all rings, but it turns out not to be. Rings in which this is true are called integral domains.[57]

Definition 2.3.12. A commutative ring R is an *integral domain*[58] if for any $x, y \in R$, if $xy = 0$ than at least one of x and y is equal to 0.

Proposition 2.3.13. *A commutative ring R is an integral domain if and only if it can be embedded into a field, i.e. there exists some field F and an injective ring homomorphism $f : R \hookrightarrow F$.*

This justifies our claim that this property, sometimes called the zero product property, is an important and "nice" property of the integers.

Before we can prove this proposition, we need the following definition.

Definition 2.3.14. Let R be an integral domain. Then its *field of fractions* (or fraction field or quotient field) is the following ring: consider the set of pairs $(a, b) \in R \times R$ where

[53] Since we are concerned with unital rings, i.e. rings with a multiplicative identity 1, we also require that $f(1) = 1$. The requirement that $f(xy) = f(x)f(y)$ implies this unless $f(x) = 0$ for all x, so really this is just saying that the "zero homomorphism" sending everything to zero doesn't count (unless the target ring S is the zero ring, in which case $0 = 1$ and so $f(1) = 1$ is still satisfied).

[54] Which we have not defined, but is defined in the natural way: a map $f : M \to N$ for monoids M and N such that $f(xy) = f(x)f(y)$ for every $x, y \in M$ and $f(1_M) = 1_N$ for 1_M and 1_N the identities of each of M and N.

[55] By "nice" we usually mean having good properties that make it easy to work with. For example, one might say that the nicest groups are the cyclic groups.

[56] An embedding of rings is the same thing as an injective homomorphism; if R embeds into S, $R \hookrightarrow S$, then since R is isomorphic to its image under this injection we can think of R as a subring of S.

[57] The name, in particular the word "integral," is intended to denote that this is the fundamental property that makes a ring similar to the integers, though in fact $\mathbb{Z}$ has a variety of other special properties.

[58] Sometimes also referred to just as a domain; technically the difference is that a domain does not have to be commutative, but we'll use them interchangeably.

the second element of the pair is nonzero. We'll define the following equivalence relation: $(a,b) \sim (c,d)$ if $ad = bc$. Then the field of fractions $\mathrm{Frac}(R)$ is defined to be the set of equivalence classes under this relation, with two operations: $(a,b)+(c,d) = (ad+bc, bd)$ and $(a,b) \times (c,d) = (ac, bd)$.

This might seem like a strange and unmotivated definition, and it's not immediately clear that the two operations defined above are well-defined, much less well-behaved. Much becomes clear upon considering an example: let $R = \mathbb{Z}$, as $\mathbb{Z}$ is indeed an integral domain. For some suggestive notation, write a pair (a,b) as $\frac{a}{b}$. Then the equivalence relation is

$$\frac{a}{b} \sim \frac{c}{d} \quad \text{if and only if} \quad ad = bc.$$

But this is just the usual rule for fractions: two fractions $\frac{a}{b}$ and $\frac{c}{d}$ represent the same rational number if and only if $ad = bc$, as can easily be seen by multiplying both sides by bd. Further, the addition relation is exactly what you expect from fractions:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad}{bd} + \frac{bc}{bd} = \frac{ad+bc}{bd}$$

and

$$\frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd},$$

so we have a canonical ring isomorphism $\mathrm{Frac}(\mathbb{Z}) \simeq \mathbb{Q}$: the rational numbers $\mathbb{Q}$ are the field of fractions of $\mathbb{Z}$. Generally speaking, we have the following proposition, which gives us a way of thinking about $\mathrm{Frac}\, R$.

Proposition 2.3.15. *Let R be an integral domain. Then as the name suggests* $\mathrm{Frac}\, R$ *is a field, and there is a natural injection*[59] $R \hookrightarrow \mathrm{Frac}\, R$.

Proof. First, we need to show that $\mathrm{Frac}\, R$ is a ring. If we think of the elements of $\mathrm{Frac}\, R$ as fractions $\frac{a}{b}$ with a and b in R (b nonzero), this is not hard to see: verify that the rules for addition and multiplication satisfy the distributive laws and that such pairs (a,b) are closed under them, and they satisfy the group and monoid requirements; check that $\frac{1}{1}$ (and its equivalence class; anything of the form $\frac{a}{a}$ is equivalent to $\frac{1}{1}$) is a multiplicative identity, and that $\frac{0}{1}$ (and its equivalence class; anything of the form $\frac{0}{a}$ is equivalent to $\frac{0}{1}$) is an additive identity. From the definition of multiplication it's also clear that $\mathrm{Frac}\, R$ is a commutative ring, since R is.

Next, we need to show that any nonzero element $\frac{a}{b}$ of $\mathrm{Frac}\, R$ has an inverse. From the rational case, we expect that the inverse of $\frac{a}{b}$ should be $\frac{b}{a}$; and indeed $\frac{a}{b} \times \frac{b}{a} = \frac{ab}{ab} \sim \frac{1}{1}$. The only case in which this is impossible is if $a = 0$, so that $\frac{b}{a}$ is not defined. But in this case $\frac{a}{b} = \frac{0}{b} \sim \frac{0}{1}$ is the zero element of $\mathrm{Frac}\, R$, and so it is the one element that does not have to have an inverse for $\mathrm{Frac}\, R$ to be a field. Thus $\mathrm{Frac}\, R$ is indeed a field.

Finally, we have the natural embedding $R \hookrightarrow \mathrm{Frac}\, R$ sending $r \in R$ to $\frac{r}{1} \in \mathrm{Frac}\, R$, as in the case of the integers embedding into the rationals. Verify that this is indeed a homomorphism. □

[59] Really what we mean here is an injective homomorphism, but generally speaking when we know we're in the setting of rings and we have a map between rings we assume that we want it to be a homomorphism unless explicitly stated otherwise; homomorphisms are just the "right kind" of maps between rings. Similarly if the setting is groups we assume that we want group homomorphisms, and so on.

In light of this embedding we'll typically write an element r of R just as r, whether being thought of as an element of R or of the larger ring $\operatorname{Frac} R$.

Proof of Proposition 2.3.13. First, suppose that R has an injective homomorphism $f : R \hookrightarrow F$ to some field F. Let x, y be elements of R such that $xy = 0$. Since f is a homomorphism, it follows that $f(x)f(y) = f(xy) = f(0) = 0$. Since F is a field, if both $f(x)$ and $f(y)$ are nonzero then they both have inverses in F, so that we have $1 = f(x)^{-1}f(y)^{-1}f(x)f(y) = f(x)^{-1}f(y)^{-1} \cdot 0 = 0$, which is impossible unless R is the zero ring. (If R is the zero ring, then it is certainly an integral domain and embeds into any field by $0 \mapsto 0$, so the proposition is also true in this case.) Therefore, if R is not the zero ring, at least one of $f(x)$ and $f(y)$ must be equal to 0. But since f is injective the only element which maps to 0 under f is $0 \in R$, so at least one of x and y must be 0. Therefore R must be a domain.

On the other hand, suppose that R is a domain. Then by Proposition 2.3.15 it embeds naturally into $\operatorname{Frac} R$, which is a field. □

Incidentally, this gives an easy proof that every field is an integral domain: the identity function $F \to F$, taking $x \in F$ to itself, is an isomorphism and therefore also an injection, so the proposition shows that F must therefore be an integral domain.

The fraction field of an integral domain R is the field generated by R, in the sense that if we take R and add inverses of every nonzero element (in the sense discussed previously of adjoining elements) then we get $\operatorname{Frac} R$. For example, we could think of $\mathbb{Q}$ as

$$\mathbb{Z}\left[\ldots, -\frac{1}{4}, -\frac{1}{3}, -\frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right].$$

Of course, we could do this more efficiently: for example, there's no need to adjoin both $\frac{1}{2}$ and $-\frac{1}{2}$, since if we have one we could multiply it by -1 to get the other. Similarly, there's no need to adjoin $\frac{1}{4}$, since $\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$, so adjoining $\frac{1}{2}$ is enough; and adjoining $\frac{1}{2}$ and $\frac{1}{3}$ gives us for example $\frac{1}{6} = \frac{1}{2} \cdot \frac{1}{3}$ for free. Thus a simpler way of writing this would be to take $\mathbb{Z}$ and adjoin the inverses of every prime:

$$\mathbb{Q} \simeq \mathbb{Z}\left[\frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \ldots\right].$$

We can also do things like this in greater generality. In particular, we can define a more general form of the field of fractions which is not necessarily a field but can be applied to all rings.

Definition 2.3.16. Let R be a commutative ring, and $S \subseteq R$ be a multiplicative subset, i.e. a subset of R not containing 0 such that for any $a, b \in S$ their product ab is also in S.[60] Then the *localization* of R at S, written $S^{-1}R$, is the set of equivalence classes of pairs $(a, s) \in R \times S$ under the relation that $(a, s) \sim (b, t)$ if and only if there exists $u \in S$ such that $u(at - bs) = 0$.

Example 2.3.17. Suppose that R is an integral domain. Then if $u(at - bs) = 0$ then either $u = 0$ or $at - bs = 0$. Since $u \in S$ and S does not contain 0, s cannot be 0, so we must

[60] Equivalently, S is a submonoid of $(R, \times)$.

have $at = bs$: that is, if R is an integral domain then the equivalence relation on $S^{-1}R$ is the same as that for the fraction field, except that here the denominator in $\frac{a}{s}$ is required to be an element of S. Thus $S^{-1}R$ is in this case the subset of $\operatorname{Frac} R$ consisting of fractions whose denominator is an element of S.

For example, let $R = \mathbb{Z}$ and S be the submonoid of $(\mathbb{Z}, \times)$ generated by 2, i.e. $S = \{1, 2, 4, 8, \ldots\}$. Then $S^{-1}\mathbb{Z}$ is the set of fractions $\frac{a}{2^n}$ for some n, where a is any integer. Equivalently, $S^{-1}\mathbb{Z} \simeq \mathbb{Z}[\frac{1}{2}]$.

Alternatively, let R be any integral domain, and S be every nonzero element of R (in fact, the condition that R is an integral domain is precisely the condition that this choice for S is in fact a multiplicative set). Then $S^{-1}R \simeq \operatorname{Frac}(R)$.

Example 2.3.18. Let $R = \mathbb{R}[x]$ by the set of polynomials with real coefficients in one variable; think of these polynomials as functions $\mathbb{R} \to \mathbb{R}$. Let t be any real number, and let S be the set of polynomials $f : \mathbb{R} \to \mathbb{R}$ in R such that $f(t) \neq 0$. (This is a multiplicative set because $\mathbb{R}$ is an integral domain (and in fact a field).) Then $S^{-1}R$ is the set of *rational* functions[61] which are well-defined at t. For example, if $t = 0$, then $\frac{x}{x-1}$ is in $S^{-1}R$, since its value at $t = 0$ is $\frac{0}{0-1} = 0$, but $\frac{1}{x}$ is not, since it has a pole at 0.

This latter example shows the geometric meaning of localization: if we have a ring of functions which have to be well-defined everywhere, then localizing at a particular point means requiring that those functions be well-defined at that point, so that if we zoom in on it enough it looks like they're well-defined everywhere.

We've defined ring homomorphisms, and just as in the group case we can also look that the kernel and image of such a homomorphism.

Definition 2.3.19. Let $f : R \to S$ be a ring homomorphism. Then the *kernel* of f, written $\ker(f)$, is the subset of R consisting of elements $x \in R$ such that $f(x) = 0$.[62]

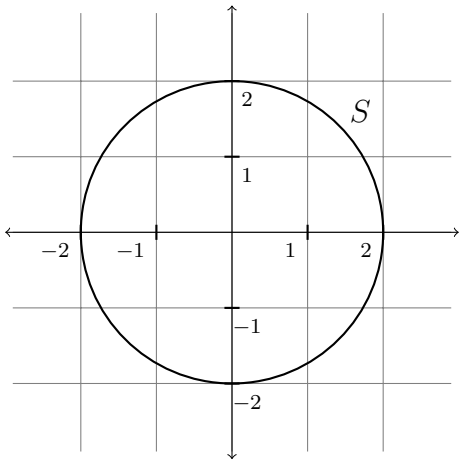
Again, it will turn out that these are subrings of the domain and codomain, and that f is injective if and only if $\ker(f) = 0$; but just as in the group case, they (especially the kernel) will have other nice properties. For groups, the extra property that a kernel has is being a normal subgroup; but since for the most part we only care about commutative rings, the idea of normality is no longer very relevant. Instead, the special property here is motivated by a geometric example.

Example 2.3.20. Let R be the ring of real-valued functions on the plane $\mathbb{R}^2$, i.e. the set of functions $\mathbb{R}^2 \to \mathbb{R}$, which is a ring by Example 2.3.6. Let S be some subset of $\mathbb{R}^2$; for example, maybe S is the unit circle.

Let $I \subseteq R$ be the set of functions $f : \mathbb{R}^2 \to \mathbb{R}$ such that for any $(a, b) \in S$ we have $f(a, b) = 0$. Then I satisfies a number of nice properties. First, it is an abelian group under addition: if f and g are both in I, i.e. $f(a, b) = g(a, b) = 0$ for any $(a, b) \in S$, then $(f + g)(a, b) = f(a, b) + g(a, b) = 0 + 0 = 0$ for any $a, b \in S$, so I is closed under addition; the zero function $f(a, b) = 0$ for *all* (a, b), not just in S, is certainly also zero on S, so it is in I as well; and if $f \in I$ then $-f \in I$, since $-0 = 0$.

[61] Recall that a rational function is one which can be written as the ratio of two polynomials $\frac{f(x)}{g(x)}$.

[62] You might wonder why the kernel is taken to be the preimage of 0 rather than of 1. The answer is that 0, by itself, forms a ring, the zero ring $\{0\}$, but 1 by itself does not, since $1 + 1 \neq 1$ unless our original ring R is itself the zero ring.



Next, I is closed under multiplication, by the same argument as above.[63] But in fact, more is true. If $f \in I$ and g is any function in R, then $f \cdot g$ is in I: $(fg)(a,b) = f(a,b)g(a,b) = 0 \cdot g(a,b) = 0$.

This motivates the following definition.

Definition 2.3.21. Let R be a commutative ring. A subset $I \subseteq R$ is an *ideal* of R if it satisfies the following two conditions:

(1) I is an additive subgroup of R;

(2) For any $x \in I$ and $y \in R$, we have $xy \in I$.

Let's first check that ideals satisfy the claimed property that motivated us to look for them.

Proposition 2.3.22. *Let $f : R \to S$ be a homomorphism of commutative rings. Then $\ker(f)$ is an ideal.*

Proof. Since by definition $f : (R,+) \to (S,+)$ is a homomorphism of abelian groups, by the theory of group kernels we have that $\ker(f)$ is a subgroup of $(R,+)$, i.e. $\ker(f)$ is an additive subgroup of R. Let $x \in \ker(f)$ and $y \in R$. Then $f(xy) = f(x)f(y) = 0 \cdot y = 0$, so $xy \in \ker(f)$. Therefore $\ker(f)$ is an ideal of R. □

There is a great deal more that can be said about ideals. For example, given some collection $C = \{x_1, x_2, \ldots\}$ of elements of a ring R, we can consider the ideal *generated* by C, written $(x_1, x_2, \ldots)$, which is formed by taking the abelian subgroup of $(R,+)$ generated by $x_1, x_2, \ldots$, and then for every x in this abelian subgroup and every $y \in R$ adding in xy until we have an ideal. If an ideal is generated by finitely many elements, it is called *finitely generated*; if it is generated by exactly one (nonzero) element, it is called a *principal ideal*.

Example 2.3.23. Consider $R = \mathbb{Z}$. What are all the ideals of $\mathbb{Z}$?

[63] In fact, this is enough to show that I is a subring of R, except that it is not unital: the identity 1, i.e. $f(a,b) = 1$ for all (a,b), is not necessarily in I, and in fact cannot be in I unless S is the empty set.

Well, we always have the simplest ideal (0), whose only element is 0. We also always have one more ideal: R itself, here $\mathbb{Z}$. (Convince yourself that every ring is always an ideal of itself.) This can be written (1), since for every $y \in R$ we have $1 \cdot y = y$ in our ideal. This also shows that if an ideal contains 1, then it is the entire ring.

What else is there? Well, we know what the ideals generated by 0 and 1 look like. What next? What about -1? Well, since -1 is invertible, for every $y \in R$ we have $y = (-1) \cdot (-y)$ and so $y \in (-1)$, and so (-1) is *also* equal to all of R. This is indicative of a general phenomenon: ideals do not have unique generators, and for every unit $u \in R^\times$ we have $(u) = (1) = R$.

Okay, let's try (2). The abelian group generated by 2 is $\{\ldots, -4, -2, 0, 2, 4, \ldots\}$, i.e. the set of even numbers; and indeed this set is closed under multiplication by any other element of $\mathbb{Z}$, i.e. for any even integer $2n$ and any integer m the product $2mn$ is also even. Great: the set of even numbers (2) is an ideal.

Alright, what about (-2)? Well, the abelian group generated by -2 is the same one as that generated by 2, and again forms an ideal, so $(-2) = (2)$. Again, this is a general phenomenon: if $u \in R^\times$ is a unit and $x \in R$ is any element, then $(ux) = (x)$.

What's next? We can form the ideal (3) in the same way as 2: it is the set of integers, positive or negative (or zero), which are divisible by 3. (Check that this is an ideal.) We can repeat this for any positive integer n, so so far we have the ideals (0), (1), (2), (3), ...

Okay, we've covered all of the principal ideals of $\mathbb{Z}$. What about the others?

We know that 0 is already in any ideal, and if 1 is in an ideal than that ideal must just be (1), so let's take the two next generators: what is the ideal $(2, 3)$? Well, it's the set of integers we can get by adding together multiples of 2 and 3. Now, $\gcd(2, 3) = 1$; and from Proposition 1.9.3 we know that there exist some integers x and y such that $2x + 3y = \gcd(2, 3) = 1$.[64] Therefore $1 \in (2, 3)$, and so $(2, 3) = (1)$. In fact, this shows that $(a, b) = (1)$ for any relatively prime integers a, b.

Hmm, okay; what's next? What about $(2, 4)$? Well, that's just inefficient: 4 is already in (2), so adding 4 as a generator doesn't change anything: $(2, 4) = (2)$. Similarly, if a divides b then in general we always have $(a, b) = (a)$.

Well, we've ruled out two big classes of pairs; what's left? Consider for example $(4, 6)$. Neither is generated by the other, nor are 4 and 6 relatively prime: instead, $\gcd(4, 6) = 2$. Therefore by Proposition 1.9.3[65] there exist x and y such that $4x + 6y = 2$, so $2 \in (4, 6)$, and therefore so is everything generated by 2, i.e. $(2) \subseteq (4, 6)$. On the other hand, since 2 divides both 4 and 6, anything that can be written as a linear combination of 4 and 6 is going to be even: $4x + 6y = 2(2x + 3y)$ will always be even for integers x and y. Therefore $(4, 6) \subseteq (2)$, and together with the above this shows that $(4, 6) = (2)$.

The same argument holds in the general case: for any two integers a, b, we will always have $(a, b) = (\gcd(a, b))$.

Okay, so every ideal of $\mathbb{Z}$ which can be generated by two elements can also be generated by one, i.e. every such ideal is principal. Can we get any more ideals with more generators?

No! An ideal with three generators, say, (a, b, c), is the same thing as adjoining c to the ideal (a, b), and then completing the ideal. But we know that $(a, b) = (\gcd(a, b))$, so

[64] Of course, this is clear in this case: take $x = -1$ and $y = 1$.

[65] Or taking $x = -1$, $y = 1$ again.

$(a, b, c) = (\gcd(a, b), c)$, which then is an ideal generated by two elements and so is also principal. Repeating this argument shows that any ideal of $\mathbb{Z}$ generated by finitely many elements must be principal.

What about infinitely many generators? By the argument above, we can replace any two pairs of generators g_1 and g_2 with their greatest common divisor $\gcd(g_1, g_2)$ without changing the ideal. If this is equal to 1, then the ideal is just (1); assume that this does not happen. Then for a third generator g_3 we can replace $\gcd(g_1, g_2)$ and g_3 with $\gcd(\gcd(g_1, g_2), g_3) = \gcd(g_1, g_2, g_3)$, where gcd is extended to be the greatest common divisor of arbitrarily many integers. Assume again that this is not 1, and iterate this process. Either eventually we get to 1 or there is some integer m such that every generator of our ideal is divisible by m. In the former scenario, the ideal is just (1); in the latter scenario, since every generator is in (m) the ideal is a subset of (m). But on the other hand since $\gcd(g_1, g_2), \gcd(g_1, g_2, g_3), \gcd(g_1, g_2, g_3, g_4), \ldots$ can never increase but only decrease or remain constant, as we iterate infinitely many times it must eventually drop to m, the greatest common divisor of all of the generators, i.e. m is in our ideal and so (m) is contained in our ideal. Therefore the ideal is precisely (m), and so we conclude that *every* ideal of $\mathbb{Z}$ is principal.

So in fact we have found another special property of the integers, besides being an integral domain.

Definition 2.3.24. A commutative ring R is called a *principal ring* if every ideal of R is principal. It is called a *principal ideal domain*, or PID, if it is a principal ring and an integral domain.

We could also relax the condition to finitely generated rather than principal:

Definition 2.3.25. A commutative ring R is called *Noetherian* if every ideal of R is finitely generated.

There are a number of equivalent conditions, which we may get into later. Almost every ring we will encounter will be Noetherian; non-Noetherian rings are generally in some sense very "big," almost impractically so. For example, the simplest example of a non-Noetherian ring is a polynomial ring in infinitely many variables $k[x_1, x_2, x_3, \ldots]$, where k is a field for simplicity. Then the ideal $(x_1, x_2, x_3, \ldots)$ cannot be finitely generated. But this ring is already so big as to be impractical, and most of the time we can safely assume all our rings to be Noetherian.

The example above showed that $\mathbb{Z}$ is a principal ideal domain, since we already knew it was an integral domain.

Example 2.3.26. Another, even simpler case is that of fields: let k be a field. We know that it has two ideals, (0) and (1). What else? Well, we know that $(u) = (1)$ for any unit $u \in k^\times$, which for a field is just any nonzero element; so the only two principal ideals are (0) and (1). Are there any other ideals?

Let I be an ideal of k which is not equal to the zero ideal (0). Then I must contain some nonzero element u. But the I contains everything generated by u, i.e. $I \subseteq (u) = (1) = k$; and since by definition $I \subseteq k$ we conclude that $I = k = (1)$. Therefore fields only have these two ideals, so they are the simplest type of principal ideal domain.[66]

[66] This is another piece of evidence as to why fields are the simplest type of rings.

Example 2.3.27. Consider the ring $R = \mathbb{Z}[x]$ of polynomials in x with integer coefficients. This is *not* a principal ideal domain: consider the ideal $(2, x)$. If R were a principal ideal domain, then there would exist some polynomial $f(x) \in \mathbb{Z}[x]$ such that $(2, x) = (f(x))$. But then this would imply that both 2 and x were divisible by $f(x)$. Since 2 is divisible by $f(x)$, we must have $f(x) = 1$ or $f(x) = 2$. But 2 does not divide x in $\mathbb{Z}[x]$, since $\frac{x}{2}$ does not have integer coefficients, so we must have $f(x) = 1$, i.e. $(2, x) = (1) = R$. But there exist elements of R which are not in $(2, x)$: any element of $(2, x)$ can be written as $2g(x) + xh(x)$ for some polynomials $g(x)$ and $h(x)$ in $\mathbb{Z}[x]$, but this can never be equal to e.g. 3, since if

$$2g(x) + xh(x) = 3$$

then since the right-hand side has no x-term neither does the left, so $h(x) = 0$ and $g(x)$ is a constant integer c, so this reads simply $2c = 3$, which has no solutions in the integers. Therefore $(2, x) \neq (1)$, and so it is not a principal ideal.

It is also possible to define arithmetic of ideals: given two ideals I and J, their sum $I + J$ is the set of elements $i + j$ for $i \in I$ and $j \in J$, and similarly their product is the set of ij for $i \in I$ and $j \in J$. In fact, since both I and J contain 0, $I + J$ contains both I and J, and should be thought of as the ideal generated by their union. On the other hand the intersection of ideals is already an ideal (exercise!), and while related to the product is distinct.

We can also do this with single elements: for example, if I is an ideal and x an element of the ring, we can consider the set $x + I$, which is the set of elements $x + i$ for $i \in I$. If $x \in I$, then this is just I. We can define xI similarly; and in fact a common notation for ideals, in addition to (x), is xR, to denote that (x) is the set of elements xr for $r \in R$. Similarly we would then write $(x, y) = (x) + (y) = xR + yR$.

There are two further important types of ideals which will be useful going forward. An ideal of a ring R is called *proper* if it is not equal to all of R.

Definition 2.3.28. A *maximal ideal* of a commutative ring R is a proper ideal $\mathfrak{m} \subseteq R$ such that any proper ideal of R which contains $\mathfrak{m}$ is in fact equal to $\mathfrak{m}$.

Definition 2.3.29. A *prime ideal* of a commutative ring R is a proper ideal $\mathfrak{p} \subseteq R$ such that if $xy \in \mathfrak{p}$ for x and y in R then at least one of x and y is in $\mathfrak{p}$.

The definition of a maximal ideal is relatively intuitive; it's simply an ideal that is "maximal" in the sense that there are no (proper) ideals that are strictly larger than it, in the sense of strictly containing it. The definition of a prime ideal might be more confusing: it arises from the property of prime numbers that if p is prime and p divides the product xy, then it must divide at least one of x and y. This was Lemma 1.2.6.

Example 2.3.30. Let $R = \mathbb{Z}$. We know that the ideals are (n) for $n = 0, 1, 2, 3, \ldots$. Which of these are prime and/or maximal?

Well, as the name implies, every ideal (p) for p prime is a prime ideal: if $xy \in (p)$, then since (p) is the set of integers divisible by p it follows that xy is divisible by p, and thus by Lemma 1.2.6 we must have at least one of x and y divisible by p, i.e. at least one of x and y is in (p).

What about the others? Suppose that $n = ab$ for some integers $a, b > 1$, i.e. n is not prime. Then $ab = n \in (n)$, but since a and b both divide n and are both greater than 1 both

must also be smaller than n and so are not divisible by n, so neither a nor b is in (n). Thus (n) is not prime.

This leaves only the ideals (0) and (1). We know that (1) cannot be a prime ideal, since it is not maximal; what about (0)? Well, actually *yes*: the condition that (0) is a prime ideal is precisely the condition that $\mathbb{Z}$ is an integral domain!

This might be a little surprising; we would not usually think of 0 as prime. Nevertheless this turns out to be a useful definition.

What about maximal ideals? Well, again let's first look at the prime ideals (p) for p prime. Suppose that I is a proper ideal containing (p). Since $\mathbb{Z}$ is a principal ideal domain, we must have $I = (n)$ for some n; since I contains (p), every integer divisible by p is also divisible by n. In particular, p itself is divisible by n, and since I is proper n cannot be 1; therefore n must be p, so $I = (p)$ and so (p) is maximal.

On the other hand, if n is divisible by some prime p, then every integer divisible by n is also divisible by p, i.e. $(n) \subset (p)$. Therefore (n) is not maximal.

Again this leaves (0) and (1). As above, (1) cannot be maximal since it is not proper; but this time (0) is certainly not maximal, since it is contained in *every* ideal. Thus the maximal ideals are (p) for p prime.

This example suggests that every maximal ideal is prime, though not necessarily vice versa. This is true, and not too difficult to prove directly; but there is a much more elegant way to prove it, using the theory of ring quotients.

In particular: we introduced ideals by analogy with normal subgroups. But in fact our main use of normal subgroups was to define quotients: and indeed ideals are precisely the objects which allow us to define ring quotients.

Definition 2.3.31. Let R be a commutative ring, and I an ideal of R. Define an equivalence relation on R by $x \sim_I y$ if $x - y \in I$, or equivalently if $x + I = y + I$. Then the *quotient ring* R/I is the set of equivalence classes under this relation, with addition and multiplication induced by that on R.

As with quotient groups, it remains to show both that these operations are well-defined and that the quotient R/I is a ring. This is not too hard: if $i, j \in I$ and $x, y \in R$, then we want to show that $(x + i) + (y + j)$ has a well-defined value modulo I independent of the choices of i and j, i.e. that we have $(x + i) + (y + j) = (x + y) + k$ for some $k \in I$. And indeed $(x + i) + (y + j) = (x + y) + (i + j)$, and since I is an ideal $i + j$ is also in I for any $i, j \in I$. Similarly, $(x + i)(y + j) = xy + xj + iy + ij$, and since I is an ideal each of xj, iy, and ij are also in I, and so so is their sum. The other ring axioms can be checked similarly.

As in the group case, there is a natural surjection $R \to R/I$, sending $r \in R$ to the equivalence class $r + I$.

Example 2.3.32. Consider the ring of integers $\mathbb{Z}$, and let (n) be an ideal. If $n = 0$, then $\mathbb{Z}/(0)$ is just $\mathbb{Z}$, since (0) contains only one element and so $a + (0) = \{a\}$ for every integer a, i.e. the equivalence classes modulo (0) are just the integers.[67] On the other end of the spectrum, if $n = 1$ then $\mathbb{Z}/(1)$ is the zero ring: (1) is all of $\mathbb{Z}$, so $a + (1) = a + \mathbb{Z}$ is all of $\mathbb{Z}$, so there is only one equivalence class, which is thus an additive identity and so is labeled 0.

[67] This will be true for all rings: $R/(0) \simeq R$ for any ring R.

For any integer $n \geq 2$, an equivalence class in $\mathbb{Z}/(n)$ is $a+n\mathbb{Z}$ for some a, i.e. two integers are equivalent if their difference is divisible by n. In other words, this equivalence is just equivalence modulo n, and so $\mathbb{Z}/(n)$ can be thought of as the integers $0, 1, \ldots, n-1$ with modular addition and multiplication. Now, recalling that we can also write (n) as $n\mathbb{Z}$ with the notation as above, we realize that we've already seen this: this is just $\mathbb{Z}/n\mathbb{Z}$, which we met in Section 1.9.

Example 2.3.33. Consider the polynomial ring $\mathbb{Z}[x]$, and let I be the ideal generated by the polynomial $f(x) = x^2 + 1$. How can we think of the quotient $R/I = \mathbb{Z}[x]/(x^2+1)$?

Well, this is the set of equivalence classes of polynomials in $\mathbb{Z}[x]$ modulo (x^2+1), i.e. where any polynomial in (x^2+1) is treated as if it's 0: for example, in this ring we have $x^3 = x^3 - x \cdot 0 = x^3 - x \cdot (x^2+1) = x^3 - (x^3+x) = -x$. In fact, for any power of x we can reduce it in this way: $x^n = x^n - x^{n-2} \cdot 0 = x^n - x^{n-2}(x^2+1) = x^n - (x^n - x^{n-2}) = -x^{n-2}$, so if we take minimal representatives then we can think of $\mathbb{Z}[x]/(x^2+1)$ as having no powers of x greater than 1, since we can reduce them.

But in fact, we can do better: using the above trick with $n = 2$ gives us $x^2 = -x^0 = -1$, i.e. $\mathbb{Z}[x]/(x^2+1)$ is the ring that we get by taking $\mathbb{Z}$ and adjoining to it some element x that satisfies $x^2 = -1$, or in the presentation given $x^2 + 1 = 0$. Well, we have a different name we usually like to give this number: the imaginary unit i! So we can think of $\mathbb{Z}[x]/(x^2+1)$ as $\mathbb{Z}[i]$, which is the set of complex numbers of the form $a + bi$ where a and b are integers. This is also called the ring of Gaussian integers.

More generally, $R[x]/(f(x))$ for any ring R and polynomial $f(x) \in R[x]$ can be thought of as adjoining to R the roots of the polynomial $f(x)$. This can also occur in higher dimensions.

Example 2.3.34. Consider the ring $R = \mathbb{R}[x,y]/(x^2+y^2-1)$. What does this look like?

Well, as in Example 2.3.33, we should think of this as the set of polynomials $f(x,y)$, now in two variables, modulo the relation that $x^2 + y^2 - 1 = 0$. If we move the -1 to the other side, we get the equation $x^2 + y^2 = 1$, which might look familiar: this is the equation for a circle of radius 1. Thus this should be thought of as the ring of polynomial functions on the circle: normal polynomials, but with the restriction of the domain (x,y) to points satisfying $x^2 + y^2 = 1$.

With the idea of quotients in hand, let's connect some of the concepts we've seen so far.

Proposition 2.3.35. *Let R be a commutative ring, and I an ideal. Then R/I is a field if and only if I is maximal, and R/I is an integral domain if and only if I is prime.*

Proof. First, suppose that R/I is a field, and suppose that J is an ideal of R strictly containing I. Then there is some element x of J which is not in I, so $x + I \neq 0 + I$, since if it were then we could conclude that $x \in I$. Since $x + I$ is a nonzero element of R/I, which is a field, there exists some element $x^{-1} + I \in R/I$ such that $(x+I)(x^{-1}+I) = xx^{-1} + I = 1 + I$. Now since $x \in J$ and J is an ideal, the product xx^{-1} is also in J, and since $xx^{-1} + I = 1 + I$ we have $1 - xx^{-1} + I = 0 + I$ and therefore $1 - xx^{-1} \in I$; and since $I \subset J$, it follows that $1 - xx^{-1}$ is also in J. But then since xx^{-1} and $1 - xx^{-1}$ are both in J, so is their sum $xx^{-1} + 1 - xx^{-1} = 1$, and since $1 \in J$ we conclude that $J = (1) = R$. Therefore any ideal strictly containing I must be all of R, and so I is a maximal ideal.

Conversely, suppose that I is a maximal ideal. Let x be an element of R not in I, and consider the ideal $xR + I$. By the definition of sums of ideals, this contains I, and since I is maximal and $xR + I$ cannot be equal to I since it contains x it must be equal to all of R, i.e. $xR + I = R$. Therefore we can find some $r \in R$ and $i \in I$ such that $xr + i = 1$, and so in particular $xr + I = 1 + I$. Therefore $r + I$ is an inverse for $x + I$ in R/I, and since x was an arbitrary element of R such that $x + I \neq 0 + I$ it follows that every nonzero element of R/I has a multiplicative inverse and so R/I is a field.

So much for maximal ideals. Now, let xy be some element of I. Then $xy + I = 0 + I$ in R/I. Therefore the claim that at least one of $x + I$ and $y + I$ must be equal to $0 + I$ is exactly the claim that at least one of x and y must be in I, so the statement that R/I is an integral domain is equivalent to the statement that I is a prime ideal. □

Corollary 2.3.36. *Every maximal ideal is prime.*

Proof. Suppose that I is a maximal ideal of a ring R. Then by Proposition 2.3.35, the quotient R/I is a field, and we know (e.g. from Proposition 2.3.13) that every field is an integral domain, so R/I is also an integral domain; and then by Proposition 2.3.35 again it follows that I is prime. □

Specializing to the case $R = \mathbb{Z}$, this also gives an immediate proof of the observation we made way back in Section 1.9, which we can now phrase as follows: for $n \geq 2$, the quotient $\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is prime.

There is, unsurprisingly, a great deal more to be said about rings, and we'll come back to discuss them more later. For now, we know enough ring theory to do algebraic number theory with; but first we need to learn some linear algebra and representation theory.

2.4 Linear algebra

Linear algebra is one of the most ubiquitous parts of mathematics, with applications from scientific computing to number theory to physics. It can be thought of very concretely as the study of vectors and matrices; we will take a more abstract point of view, according to which it is fundamentally the study of our next algebraic object: vector spaces.

Definition 2.4.1. Let k be a field. Then a *vector space* over k is an abelian group V equipped with a notion of scalar multiplication by k, that is a group action of $k^\times$ on V compatible with the additive structure of both k and V: for any $a, b \in k$ and $u, v \in V$, we must have $(a + b) \cdot v = a \cdot v + b \cdot v$ and $a \cdot (u + v) = a \cdot u + a \cdot v$.[68]

Elements of V are called *vectors*, and elements of the field k are called *scalars*.

This is best explained by example.

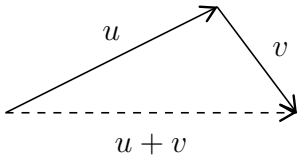
Example 2.4.2. Consider the plane $\mathbb{R}^2$, which consists of pairs (x, y) of real numbers. This is an additive group, under the addition $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$; and it is equipped with scalar multiplication by $\mathbb{R}$, with $a \cdot (x, y) = (ax, ay)$. You can check that it satisfies the desired axioms. Therefore $\mathbb{R}^2$ is a vector space over $\mathbb{R}$.

[68] Here by $k^\times$ we really mean the multiplicative monoid of k rather than the multiplicative group: the action extends to $0 \in k$, by $0 \cdot v = 0 \in V$, the zero element of V.

Example 2.4.3. Consider the set V of arrows, which start at some point a (say in $\mathbb{R}^3$, which can be thought of as a model for the physical universe) and go to some point b. These are defined by their magnitude and direction, and are a physicist's idea of vectors. (We'll consider two arrows with the same magnitude and direction but different starting points to be the same arrow: that is, we can move arrows around so long as we don't scale or rotate them and have them stay the same.) To translate to our terms, scalar multiplication by an element of $\mathbb{R}$ corresponds to scaling the magnitude of the vector appropriately: e.g. multiplying a vector by 2 keeps it pointing in the same direction but makes it twice as long, or multiplying it by -1 gives it the same magnitude but pointing in the opposite direction (which should really be thought of as the same direction but with negative length).

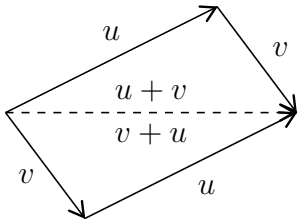
Addition is a little more complicated, but not too bad: suppose we have two arrows u and v.

To add these together, we put the start of v at the end of u and put the arrows together:



where $u + v$ is the sum of these two arrows.

We might think that since this process is not symmetric in u and v, since we put the start of v at the end of u, that this addition is not commutative, which is a problem since the addition on a vector space has to make it into an abelian group. But in fact we could do the same thing in the opposite order, and the following diagram should make it clear that the result is the same:



You can check that these notions of scalar multiplication and addition make V into a vector space.

Notice that in fact, Example 2.4.3 is secretly very similar to Example 2.4.2: rather than thinking of elements of V as arrows, we can move them such that their starting point is the origin $(0, 0, 0) \in \mathbb{R}^3$, and then an arrow from the origin is the same thing as a point $(x, y, z) \in \mathbb{R}^3$, namely the endpoint of the arrow. Explicitly, the function $\mathbb{R}^3 \to V$ taking a point (x, y, z) to the arrow from the origin to that point is bijective; it also respects the

vector space structure,[69] and so in fact this will turn out to be an isomorphism of vector spaces.[70]

In fact, this is part of a quite general observation. For any positive integer n and field k, the set k^n of tuples $(x_1, x_2, \ldots, x_n)$ with each $x_i \in k$ has the natural structure of a vector space: the abelian group structure comes from the direct product of n factors of the abelian group $(k, +)$, i.e. $(x_1, \ldots, x_n)+(y_1, \ldots, y_n) = (x_1+y_1, \ldots, x_n+y_n)$, and scalar multiplication by $\alpha \in k$ acts by $\alpha \cdot (x_1, \ldots, x_n) = (\alpha x_1, \ldots, \alpha x_n)$. Verifying that this makes k^n a vector space is an easy exercise.

These will be our main examples of vector spaces, and indeed it will turn out that every "sufficiently small" vector space will be isomorphic to one of these.[71] There is one edge case: if $n = 0$, we get the zero vector space, which is the trivial abelian group $\{0\}$ with trivial k-action. We will often neglect this case going forwards, but it is good to keep in mind.

Remark. In fact, just as we can take the direct product of groups, we can also take the direct product of rings: if R and S are rings, then $R \times S$ is the set of pairs (r, s) with $r \in R$ and $s \in S$, with the operations $(r_1, s_1)+(r_2, s_2) = (r_1+r_2, s_1+s_2)$ and $(r_1, s_1)\times(r_2, s_2) = (r_1 \times r_2, s_1 \times s_2)$. It is tempting, then, to ask why we think of k^n as just a vector space, when it seems to also come equipped with this multiplication: after all, k is a field, and so k^n is a perfectly good ring. The answer is that although the direct product of groups is fairly well behaved, the product of rings is not: for example, though k is a field, k^2 is not even an integral domain! In particular, observe that $(x, 0) \times (0, y) = (x \times 0, 0 \times y) = (0, 0)$ for any $x, y \in k$, even though $(x, 0)$ and $(0, y)$ are nonzero. Therefore since the vector space structure of k^n is very nice and the ring structure is ugly, we prefer to ignore the ring structure and view it only as a vector space.

However, there are many additional structures which can be put on vector spaces, and k^n naturally carries some of them. It is also possible to put nicer multiplications on k^n than the one induced by the direct product, and often we can even make it into a field! For $k = \mathbb{Q}$, this notion will be very important to us once we get to algebraic number theory.

One such special property of vector spaces of the form k^n is that we can equip them with something like a product, except that instead of sending a pair of vectors (u, v) to another vector, it sends them to a scalar, i.e. an element of k rather than of V.

Definition 2.4.5. Let $V = k^n$ be a vector space over k, and let $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ be two vectors in V. Then their *dot product* is

$$u \cdot v = \sum_i u_i v_i.$$

This notion satisfies a number of good properties. If u, u', v are vectors and c is a scalar, then we have

(1) $u \cdot v = v \cdot u$;

(2) $(u + u') \cdot v = u \cdot v + u' \cdot v$;

[69] Where we define the vector space structure on $\mathbb{R}^3$ in the same way as on $\mathbb{R}^2$ in Example 2.4.2.

[70] Though we haven't even defined "homomorphisms" of vector spaces yet, so we can't yet make this a rigorous statement.

[71] For a precise statement of this result, see Theorem 2.4.15.

(3) $(cu) \cdot v = c \cdot (u \cdot v)$.

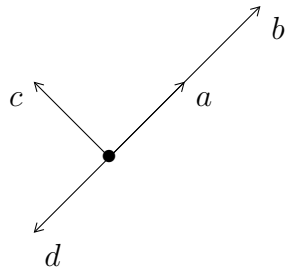
These can all be verified from the commutativity, distributivity, and associativity of ordinary multiplication on k. Note also that (1) together with (2) immediately implies the analogous statement for right-distributivity

$$u \cdot (v + v') = (v + v') \cdot u = v \cdot u + v' \cdot u = u \cdot v + u \cdot v',$$

so there is no need to verify this property independently.

The dot product can be thought of geometrically in the following way. Recall the situation of Example 2.4.2, where our vectors are points in the plane $\mathbb{R}^2$, or equivalently arrows coming from the origin $(0, 0)$. Then in a certain sense the dot product measures to what extent two vectors are pointing in the same direction.

For example, consider the vectors $a = (1, 1)$ and $b = (2, 2)$. As arrows, these are parallel, with the only difference being that b is twice the length of a; and their dot product is $a \cdot b = 1 \cdot 2 + 1 \cdot 2 = 4$. On the other hand, let $c = (-1, 1)$. Then the arrows from the origin to a and to c are perpendicular: this is reflected in that the dot product $a \cdot c = 1 \cdot (-1) + 1 \cdot 1 = 0$. Finally, if $d = (-1, -1)$, so that it is pointing in the opposite direction as a, then we get $a \cdot d = 1 \cdot (-1) + 1 \cdot (-1) = -2$.



To see this, it's helpful to define a norm: given a vector $v \in \mathbb{R}^2$, assign to it its *length* $|v| = \sqrt{v \cdot v}$. If we take a vector $v = (x, y)$, we see that this actually corresponds to the usual Euclidean distance metric: $\sqrt{v \cdot v} = \sqrt{x^2 + y^2}$, which by the Pythagorean theorem is just the distance from the origin to the point (x, y). Then, in the two-dimensional case, we have the following formula: if θ is the angle between two vectors u and v, then we have

$$u \cdot v = |u| \cdot |v| \cdot \cos\theta.$$

Thus if the two vectors are parallel, i.e. $\theta = 0°$, then since $\cos 0 = 1$ the dot product is just the product of the lengths; if pointing opposite directions, i.e. $\theta = 180°$, then since $\cos 180° = -1$ the dot product is the negative product of the lengths; and if the two vectors are orthogonal (i.e. perpendicular), then $\theta = 90°$ (or $-90°$ or equivalently $270°$) and therefore $\cos\theta = 0$, and so the dot product is 0. More generally we can think of the dot product as measuring how similar two vectors are, plus the factors coming from length.

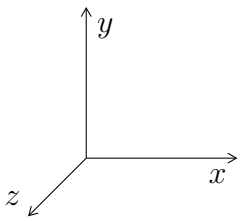
We will use this dot product structure later, to define matrix multiplication. More immediately, note that for these vector spaces k^n we can specify a vector by n elements of k, specifically $x_1, \ldots, x_n$. This is a concept worth generalizing.

Definition 2.4.6. Let V be a vector space over a field k, and $S \subseteq V$ a set of vectors in V. Then we say that S *spans* V if for every vector $v \in V$ there exists a function $c : S \to k$ such that

$$v = \sum_{s \in S} c(s) \cdot s;$$

that is, for every $s \in S$ we can choose a scalar $c(s)$ such that the formula holds. We say then that every $v \in V$ can be written as a *linear combination* of the elements of S.[72]

This might seem excessively abstract, but in fact we have a ready-made example: consider our set from Example 2.4.3 of arrows V living in $\mathbb{R}^3$, which we know corresponds to $\mathbb{R}^3$ itself. Then there is a simple set which spans V: the three unit arrows which are perpendicular to each other.



These correspond to the coordinates $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. To see that $\{x, y, z\}$ spans V, let v be the vector corresponding to some point (α, β, γ). Then define $c(x) = \alpha$, $c(y) = \beta$, and $c(z) = \gamma$: then we have

$$\begin{aligned} c(x) \cdot x + c(y) \cdot y + c(z) \cdot z &= \alpha \cdot (1, 0, 0) + \beta \cdot (0, 1, 0) + \gamma \cdot (0, 0, 1) \\ &= (\alpha, 0, 0) + (0, \beta, 0) + (0, 0, \gamma) \\ &= (\alpha, \beta, \gamma) \\ &= v. \end{aligned}$$

In fact, this is a particularly good spanning set, in that it is in a certain sense minimal. What do we mean by this?

Definition 2.4.7. Let V be a vector space over k, and $S \subseteq V$ a set of vectors. We say that the elements of S are *linearly independent if* 0 *cannot be written as a linear combination of the elements of* S *other than in the trivial manner*

$$\sum_{s \in S} 0 \cdot s = 0.$$[73]

Example 2.4.8. In our example above, the vectors $\{x, y, z\}$ are linearly independent: for any coefficients α, β, γ, we've seen that

$$\alpha x + \beta y + \gamma z = (\alpha, \beta, \gamma),$$

which is equal to the zero vector $(0, 0, 0)$ only if $\alpha = \beta = \gamma = 0$, i.e. the trivial linear combination.

[72] When the field of scalars is not clear, we may say k-linear to denote that we mean linear combinations with coefficients in k. We can use the same definition for other rings as well as fields: for example, we will often take $\mathbb{Z}$-linear combinations.

[73] Here 0 denotes the zero vector, i.e. the zero element of the abelian group V, rather than the scalar $0 \in k$.

On the other hand, let $w = (1, 1, 0)$. Then the elements $\{x, y, w\}$ are *not* linearly independent: we have $x + y = (1, 0, 0) + (0, 1, 0) = (1, 1, 0) = w$, and so

$$x + y - w = 0.$$

Thus each of x, y, and w have nontrivial coefficients, and so this is a nontrivial representation of 0.

Definition 2.4.9. Let V be a vector space, and $B \subseteq V$ a subset of V. We say that B is a *basis* for V if

(1) B spans V; and

(2) B is linearly independent.

Example 2.4.10. For V our collection of arrows in $\mathbb{R}^3$, we saw above that $\{x, y, z\}$ spans V; and Example 2.4.8 shows that x, y, z are linearly independent. Therefore $\{x, y, z\}$ is a basis for V.

On the other hand, e.g. $\{x, y\}$ is not a basis for V, since it does not span all of V: a linear combination of x and y will always have third coordinate equal to 0, so if $v = (\alpha, \beta, \gamma)$ with $\gamma \neq 0$ then γ cannot be written as a linear combination of x and y (we need z to get all of V).

Adding w does not help, since w is already in the span of x and y,[74] so that the span of x, y, and w is the same as the span of x and y. Thus $\{x, y, w\}$ is also not a basis for V.

If we add z, the situation improves: now $\{x, y, z, w\}$ spans V, since $\{x, y, z\}$ does. However, $\{x, y, z, w\}$ is not linearly independent, since $x + y + 0 \cdot z - w = 0$, so this is still not a basis for V.

This example might suggest that we expect a vector space to have a unique basis; but this is very much false. For example, consider $w = (1, 1, 0)$, $w' = (1, 0, 1)$, and $w'' = (0, 1, 1)$. This is also a basis for V. First, we can write

$$x = \frac{1}{2}(w + w' - w''),$$

$$y = \frac{1}{2}(w - w' + w''),$$

and

$$z = \frac{1}{2}(-w + w' + w''),$$

and therefore since we know that $\{x, y, z\}$ spans V so does $\{w, w', w''\}$: write an arbitrary vector v as the linear combination of x, y, and z, and then replace each by the expressions above in terms of w, w', and w''. For example,

$$(2, -1, 1) = 2x - y + z = 2 \cdot \frac{1}{2}(w + w' - w'') - \frac{1}{2}(w - w' + w'') + \frac{1}{2}(-w + w' + w'') = 2w' - w''.$$

[74] The *span* of a set of vectors is the subset of the vector space that can be written as a linear combination of those vectors.

Further, this set is also linearly independent: if

$$\alpha w + \beta w' + \gamma w'' = \alpha(1,1,0) + \beta(1,0,1) + \gamma(0,1,1) = (\alpha+\beta, \alpha+\gamma, \beta+\gamma)$$

is equal to the zero vector $(0,0,0)$, then we have

$$\alpha + \beta = 0,$$

$$\alpha + \gamma = 0,$$

and

$$\beta + \gamma = 0.$$

Let's focus on the first equation,

$$\alpha + \beta = 0.$$

Subtracting $\beta + \gamma$ from both sides, we have

$$\alpha + \beta - (\beta + \gamma) = \alpha - \gamma = -(\beta + \gamma),$$

and using the third equation we have $\beta + \gamma = 0$, so

$$\alpha - \gamma = 0$$

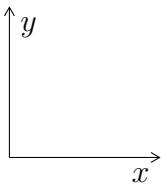
and so $\alpha = \gamma$. Now, the second equation says that

$$\alpha + \gamma = 0,$$

so since $\alpha = \gamma$ this is the statement that $2\alpha = 0$ and therefore $\alpha = 0$, so $\gamma = \alpha = 0$. Since $\alpha + \beta = 0 + \beta = 0$, it follows that $\beta = 0$, so the only solution to this system of equations is $\alpha = \beta = \gamma = 0$. Thus the only representation of 0 as a linear combination of w, w', and w'' is the trivial one. Therefore $\{w, w', w''\}$ both spans V and is linearly independent, and so is a basis for V.

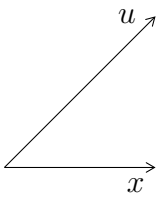
Is this basis any better or worse than our original choice of (x, y, z)? Not really, no. The original basis is the one implied by the presentation of our vector space as $\mathbb{R}^3$, but if we only care about it as a vector space then one basis is just as good as any other.

To drive this point home, let's go back to the situation of Example 2.4.2. As in this situation, there's a natural choice of basis: the unit vectors x and y in the plane



corresponding to the points $(1,0)$ and $(0,1)$. Using these, we can write down coordinates for any other vector: $v = (\alpha, \beta) = \alpha x + \beta y$. Here, these coincide with our original coordinates.

But in fact, we could use a different basis and get different coordinates. For example, consider the vector $u = (1, 1)$.



Then $\{x, u\}$ also forms a basis for this vector space: for any vector $v = (\alpha, \beta)$, we can write $v = (\alpha, \beta) = (\alpha - \beta)x + \beta u$. Thus the new coordinates of v in this basis are $(\alpha - \beta, \beta)$. *These are no better or worse than the original coordinates.* In fact, part of the general philosophy of linear algebra is that the choice of basis should not affect anything meaningful: everything "real" should be independent of the choice of basis, because the basis really only determines the direction in which we look at the vector space.

With this principle in mind, we make the following definition.

Definition 2.4.11. Let V be a vector space, and fix a basis B of V. If B is finite, then the *dimension* of V is defined to be the size of B. If B is infinite then V is said to be infinite-dimensional.

Thus for example we have shown that $\mathbb{R}^2$ has dimension 2 and $\mathbb{R}^3$ has dimension 3; more generally, the same ideas show that k^n has dimension n for any field k.

On the face of it, this definition goes against the principle outlined above: we define an important property of a vector space in terms of an arbitrarily chosen basis! Fortunately, we have the following proposition to rescue us.

Proposition 2.4.12. *Let V be a vector space with a finite basis B. Then every basis of V has the same size as B: that is, the dimension of V is independent of the choice of basis.*

Proof. First, observe that if V is not the zero vector space then no basis of V can contain 0, since if $0 \in b$ then $1 \cdot 0 = 0$ is a "nontrivial" representation of 0 in that basis. (In the zero vector space, the result is easy.)

Suppose that $B = \{b_1, b_2, \ldots, b_n\}$, and let B' be another basis of V, with elements $b'_1, b'_2, \ldots, b'_m$, where m is a positive integer which we do not yet know whether it is equal to n. First, suppose that $m < n$. Since B' spans V, there exist scalars c_i such that

$$b_1 = c_1 b'_1 + c_2 b'_2 + \cdots + c_m b'_m.$$

Since b_1 is nonzero, at least one of the c_i must be nonzero; since the order of the elements is arbitrary, we can assume that it is c_1 which is nonzero. Therefore we can write

$$b'_1 = \frac{1}{c_1}(b_1 - c_2 b'_2 - \cdots - c_m b'_m).$$

Writing $B'' = \{b_1, b'_2, b'_3, \ldots, b'_m\}$ (note that the first element is now b_1, rather than b'_1!), we conclude that b'_1 is in the span of B''. Therefore the span of B'' is equal to the span of $\{b_1, b'_1, b'_2, \ldots, b'_m\}$, which is at least as large as the span of B'; but since B' already spans V, we conclude that B'' also spans V.

Now since B'' spans V we can find coefficients c_i such that

$$b_2 = c_1 b_1 + c_2 b_2' + c_3 b_3' + \cdots + c_m b_m'.$$

Again, since b_2 is nonzero at least one of the c_i must be nonzero. But in fact, we can do better than that: if all of the c_i except c_1 are nonzero, then we would have $b_2 = c_1 b_1$ and therefore $c_1 b_1 - b_2 = 0$, which is impossible since B is linearly independent. Therefore one of the other c_i must be nonzero, and since again the order is arbitrary we can assume it is c_2. Therefore we can write

$$b_2' = \frac{1}{c_2}(b_2 - c_1 b_1 - c_3 b_3' - \cdots - c_m b_m')$$

and so as above we conclude that b_2' is in the span of $B''' = \{b_1, b_2, b_3', b_4', \ldots, b_m'\}$ and so B''' also spans V.

Iterating this process, we find that each time we must have at least one of the coefficients of the b_i' terms nonzero, since otherwise we would be expressing b_i in terms of the other b_j, which is impossible since B is linearly independent. Therefore we can continue iterating this process until we arrive at the set

$$\{b_1, b_2, \ldots, b_m\},$$

since $m < n$, which by the same argument will also span all of V. But since $m < n$, this implies that there exists some b_{m+1} not in this set: and since it is spanned by it, there exist coefficients c_i such that

$$b_{m+1} = c_1 b_1 + \cdots + c_m b_m,$$

and therefore $c_1 b_1 + \cdots + c_m b_m - b_{m+1} = 0$ gives a nontrivial representation of 0, which is impossible since B is linearly independent. Therefore the assumption that $m < n$ must be false, i.e. $m \geq n$.

But now observe that the only assumptions on B and B' which we have used in this process are first that B is linearly independent, and second that B' spans V; and from this we concluded that $m \geq n$, i.e. $|B'| \geq |B|$. But since both B and B' are bases for V, the same holds true if we reverse them, i.e. B also spans all of V, and B' is also linearly independent; and therefore we can conclude by the same argument that $|B| \geq |B'|$! Putting these two results together, we conclude that we must have $|B| = |B'|$. □

Corollary 2.4.13. *Every finite-dimensional vector space has a basis.*

Proof. Let V be a finite-dimensional vector space, and pick a random element x_1. Next, pick x_2 not a multiple of x_1; then pick an element x_3 of V which cannot be written as a linear combination of x_1 and x_2; etc. By construction all of the x_i are linearly independent and if they were to span V then they would then form a basis of V, so Proposition 2.4.12 shows that this process is always possible for $i \leq \dim V$, and then guarantees that the resulting set $\{x_1, \ldots, x_{\dim V}\}$ is a basis for V. □

The corollary is actually true for all vector spaces, not just finite-dimensional ones, but for large infinite dimensions subtleties arise.

Proposition 2.4.12 is an instantiation of a general phenomenon in linear algebra: we first define a property in terms of a basis, and then prove that in fact it is actually independent

of the chosen basis. In fact, this is a phenomenon common to all of algebra, if not more widely: the principle is that any good property should be invariant under isomorphism, but most things are easier to state for concrete objects, so we first define them concretely and then prove that they are isomorphism invariants.

In fact, the analogy between change of basis and isomorphism is not an analogy: change of basis *is* an isomorphism of vector spaces. In order to make sense of this, we need to first define what a map of vector spaces is.

Definition 2.4.14. Let V and W be vector spaces over a field k. A *linear transformation* is a function $T : V \to W$ which is *linear*, i.e. it satisfies the following two properties:

(1) For any two vectors $v_1, v_2 \in V$ we have $T(v_1 + v_2) = T(v_1) + T(v_2)$; and

(2) For any vector $v \in V$ and scalar $c \in k$ we have $T(cv) = cT(v)$.

Linear transformations play the same role for vector spaces as homomorphisms do for groups and rings. Looking back to the definition of vector spaces, linear transformations can be thought of as functions that respect the vector space structure: the first property above is the statement that T respects the abelian group structure, i.e. T is a homomorphism of abelian groups $V \to W$, and the second is the statement that T respects the k-action on V and W. Note that in order for this second property to make sense it's important that V and W be vector spaces over the same field, since c acts on $v \in V$ on the one hand and on $T(v) \in W$ on the other.

As usual, since linear transformations are functions with extra structure, we can define them to be surjective or injective if the underlying functions are; and a linear transformation is an isomorphism if it is invertible, or equivalently if it is bijective. Similarly, as linear transformations are also homomorphisms of abelian groups, we can define their images and kernels, which will be *subspaces* of the domain and codomain: a subspace of a vector space V, naturally, is a subset of V which is itself a vector space (over the same field).

There are some immediate remarks to make. Let V be a vector space of dimension n over a field k. Then choosing a basis $(b_1, \ldots, b_n)$ is equivalent to defining an isomorphism $\varphi : k^n \to V$, sending a tuple $(c_1, \ldots, c_n) \in k^n$ to

$$c_1 b_1 + c_2 b_2 + \cdots + c_n b_n$$

in V: the condition that the b_i are linearly independent corresponds to the requirement that φ be injective, and the condition that the b_i span V corresponds to the condition that φ be surjective. This leads immediately to the following theorem.

Theorem 2.4.15. *Every finite-dimensional vector space V over k is isomorphic to k^n, where n is the dimension of V.*

Proof. By Corollary 2.4.13 and Proposition 2.4.12, we can find a basis $\{x_1, \ldots, x_n\}$ for V. By the above discussion, this is equivalent to finding an isomorphism with k^n. □

The idea of subspaces will also be a useful one: for example, we can more rigorously define the span of a set of vectors to be the smallest subspace containing them.

However, due to the idea of bases there's somewhat more to say about transforms themselves.

Definition 2.4.16. A *matrix* over a field k is a rectangular array of elements of k.

Given two matrices of the same dimensions, say $m \times n$,[75], we can add them in the natural way: if our matrices A and B have coordinates A_{ij} and A_{ij} in the ith row and jth column, then $A + B$ has (i, j)th coordinate $A_{ij} + A_{ij}$. For example, we have

$$\begin{pmatrix} 3 & 4 & -1 \\ 0 & \frac{1}{2} & 2 \end{pmatrix} + \begin{pmatrix} 0 & -1 & -\frac{3}{4} \\ 1 & 4 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 3 & -\frac{7}{4} \\ 1 & \frac{9}{2} & 4 \end{pmatrix}.$$

We can also define scalar multiplication in the natural way: for $c \in k$ and M a matrix with coordinates M_{ij}, the product cM has coordinates cM_{ij}. For example,

$$2 \cdot \begin{pmatrix} 3 & 4 & -1 \\ 0 & \frac{1}{2} & 2 \end{pmatrix} = \begin{pmatrix} 6 & 8 & -2 \\ 0 & 1 & 4 \end{pmatrix}.$$

Thus the set of $m \times n$-dimensional matrices can be thought of as a vector space over k; since there are $m \cdot n$ entries, each of which can be chosen independently, it is in fact an $m \cdot n$-dimensional vector space.

However, matrices have additional structure. Let A be an $m \times n$-dimensional matrix with coefficients A_{ij}, and let B be an $n \times p$-dimensional matrix with coefficients A_{ij}: for example,

$$A = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 4 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} 6 & -2 & -1 & 0 \\ 0 & 1 & 8 & -3 \end{pmatrix},$$

where A is 3×2-dimensional and B is 2×4-dimensional (and for example $A_{1,2} = 0$, and $A_{2,3} = 8$). Then we can define their *product* AB to be the matrix whose (i, j)th coordinate is

$$(AB)_{ij} = \sum_t A_{it} B_{tj}.$$

Equivalently, $(AB)_{ij}$ is the dot product of the ith row of A with the jth column of B.

In our example, the $(1, 1)$ coordinate of AB would be the dot product of the first row of A with the first column of B, i.e.

$$(AB)_{1,1} = 1 \cdot 6 + 0 \cdot 0 = 6.$$

Doing this for all three rows of A and all four columns of B gives us a 3×4-dimensional matrix, whose values you can verify to be

$$AB = \begin{pmatrix} 6 & -2 & -1 & 0 \\ 12 & -5 & -10 & 3 \\ 24 & -5 & 20 & -9 \end{pmatrix}.$$

In general, the product of an $m \times n$ matrix and an $n \times p$ matrix will always be $m \times p$-dimensional.

Notice that the requirement that the number of columns of A and the number of rows of B be the same is an important one: without it, the multiplication is not well-defined,

[75] That is, with m rows and n columns.

because in order for the dot product of a row of A with a column of B to make sense they have to be the same length. This shows us, by the way, that matrix multiplication is very noncommutative: not only is BA not necessarily the same thing as AB, it (as in this example) does not even necessarily exist!

This notion of multiplication might seem unmotivated and unnecessarily complicated. After all, we could just as well define multiplication element-wise, as we did with addition: $(AB)_{ij} = A_{ij}B_{ij}$. So why not?

The answer, which is also the reason we care about matrices, is that matrices are supposed to be our way of writing down linear transformations!

To see this, recall that the set of $m \times n$ matrices is itself a vector space, since it's equipped with a notion of addition (and is an abelian group under it) and a compatible scalar multiplication. Thus in particular the set of $n \times 1$ matrices is a vector space of dimension n, whose elements are vectors of the form

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

These are called *column vectors*, and since there is a clear isomorphism to k^n, sending a column vector of this form to $(x_1, \ldots, x_n) \in k^n$, we will usually think of k^n as this vector space.[76]

Now, let A be an $m \times n$ matrix, and v be the column vector above with entries $x_1, \ldots, x_n$. Then we can compute the product Av; since A is $m \times n$ and v is $n \times 1$, we know that the output should be an $m \times 1$ matrix, i.e. an m-dimensional column vector, or equivalently an element of k^m. I claim that in this case A defines a linear transformation $A : k^n \to k^m$.

Well, what are the requirements for this transformation to be linear? We must have $A(v_1 + v_2) = Av_1 + Av_2$, and $A(cv) = cAv$ for any vectors v, v_1, v_2 and scalar c. Both of these follow from the more general proposition.

Proposition 2.4.17. *Let A, A' be $m \times n$ matrices, and B, B' be $n \times p$ matrices. Then*

$$A(B + B') = AB + AB'$$

and

$$(A + A')B = AB + A'B,$$

and for any scalar c we have

$$(cA)B = A(cB).$$

Proof. Let A_{ij} be the coefficients of A, and similarly for the other matrices. Then

$$A(B + B')_{ij} = \sum_t A_{it}(B + B')_{tj} = \sum_t A_{it}B_{tj} + \sum_t A_{it}B'_{tj} = (AB)_{ij} + (AB')_{ij},$$

[76] There are also *row vectors*, which are $1 \times n$ matrices, but column vectors will usually be more convenient for our notation.

and since addition is element-wise it follows that

$$A(B + B') = AB + AB'.$$

The verification for the second statement is similar.

For the scalar statement, we have

$$((cA)B)_{ij} = \sum_t (cA)_{it} B_{tj} = \sum_t cA_{it} B_{tj} = \sum_t A_{it}(cB_{tj}) = (A(cB))_{ij},$$

and so

$$(cA)B = A(cB).$$

□

Applying Proposition 2.4.17 with $p = 1$ immediately shows that matrices act linearly on vectors, i.e. matrices define linear transformations.

What if we had used a different notion of multiplication, e.g. the element-wise multiplication mentioned above? Then matrices could not act on vectors, since multiplication (like addition) would only be defined on matrices of the same dimensions. Thus the only multiplication we would have would be multiplication of vectors by other vectors, which is exactly the multiplication we decided we didn't want in Remark 2.4.4.

This still doesn't completely justify this notion of multiplication, since we don't know that some other notion wouldn't work equally well or better. This is fixed by the following theorem.

Theorem 2.4.18. *Let V and W be finite-dimensional vector spaces. Then, after choosing bases for V and W, every linear transformation $V \to W$ is given by a matrix. Further, if U is another finite-dimensional vector space, then after choosing a basis for U the composition of two linear transforms $T_1 : V \to W$ and $T_2 : W \to U$, each given by a matrix A_1 and A_2 respectively, is given by the matrix product $A_2 A_1$.*

In particular, not only is every linear transform given by a matrix (after choosing a basis), but also our above definition of the product of matrices is the "correct" one: it corresponds to the composition of linear transforms.[77] This also leads us to think of linear transforms, and operators more generally, as "multiplication" by some operator, and conversely to think of multiplication as the composition of functions. This helps explain why matrix multiplication is not commutative.

Proof of Theorem 2.4.18. First, choosing bases for V, W, and U amounts to choosing isomorphisms with $k^{\dim V}$, $k^{\dim W}$, and $k^{\dim U}$ respectively, so it suffices to show that every linear transform $k^m \to k^n$ for all m and n is given by a matrix, and that the composition law holds.

Let $T : k^m \to k^n$ be some linear transform. Write $e_1 = (1, 0, 0, \ldots, 0)$, $e_2 = (0, 1, 0, 0, \ldots, 0)$, and so on up to $e_m = (0, 0, \ldots, 0, 1)$; in particular the e_i form a basis for k^m. Then since every $x \in k^m$ can be written as a linear combination of the e_i, by

$$x = x_1 e_1 + \cdots + x_m e_m,$$

[77] Note that the product is $A_2 A_1$ rather than $A_1 A_2$ because A_1 acts first: if we put in a vector v, we first find $A_1 v$, and then compute $A_2 A_1 v$, analogous to taking $T_2(T_1(v))$. This is a result of our notation of operators acting on the left.

we have

$$T(x) = T(x_1 e_1 + \cdots + x_m e_m) = x_1 T(e_1) + \cdots + x_m T(e_m)$$

by the linearity of T.

Now, each $T(e_i)$ is itself a vector in k^n: write T_{ij} for the jth coordinate of $T(e_i)$. Then

$$T(x) = \sum_i x_i T(e_i)$$

has jth coordinate

$$T(x)_j = \sum_i T_{ij} x_i.$$

In particular, if we think of x and $T(x)$ as column vectors, i.e. $m \times 1$- and $n \times 1$-dimensional matrices respectively, then we see that

$$T(x)_{j,1} = \sum_i T_{ij} x_{i,1},$$

i.e. for some matrix T with coordinates T_{ij} we have

$$T(x) = Tx.$$

Thus the linear transform T is given by a matrix.

Now, let $S : k^n \to k^p$ be another linear transform, which by the above is also given by a matrix. Then we can similarly describe the composition $S \circ T : k^m \to k^p$ by computing each of the $S(T(e_i))$. Explicitly, writing $E_j = (0, \ldots, 0, 1, 0, \ldots, 0) \in k^n$ for the nonzero coordinate in the jth position, we have

$$T(e_i) = \sum_j T_{ij} E_j$$

and therefore

$$S(T(e_i)) = \sum_j T_{ij} S(E_j).$$

Since S is itself given by a matrix, the rth coordinate of $S(E_j)$ is just S_{jr}, and so the rth coordinate of $S(T(e_i))$ is

$$S(T(e_i))_r = \sum_j S_{jr} T_{ij}.$$

But this is exactly the formula for the (i, r)th coordinate of the matrix product ST; and so we conclude that

$$S(T(x)) = S\left(T\left(\sum_i x_i e_i\right)\right) = \sum_i x_i S(T(e_i))$$

has jth coordinate

$$S(T(x))_j = \sum_i x_i S(T(e_i))_j = \sum_i x_i (ST)_{ij}.$$

But this is precisely the formula for the jth coordinate of the matrix product

$$STx,$$

and so we conclude that the composition of the linear transforms S and T is given by the matrix product ST. □

Great! So now we know what matrices are, how to work with them, and why they are the "correct" notion. More abstractly, we know how to write down linear transformations in the form of matrices.

Actually, that's not quite true—we know how to write down a linear transformation as a matrix, once we've chosen a basis for each of the vector spaces involved. This makes a difference.

For example, recall the situation of Example 2.4.2, where we had two different bases: the "obvious" one $B = \{(0,1),(1,0)\}$, and a "tilted" one $B' = \{1,0),(1,1)\}$. Consider the linear transformation $F : \mathbb{R}^2 \to \mathbb{R}^2$ which reflects each point around the y-axis, so that for example $(2,4)$ is sent to $(-2,4)$. (Convince yourself that this is a linear transform!)

What is the matrix for F? First, let's work in the basis B. In this form, our transform sends the coordinates $(0,1)$ to $(0,1)$ and $(1,0)$ to $(-1,0)$. Since this is a linear transform $\mathbb{R}^2 \to \mathbb{R}^2$, it is represented by a 2×2 matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

which we know satisfies the properties

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Well, we know how matrix multiplication works, so we can reformulate these: the first becomes $a \cdot 0 + b \cdot 1 = b = 0$ and $c \cdot 0 + d \cdot 1 = d = 1$, and the second becomes $a \cdot 1 + b \cdot 0 = a = -1$ and $c \cdot 1 + d \cdot 0 = c = 0$. Thus our matrix must be

$$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

and we can verify that in fact this acts as desired on our basis vectors and therefore, by linearity, acts by F on all of $\mathbb{R}^2$.

Great. Now let's work with our second basis B'. Write $x = (1,0)$ and $u = (1,1)$ as before. We have $F(x) = -x$ and $F(u) = (-1,1) = u - 2x$. Therefore if our vector is

$$v = \alpha x + \beta u$$

then F acts by

$$F(v) = \alpha F(x) + \beta F(u) = -\alpha x + \beta(u - 2x) = -(\alpha + 2\beta)x + \beta u.$$

In column vector notation, this is

$$v = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

and

$$Fv = \begin{pmatrix} -\alpha - 2\beta \\ \beta \end{pmatrix},$$

so writing F as a matrix and applying with (α, β) equal to $(0, 1)$ and $(1, 0)$ as above gives

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Calculating the products as above, the first of these becomes $b = -2$, $d = 1$, and the second becomes $a = -1$, $c = 0$, so in this basis we have

$$F = \begin{pmatrix} -1 & -2 \\ 0 & 1 \end{pmatrix}.$$

This is notably different from our previous result: changing the basis we're working in genuinely does change the matrix, even when the transform is the same.

How can we go between these two bases? We actually already kind of know how to do this. Suppose that $b = \{b_i\}$ and $b' = \{b'_i\}$ are two bases for a vector space V of dimension say n. If

$$v = c_1 b_1 + \cdots + c_n b_n$$

is the representation of v in b and we want to write it in b', this isn't too hard: write each of the b_i in b' as $b_i = \beta_{1,i} b'_1 + \cdots + \beta_{n,i} b'_n$, and let β be the matrix with each of these as rows:

$$\beta = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,n} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n,1} & \beta_{n,2} & \cdots & \beta_{n,n} \end{pmatrix}.$$

Then

$$\beta v = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,n} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n,1} & \beta_{n,2} & \cdots & \beta_{n,n} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

has ith coordinate

$$(\beta v)_i = \sum_j \beta_{ij} c_j$$

and so summing against the b'_i gives

$$(\beta v)_1 b'_1 + \cdots + (\beta v)_n b'_n = \sum_i \sum_j \beta_{ij} c_j b'_i = \sum_j c_j \sum_i \beta_{ij} b'_i = \sum_j c_j b_i = v$$

by the definition of the β_{ij} and the c_i. Therefore the $(\beta v)_i$ are the coordinates of v in b', and so multiplication by β is how we get from the basis b to the basis b'.

Note that this is just the matrix of the isomorphism which we identified with change of basis above.

In our example above, if we want to change basis from B to B', by this calculation our matrix is

$$\beta = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Indeed, this takes our original vector

$$\begin{pmatrix} x \\ y \end{pmatrix}$$

to

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ x+y \end{pmatrix}$$

as desired. To see how this acts on matrices, follow through the proof of Theorem 2.4.18 to see that a matrix A in the original basis is sent to

$$\beta^{-1} A \beta$$

where β^{-1} is the matrix of the inverse transform to that of β (recall that the transform associated to β is an isomorphism, and so has an inverse).

This is all very well, but doesn't tell us very much about how to compute the matrix in a new basis unless we know how to invert β. There are plenty of algorithms for computing the inverse of a matrix which we won't go into, but in this case it's not two hard. We need a few observations first.

First, what are we looking for? The inverse should have the property that $\beta\beta^{-1} = \beta^{-1}\beta =$ id, like any good inverse; but what is the identity here? Well, in order for $\beta\beta^{-1}$ and $\beta^{-1}\beta$ to both make sense, β must be square (which of course is true in our case, but we're thinking more broadly for the moment). Thus if β is $n \times n$ then id will also be. What then is the matrix of the identity $V \to V$?

Well, if $V = k^n$, then id sends each e_i to itself; doing out the calculation as above, we find that the identity $n \times n$ matrix, written I_n, will always be given by

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Thus what we're looking for is a matrix β^{-1} such that $\beta\beta^{-1} = \beta^{-1}\beta = I_n$.

In the 2×2 case, there's a relatively simple formula. Specifically, if

$$\beta = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then

$$\beta^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

This can be verified just by multiplying out the matrices; we'll come back to the factor $\frac{1}{ad-bc}$ shortly.

Meanwhile, this enables us to compute base changes. In our example, we have

$$\beta = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

and thus

$$\beta^{-1} = \frac{1}{1\cdot 1 - 1\cdot 1}\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix},$$

and so in our new basis we compute that F is

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -1 & -2 \\ 0 & 1 \end{pmatrix}$$

as we found above.

Recall from group theory that this operation $F \mapsto \beta^{-1}F\beta$ is conjugation by β (or, depending on your convention, by β^{-1}). Thus what we conclude is that any information about a linear transform T that is basis-independent should be reflected in information about its matrix which is invariant under conjugation.

Let's restrict to square matrices, as these are the most important case and allow us to not worry about whether the product of matrices exists. What properties of matrices are invariant under conjugation?

There are quite a few, but let's just give two important ones. The first is relatively simple.

Definition 2.4.19. Let A be a square matrix. Its *trace*, denoted $\operatorname{tr}(A)$, is the sum of its diagonal entries.

It might be surprising that the trace is conjugation-invariant; observe that for example the trace of F in the example above is 0 in both bases, though the matrices are otherwise different. Verifying this is algebraically straightforward but tedious, so let's not.

The trace is also additive, which is immediate from the definition and the element-wise addition of matrices: $\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B)$. It is not, however, multiplicative: $\operatorname{tr}(AB)$ is not necessarily equal to $\operatorname{tr}(A)\operatorname{tr}(B)$.

The second invariant is more complicated, and we give a definition only for completeness.

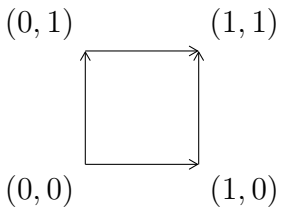
Definition 2.4.20. Let A be a square matrix. If A is 1×1, i.e. just a scalar $A_{1,1}$, then its *determinant*, denoted $\det(A)$, is just $A_{1,1}$. Otherwise, suppose that A is $n\times n$. For each i from 1 to n, define $A^{(i)}$ to be the submatrix of A consisting of A absent the top row and ith column; this is called the $(1,i)$th minor of A, and is an $(n-1)\times(n-1)$-dimensional matrix. Then define

$$\det(A) = \sum_i (-1)^{i-1}\det A^{(i)},$$

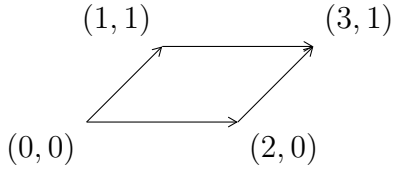
and repeat the definition on $A^{(i)}$ for each i until we get down to 1×1-dimensional matrices.

This is one of many definitions, and is not terribly important. The important property of determinants is that $\det(AB) = \det(A)\det(B)$ for any $n \times n$ matrices A and B. Since I_n satisfies $I_n A = A$, we therefore have for example $\det(A) = \det(I_n A) = \det(I_n)\det(A)$, and so since this is true for any A we conclude that $\det(I_n) = 1$.

Generally speaking, $\det(A)$ should be thought of as measuring how much A scales vectors. For example, let A be a 2×2 matrix. Thinking of $(0,1)$ and $(1,0)$ as arrows, we can form their sum and take the resulting square:



and take the area of this square, which of course is just 1. If we act on the vectors by some matrix A, they will still form a parallelogram in this way, and we can take its area again.



The area of the new parallelogram will be $\det(A)$.

In the $n = 2$ case, there is also a simple formula. If

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then

$$\det(A) = ad - bc.$$

This should look familiar: it cropped up in the formula

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

In particular, since k is a field A is invertible if and only if $\det(A)$ is nonzero. This is true more generally: any square matrix is invertible if and only if its determinant is nonzero.

For example,

$$\begin{pmatrix} 6 & 4 \\ 3 & 2 \end{pmatrix}$$

has determinant $6 \cdot 2 - 4 \cdot 3 = 0$, and so is not invertible. Meanwhile

$$\begin{pmatrix} 1 & 0 \\ 2 & 2 \end{pmatrix}$$

has determinant $1 \cdot 2 - 0 \cdot 2 = 2$, and so it is invertible—but notice that, though 2 is certainly invertible in, say, $\mathbb{Q}$, it is *not* invertible in $\mathbb{Z}$, and so A^{-1} cannot be a matrix over the integers![78] Indeed, we can compute

$$A^{-1} = \begin{pmatrix} 1 & 0 \\ -1 & \frac{1}{2} \end{pmatrix}$$

to find that it exists but is not integral.

We can think of the set of $n \times n$ matrices as a ring, with matrix addition and multiplication as the operations and identity element I_n. This is written $\mathrm{Mat}_{n \times n}$. Note that, unlike the rings we have encountered before, $\mathrm{Mat}_{n \times n}$ is not commutative.

As usual with rings, we can take its group of units $\mathrm{Mat}_{n \times n}^{\times}$, which is the group of invertible $n \times n$ matrices (implicitly over our field k). This has a special name: it is the *general linear group*, written $\mathrm{GL}_n(k)$ to signify that its entries must be elements of k. It is one of the classical examples of a nonabelian group.

Since $\mathrm{Mat}_{n \times n}$ can be thought of as the set of linear transformations $k^n \to k^n$, the subset $\mathrm{GL}_n(k)$ of *invertible* linear transformations is the set of *isomorphisms* $k^n \to k^n$. By the discussion above, for k a field these are precisely the matrices with nonzero determinant.[79]

There are various other groups we can define, but GL_n is the most important one. Others include $\mathrm{SL}_n(k)$, the *special linear group*, which is the set of $n \times n$ matrices over k whose determinant is precisely 1; the *orthogonal group*, which is the set $O_n(k)$ of $n \times n$ matrices A such that $A^{\mathsf{T}} A = A A^{\mathsf{T}} = I_n$, where A^{T} is the *transpose* of A, given by exchanging the rows and columns of A, so that for example if

$$A = \begin{pmatrix} 2 & 3 \\ 0 & -1 \end{pmatrix}$$

then

$$A^{\mathsf{T}} = \begin{pmatrix} 2 & 0 \\ 3 & -1 \end{pmatrix};$$

and the *special orthogonal group*, which is the intersection of $\mathrm{SL}_n(k)$ and $O_n(k)$.[80]

The orthogonality condition is worth saying a little more about. In particular, the product AA^{T} has (i,j)th coordinate the dot product of the ith row of A with the jth column of A^{T}; but by the definition of A^{T}, the jth column of A^{T} is the same thing as the jth row of A, so $(AA^T)_{ij}$ is the dot product of the ith row and the jth row of A. Then the requirement that $AA^T = I_n$, since I_n is 1 on the diagonal entries and 0 everywhere else, is just the requirement that the dot product of the ith row and jth row of A is 1 if $i = j$ and 0 otherwise. Recalling that the dot product measures the "similarity" of two vectors, and in particular is 0 if and only if they are orthogonal, explains the name of this group: if A is an orthogonal matrix, i.e. an element of $O_n(k)$, then all of its rows are orthogonal. To get the reverse direction, there is another required condition: if v is a row of A, then in addition to being orthogonal

[78] We've only been talking about matrices over fields, but the theory works just as well over rings.

[79] If k is not a field, then we can still define $\mathrm{GL}_n(k)$, and now it is characterized as the set of matrices whose determinant lies in $k^{\times}$, i.e. $A \in \mathrm{Mat}_{n \times n}$ is invertible if and only if $\det(A) \in k$ is.)

[80] Sometimes you'll see the notation $\mathrm{GL}(n,k)$ instead of $\mathrm{GL}_n(k)$, and similarly for the other groups; they mean the same thing.

to all the other rows we also need to have $v \cdot v = 1$, or equivalently $|v| = 1$. Sets of vectors which are mutually orthogonal and which all have length 1 are called *orthonormal*.

Notice also that the above is all in terms of matrices, i.e. linear transformations with a chosen base. However, we can also think of it more abstractly: given a vector space V over k, we write $\mathrm{GL}(V)$ (or $\mathrm{GL}(V, k)$ if we want to specify the field) for the set of isomorphisms $V \to V$ as a k-vector space, and similarly for the other groups.

Our last topic before we can move on to representation theory is eigenvalues and eigenvectors.

Definition 2.4.21. Let A be an $n \times n$ matrix over k. We say that a vector v is an *eigenvector* of A if there exists a scalar $\lambda \in k$ such that $Av = \lambda v$. The scalar λ is called the *eigenvalue* of v with respect to A.

Example 2.4.22. Consider the matrix

$$A = \begin{pmatrix} 2 & 1 \\ 0 & -1 \end{pmatrix}.$$

If v is an eigenvector for A, with eigenvalue λ, then writing

$$v = \begin{pmatrix} x \\ y \end{pmatrix}$$

we have

$$Av = \begin{pmatrix} 2 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix},$$

which multiplying out becomes $2x + y = \lambda x$ and $-y = \lambda y$. The second equation tells us that either $y = 0$ or $\lambda = -1$. If $y = 0$, then the first equation reads $2x = \lambda x$, and so either $x = 0$ or $\lambda = 2$. But if both x and y are 0, then v is just the zero vector, and so Av is going to be equal to v for *any* matrix A, i.e. v is an eigenvector for every matrix—and in fact is an eigenvector with every eigenvalue, since $\lambda v = v$. This is a "degenerate" case and so we'll disregard it.

In fact, thinking about this a little more shows that we don't really care about specific eigenvectors: after all, if v is an eigenvector for A, i.e. $Av = \lambda v$ for some λ, then so is any multiple of v: that is, $A(cv) = cAv = c\lambda v = \lambda(cv)$. Similarly, if there are two (or more!) eigenvectors with the same eigenvalue λ, then any linear combination of those eigenvectors is also an eigenvector with eigenvalue λ. Thus what we really care about is the *subspace* associated to each eigenvalue λ: that is, the subspace $S \subseteq V$ such that A acts on S by scalar multiplication by λ.

Back to our current example: we've discounted the case where x and y are both 0, since this lies in every such subspace. Under our assumption that $y = 0$, this only leaves the case where $\lambda = 2$ and x is anything: that is, the subspace corresponding to $\lambda = 2$ is the set of vectors of the form $(x, 0)$ for all $x \in k$. (Check that this is a subspace, i.e. is itself a vector space.)

Great! We've found a whole subspace of eigenvectors corresponding to the eigenvalue 2, called the *eigenspace*. Are there any more? Also, our subspace is one-dimensional, since there is one free parameter x; our overall subspace is two-dimensional. Have we found all of the vectors in the eigenspace of 2?

This second question is easier: yes, we have, because the only 2-dimensional subspace of our 2-dimensional vector space is the whole vector space, and there are certainly vectors on which A does not act by scalar multiplication because if there were not then A would be equal to twice the identity,

$$2I_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

which it is not. For example,

$$A\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

What about the first question, are there any more eigenvectors? Well, we haven't finished our calculation from above: we never moved beyond the assumption that y was zero! If y is not zero, then $\lambda = -1$, and so the first equation becomes $2x + y = -x$, or $y = -3x$. Therefore the eigenvalues of A are 2 and -1, and the eigenspace associated to -1 is the set of vectors of the form $(x, -3x)$ for all x.

In the example above, how do we know that we have found all of the eigenvalues? In general, if A is an $n \times n$ matrix (or more abstractly, $A : V \to V$ is a linear transformation on an n-dimensional vector space), then it has at most n eigenvalues. To see this, we need a little more notation: in particular, we need to be able to create new vector spaces from old ones.

Definition 2.4.23. Let V_1 and V_2 be vector spaces. Then their *direct sum*, written $V_1 \oplus V_2$, is the vector space whose elements are pairs (v_1, v_2) with addition induced by the direct sum of abelian groups[81] and scalar multiplication by $c(v_1, v_2) = (cv_1, cv_2)$.

The direct sum of two vectors spaces should be thought of as the elements which can be written as linear combinations of elements of the two spaces. Thus for example if V_1 has dimension n_1 and V_2 has dimension n_2 then $V_1 \oplus V_2$ has dimension $n_1 + n_2$.

Proposition 2.4.24. *A linear transform between n-dimensional vector spaces has at most n eigenvalues.*

Proof. To each eigenvalue is associated an eigenspace, which has dimension at least 1. Since all of these eigenspaces are contained within the overall vector space V, all linear combinations of their elements lie within the vector space; in other words, if V_λ is the subspace of V associated to λ, i.e. our operator A acts on vectors in V_λ by $Av = \lambda v$ then

$$\bigoplus_\lambda V_\lambda \subseteq V.$$

Taking dimensions of each size and using the formula that the dimension of a direct sum is the sum of dimensions, we have

$$\dim\left(\bigoplus_\lambda V_\lambda\right) = \sum_\lambda \dim V_\lambda \leq \dim V = n.$$

[81] Previously, we've referred to this as the direct product of groups; when there are only finitely many terms, they are the same thing. Here, the addition being induced by the direct sum means that $(v_1, v_2) + (v_1', v_2') = (v_1 + v_1', v_2 + v_2')$ just as for the direct sum/direct product composition laws.

Since each $\dim V_\lambda$ is at least 1, the sum

$$\sum_\lambda \dim V_\lambda$$

is at least the number of eigenvalues, call it E. Therefore we conclude that

$$E \leq \sum_\lambda \dim V_\lambda \leq \dim V = n$$

and therefore $E \leq n$. □

For generic operators A, we will usually have $\dim V_\lambda = 1$ for all λ and thus the leftmost inequality is actually an equality:

$$E = \sum_\lambda \dim V_\lambda.$$

However this will fail reasonably often: for example, the only eigenvalue of the identity I_n is 1, since $I_n v = v$ for every v, and the entire vector space k^n is the eigenspace of 1.

Similarly, if our field k is sufficiently nice[82] then the second inequality holds, i.e.

$$\sum_\lambda \dim V_\lambda = \dim V.$$

This can be lifted to a statement about direct sums:

$$\bigoplus_\lambda V_\lambda = V,$$

i.e. every element of V can be written as a linear combination of eigenvectors of V. Thus in good cases we should expect A to have exactly n eigenvalues; but this may well fail for one reason or the other.

Since the eigenvalues can be defined using only the linear transformation interpretation of a matrix, they are basis-independent; in fact they encode a great deal of the other basis-independent information we've seen. For example, we can ask if a linear transformation A is invertible. Well, assuming we're in the nice case where

$$\bigoplus_\lambda V_\lambda = V,$$

we only have to worry about the restriction of A to each V_λ, since we can reconstruct the action of A on all of V by adding them together. Thus the question is just whether A is invertible on all of the V_λ. Since A acts on V_λ by multiplication by λ, this is just the question of whether any of the λ are zero; if so, then A is not invertible, while if not then A is invertible. To only have to test whether a single number is zero, take the product of all the eigenvalues; then it is nonzero if and only if all of the eigenvalues are, and therefore is nonzero if and only if A is invertible.

[82] In particular algebraically closed, which is a term we'll get to eventually.

Recall that we had another object that was nonzero if and only if A was nonzero: the determinant of A. And in fact, the basis-free definition of A is that $\det(A)$ is the product of the eigenvalues of A!

The trace can also be defined in this way: it is the *sum* of the eigenvalues. This is easier to see, at least in the nicest case: if A has n distinct eigenvalues, then choosing a nonzero vector v_λ for each λ gives a basis for V. Writing A in this basis, the resulting matrix will then be diagonal, with diagonal entries the eigenvalues; and so the trace, the sum of the diagonals, is just the sum of the eigenvalues.

The eigenvalues can also be encoded fully in a single object.

Definition 2.4.25. Let A be a linear transformation between n-dimensional vector spaces over k. Its *characteristic polynomial* is the polynomial $P(x)$ in $k[x]$ defined by

$$P(x) = \det(xI_n - A).$$

This gives a polynomial

$$P(x) = x^n - a_{n-1}x^{n-1} + a_{n-2}x^{n-1} - \cdots + (-1)^{n-1}a_1 x + (-1)^n a_0$$

of degree n whose coefficients encode information about A. For example, $a_n = \operatorname{tr}(A)$, and $a_n = \det(A)$.

In certain cases, a linear transformation $V \to V$ gives a decomposition of V as a direct sum.

Lemma 2.4.26. *Let V be a vector space, and $T : V \to V$ a linear transformation such that there exists a nonzero scalar c such that for any $v \in V$ we have $T(T(v)) = cT(v)$. Then $V = \ker(T) \oplus \operatorname{im}(T)$.*

Proof. Let v be a vector. Then $T(v)$ is in the image of T, by definition. On the other hand $v - \frac{1}{c}T(v)$ is in the kernel of T, since

$$T\left(v - \frac{1}{c}T(v)\right) = T(v) - \frac{1}{c}T(T(v)) = T(v) - T(v) = 0.$$

Therefore we can write v as a linear combination of an element of $\ker(T)$ and an element of $\operatorname{im}(T)$; and this decomposition is unique, since if $v \in \ker(T) \cap \operatorname{im}(T)$, i.e. there exists some w such that $v = T(w)$ and $T(v) = T(T(w)) = 0$, then $T(T(w)) = cT(w) = 0$ and so $T(w) = v = 0$. □

The direct sum is not the only way of generating a new vector space from hold ones. In fact, we have already seen one of the main other constructions.

Proposition 2.4.27. *Let V and W be vector spaces over k. Then the set of k-linear transformations[83] from V to W, written $\operatorname{Hom}_k(V, W)$, is itself a k-vector space.*

[83]This just means linear transformations as k-vector spaces, i.e. commuting with scalar multiplication by k. For example, there are $\mathbb{R}$-vector spaces that are not $\mathbb{C}$-vector spaces, such as $\mathbb{R}$ itself; and there are vector spaces, such as $\mathbb{C}^2$, which are vector spaces over both $\mathbb{C}$ and $\mathbb{R}$ but have different behavior; for example, $\mathbb{C}^2$ is two-dimensional over $\mathbb{C}$, but four-dimensional over $\mathbb{R}$, and $\operatorname{Hom}_{\mathbb{R}}(\mathbb{C}^2, \mathbb{C}) \simeq \operatorname{Mat}_{2\times 4}(\mathbb{R})$ is very different from $\operatorname{Hom}_{\mathbb{C}}(\mathbb{C}^2, \mathbb{C}) \simeq \operatorname{Mat}_{1\times 2}(\mathbb{C})$.

Proof. The set of k-linear transformations $V \to W$, upon choosing a basis, is just the set of $(\dim W) \times (\dim V)$ matrices over k, which we have already seen to be a vector space. More abstractly, we can add linear transformations by $(T_1 + T_2)(v) = T_1(v) + T_2(v)$, and scale them by $(cT)(v) = cT(v)$. These are compatible by the distributive law on k. □

A particularly important example of this is when W is the one-dimensional vector space $W = k$. In this case $\mathrm{Hom}_k(V, k)$ is called the *dual* vector space of V, written $V^\vee$, and consists of linear functions $V \to k$. For example, if the elements of V are n-dimensional column vectors, then the theory of matrices tells us that linear transforms $V = k^n \to k$ are given by $1 \times n$ matrices, i.e. *row vectors*

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}.$$

Thus here $V^\vee$ can be thought of as the set of n-dimensional row vectors.

Why is this called the dual? Well, what is the dual $(V^\vee)^\vee$ of $V^\vee$? In our example, this is the set of linear transformations from the set of n-dimensional row vectors to scalars, i.e. 1×1 matrices, and again the theory of matrices tells us that these maps correspond to $n \times 1$ matrices, i.e. column vectors. Thus this double dual is the same thing as what we started with! Thus if we were to keep taking the dual, e.g. $(((V^\vee)^\vee)^\vee)^\vee$, we will always get essentially the same thing as either V or $V^\vee$; there are really only these two options.

Of course, V and $V^\vee$ are both n-dimensional vector spaces over k, so they are both isomorphic to k^n and therefore isomorphic to each other. Why then do we say that $(V^\vee)^\vee$ is the "same thing" as V, but $V^\vee$ is different?

The answer is that although V and $V^\vee$ are isomorphic, they are not *canonically* isomorphic: in order to find an isomorphism, we have to choose bases for both V and $V^\vee$, and there is no canonical way to make these choices. Therefore we don't expect this isomorphism to be "natural".

However, there *is* a canonical isomorphism $V \to (V^\vee)^\vee$. Specifically, recall that $(V^\vee)^\vee = \mathrm{Hom}(\mathrm{Hom}(V, k), k)$ is the set of functions $\mathrm{Hom}(V, k) \to k$. Then given a vector v, we can define such a function $\phi_v : \mathrm{Hom}(V, k) \to k$ sending a function $f : V \to k$ to $\phi_v(f) := f(v)$, i.e. we send a linear function f to its value at v. This is a linear map, as is the assignment $v \mapsto \phi_v$. We want to show that $v \mapsto \phi_v$ is an isomorphism. First, it is injective, since if ϕ_v is the zero function, i.e. $\phi_v(f) = f(v) = 0$ for every linear function $f : V \to k$, then v must be zero: if not then we can define a function f to be 1 at v and extend linearly to get a counterexample. But then since $v \mapsto \phi_v$ is injective and both V and $(V^\vee)^\vee$ are n-dimensional k-vector spaces, it must be an isomorphism: since $v \mapsto \phi_v$ is injective, its dimension as a subspace of $(V^\vee)^\vee$ is n-dimensional, and therefore must be all of $(V^\vee)^\vee$.

There is one more way of forming new vector spaces from old ones: the tensor product.

Definition 2.4.28. Let V and W be two vector spaces over k. Their *tensor product*, written

$$V \otimes W$$

or

$$V \otimes_k W$$

if the field is potentially unclear, is the vector space consisting of linear combinations of elements of the form

$$v \otimes w$$

for $v \in V$ and $w \in W$, with the relations that

$$(v_1 + v_2) \otimes w = v_1 \otimes w + v_2 \otimes w,$$

$$v \otimes (w_1 + w_2) = v \otimes w_1 + v \otimes w_2,$$

and

$$(cv) \otimes w = v \otimes (cw),$$

so that we can unambiguously write $cv \otimes w$ and not worry about which factor c is acting on. This addition and scalar multiplication makes $V \otimes W$ into a vector space over k.

Example 2.4.29. Let $V = \mathbb{R}^2$ and $W = \mathbb{R}^3$ as vector spaces over $\mathbb{R}$. We have a natural basis for each of these consisting of $v_1 = (1, 0)$, $v_2 = (0, 1)$ for V and $w_1 = (1, 0, 0)$, $w_2 = (0, 1, 0)$, and $w_3 = (0, 0, 1)$. Then the basis vectors for $V \otimes W$ are

$$v_1 \otimes w_1, \qquad v_1 \otimes w_2, \qquad v_1 \otimes w_3, \qquad v_2 \otimes w_1, \qquad v_2 \otimes w_2, \qquad v_2 \otimes w_3.$$

Thus $V \otimes W$ is the six-dimensional $\mathbb{R}$-vector space consisting of linear combinations of these vectors.

In general, the same idea shows that $\dim(V \otimes W) = \dim V \cdot \dim W$.

2.5 Representation theory

Now that we've introduced fields and vector spaces over them, we can reveal our secret goal: this is all actually about the study of groups!

Specifically, we've encountered the idea of a group action, where a group G acts on a set S. If we look at all G-sets, i.e. sets S with an action of G, we can recover some information about G.

Unfortunately, since sets themselves have relatively little structure, it's not much easier to study G-sets than it is to study G directly. This suggests the idea of studying some other class of algebraic objects equipped with an action of the group G, in a way compatible with their algebraic structure.

What is the appropriate structure? Well, our options so far are groups, rings, and vector spaces. Let's say we have a group action of G on another group H. In order for the action to be compatible with the group structure on H, what this really means is that to each $g \in G$ we assign a group homomorphism $g : H \to H$, with the usual compatibility rules on these actions, i.e. $(g_1 g_2) \cdot h = g_1 \cdot (g_2 \cdot h)$ for any $g_1, g_2 \in G$ and $h \in H$, and the group identity e_G acts by the identity $e_G \cdot h = h$ for all $h \in H$.

A point of notation I don't think we've introduced yet: if A is some algebraic object equipped with a notion of homomorphisms (e.g. groups or rings, or vector spaces where the homomorphisms are linear transformations, or even sets where the "homomorphisms" are just functions)[84] then a homomorphism $A \to A$ is called an *endomorphism*, and an isomorphism $A \to A$, i.e. an invertible endomorphism, is called an *automorphism*. Thus a

[84] We'll be able to formalize this notion when we talk about category theory.

group action of G on A is an assignment of an endomorphism $A \to A$ for each group element $g \in G$; since every element of g is invertible, so is the corresponding endomorphism, since $g^{-1} \cdot (g \cdot h) = (g^{-1}g) \cdot h = e_G \cdot h = h$. Thus a group action of G on A is really an assignment of an automorphism of A to each element of G in a compatible way. Writing $\mathrm{Aut}(A)$ for the set of automorphisms of A, we can therefore think of a group action of G on A as a map $\rho : G \to \mathrm{Aut}(A)$ satisfying certain compatibility properties.

What are these compatibility properties? Well, the condition that e_G acts by the identity can be written in this way as $\rho(e_G) = \mathrm{id} \in \mathrm{Aut}(A)$, and the condition that $g_1 \cdot (g_2 \cdot h) = (g_1 g_2) \cdot h$ is equivalent to the condition that $\rho(g_1)\rho(g_2) = \rho(g_1 g_2)$. This should remind us of a group homomorphism: and in fact $\mathrm{Aut}(A)$ is a group under composition, i.e. if σ and θ are two automorphisms of A then we form the product $\sigma \circ \theta$ for the automorphism of A given by first doing θ and then σ. Since automorphisms are invertible and we have an identity element $\mathrm{id} : A \to A$, this shows that $\mathrm{Aut}(A)$ is a group, and therefore a group action on A is precisely a group homomorphism $G \to \mathrm{Aut}(A)$.

Okay. If A is a group, as above, this definition makes perfect sense, but as in the case of sets we don't really know how to say much about $\mathrm{Aut}(A)$; it can be defined, but isn't necessarily any easier.

What about rings? In general, rings can be pretty complicated objects; let's restrict to fields for simplicity, but even if we choose a "nice" field like the complex numbers $\mathbb{C}$ the automorphism group $\mathrm{Aut}(\mathbb{C})$ is still extremely complicated.[85]

How can we make it simpler? Instead of thinking of $\mathbb{C}$ as a field, what if we consider it as a one-dimensional vector space over itself? This makes sense: $\mathbb{C}$ has good notion of multiplication and scalar multiplication by itself, so it is in fact a vector space over $\mathbb{C}$. What are the vector space automorphisms?

These are the invertible linear transformations $T : \mathbb{C} \to \mathbb{C}$. Theorem 2.4.18 tells us that T is given by a matrix; since $\mathbb{C}$ is a one-dimensional vector space, T is given by a 1×1 matrix acting by

$$\begin{pmatrix} t \end{pmatrix} \begin{pmatrix} x \end{pmatrix} = \begin{pmatrix} tx \end{pmatrix},$$

i.e. this is just scalar multiplication by some complex number t. Thus the endomorphisms of $\mathbb{C}$ *as a* $\mathbb{C}$*-vector space*, written $\mathrm{End}_{\mathbb{C}}(\mathbb{C})$, can be canonically identified with the complex numbers $\mathbb{C}$: $\mathrm{End}_{\mathbb{C}}(\mathbb{C})$. Such an endomorphism given by scalar multiplication by t is invertible if and only if t is, i.e. $\mathrm{Aut}_{\mathbb{C}}(\mathbb{C})$, the automorphisms of $\mathbb{C}$ as a $\mathbb{C}$-vector space, is just $\mathbb{C}^\times$.

This is something we can get our hands on: $\mathbb{C}^\times$ is just the set of nonzero complex numbers, made into a group by the operation of multiplication. Of course, we can also consider other vector spaces. Theorem 2.4.18 tells us that the endomorphisms of $\mathbb{C}^n$ as a $\mathbb{C}$-vector space are precisely $n \times n$ matrices over $\mathbb{C}$, i.e. $\mathrm{End}_{\mathbb{C}}(\mathbb{C}^n) = \mathrm{Mat}_{n\times n}(\mathbb{C})$; by definition, the invertible matrices and thus the automorphisms of $\mathbb{C}^n$ are then given by

$$\mathrm{Aut}_{\mathbb{C}}(\mathbb{C}^n) = \mathrm{GL}_n(\mathbb{C}).$$

Thus what we have shown above is that $\mathrm{GL}_1(\mathbb{C}) = \mathbb{C}^\times$. Of course, we can do the same thing over any field k, though in this section we'll primarily work with $k = \mathbb{C}$.

[85] Though field automorphisms are certainly worth studying, and will be the focus of the next section.

Again, these are objects we can get our hands on and understand reasonably well, while still giving a lot of information about the group we want to study; and so we arrive at the following definition.

Definition 2.5.1. A *representation* of a group G over a field k is a vector space V over k with an action by G, i.e. for each $g \in G$ a map $V \to V$ satisfying the compatibility conditions

(1) each map $g : V \to V$ is a linear transformation;

(2) the identity e_G acts by the identity, i.e. $e_G \cdot v = v$ for all $v \in V$;

(3) and for all $g_1, g_2 \in G$ and $v \in V$ we have $(g_1 g_2) \cdot v = g_1 \cdot (g_2 \cdot v)$.

Equivalently, a representation is a vector space V together with a group homomorphism $\rho : G \to \mathrm{Aut}(V) = \mathrm{GL}(V) = \mathrm{GL}(V, k)$. If we only care about the isomorphism class of V, we can write $V \simeq k^n$ and so a representation is just a homomorphism $\rho : G \to \mathrm{GL}_n(k)$; we will write any of (V, ρ), just V, or just ρ for the representation depending on context.

The techniques of representation theory can be applied to any group, but for simplicity we'll mostly be focusing in this section on finite groups.

Let's start even simpler, with finite abelian groups. In fact, let's see how simple we can get: the smallest group $\mathbf{1} = \{1\}$ is trivial, and so doesn't have any exciting representations: any map $\rho : \mathbf{1} \to \mathrm{GL}_n(\mathbb{C})$ must take 1 to the identity matrix I_n, and so nothing interesting happens. In general there is always a representation that looks like this: if G is a group, we can always map every element of G to the identity I_n. This is called the *trivial representation*, and though important is the most boring representation. In this case it is the only one.

Let's try the next-smallest group, the simplest nontrivial group: the cyclic group C_2 of order 2, which we can think of as $\{1, -1\}$ with group operation multiplication, so that 1 is the identity and -1 has order 2, since $(-1)^2 = 1$. What are its representations?

Let's start with the one-dimensional representations, i.e. actions of C_2 on $\mathbb{C}$ as a $\mathbb{C}$-vector space, or equivalently homomorphisms $\rho : C_2 \to \mathbb{C}^\times$. Since the identity of $\mathbb{C}^\times$ is 1, we must have $\rho(1) = 1$; and $\rho(-1)$ must satisfy $\rho(-1)^2 = \rho((-1)^2) = \rho(1) = 1$. There are two solutions to the equation $\rho(-1)^2 = 1$: either $\rho(-1) = 1$ or $\rho(-1) = -1$.

This first of these is the trivial representation, since $\rho(g) = 1$ for every $g \in C_2$. The second representation, $\rho(1) = 1$ and $\rho(-1) = -1$, is a nontrivial representation: this is exciting, we've gotten nontrivial information!

Note that this second representation is *not* the identity map, because its domain (C_2) and codomain ($\mathbb{C}^\times$) are different. They only look like the identity because we chose to write C_2 in this particular way; thus we see that even our way of writing down groups is in some cases actually choosing a representation.

Since the group structure of C_2 required that $\rho(-1)^2 = 1$, these are the only two one-dimensional representations. Let's keep going up: what are the two-dimensional representations, i.e. homomorphisms $\rho : C_2 \to \mathrm{GL}_2(\mathbb{C})$?

Well, we have the same two requirements: $\rho(1)$ always has to be the identity

$$\rho(1) = I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and $\rho(-1)$ must satisfy the identity $\rho(-1)^2 = \rho(1) = I_2$. Thus we are looking for matrices

$$\rho(-1) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

such that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^2$$

is equal to I_2. Well, we can multiply this out: we get

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^2 = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2+bc & ab+bd \\ ac+cd & bc+d^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which gives us four equations:

$$\begin{aligned} a^2 + bc &= 1, \\ ab + bd &= 0, \\ ac + cd &= 0, \\ bc + d^2 &= 1. \end{aligned}$$

Now, these equations do not have a unique solution, or even a finite number of solutions. However, we shouldn't expect them to: for $n > 1$, there are infinitely many choices of basis even if we don't count scaling,[86] and we only care about representations up to isomorphism, i.e. up to basis change. We know that changing the basis acts by conjugacy on our matrix, and so what we really want to count is conjugacy classes of matrices $\rho(-1)$ satisfying $\rho(-1)^2 = I_2$.[87] In the two-dimensional case, it is actually quite easy to check if two matrices are conjugate: we know two conjugacy invariants of matrices, the trace and the determinant, and for 2×2 matrices these completely determine the conjugacy class, i.e. A and B are conjugate if and only if $\operatorname{tr}(A) = \operatorname{tr}(B)$ and $\det(A) = \det(B)$.[88] Thus it's enough to determine the possible values of $\operatorname{tr}(\rho(-1)) = a + d$ and $\det(\rho(-1)) = ad - bc$.

Let's take another look at our equations. Factoring the second and third equations, we get $b(a + d) = 0$ and $c(a + d) = 0$, so either $a + d = 0$ or $b = c = 0$. In the first case, this immediately tells us that $\operatorname{tr}(\rho(-1)) = 0$, and therefore $a = -d$ and so $a^2 + bc = -ad + bc = -\det(\rho(-1))$. But then the first (or equivalently fourth) equation tells us

[86] That is, if we count v and cv as the same basis vector. In the one-dimensional case, strictly speaking there are infinitely many bases, since we can take any nonzero complex number to be a basis vector; but they are all the same up to scaling, and since scaling e.g. -1 by any nontrivial factor makes it no longer true that $(-1)^2 = 1$ we can ignore this particular kind of basis change. However, in larger dimensions we can have nontrivial basis change, and so we need to account for it.

[87] A conjugacy class is just an equivalence class where the equivalence relation is whether two matrices are conjugate: A and B are conjugate if there exists some matrix M such that $A = MBM^{-1}$.

[88] We won't prove this, but it's not difficult: just write down general matrices A, B, and M, compute MBM^{-1}, and verify that it's possible to choose M such that $A = MBM^{-1}$ so long as B has the same trace and determinant as A. This is tedious but not too difficult. (The reverse direction is easier: if A and B have different trace or determinant, then they cannot be conjugate, since trace and determinant are conjugacy-invariant. This part is true for all dimensions, but the other direction is special for 2×2 matrices.

that $\det(\rho(-1)) = -1$. A representative of this conjugacy class, i.e. isomorphism class of representations, is

$$\rho(-1) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

or equivalently

$$\rho(-1) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

On the other hand, suppose that instead $b = c = 0$. Then the first and fourth equations tell us that $a^2 = d^2 = 1$, and so a and d are each ± 1. This leaves four possibilities:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The first of these is the identity matrix, and corresponds to the trivial representation, since then $\rho(1) = \rho(-1) = I_2$. The second and third matrices are the case above, and are in the same isomorphism class. The final matrix is a distinct isomorphism class, since it has trace -2 and determinant 1.

Thus we have three isomorphism classes of two-dimensional representations of C_2. But notice something: in every case, we can choose the matrix to be diagonal. A diagonal matrix A acts on a vector by

$$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax \\ by \end{pmatrix},$$

i.e. it acts separately by scalar multiplication on each coordinate. Thus if we write our vector space $V = \mathbb{C}^2$ as the direct sum $V = \mathbb{C} \oplus \mathbb{C}$ of one-dimensional representations, then A decomposes as the direct sum of two 1×1 matrices:

$$A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} = (a) \oplus (b)$$

acting on $\mathbb{C} \oplus \mathbb{C}$ by multiplication by a on the first factor and multiplication by b on the second factor. More generally, if $A_1 : V_1 \to W_1$ and $A_2 : V_2 \to W_2$ are linear transformations, then $A_1 \oplus A_2 : V_1 \oplus V_2 \to W_1 \oplus W_2$ acts by $(v_1, v_2) \mapsto (A_1 v_1, A_2 v_2) \in W_1 \oplus W_2$. In terms of matrices, this results in block matrices on the diagonals: for example,

$$\begin{pmatrix} 2 & 0 \\ 3 & -1 \end{pmatrix} \oplus \begin{pmatrix} 4 & 5 & 0 \\ 0 & 1 & -1 \\ 2 & 2 & -3 \end{pmatrix} = \left(\begin{array}{cc|ccc} 2 & 0 & 0 & 0 & 0 \\ 3 & -1 & 0 & 0 & 0 \\ \hline 0 & 0 & 4 & 5 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 2 & 2 & -3 \end{array}\right).$$[89]

In our case, the fact that all our matrices are diagonal should make us think that they will split as direct sums. Let's go back to the one-dimensional case: write $\chi_0 : C_2 \to \mathbb{C}^\times$ for the trivial representation $\chi_0(1) = \chi_0(-1) = 1$, and write $\chi_1 : C_2 \to \mathbb{C}^\times$ for the nontrivial

[89] The vertical and horizontal lines are just to demonstrate how the matrices are split up, and do not have any other mathematical significance.

representation with $\chi_1(-1) = -1$.[90] Since $\mathbb{C}^\times$ is the same thing as $\mathrm{GL}_1(\mathbb{C})$, think of $\chi_0(1)$, $\chi_0(-1)$, $\chi_1(1)$, and $\chi_1(-1)$ as 1×1 matrices. We can decompose all of the matrices above as a direct sum of some number of copies of (1) and (−1). But in fact, we can do better than that.

Define the direct sum of representations by $(\rho_1 \oplus \rho_2)(g) = \rho_1(g) \oplus \rho_2(g)$. Then we have for example ρ_0, the two-dimensional trivial representation, equal to $\chi_0 \oplus \chi_0$, since for each $g \in C_2$ we have $\rho_0(g) = I_2 = \chi_0(g) \oplus \chi_0(g) = (1) \oplus (1)$.

For the other representations, the value at 1 is irrelevant, since for any representation we have $\rho(1) = I_2 = (1) \oplus (1)$ and the value of every one-dimensional representation at 1 is (1). Thus the only value that matters for the classification is $\rho(-1)$; and using this we see that our three two-dimensional representations are

$$\chi_0 \oplus \chi_0, \qquad \chi_0 \oplus \chi_1 \simeq \chi_1 \oplus \chi_0, \qquad \chi_1 \oplus \chi_1.$$

In particular, every two-dimensional representation is the direct sum of one-dimensional representations!

We could go ahead and compute the three-dimensional representations in the same way, and we would find that up to isomorphism they are

$$\chi_0 \oplus \chi_0 \oplus \chi_0, \qquad \chi_0 \oplus \chi_0 \oplus \chi_1, \qquad \chi_0 \oplus \chi_1 \oplus \chi_1, \qquad \chi_1 \oplus \chi_1 \oplus \chi_1.$$

This pattern will continue: all finite-dimensional representations of C_2 are (isomorphic to) direct sums of one-dimensional representations!

In general, we want the following definition.

Definition 2.5.2. Let G be a group, and V be a representation of G over $\mathbb{C}$, with action given by $\rho : G \to \mathrm{GL}(V)$. Let W be a subspace of V. We say that W is a *subrepresentation* of V if for every $g \in G$, the automorphism $\rho(g)$ fixes W: that is, for any $w \in W$, we have $\rho(g)(w) \in W$.

This is the same thing as requiring that W itself be a representation of G: the action $W \to W$ is induced by the action on the larger space V, but we need the image of W to actually land in W to make sure that we have a map $W \to W$ rather than just $W \to V$.

We can now give the main theorem explaining why representations should split as direct sums.

Theorem 2.5.3. *Let V be a complex representation[91] of a finite group G, and let W be a subrepresentation of V. Then there exists another subrepresentation U of V such that*

$$V \simeq W \oplus U.$$

Proof. Let $\{b_1, \ldots, b_{\dim W}\}$ be a basis for W. Then using the usual method from the previous section for constructing bases, we can build a basis for V extending this basis for W: $\{b_1, \ldots, b_{\dim W}, b_{\dim W+1}, \ldots, b_{\dim V}\}$. Let U be the subspace of V spanned by $b_{\dim W+1}, \ldots, b_{\dim V}$.[92] Then by the definition of the direct sum and by construction $W \oplus U = V$ as vector spaces (though not necessarily as representations).

[90] It is traditional to use χ for one-dimensional representations, for reasons that will become clear shortly.

[91] That is, a representation over $\mathbb{C}$.

[92] This is called the complement of W in V, and is a perfectly good linear algebra concept independent of the representation-theoretic structure.

Given a vector $v \in V$, we can write $v = w + u$ for $w \in W$ and $u \in U$, since $V = W \oplus U$ as vector spaces. Write $p_W(v)$ for the function that takes v to w in this decomposition; this is called the *projection* of V onto W, and is a linear transformation $V \to W$. For each $g \in G$, we can act on v by g and then project down to W to get $p_W(g \cdot v)$; then act by g^{-1} to get a vector $g^{-1} \cdot p_W(g \cdot v)$. The map t_g defined by

$$v \mapsto g^{-1} \cdot p_W(g \cdot v)$$

is a linear transformation since p_W is and all of the group elements act by linear transformations, and the composition of two linear transformations is also linear,[93] and since W is a subrepresentation of V and $p_W(g \cdot v)$ is in W by the definition of p_W the resulting vector

$$g^{-1} \cdot p_W(g \cdot v)$$

is also in W. Now let T be the linear transformation $V \to V$ defined by

$$v \mapsto \sum_g g^{-1} \cdot p_W(g \cdot v).$$

For each $w \in W$ we have $g \cdot w \in W$ since W is a subrepresentation of V, and so $p_W(g \cdot w) = g \cdot w$ and therefore $g^{-1} \cdot p_W(g \cdot w) = g^{-1} \cdot (g \cdot w) = (g^{-1}g) \cdot w = e_G \cdot w = w$, and so

$$T(w) = \sum_g w = |G| \cdot w.$$

As each $g^{-1} \cdot p_W(g \cdot v)$ is in W, so is their sum, so $T(v)$ is in W for every v; and therefore $T(T(v)) = |G|T(v)$ by the above. Therefore Lemma 2.4.26 tells us that $V = \ker(T) \oplus \mathrm{im}(T)$. Since T is linear we have

$$T(w/|G|) = w$$

for every $w \in W$, so in fact the image of T is all of W; so by Lemma 2.4.26 we have $V = \ker(T) \oplus W$, so if we can show that $\ker(T)$ is a subrepresentation of V then we're done.

Let $v \in \ker(T)$, and $g' \in G$. Then

$$T(g' \cdot v) = \sum_g g^{-1} \cdot p_W(g \cdot (g' \cdot v) = \sum_g g^{-1} \cdot p_W((gg') \cdot v),$$

and letting $h = gg'$ we can rewrite this as

$$\sum_h (g'h^{-1}) \cdot p_W(h \cdot v) = g' \cdot \sum_h h^{-1} \cdot p_W(h \cdot v) = g' \cdot T(v) = g' \cdot 0 = 0,$$

since $v \in \ker(T)$. Therefore $g' \cdot v$ is also in $\ker(T)$ for every $g' \in G$ and $v \in \ker(T)$, and so $\ker(T)$ is stable under the group action: that is, $\ker(T)$ is a subrepresentation of V such that $V = W \oplus \ker(T)$ as representations of G. □

[93] e.g. because it is given by the product of matrices.

If we have a representation V of our group G which can be written as $V \simeq W \oplus U$ for subrepresentations W and U, then we can understand V by understanding U and W. This is generally a useful approach because smaller-dimensional representations are generally simpler (recall that $\dim V = \dim U + \dim W$). What Theorem 2.5.3 tells us is that it is enough to understand the *irreducible* representations, i.e. those representations V whose only subrepresentation is V itself.

Corollary 2.5.4 (Maschke's theorem)**.** *Let G be a finite group. Then any finite-dimensional complex representation of G can be written as a direct sum of irreducible representations.*

Proof. Let V be a complex representation of G. If V is irreducible, then the result is immediate: V can be written as the direct sum with one term consisting of V itself, and since V is irreducible this satisfies the statement of the corollary.

In the case where V is not irreducible, the proof proceeds by induction. Let $n = \dim V$. If $n = 1$, then V must be irreducible, since if $V \simeq U \oplus W$ then $\dim V = 1 = \dim U + \dim W$ and each of $\dim U$ and $\dim W$ must be at least 1, so this is impossible; therefore we are in the case above and the result holds. If $n = 2$, then either V is irreducible or it has some proper[94] subrepresentation U, and then by Theorem 2.5.3 there also exists some other subrepresentation W such that $V \simeq U \oplus W$. But since both U and V have dimension 1, we know that they are irreducible and so the result still holds.

For $n = 3$, the method is similar: if V is not irreducible, then by Theorem 2.5.3 it can be written as $U \oplus W$ for some subrepresentations U and W, each of which has dimension at least 1; since $\dim U + \dim W = \dim V = 3$, it follows that they both have dimension at most 2. Since we know that the result holds for 2-dimensional representations, we can then replace each of U and V by a direct sum of irreducible representations, and putting these together gives V as a direct sum of irreducible representations.

We can continue like this: if we know that all representations of dimension at most $n-1$ can be written as a direct sum of irreducible representations, then if V is a reducible[95] representation with $V \simeq U \oplus W$ then applying the claim to U and W shows that V can also be written as a direct sum of irreducible representations. Doing this for all n gives the result for all finite-dimensional representations. □

Remark. The technique used in this proof is known as *induction*, and is frequently useful. The idea is the following: suppose we have some assertion which we wish to prove for all natural numbers n; in this case, the assertion is that all n-dimensional complex representations of a finite group G can be written as the direct sum of irreducible representations. Then there are two things we have to prove: first, we need to prove that the result holds for some base case, say $n = 1$, and then we need to prove that if the assertion holds for everything up to $n-1$, then it also holds for n. Putting these together, since the assertion holds for $n = 1$, it also holds for $n = 2$; and then it holds for $n = 3$ as well, and so on.

Thus it suffices to study *irreducible* representations of G. This should remind you of how in order to study the natural numbers $\mathbb{N}$, we end up studying the "indecomposable objects," i.e. the primes.

[94] I.e. not equal to all of V.

[95] I.e. not irreducible.

Given a finite group G, we might then ask: how many irreducible representations does it have? What are their dimensions? What do they look like?

Before we can answer these questions, we need to understand what maps of representations look like.

Definition 2.5.6. Let G be a finite group, and V and W be representations of G over k. A *homomorphism* of representations $\varphi : V \to W$ is a linear transformation $V \to W$ as k-vector spaces which respects the action of G, i.e. for any $g \in G$ and $v \in V$ we have

$$\varphi(g \cdot v) = g \cdot \varphi(v).$$

In other words, φ *commutes* with the action of G: it does not matter which we apply first.

Lemma 2.5.7 (Schur's lemma)**.** *Let V and W be complex irreducible representations of a finite group G. Either V and W are isomorphic as representations, in which case the only homomorphisms of representations $V \to W$ are given by multiplication by some scalar, or they are not isomorphic, in which case the only homomorphism of representations $V \to W$ is the zero map $v \mapsto 0$.*[96]

Proof. Let $\varphi : V \to W$ be a homomorphism of representations, and assume that it is nonzero, i.e. there exists some $v \in V$ such that $\varphi(v) \neq 0$. Let U be the kernel of φ, i.e. the subspace of V consisting of vectors $u \in V$ such that $\varphi(u) = 0$. Since

$$\varphi(g \cdot u) = g \cdot \varphi(u),$$

we have

$$\varphi(g \cdot u) = 0$$

for all $u \in U$ and $g \in G$, and so $g \cdot u$ is also in the kernel of φ if u is. In particular, g takes elements of U to elements of U for all $g \in G$, and so U is a subrepresentation of V. But since there exists some v such that $\varphi(v) \neq 0$, i.e. $v \notin U$, the subrepresentation U is not equal to all of V; since V is irreducible, it follows that U must be the trivial subspace $\{0\}$. Since $U = \ker \varphi$, this implies that φ is an isomorphism.

Thus we've shown that either φ is the zero map or it is an isomorphism. Suppose it is an isomorphism, so that (since we are concerned only with isomorphism classes) we can think of φ as a homomorphism of representations $V \to V$. We want to show that it is given by scalar multiplication.

Let λ be an eigenvalue of φ.[97] Then we can form another linear transformation

$$(\varphi - \lambda \cdot \mathrm{id}_V) : V \to V$$

given by

$$v \mapsto \varphi(v) - \lambda v.$$

[96] This could also be phrased as saying that the only maps between irreducible representations are multiplication by scalars; either that scalar is 0, leading to the zero map, or it is not, in which case it is an isomorphism.

[97] That is, a scalar such that there exists some nonzero $v \in V$ such that $\varphi(v) = \lambda v$.

In fact this is also a homomorphism of representations:

$$(\varphi - \lambda \operatorname{id}_V)(g \cdot v) = \varphi(g \cdot v) - \lambda g \cdot v = g \cdot \varphi(v) - g \cdot \lambda v = g \cdot (\varphi(v) - \lambda v).$$

Therefore we can apply the above result to it: either it is an isomorphism or it is the zero map.

But wait a moment: it cannot possibly be an isomorphism, because by definition there exists some nonzero $v \in V$ such that $\varphi(v) = \lambda v$, and so

$$(\varphi - \lambda \operatorname{id}_V)(v) = \varphi(v) - \lambda v = \lambda v - \lambda v = 0.$$

Therefore $\varphi - \lambda \operatorname{id}_V$ has nontrivial kernel and so is not an isomorphism; so by the above it must be the zero map, $\varphi - \lambda \operatorname{id}_V = 0$. Thus we conclude that $\varphi = \lambda \operatorname{id}_V$, i.e. $\varphi(v) = \lambda v$ for *all* $v \in V$. □

Thinking of a representation as a map $\rho : G \to \mathrm{GL}_n(\mathbb{C})$ is pleasantly concrete—it's nice to be able to write down the action of each g as $\rho(g)$—but somewhat inconvenient: a whole matrix for each group element is a lot to keep track of. It would be nicer if we could reduce this somehow: say, assign a number to each group element. This is already the case for one-dimensional representations, for which $\rho(g) \in \mathrm{GL}_1(\mathbb{C}) = \mathbb{C}^\times$ is just a nonzero complex number for each g, but how could we do this for higher-dimensional representations?

Well, recall that we have two natural functions that take in matrices and spit out scalars: trace and determinant. To decide which to use, let's go back to our C_2 example, the cyclic group of order 2. Recall that up to isomorphism the three two-dimensional representations are ρ_0, ρ_1, ρ_2, each of which send the identity $1 \in C_2$ to the 2×2 identity matrix $1 \mapsto I_2$ and which act on the nontrivial element -1 to

$$\rho_0(-1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_2, \qquad \rho_1(-1) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \rho_2(-1) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I_2.$$

These have determinants 1, -1, and 1 respectively, and traces 2, 0, and -2. Since all of the traces are distinct while the determinants are not, this suggests that taking the trace is the better choice.

Definition 2.5.8. A *character* χ of a group G is the trace of a finite-dimensional representation $\rho : G \to \mathrm{GL}_n(k)$, i.e.

$$\chi(g) = \operatorname{tr}(\rho(g)).$$

It is thus a function $G \to k$.

We say that a character is irreducible if it comes from an irreducible representation.

Characters satisfy a number of good properties: for example, $\operatorname{tr}(\rho_1(g) \oplus \rho_2(g)) = \operatorname{tr}(\rho_1(g)) + \operatorname{tr}(\rho_2(g))$, and so we can detect whether a representation is the direct sum of two others by checking if its character is the sum of two other characters. Since the trace is an isomorphism invariant, unlike the presentation of the matrix $\rho(g)$, the character χ is well-defined, not just an isomorphism class, and so checking this may be much easier than checking whether ρ is

isomorphic to a direct sum. Observe also that since $\rho(1)$ is the $n \times n$ identity matrix

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

where n is the dimension of the representation ρ, its trace $\chi(1)$ is the sum of n copies of 1, i.e. $\chi(1) = n$. Thus $\chi(1)$ gives the dimension of the representation χ comes from.

Another important property of characters comes from the following (purely linear algebra) property of the trace.

Proposition 2.5.9. *Let A and B be square matrices. Then*

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

Proof. The trace of AB is the sum of the diagonal entries $(AB)_{i,i}$. By the definition of matrix multiplication, we have

$$(AB)_{i,i} = \sum_j A_{ij} B_{ji},$$

and so

$$\mathrm{tr}(AB) = \sum_i (AB)_{i,i} = \sum_i \sum_j A_{ij} B_{ji}.$$

If we rearrange the order of summation, then we see that this is equal to

$$\sum_j \sum_i A_{ij} B_{ji} = \sum_j \sum_i B_{ji} A_{ij}.$$

But now the same formula shows that

$$(BA)_{j,j} = \sum_i B_{ji} A_{ij},$$

and so we have

$$\mathrm{tr}(AB) = \sum_i \sum_j A_{ij} B_{ji} = \sum_j \sum_i B_{ji} A_{ij} = \sum_j (BA)_{j,j} = \mathrm{tr}(BA).$$

□

Remark. Note that this does *not* imply more elaborate commutativity relations: for example, $\mathrm{tr}(ABC)$ is not necessarily equal to $\mathrm{tr}(ACB)$. However, it is true that $\mathrm{tr}(A_1 A_2 \cdots A_n) = \mathrm{tr}(A_n A_1 A_2 \cdots A_{n-1}) = \mathrm{tr}(A_2 A_3 \cdots A_n A_1)$ for square matrices $A_1, \ldots, A_n$, as can be deduced from the above.

Corollary 2.5.11. *Let $\chi : G \to k$ be a character of G. Then for any $g, h \in G$ we have $\chi(gh) = \chi(hg)$. In particular, $\chi(h) = \chi(ghg^{-1})$ for any $g, h \in G$.*

Proof. We have

$$\chi(gh) = \mathrm{tr}(\rho(gh)) = \mathrm{tr}(\rho(g)\rho(h))$$

since ρ is a group homomorphism, where ρ is the representation that χ comes from. By Proposition 2.5.9, this is equal to

$$\mathrm{tr}(\rho(h)\rho(g)) = \mathrm{tr}(\rho(hg)) = \chi(hg),$$

so $\chi(gh) = \chi(hg)$. By the same logic,

$$\begin{aligned}\chi(ghg^{-1}) &= \mathrm{tr}(\rho(ghg^{-1})) = \mathrm{tr}(\rho(gh)\rho(g^{-1})) = \mathrm{tr}(\rho(g^{-1})\rho(gh)) \\ &= \mathrm{tr}(\rho(g^{-1}gh)) = \mathrm{tr}(\rho(h)) = \chi(h).\end{aligned}$$

□

In other words, every character is conjugation-invariant: the value of χ at h is the same as its value at every conjugate of h, i.e. every element ghg^{-1} for all $g \in G$.

More generally, we say that a function $f : G \to k$ is a *class function* if it is conjugacy-invariant, i.e. if $f(h) = f(ghg^{-1})$ for all $g, h \in G$. Thus Corollary 2.5.11 states that characters are class functions.

Now, the set of functions $S \to k$ for any finite set S forms a vector space over k, with addition given by $(f+g)(s) = f(s) + g(s)$ and scalar multiplication by $(c \cdot f)(s) = c \cdot f(s)$. Its dimension is $|S|$, with a basis given by the functions $\mathbf{1}_s$ for every $s \in S$, which are defined by $\mathbf{1}_s(t) = 0$ if $t \neq s$ and $\mathbf{1}_s(s) = 1$.

In our case, we can take S to be the set of conjugacy classes of G: each class function assigns a value to a conjugacy class. To get the full function $G \to k$, we compose with the function $G \to S$ sending an element g to its conjugacy class, which we will write $[g]$.

Therefore the set of class functions forms a k-vector space with dimension equal to the number of conjugacy classes of G. Since the characters are all class functions, i.e. elements of this vector space, it is tempting to assert that they should have some special role: for example, that they should form a basis for this space. In fact, after restricting to irreducible characters, more is true. Before giving the main result, though, let's compute some characters.

We've already computed the irreducible representations of the cyclic group C_2 (though strictly speaking we have not yet proven that they are the only ones). Since they are one-dimensional, taking the trace just gives back the same number, and so we have the character table

	1	-1
χ_0	1	1
χ_1	1	-1

.

Here the columns correspond to conjugacy classes; but since C_2 is abelian, every conjugacy class is just one element ($ghg^{-1} = gg^{-1}h = h$) and so conjugacy classes are the same thing as elements.

Now, suppose that $\rho : C_2 \to \mathrm{GL}_n(\mathbb{C})$ is an n-dimensional representation of C_2. We know that $\rho(1) = I_n$, so the question is what $\rho(-1)$ is. Suppose that $\mathrm{tr}(\rho(-1)) = m$, so that the corresponding character $\chi = \mathrm{tr}(\rho)$ has $\chi(1) = n$ and $\chi(-1) = m$. From our computations before, we think that χ_0 and χ_1 are the only two irreducible representations of C_2, which by

Corollary 2.5.4 would imply that ρ is a direct sum of χ_0 and χ_1, and so χ can be written as a direct sum of χ_0 and χ_1. Is this true?

Well, this is really the statement that there exist some positive integers a and b such that $\chi = a\chi_0 + b\chi_1$. Evaluating at the two points, this gives two equations:

$$n = a + b$$

and

$$m = a - b.$$

Now, what is the trace? Well, from the previous section we know that it is the sum of the eigenvalues. For $\rho(1)$, all of the eigenvalues are 1, and so $\mathrm{tr}(\rho(1)) = \chi(1) = n$. For $\rho(-1)$, the only thing we know is that $\rho(-1)^2 = \rho((-1)^2) = \rho(1) = I_n$. But that tells us something: let λ be an eigenvalue of $\rho(-1)$. Then there exists some vector v such that $\rho(-1)v = \lambda v$; and then on the one hand $\rho(-1)^2 v = I_n v = v$, while on the other hand

$$\rho(-1)^2 v = \rho(-1)(\rho(-1)v) = \lambda\rho(-1)v = \lambda^2 v.$$

Therefore $\lambda^2 v = v$, and so $\lambda^2 = 1$, so $\lambda = \pm 1$.

Therefore the trace $m = \chi(-1)$ is the sum of n values, each of which is either 1 or -1. Suppose that there are s copies of 1 and t copies of -1. Then $\chi(-1) = m = s - t$. But on the other hand the number of total copies $s + t$ is n, so we have

$$n = s + t$$

and

$$m = s - t.$$

These are precisely the equations that a and b are supposed to satisfy, so letting $a = s$ and $b = t$ gives a solution: $\chi = a\chi_0 + b\chi_1$.

Note, though, that this doesn't actually yet prove that ρ is a direct sum of copies of χ_0 and χ_1: we only know this for the trace, which we expect to contain less information, since it is only numbers rather than matrices. We'll come back to this point.

Let's try a more complicated, but still abelian, group: $(\mathbb{Z}/5\mathbb{Z})^\times$. This is isomorphic to the cyclic group of order 4, with identity 1 and generator 2: we have $2^2 = 4$, $2^3 = 8 \equiv 3 \pmod 5$, and $2^4 = 16 \equiv 1 \pmod 5$. Let $\chi : (\mathbb{Z}/5\mathbb{Z})^\times$ be a one-dimensional representation. Then $\chi(1) = 1$ and $\chi(2^4) = \chi(1) = 1$, so $\chi(2)^4 = 1$. Every other value of χ is determined by $\chi(2)$, since $\chi(4) = \chi(2^2) = \chi(2)^2$ and $\chi(3) = \chi(2^3) = \chi(2)^3$, so the one-dimensional characters correspond to choosing $\chi(2)$ any of the fourth roots of 1, i.e. $1, i, -i, -1$. Listing these gives

	1	2	3	4
χ_0	1	1	1	1
χ_1	1	i	$-i$	-1
χ_2	1	$-i$	i	-1
χ_3	1	-1	-1	1

.

Observe that this is precisely the table of Dirichlet characters modulo 5 from Section 1.9. Indeed, in general Dirichlet characters modulo q are just the one-dimensional characters of

$(\mathbb{Z}/q\mathbb{Z})^\times$. Using essentially the same argument as above, one can show that every character of $(\mathbb{Z}/5\mathbb{Z})^\times$ is a sum of these four one-dimensional characters, and thus that they are the only irreducible ones (up to the missing piece relating characters back to their representations): thus what this is really saying is that Dirichlet characters are the irreducible representations of $(\mathbb{Z}/q\mathbb{Z})^\times$.

One more example before we get to more abstract results: a nonabelian group, the symmetric group S_3. This group is generated by two elements x and y, with relations $x^2 = 1$, $y^3 = 1$, and $yx = xy^2$. This has six elements (1, x, y, y^2, xy, and xy^2) but only three conjugacy classes: 1 is its own conjugacy class, but we have $yxy^{-1} = xy^2y^{-1} = xy$ and $y^2xy^{-2} = yyxy^{-2} = yxy^2y^{-2} = yx = xy^2$, so x, xy, and xy^2 are conjugate; and $xyx^{-1} = xyx^{-1}x^2$, since $x^2 = 1$, and so $xyx^{-1} = xyx = xxy^2 = y^2$, so y and y^2 are conjugate. Thus the three conjugacy classes are 1, $[x] = \{x, xy, xy^2\}$, and $[y] = \{y, y^2\}$.

Let $\chi : S_3 \to \mathbb{C}^\times$ be a one-dimensional representation. Then $\chi(x)^2 = \chi(x^2) = \chi(1)$, so $\chi(x) = \pm 1$; and $\chi(y)^3 = \chi(y^3) = 1$. But by the above we have that y and y^2 are conjugate, and we know that χ is conjugacy-invariant; so $\chi(y) = \chi(y^2) = \chi(y)^2$. Since $\chi(y)$ must be nonzero, we can solve this to get $\chi(y) = 1$. This gives us our first two characters:

	1	$[x]$	$[y]$
χ_0	1	1	1
χ_1	1	-1	1

.

What about higher-dimensional representations? Rather than derive it rigorously, I'll just tell you that there is one more irreducible representation, which can be written as

$$\rho(1) = I_2, \qquad \rho(x) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \rho(y) = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}.$$

(We'll explain how we know this presently.) This completes our character table:

	1	$[x]$	$[y]$
χ_0	1	1	1
χ_1	1	-1	1
χ_2	2	0	-1

.

Looking at all of these tables, we should notice some things. First of all, they are square: there are as many irreducible representations as conjugacy classes. Second, the first two, corresponding to abelian groups, have a pleasing symmetry. In fact, they have even nicer properties: if we were to sum along any row or column except the first, we would get 0, while summing along the first row or column gives the order of the group.

These latter properties don't seem to hold for the nonabelian case S_3, but we can fix that. First, let's look at rows. Summing along the second or third row gives 1, not 0; but recall that the columns actually correspond to conjugacy classes, not elements, and if we were to write them out then the rows are better behaved:

	1	x	xy	xy^2	y	y^2
χ_0	1	1	1	1	1	1
χ_1	1	-1	-1	-1	1	1
χ_2	2	0	0	0	-1	-1

.

Now summing along any row except the first gives 0, while summing along the first row gives 6, the order of the group.

The columns are still problematic, though. How can we fix this? Well, χ_2 is different than the first two characters, since it comes from a two-dimensional representation. What if we were to count it twice?

	1	x	xy	xy^2	y	y^2
χ_0	1	1	1	1	1	1
χ_1	1	-1	-1	-1	1	1
χ_2	2	0	0	0	-1	-1
χ_2	2	0	0	0	-1	-1

.

Now the sum along any column except the first is 0, and along the first column gives 6, as desired.

The underlying results these are hinting at are called the orthogonality relations.

Theorem 2.5.12 (Orthogonality relations)**.** *Let G be a finite group, with n elements and m conjugacy classes.*

(1) There are exactly m finite-dimensional complex irreducible representations of G, whose characters form a basis for the space of class functions $G \to \mathbb{C}$.

(2) Let χ_1, χ_2 be irreducible characters of G. Then

$$\sum_{g \in G} \chi_1(g)\overline{\chi_2(g)}$$

is equal to 0 unless $\chi_1 = \chi_2$, in which case it is n.[98]

(3) Let g and h be elements of G. Then

$$\sum_{\chi} \chi(g)\overline{\chi(h)}$$

is equal to 0 unless g and h are conjugate, in which case it is n. Here the sum is taken over all irreducible characters χ of G.

I'm going to skip the proof for now because it's a pain. (Mess around with matrices and apply Schur's lemma.) I may come back and fill this in later.

The orthogonality part of this theorem, i.e. parts (2) and (3), is very useful when working with characters. However for applications part (1) is often the most useful: it allows us to write any class function on G as a linear combination of the irreducible characters χ.

Let's specialize to the case where G is abelian, of order n. Then there are n conjugacy classes, since $ghg^{-1} = gg^{-1}h = h$, i.e. h is the only element in its conjugacy classes; therefore by part (1) of Theorem 2.5.12 there are n irreducible characters. For each character χ,

[98] Here $\overline{\chi_2(g)}$ means the complex conjugate of $\chi_2(g)$.

write $\dim \chi$ for the dimension of the representation of which χ is a character; recall that $\dim \chi = \chi(1)$. Then choosing $g = h = 1$ in part (3) gives

$$\sum_{\chi} \chi(1)\overline{\chi(1)} = \sum_{\chi} (\dim \chi)^2 = n,$$

which is true for any group, abelian or not. In our case there are n characters, and so since $\dim \chi \geq 1$ if any of the χ have dimension greater than 1 then the left-hand side is greater than n, which is impossible since the right-hand side is n. Therefore we have proven the following.

Corollary 2.5.13. *Every irreducible character of a finite abelian group G has dimension 1.*[99]

In fact, we can say more. Recall that given two vector spaces U and V, there are various operations we can do with them: for example, we can take their direct sum $U \oplus V$, or their tensor product $U \otimes V$.

If U and V are representations of G, i.e. each carries a G-action, then we can do the same thing. We've already seen how the direct sum of representations works: that always gives us something reducible, which we're not very interested in. How about the tensor product of representations?

Well, for each $g \in G$, we get linear maps $g_U : U \to U$ and $g_V : V \to V$, since U and V are representations of G, taking $u \in U$ to $g \cdot u$ and $v \in V$ to $g \cdot v$. Then for $u \otimes v \in U \otimes V$ we define

$$g \cdot (u \otimes v) = (g \cdot u) \otimes (g \cdot v).$$

Since the action of g on each factor is linear, we can extend this to a linear map $U \otimes V \to U \otimes V$ by defining

$$g \cdot (u_1 \otimes v_1 + u_2 \otimes v_2) = g \cdot (u_1 \otimes v_1) + g \cdot (u_2 \otimes v_2) = (g \cdot u_1) \otimes (g \cdot v_1) + (g \cdot u_2) \otimes (g \cdot v_2).$$

You can check that for various $g \in G$ these are compatible, and so this action makes $U \otimes V$ into a G-representation.

How about its character? We need another fact from linear algebra we forgot to prove before: for two matrices A and B,

$$\operatorname{tr}(A \otimes B) = \operatorname{tr}(A) \cdot \operatorname{tr}(B),$$

analogous to $\operatorname{tr}(A \oplus B) = \operatorname{tr}(A) + \operatorname{tr}(B)$. We won't prove this fully here, but here's a sketch: recall that the trace is the sum of the eigenvalues, and then work out that if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A and $\mu_1, \ldots, \mu_n$ are the eigenvalues of B, then the eigenvalues of $A \otimes B$ are the products $\lambda_i \mu_j$ for $i, j \in \{1, 2, \ldots, n\}$, and summing all of these we get

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \mu_j = \left(\sum_{i=1}^{n} \lambda_i \right) \left(\sum_{j=1}^{n} \mu_j \right) = \operatorname{tr}(A) \cdot \operatorname{tr}(B).$$

[99] From now on when we're talking about finite abelian groups since we know that the only "interesting" characters are one-dimensional we'll assume that any characters we're talking about are one-dimensional, i.e. just say "character" instead of 'irreducible character" or "one-dimensional character." Representations may still be higher-dimensional, though they will not be irreducible for G finite abelian. (For more general G we'll keep the modifiers "irreducible" and "one-dimensional," since they can no longer be assumed.)

If we write the action of g on U and V above as matrices $\rho_U(g)$ and $\rho_V(g)$, then the action we described is that of $\rho_U(g) \otimes \rho_V(g)$, whose trace is therefore

$$\chi_{U\otimes V}(g) = \operatorname{tr}(\rho_U(g) \otimes \rho_V(g)) = \operatorname{tr}(\rho_U(g)) \cdot \operatorname{tr}(\rho_V(g)) = \chi_U(g) \cdot \chi_V(g).$$

In particular,

$$\dim U \otimes V = \dim \chi_{U\otimes V} = \chi_{U\otimes V}(1) = \chi_U(1) \cdot \chi_V(1) = \dim \chi_U \cdot \dim \chi_V = \dim U \cdot \dim V.$$

This is a useful way of building representations in general, as $U \otimes V$ may be irreducible, unlike $U \oplus V$. However, in our special case where G is abelian, it's even better: assume that U and V are irreducible, and therefore one-dimensional. Then $U \otimes V$ is also one-dimensional, and therefore also irreducible. That is: if we write $\widehat{G}$ for the set of irreducible characters of G, then the tensor product gives a binary operation on this set. In the one-dimensional case, we can think of this as just multiplication: $(\chi_1, \chi_2) \mapsto \chi_1 \cdot \chi_2$, which the above shows is the irreducible character of the tensor product of the corresponding representations. Since this operation is just multiplication of functions, it is associative and commutative.

In fact, this operation makes $\widehat{G}$ into an abelian group! We have an identity, given by the trivial character $\chi_0(g) = 1$ for every $g \in G$, since $(\chi_0 \cdot \chi)(g) = \chi_0(g) \cdot \chi(g) = 1 \cdot \chi(g) = \chi(g)$; and any character χ has an inverse given by its complex conjugate $\overline{\chi}$. To see why this is its inverse, observe that $|\chi(g)| = 1$ for any g, since $\chi(g)^n = \chi(g^n) = \chi(1) = 1$ where n is the order of the group G, and so $|\chi(g)|^n = |\chi(g)^n| = |1| = 1$; since $|\chi(g)|$ is a nonnegative real number, the only way for its nth power to be 1 is if $|\chi(g)| = 1$. But then $\chi(g)\overline{\chi(g)} = |\chi(g)|^2 = 1^2 = 1$ for every g, and so $\overline{\chi}$ is the inverse of χ in $\widehat{G}$.

This group of characters $\widehat{G}$ is called the Pontryagin dual group of G. We need the requirement that G is abelian for it to make sense, but the requirement that G is finite is not necessary; however, if we allow arbitrary abelian groups G then the dual group can be poorly behaved.

In the finite case, though,[100] it is very well-behaved indeed.

Theorem 2.5.14. *Let G be a finite abelian group. Then its Pontryagin dual $\widehat{G}$ is isomorphic to G.*

Proof. First, let's consider the special case where G is cyclic, say of order n. Let g be a generator of G, so that the elements of G are $\{1, g, g^2, \ldots, g^{n-1}\}$. What are the characters?

Well, we have the trivial character, $\chi_0(h) = 1$ for every $h \in G$. What else?

Let's think about our generator g. If χ is a character of G, then $\chi(g)^n = \chi(g^n) = \chi(1) = 1$, i.e. $\chi(g)$ is an nth root of unity. Fix some nth root of unity ζ, i.e. a complex number such that $\zeta^n = 1$. Then we can define a character by $\chi(g) = \zeta$, and then $\chi(g^k) = \chi(g)^k = \zeta^k$ for every k.

That's all well and good, but it's only one character. But we can generalize this: for every $j \in \{0, 1, \ldots, n-1\}$, write

$$\chi_j(g^k) = \zeta^{jk},$$

[100] And in fact more generally; this theorem holds for a class of groups called "locally compact abelian groups," of which finite abelian groups are a special case. We'll come back to this.

so that $j = 0$ gives the trivial character and $j = 1$ gives the character defined above. This is a group homomorphism $G \to \mathrm{GL}_1(\mathbb{C}) = \mathbb{C}^\times$, and so is a character. Since this defines n distinct characters, of which there are precisely n, all of the characters are of this form.

We've computed the set $\widehat{G}$; what does this look like under multiplication? Well, the identity is the trivial character; and then $\chi_j = \chi_1^j$, i.e. $\chi_j(g^k) = \zeta^{kj} = \chi_1(g^k)^j$. Therefore $\widehat{G}$ is the cyclic group of order n generated by χ_1, and so it is isomorphic to G under the map $\chi_j \mapsto g^j$.[101]

This proves the theorem for cyclic groups. What about more generally?

Suppose that our group G is the product of two groups A and B, i.e. $G \simeq A \times B$. What are the characters of G?

Well, suppose that A is of order a and B is of order B, so that $n = |G| = |A \times B| = |A| \cdot |B| = ab$. Then A has a characters $\chi_0^A, \chi_1^A, \dots, \chi_{a-1}^A$ and B has b characters $\chi_0^B, \chi_1^B, \dots, \chi_{b-1}^B$. Taking a pair (χ_i^A, χ_j^B) gives a character χ_{ij} of G: since $G \simeq A \times B$, we can think of elements of G as pairs (g, h) for $g \in A$ and $h \in B$, and then define

$$\chi_{ij}(g, h) = \chi_i^A(g) \cdot \chi_j^B(h).$$

This defines $|\widehat{A}| \cdot |\widehat{B}| = ab = n$ characters, and since there are exactly n characters of G all of them are of this form. Therefore the characters of G are pairs of characters of A and B, i.e. elements of $\widehat{A} \times \widehat{B}$; and their multiplication is given by $\chi_{ij}\chi_{i'j'} = \chi_{ii',jj'}$, i.e. $(\chi_i^A, \chi_j^B) \cdot (\chi_{i'}^A, \chi_{j'}^B) = (\chi_i^A \chi_{i'}^A, \chi_j^B \chi_{j'}^B)$, which is the same multiplication as in $\widehat{A} \times \widehat{B}$. Therefore $\widehat{G} \simeq \widehat{A} \times \widehat{B}$.

And in fact this is enough. Recall the classification of finite abelian groups (Theorem 2.2.18): every finite abelian group G is isomorphic to the direct product of cyclic groups of prime power order; that is, we can write $G \simeq C_1 \times C_2 \times \cdots \times C_r$ for cyclic groups $C_1, \dots, C_r$. By the above,

$$\widehat{G} \simeq (C_1 \times \widehat{\cdots \times C_{r-1}}) \times \widehat{C}_r \simeq (C_1 \times \widehat{\cdots \times C_{r-1}}) \times C_r.$$

Repeating the process for all the factors, we conclude

$$\widehat{G} \simeq C_1 \times \cdots \times C_r \simeq G,$$

which concludes the proof.[102] □

We have called this group a "dual"; is this justified? As in the case of vector spaces, we would like to be able to identify the double dual $\widehat{\widehat{G}}$ with G, so that if we keep taking duals there are only two "distinct" possibilities. We can do roughly the same thing as in that case: consider the map $G \to \widehat{\widehat{G}}$ given by sending g to the character $\psi_g : \widehat{G} \to \mathbb{C}^\times$ of $\widehat{G}$ defined by $\psi_g(\chi) = \chi(g)$. This is a character, since $\psi_g(\chi_1\chi_2) = \chi_1(g)\chi_2(g) = \psi_g(\chi_1)\psi_g(\chi_2)$; and it is distinct for every g, so there are $|G|$ possible such characters, and since there are $|\widehat{G}| = |G|$ characters of $\widehat{G}$ total all of them are of this form. Further this map $g \mapsto \psi_g$ is itself a group homomorphism: ψ_{gh} is the character sending $\chi \mapsto \chi(gh) = \chi(g)\chi(h) = \psi_g(\chi)\psi_h(\chi)$, so $\psi_{gh} = \psi_g\psi_h$. Therefore this is a group isomorphism, and so we have proved the following.

[101] Observe that this is not a *canonical* isomorphism: we had to pick an nth root of unity ζ, and there is no canonical way of doing this. Picking a different value would reorder the χ_j and therefore give a different isomorphism.

[102] Again, this isomorphism is not canonical, since none of the isomorphisms for the cyclic factors are.

Proposition 2.5.15. *Let G be a finite abelian group. Then there is a* canonical *isomorphism $G \to \widehat{\widehat{G}}$.*[103]

As mentioned, this is very similar to the theory of vector space duals. Indeed, we can make this explicit. Consider the vector space V generated by elements of G; that is, the space of elements of the form

$$a_1 g_1 + \cdots + a_n g_n$$

for g_i elements of g and $a_i \in \mathbb{C}$.[104] Thus G is a basis for V. The dual space $V^\vee = \mathrm{Hom}(V, \mathbb{C})$ of linear functions on V then has a natural basis consisting of the characters of G, i.e. $\widehat{G}$ is a basis for $V^\vee$. Taking duals again, $\widehat{\widehat{V}}$ is identified with V, and its natural basis is again G.

We've said quite a bit about the characters of a finite group G, but not all that much about its representations. Fortunately, this is just as good.

Theorem 2.5.16. *Every complex finite-dimensional representation of a finite group G is determined by its character.*

Thus classifying the characters automatically gives us a classification of the representations.

Proof. First, we show that it suffices to prove the theorem for irreducible representations. Suppose that we know that all irreducible representations are determined by their characters, and let ρ be an arbitrary finite-dimensional complex representation of G. By Corollary 2.5.4, there exist irreducible representations $\rho_1, \ldots, \rho_k$ such that

$$\rho = \rho_1 \oplus \cdots \oplus \rho_k,$$

so that taking traces we get

$$\mathrm{tr}\, \rho = \mathrm{tr}\, \rho_1 + \cdots + \mathrm{tr}\, \rho_k.$$

Writing $\chi = \mathrm{tr}\, \rho$ and $\chi_i = \mathrm{tr}\, \rho_i$ for the characters, this is $\chi = \chi_1 + \cdots + \chi_k$. We want to show that given χ we can recover ρ, or in other words if ρ' is another finite-dimensional complex representation of G with $\mathrm{tr}\, \rho' = \chi$ then ρ and ρ' must be isomorphic. But since all of the χ_i are irreducible by assumption we can recover the representations, i.e. for any ρ_i' such that $\mathrm{tr}\, \rho_i' = \chi_i$ we have $\rho_i \simeq \rho_i'$, and so we conclude that since χ splits into the sum of the χ_i we must have

$$\rho' = \rho_1' \oplus \cdots \oplus \rho_k' \simeq \rho_1 \oplus \cdots \oplus \rho_k = \rho$$

and so ρ' and ρ are isomorphic as desired.

It remains to prove the theorem for irreducible representations. Suppose that G has m conjugacy classes, so that the space of class functions on G is m-dimensional, since the set of functions $f_c(x)$ for each conjugacy class c of g, defined to be 1 for $x \in c$ and 0 otherwise, is a basis for the space of class functions. Then since by Theorem 2.5.12 there

[103] This isomorphism, unlike the case in the regular dual case, is canonical, since we don't have to make an arbitrary choice of a root of unity or anything else.

[104] These are *formal* linear combinations; it doesn't really mean anything to multiply an element of G by a complex number, but what this really means is that we're assigning to each $g \in G$ a vector v_g such that the v_g give a basis for V.

are m finite-dimensional complex irreducible representations and their characters are a basis for this space, there must also be exactly m distinct characters, i.e. all of the irreducible representations have different characters. Therefore each one is determined by its character. By the above, it follows that all representations are determined by their characters. $\square$

Bibliography

[1] Michael Artin. *Algebra*. Pearson Education, second edition, 2011.

[2] Adolf J Hildebrand. Introduction to Analytic Number Theory Math 531 Lecture Notes, Fall 2005. *URL:* *https://faculty.math.illinois.edu/~hildebr/ant/main.pdf*, 2006.

[3] Andrew V Sutherland. 18.785: Number Theory I lecture notes. *URL:* *https://math.mit.edu/classes/18.785/2019fa/lectures.html*, 2019.