April 30, 2023

An Introduction to Number Theory

J. J. P. Veerman

Preface

This work contains a one year (three terms or two semesters) first course in number theory at the advanced undergraduate to intermediate graduate level. No prior knowledge of number theory is assumed. We do require knowledge and familiarity with the notions of basic linear algebra (matrices, determinants), calculus (integrals), and proof-writing. But beyond that, this text aims to be completely self-contained. Part I covers what students might learn in an advanced undergraduate first course on number theory. Part II discusses the basics of the different branches of number theory (algebraic, analytic, and ergodic or probabilistic) and is pitched at a slightly higher level. Part III discusses some more advanced topics.

Too often, in our opinion, it seems that introductory texts skimp on either the algebraic aspects or the geometric aspects of number theory (sometimes both). We have tried to cover all areas of introductory number theory. In contrast, algebraic texts are often very hard to read for the non-algebraically inclined students due to the barrage of definitions one needs to digest to get to the "meat". Our approach is pragmatic: we want to understand a broad outline of number theory, so how do we do that most efficiently? For example, when dealing with continued fractions, a geometrical approach greatly clarifies the subject by making it easy to visualize results, and so we adopt that approach. In the study of algebraic integers, we skip or summarize many of the commonly used definitions. For instance, we do not dwell very long on the fact that the integers form a ring. We merely note that the crucial part is the *distinction* between ring (where multiplication does not have an inverse) and field (where it does). This — together with some other observations — is sufficient to arrive in very few pages at accurate definitions of, say, Euclidean domains. Our approach, then, is to try to uncover number theory by any means, regardless of whether the subject at hand is traditionally seen as part of algebra, analysis, or geometry.

Each chapter consists of some basic material followed by exercises. The exercises are of two types. Some are simple computational routines, to get the students used to the notions described in the text. Others develop the theory a little further. All exercises, including the ones of the latter type, should be easily doable for graduate students. Nonetheless, my recommendation is to have the students work out all the exercises in groups to lessen even more the chance of getting stuck. Nothing discourages self-study more than being stuck in some exercise. If students tend to get stuck in some particular exercise, I hope I will be notified, so I can add more hints in the next edition.

You can't really learn mathematics without doing it. But on the other hand, you can't possible invent everything from scratch either. So a good way to design a course, is to give the lay of the land in each chapter, and let students figure out some of the details and corollaries through exercises at the end of the chapter. I wrote these notes aiming for a division of roughly equal parts between blackboard type lecturing and interactive practice for the student of about half-half. This has the additional advantage that the basic and most important material is concise and contained in few pages, and therefore can be consulted with great ease to find important results.

A word on alternate uses of this work. Because of the crossdisciplinary (within mathematics) approach taken, parts of the book can easily be used in other courses. For example, a graduate course on dynamical systems or ergodic theory could contain an 8 week segment consisting of Chapters 6, 9, 10, and 14. Chapters 4, 2, 11, 12, and 13 could easily constitute a short analytic number theory course. Chapters 5, 7, and 8 could start off a course in algebraic number theory.

This work *can* be used at the intermediate undergraduate level, where students still need to train how to write proofs. In that case, I recommend a modest program consisting roughly of the following content. Start with Chapters 1 and 2, followed by Sections 3.1 through 3.4 and 3.6, 4.1 and

4.2, and finally Chapter 5. If this is the course taken, then the instructor would probably want to gather some extra material on how to write proofs, starting with mathematical induction. Care should be taken that some of the exercises may be a little too advanced for the students at this level.

We advise the reader that underlined words and clauses are indexed.

J. J. P. Veerman
Portland, Oregon
September, 2022.

List of Figures

Contents

Part 1

Introduction to Number Theory

Chapter 1

A Quick Tour of Number Theory

Overview. We give definitions of the following concepts of congruence and divisor in the integers, of rational and irrational number, and of countable versus uncountable sets. We also discuss some of the elementary properties of these notions.

Before we start, a general comment about the structure of this book may be helpful. Each chapter consists of a "bare bones" outline of a piece of the theory followed by a number of exercises. These exercises are meant to achieve two goals. The first is to get the student used to the mechanical or computational aspects of the theory. For example, the division algorithm in Chapter 2 comes back numerous times in slightly different guises. In Chapter 3, we use solve equations of the type $ax+by=c$ for given a, b, and c, and in Chapter 6, we take that even further to study continued fractions. To recognize and understand the use of the algorithm in these different contexts, it is therefore crucial that the student sufficient practice with elementary examples. Thus, even if the algorithm is "more or less" clear or familiar, a wise student will carefully do all the computational problems in order for it to become "thoroughly" familiar. The second goal of the exercises is to extend the bare bones theory, and fill in some details covered in most textbooks. For instance, in this Chapter we explain what rational and irrational numbers are. However, the proof that the number e is irrational is left to the exercises. In summary, as a rule the student should spend at least as much time on the exercises as on the theory.

The natural numbers starting with 1 are denoted by $\mathbb{N}$, and the collection of all integers (positive, negative, and 0) by $\mathbb{Z}$. Elements of $\mathbb{Z}$ are also called *integers* .

1.1. Divisors and Congruences

Definition 1.1. *Given two numbers a and b. A multiple b of a is a number that satisfies $b = ac$. A divisor a of b is an integer that satisfies $ac = b$ where c is an integer. We write $a \mid b$. This reads as a* divides *b or a* is a divisor of *b.*

Definition 1.2. *Let a and b non-zero. The greatest common divisor of two integers a and b is the maximum of the numbers that are divisors of both a and b. It is denoted by $\gcd(a,b)$. The least common multiple of a and b is the least of the positive numbers that are multiples of both a and b. It is denoted by $\operatorname{lcm}(a,b)$.*

Note that for any a and b in $\mathbb{Z}$, $\gcd(a,b) \geq 1$, as 1 is a divisor of every integer. Similarly $\operatorname{lcm}(a,b) \leq |ab|$.

Definition 1.3. *A number $p > 1$ is prime*[1] *in $\mathbb{N}$ if its only divisors in $\mathbb{N}$ are a and 1 (the so-called trivial divisors). A number $a > 1$ is composite or reducible if it has more than 2 divisors in $\mathbb{N}$. (The number 1 is neither.)*

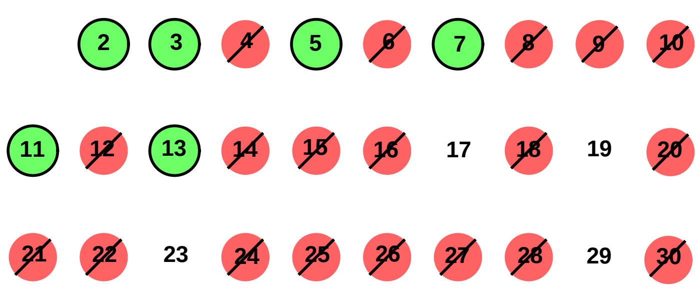


Figure 1. Eratosthenes' sieve up to $n = 30$. All multiples of a less than $\sqrt{31}$ are cancelled. The remainder are the primes less than $n = 31$.

[1] In a more general context — see Chapter 8 — these are called *irreducible numbers*, while the term prime is reserved for numbers satisfying Corollary 2.9.

An equivalent definition of prime is a natural number with precisely two (distinct) divisors. *Eratosthenes' sieve* is a simple and ancient method to generate a list of primes for all numbers less than, say, 225. First, list all integers from 2 to 225. Start by circling the number 2 and crossing out all its remaining multiples: 4, 6, 8, etcetera. At each step, circle the smallest unmarked number and cross out all its remaining multiples in the list. It turns out that we need to sieve out only multiples of $\sqrt{225} = 15$ and less (see exercise 2.5). This method is illustrated if Figure 1. When done, the primes are those numbers that are circled or unmarked in the list.

It will turn out that it is more natural to work in $\mathbb{Z}$ where all elements have an additive inverse. We therefore introduce extend the definition of primes to $\mathbb{Z}$ and introduce units.

Definition 1.4. *A (multiplicative) unit in $\mathbb{Z}$ is a number a such that there is $b \in \mathbb{Z}$ with the property that $ab = 1$. The only units in $\mathbb{Z}$ are 1 and -1. All other numbers are non-units. A number $n \neq 0$ in $\mathbb{Z}$ is called composite or reducible if it can be written as a product of two non-units. If n is not 0, not a unit, and not composite, it is a prime or irreducible .*

Remark 1.5. A concise way to characterize a unit is saying that it is an *invertible* element.

Definition 1.6. *Let a and b in $\mathbb{Z}$. Then a and b are relatively prime if* $\gcd(a,b) = 1$.

Definition 1.7. *Let a and b in $\mathbb{Z}$ and $m \in \mathbb{N}$. Then a is congruent to b modulo m if $a + my = b$ for some $y \in \mathbb{Z}$ or $m \mid (b-a)$. We write*

$$a =_m b \quad \text{or} \quad a = b \mod m \quad \text{or} \quad a \in b + m\mathbb{Z}\,.$$

Definition 1.8. *The residue of a modulo m is the (unique) integer r in $\{0, \cdots m-1\}$ such that $a =_m r$. It is denoted by* $\mathrm{Res}_m(a)$.

These notions are cornerstones of much of number theory as we will see. But they are also very common in all kinds of applications. For instance, our expressions for the time on the clock are nothing but counting modulo 12 or 24. To figure out how many hours elapse between 4pm and 3am next morning is a simple exercise in working with modular arithmetic, that is: computations involving congruences.

1.2. Rational and Irrational Numbers

We start with a few results we need in the remainder of this subsection.

Theorem 1.9 (**well-ordering principle**). *Any non-empty set S in $\mathbb{N} \cup \{0\}$ or in $\mathbb{N}$ has a smallest element.*

Proof. Suppose this is false. Pick $s_1 \in S$. Then there is another natural number s_2 in S such that $s_2 \leq s_1 - 1$. After a finite number of steps, we pass zero, implying that S has elements less than 0 in it. This is a contradiction and so the supposition is false. ■

Note that any non-empty set S of integers with a lower bound can be transformed by addition of a integer $b \in \mathbb{N}_0$ into a non-empty $S + b$ in $\mathbb{N}_0$. Then $S + b$ has a smallest element, and therefore so does S. Furthermore, a non-empty set S of integers with a upper bound can also be transformed into a non-empty $-S + b$ in $\mathbb{N}_0$. Here, $-S$ stands for the collection of elements of S multiplied by -1. Thus we have the following corollary of the well-ordering principle.

Corollary 1.10. *Let be a non-empty set S in $\mathbb{Z}$ with a lower (upper) bound. Then S has a smallest (largest) element.*

Definition 1.11. *i) An element $x \in \mathbb{R}$ is called rational if it a root of a degree 1 polynomial with integer coefficients, that is: $qx - p = 0$ where p and $q \neq 0$ are integers.*
ii) Otherwise it is called an irrational number.

Remark 1.12. Note that the integers themselves are also considered to be rational numbers: they satisfy part (i) above with the leading coefficient, q, equal to 1.

The set of integers is denoted by $\mathbb{Z}$, and the rational numbers are denoted by $\mathbb{Q}$. The usual way of expressing a rational number is that it can be written as $\frac{p}{q}$. The advantage of expressing a rational number as the solution of a degree 1 polynomial, however, is that it naturally paves the way to Definitions 1.16 and 1.17.

Theorem 1.13. *Any interval in $\mathbb{R}$ contains an element of $\mathbb{Q}$. We say that $\mathbb{Q}$ is dense in $\mathbb{R}$.*

The crux of the following proof is that we take an interval and scale it up until we know there is an integer in it, and then scale it back down.

Proof. Let $I = (a,b)$ with $b > a$ any interval in $\mathbb{R}$. From Corollary 1.10 we see that there is an n such that $n > \frac{1}{b-a}$. Indeed, if that weren't the case, then $\mathbb{N}$ would be bounded from above, and thus it would have a largest element n_0. But if $n_0 \in \mathbb{N}$, then so is $n_0 + 1$. This gives a contradiction and so the above inequality must hold.

It follows that $nb - na > 1$. Thus the interval (na, nb) contains an integer, say, p. So we have that $na < p < nb$. The theorem follows upon dividing by n. ■

Theorem 1.14. *$\sqrt{2}$ is irrational.*

Proof. Suppose $\sqrt{2}$ can be expressed as the quotient of integers $\frac{r}{s}$. We may assume that $\gcd(r,s) = 1$ (otherwise just divide out the common factor). After squaring, we get

$$2s^2 = r^2 \ .$$

The right-hand side is even, therefore the left-hand side is even. But the square of an odd number is odd, so r is even. But then r^2 is a multiple of 4. Thus s must be even. This contradicts the assumption that $\gcd(r,s) = 1$. ■

It is pretty clear who the rational numbers are. But who or where are the others? We just saw that $\sqrt{2}$ is irrational. It is not hard to see that the sum of any rational number plus $\sqrt{2}$ is also irrational. Or that any rational non-zero multiple of $\sqrt{2}$ is irrational. The same holds for $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, etcetera. We look at this in exercise 1.7. From there, is it not hard to see that the irrational numbers are also dense (exercise 1.7). In exercise 1.13, we prove that the number e is irrational. The proof that π is irrational is a little harder and can be found in [**30**][section 11.17]. In Chapter 2, we will use the fundamental theorem of arithmetic, Theorem 2.11, to construct other irrational numbers. In conclusion, whereas rationality is seen at face value, irrationality of a number may take some effort to prove, even though they are much more numerous as we will see in Section 1.4.

If you think about it, we cannot express the exact numerical value of an irrational number! The only way to do that would be in a decimal (or any other base) expansion. But if such an expansion were finite, of course, the number would be rational! Thus the question of how well we can approximate irrational numbers by rational ones arises (see exercise 1.16). Here is an important general result which we will have occasion to prove in Chapter 6.

Theorem 1.15. *Let $\rho \in \mathbb{R}$ be irrational. Then there are infinitely many $\frac{p}{q} \in \mathbb{Q}$ such that $\left|\rho - \frac{p}{q}\right| < \frac{1}{q^2}$.*

1.3. Algebraic and Transcendental Numbers

The set of polynomials with coefficients in $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ is denoted by $\mathbb{Z}[x]$, $\mathbb{Q}[x]$, $\mathbb{R}[x]$, and $\mathbb{C}[x]$, respectively.

Definition 1.16. *An element $x \in \mathbb{C}$ is called an algebraic integer if it satisfies $p(x) = 0$, where p is a non-zero polynomial in $\mathbb{Z}[x]$ with leading coefficient 1.*

Definition 1.17. *An element $x \in \mathbb{C}$ is called an algebraic number if it satisfies $p(x) = 0$, where p is a non-zero polynomial in $\mathbb{Z}[x]$. Otherwise it is called a transcendental number.*

The transcendental numbers are even harder to pin down than the general irrational numbers. We do know that e and π are transcendental, but the proofs are considerably more difficult (see [**32**]). We'll see below that the transcendental numbers are far more abundant than the rationals or the algebraic numbers. In spite of this, they are harder to analyze and, in fact, even hard to find. This paradoxical situation where the most prevalent numbers are hardest to find, is actually pretty common in number theory.

The most accessible tool to construct transcendental numbers is Liouville's Theorem. The setting is the following. Given an algebraic number y, it is the root of a polynomial with integer coefficients $f(x) = \sum_{i=0}^{d} a_i x^i$, where we always assume that the coefficient a_d of the highest power is non-zero. That highest power is called the degree of the polynomial and is denoted by $\deg(f)$. Note that we can always find a polynomial of higher degree that has y as a root. Namely, multiply f by any other polynomial g.

Definition 1.18. *We say that $f(x) = \sum_{i=0}^{d} a_i x^i$ in $\mathbb{Z}[x]$ is a minimal polynomial in $\mathbb{Z}[x]$ for ρ if f is a non-zero polynomial in $\mathbb{Z}[x]$ of minimal degree, say d, such that $f(\rho) = 0$. We say that the degree of ρ is d.*

Theorem 1.19 (Liouville's Theorem). *Let f be a minimal polynomial of degree $d \geq 2$ for $r \in \mathbb{R}$. Then*

$$\exists\, c(r) > 0 \ \text{ such that } \forall\, \frac{p}{q} \in \mathbb{Q} : \left| r - \frac{p}{q} \right| > \frac{c(r)}{q^d} .$$

Proof. Clearly, if $\left| r - \frac{p}{q} \right| \geq 1$, the inequality is satisfied. So assume that $\left| r - \frac{p}{q} \right| < 1$. Now let f be a minimal polynomial for r (see Figure 2), and set

$$K = \max_{t \in [r-1, r+1]} \left| f'(t) \right| .$$

We know that $f(p/q)$ is not zero, because otherwise f would have a factor $(x - p/q)$. In that case, the quotient g of f and $(x - p/q)$ would not necessarily have integer coefficients, but some integral multiple mg of g would. However, mg would be of lower degree, thus contradicting the minimality of f. This gives us that $q^d f(p/q)$ is an integer, because

$$\left| q^d f\left(\frac{p}{q}\right) \right| = \left| \sum_{i=0}^{d} a_i p^i q^{d-i} \right| \geq 1 ,$$

because it is a non-zero integer. Thus $|f(p/q)| \geq q^{-d}$. Finally, we use the mean value theorem which tells us that for K as above, there is a t between r and $\frac{p}{q}$ such that

$$K \geq \left| f'(t) \right| = \left| \frac{f\left(\frac{p}{q}\right) - f(r)}{\frac{p}{q} - r} \right| \geq \frac{q^{-d}}{\left| \frac{p}{q} - r \right|} ,$$

since $f(r) = 0$. This gives us the desired inequality. ■

Definition 1.20. *A real number ρ is called a* <u>*Liouville*</u> <u>*number*</u> *if for all $n \in \mathbb{N}$, there is a rational number $\frac{p}{q}$ such that*

$$\left| \rho - \frac{p}{q} \right| < \frac{1}{q^n} .$$

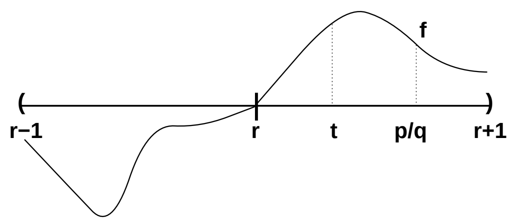


Figure 2. f is a *minimal* polynomial for the irrational number r. By minimality $f'(p/q)$ is not zero. On the interval $(r-1,r)$, the absolute value of the derivative of f attains its maximum at t.

It follows directly from Liouville's theorem that such numbers must be transcendental. Liouville numbers can be constructed fairly easily. The number

$$\rho = \sum_{k=1}^{\infty} 10^{-k!}$$

is an example. If we set $\frac{p}{q}$ equal to $\sum_{k=1}^{n} 10^{-k!}$, then $q = 10^{n!}$. Then

$$\left|\rho - \frac{p}{q}\right| = \sum_{k=n+1}^{\infty} 10^{-k!}. \tag{1.1}$$

It is easy to show that this is less than q^{-n} (exercise 1.15).

It is worth noting that there is an optimal version of Liouville's Theorem. We record it here without proof.

Theorem 1.21 (**Roth's Theorem**)**.** *Let $\rho \in \mathbb{R}$ be algebraic. Then for all $\varepsilon > 0$*

$$\exists\, c(\rho,\varepsilon) > 0 \ \text{ such that } \ \forall\, \frac{p}{q} \in \mathbb{Q} \,:\, \left|\rho - \frac{p}{q}\right| > \frac{c(\rho,\varepsilon)}{q^{2+\varepsilon}},$$

where $c(\rho,\varepsilon)$ depends only on ρ and ε.

This result is all the more remarkable if we consider it in the context of Theorem 1.15.

1.4. Countable and Uncountable Sets

Definition 1.22. *i) A set S is finite if there is a bijection $f : \{1,\cdots,n\} \to S$ for some $n > 0$.*
ii) A set S is countably infinite if there is a bijection $f : \mathbb{N} \to S$.

iii) A set S is countable if it is finite or if it is countably infinite. iv) An infinite set for which there is no bijection as in (ii) is called uncountable.

Proposition 1.23. *Every infinite set S contains a countable subset.*

Proof. Choose an element s_1 from S. Now $S-\{s_1\}$ is not empty because S is not finite. So, choose s_2 from $S-\{s_1\}$. Then $S-\{s_1,s_2\}$ is not empty because S is not finite. In this way, we can remove s_{n+1} from $S-\{s_1,s_2,\cdots s_n\}$ for all n. The set $\{s_1,s_2,\cdots\}$ is countable and is contained in S. ∎

So countable sets are the *smallest* infinite sets in the sense that there are no infinite sets that contain no countable set. But there certainly are larger sets, as we will see next.

Theorem 1.24. *The set* $\mathbb{R}$ *is uncountable.*

Proof. The proof is one of mathematics' most famous arguments: Cantor's diagonal argument [**19**]. The argument is developed in two steps .

Let T be the set of semi-infinite sequences formed by the digits 0 and 2. An element $t\in T$ has the form $t=t_1t_2t_3\cdots$ where $t_i\in\{0,2\}$. The first step of the proof is to prove that T is uncountable. So suppose it is *countable*. Then a bijection t between $\mathbb{N}$ and T allows us to uniquely define the sequence $t(n)$, the unique sequence associated to n. Furthermore, they form an exhaustive list of the elements of T. For example,

$$\begin{array}{rcl}
t(1) & = & \mathbf{0},0,0,0,0,0,0,0,0,0,0\cdots \\
t(2) & = & 2,\mathbf{0},2,0,2,0,2,0,2,2,2\cdots \\
t(3) & = & 0,0,\mathbf{0},2,2,2,2,2,2,2,2\cdots \\
t(4) & = & 2,2,2,\mathbf{2},2,2,0,0,0,0,0\cdots \\
t(5) & = & 0,0,0,2,\mathbf{0},0,2,0,0,2,0\cdots \\
t(6) & = & 2,0,0,0,0,\mathbf{2},0,0,0,2,2\cdots \\
\vdots & \vdots & \vdots
\end{array}$$

Construct t^* as follows: for every n, its nth digit differs from the nth digit of $t(n)$. In the above example, $t^*=\mathbf{2},\mathbf{2},\mathbf{2},\mathbf{0},\mathbf{2},\mathbf{0},\cdots$. But now we have a contradiction, because the element t^* cannot occur in the list. In other words, there is no *surjection* from $\mathbb{N}$ to T. Hence there is no bijection between $\mathbb{N}$ and T.

The second step is to show that there is a subset K of $\mathbb{R}$ such that there is no surjection (and thus no bijection) from $\mathbb{N}$ to K. Let t be a sequence with digits t_i. Define $f : T \to \mathbb{R}$ as follows

$$f(t) = \sum_{i=1}^{\infty} t_i 3^{-i} \,.$$

If s and t are two distinct sequences in T, then for some k they share the first $k-1$ digits but $t_k = 2$ and $s_k = 0$. So

$$f(t) - f(s) = 2 \cdot 3^{-k} + \sum_{i=k+1}^{\infty} (t_i - s_i) 3^{-i} \geq 2 \cdot 3^{-k} - 2 \sum_{i=k+1}^{\infty} 3^{-i} = 3^{-k} \,.$$

Thus f is injective. Therefore f is a bijection between T and the subset $K = f(T)$ of $\mathbb{R}$. If there is a surjection g from $\mathbb{N}$ to $K = f(T)$, then,

$$\mathbb{N} \xrightarrow{g} K \xleftarrow{f} T \,.$$

And so $f^{-1}g$ is a surjection from $\mathbb{N}$ to T. By the first step, this is impossible. Therefore, there is no surjection g from $\mathbb{N}$ to K, much less from $\mathbb{N}$ to $\mathbb{R}$. ■

The crucial part here is the diagonal step, where an element is constructed that *cannot* be in the list. This really means the set T is *strictly larger* than $\mathbb{N}$. The rest of the proof seems an afterthought, and perhaps needlessly complicated. You might think that it is much more straightforward to just use the digits 0 and 1 and the representation of the real numbers on the base 2, as opposed to the digits 0 and 2 and the base 3. But if you do that, you run into a problem that has to be dealt with. The sequence t^* might end with an infinite all-ones subsequence such as $t^* = 1, 1, 1, 1, \cdots$. This corresponds to the real number $x = 1.0...$ which *might* be in the list. To circumvent that problem leads to slightly more complicated proofs (see exercise 1.9).

Meanwhile, this gives us a very nice corollary which we will have occasion to use in later chapters. For b an integer greater than 1, denote by $\{0, 1, 2, \cdots b-1\}^{\mathbb{N}}$ the set of sequences $a_1 a_2 a_3 \cdots$ where each a_i is in $\{0, 1, 2, \cdots b-1\}$. Such sequences are often called words.

Corollary 1.25. *(i) The set of infinite sequences in* $\{0, 1, 2, \cdots b-1\}^{\mathbb{N}}$ *is uncountable. (ii) The set of finite sequences (but without bound) in* $\{0, 1, 2, \cdots b-1\}^{\mathbb{N}}$ *is countable.*

Proof. The proof of (i) is the same as the proof that T is uncountable in the proof of Theorem 1.24. The proof of (ii) consists of writing first all b words of length 1, then all b^2 words of length 2, and so forth. Every finite string will occur in the list. ∎

Theorem 1.26. *(i) The set $\mathbb{Z}^2$ is countable. (ii) $\mathbb{Q}$ is countable.*

Proof. (i) The proof relies on Figure 3. In it, a directed path γ is traced out that passes through all points of $\mathbb{Z}^2$. Imagine that you start at $(0,0)$ and travel along γ with unit speed. Keep a counter $c \in \mathbb{N}$ that marks the point $(0,0)$ with a "1". Up the value of the counter by 1 whenever you hit a point of $\mathbb{Z}^2$. This establishes a bijection between $\mathbb{N}$ and $\mathbb{Z}^2$.

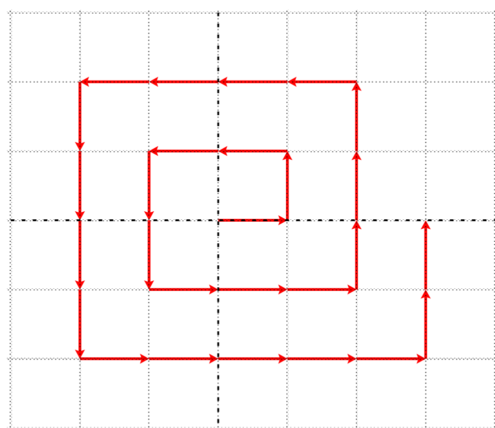


Figure 3. A directed path γ passing through all points of $\mathbb{Z}^2$.

(ii) Again travel along γ with unit speed. Keep a counter $c \in \mathbb{N}$ that marks the point $(0,1)$ with a "1". Up the value of the counter by 1. Continue to travel along the path until you hit the next point (p,q) that is not a multiple of any previous and such q is not zero. Mark that point with the value of the counter. $\mathbb{Q}$ contains $\mathbb{N}$ and so is infinite. Identifying each marked point (p,q) with the rational number $\frac{p}{q}$ establishes the countability of $\mathbb{Q}$. ∎

Notice that this argument really tells us that the product $(\mathbb{Z} \times \mathbb{Z})$of a countable set $(\mathbb{Z})$ and another countable set is still countable. The same holds for any finite product of countable set. Since an uncountable set is strictly larger than a countable, intuitively this means that an uncountable set must be *a lot larger* than a countable set. In fact, an extension of the above argument shows that the set of algebraic numbers numbers is countable (see exercises 1.8 and 1.24). And thus, in a sense, it forms small subset of all reals. All the more remarkable, that almost all reals *that we know anything about* are algebraic numbers, a situation we referred to at the end of Section 1.4.

It is useful and important to have a more general definition of when two sets "have the same number of elements".

Definition 1.27. *Two sets A and B are said to have the same* <u>*cardinality*</u> *if there is a* bijection $f : A \to B$*. It is written as* $|A| = |B|$*. If there is an* injection $f : A \to B$*, then* $|A| \leq |B|$*.*

Definition 1.28. *An* <u>*equivalence relation*</u> *on a set A is a (sub)set R of ordered pairs in* $A \times A$ *that satisfy three requirements.*
- $(a,a) \in R$ *(reflexivity).*
- If $(a,b) \in R$*, then* $(b,a) \in R$ *(symmetry).*
- If $(a,b) \in R$ *and* $(b,c) \in R$*, then If* $(a,c) \in R$ *(transitivity).*
Usually $(a,b) \in R$ *is abbreviated to* $a \sim b$*. The mathematical symbol "=" is an equivalence.*

It is easy to show that having the same cardinality is an equivalence relation on sets (exercise 1.22). Note that the *cardinality of a finite set* is just the number of elements it contains. An excellent introduction to the cardinality of infinite sets in the context of naive set theory can be found in [**36**].

1.5. Exercises

Exercise 1.1. Apply Eratosthenes' Sieve to get all prime numbers between 1 and 200. *(Hint: you should get 25 primes less than 100, and 21 between 100 and 200.)*

Exercise 1.2. Factor the following into prime numbers (write as a product of primes): 393, 16000, 5041, 1111, 1763, 720.

Exercise 1.3. Find pairs of primes that differ by 2. These are called *twin primes*. Do you think there there infinitely many such pairs? See Figure 4.

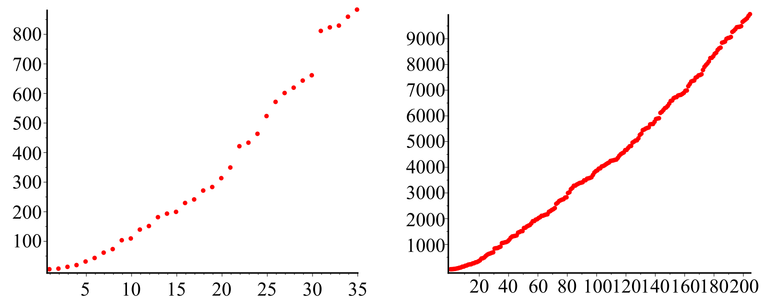


Figure 4. Lowest member of nth twin prime pair less than 1000 (left) and less than 10000 (right), horizontal axis is n.

Conjecture 1.29 (**Twin Prime Conjecture**). *There are infinitely many twin prime pairs*[2].

Exercise 1.4. Show that all even integers greater than 3 but smaller than 21 can be written as the sum of two primes. Is this always true?

Conjecture 1.30 (**Goldbach Conjecture**). *Every even number greater than two is a sum of two (positive) primes*[3].

Exercise 1.5. Comment on the types of numbers (rational, irrational, transcendental) we use in daily life.
a) What numbers do we use to pay our bills?
b) What numbers do we use in computer simulations of complex processes?
c) What numbers do we use to measure physical things?
d) Give examples of the usage of the "other" numbers.

Exercise 1.6. Let a and b be rationals and x and y irrationals.
a) Show that ax is irrational if and only if $a \neq 0$.
b) Show that $b+x$ is irrational.
c) Show that $ax+b$ is irrational if and only if $a \neq 0$.
d) Conclude that $a\sqrt{2}+b$ is irrational if and only if $a \neq 0$.

Exercise 1.7. a) Show that $\sqrt{3}$, $\sqrt{5}$, et cetera (square roots of primes) are irrational. (*Hint: use Corollary 2.9.*)
b) Show that for p prime, the numbers $\{a\sqrt{p}+b \,:\, a,b \in \mathbb{Z}\}$ are dense in the reals.

[2] Still unsolved in 2022.
[3] Still unsolved in 2022.

Lemma 1.31. *The countable union of countable sets is countable.*

Exercise 1.8. a) Use an pictorial argument similar to that of Figure 3 to show that $\mathbb{N} \times \mathbb{N}$ (the set of lattice points (n,m) with n and m in $\mathbb{N}$) is countable.
b) Suppose A_i are countable sets where $i \in I$ and I countable. Show that there is a bijection $\{1,\cdots,n\} \to I$ or $\mathbb{N} \to I$.
c) Define $A'_1 = A_1$, $A'_2 = A_2 \backslash A'_1$, $A'_3 = A_3 \backslash \{A'_1 \cup A'_2\}$, et cetera. Show that there is a bijection $f_i : \{1,\cdots,n_i\} \to A_i$ or $f_i : \mathbb{N} \to A_i$ for each i.
c) Show there is a bijection $F : \mathbb{N} \times \mathbb{N} \to \bigcup_{i \in I} A_i$ given by $F(n,m) = f_n(m)$. (*Hint: place the elements of* A'_1 *on* $(1,1)$, $(1,2)$, $(1,3)$, *...; the elements of* A'_2 *on* $(2,1)$, $(2,2)$, $(2,3)$, *... and so on. Now use the argument in item (a).*)
d) Conclude that Lemma 1.31 holds.

Exercise 1.9. What is wrong in the following attempt to prove that $[0,1]$ is uncountable?
Assume that $[0,1]$ is countable, that is: there is a bijection f between $[0,1]$ and $\mathbb{N}$. Let $r(n)$ be the unique number in $[0,1]$ assigned to n. Thus the infinite array $(r(1), r(2), \cdots)$ forms an exhaustive list of the numbers in $[0,1]$, as follows:

$$\begin{array}{rcl} r(1) & = & 0.\mathbf{0}0000000000\cdots \\ r(2) & = & 0.1\mathbf{0}101010111\cdots \\ r(3) & = & 0.00\mathbf{0}11111111\cdots \\ r(4) & = & 0.111\mathbf{1}1100000\cdots \\ r(5) & = & 0.0001\mathbf{0}010010\cdots \\ r(6) & = & 0.10000\mathbf{1}00011\cdots \\ \vdots & \vdots & \vdots \end{array}$$

(Written as number on the base 2.) Construct r^* as the string whose nth digit differs from that of $r(n)$. Thus in this example:

$$r^* = 0.\mathbf{111010}\cdots ,$$

which is different from all the other listed binary numbers in $[0,1]$.
(Hint: what if r^* *ends with an infinite all ones subsequence?)*

Exercise 1.10. The set $f(T)$ in the proof of Theorem 1.24 is called the <u>middle</u> <u>third</u> <u>Cantor</u> <u>set</u>. Find its construction. What does it look like? (*Hint: locate the set of numbers whose first digit (base 3) is a 1; then the set of numbers whose second digit is a 1.*)

Exercise 1.11. The integers exhibit many, many other intriguing patterns. Given the following function:

$$\begin{cases} n \text{ even:} & f(n) = \frac{n}{2} \\ n \text{ odd:} & f(n) = \frac{3n+1}{2} \end{cases} .$$

a) (Periodic orbit) Show that f sends 1 to 2 and 2 to 1.
b) (Periodic orbit attracts) Show that if you start with any positive integer less than 18 and apply f repeatedly, eventually you fall on the orbit in (a). See Figure 5.
c) Replace "$+1$" by "-1" and show that now 1 is a fixed point.
d) Show that the system is (c) has another periodic orbit.

Conjecture 1.32 (**Collatz conjecture or** $3n+1$ **problem).** *The orbit of every positive integer under the map f defined in exercise 1.11 ends in* $2 \leftrightarrow 1$[4].

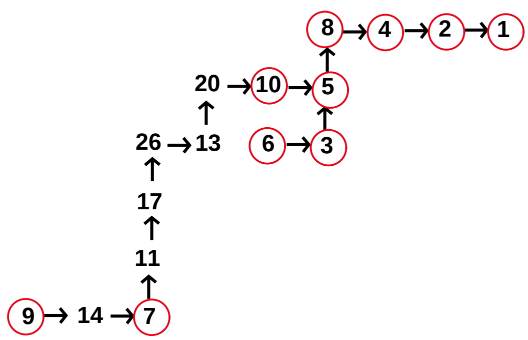


Figure 5. The orbits of n under f of exercise 1.11 for $n \in \{1, \cdots, 10\}$.

Exercise 1.12. This exercise prepares for Mersenne and Fermat primes, see Definition 5.13.
a) Use $\sum_{i=0}^{a-1} 2^{ib} = \frac{2^{ab}-1}{2^a-1}$ to show that if $2^p - 1$ is prime, then p must be prime.
b) Use $\sum_{i=0}^{a-1} (-2^b)^i = \frac{(-2^b)^a-1}{(-2)^a-1}$ to show that if $2^p + 1$ is prime, then p has no odd factor. (*Hint: assume a is odd.*)

[4]Still unsolved in 2022. For an interesting survey, see M. Chamberland, *A 3x+ 1 survey: Number theory and dynamical systems* in [**24**]

Exercise 1.13. In what follows, we assume that $e-1=\sum_{i=1}^{\infty}\frac{1}{i!}=\frac{p}{q}$ is rational and show that this leads to a contradiction.
a) Show that the above assumption implies that

$$\sum_{i=1}^{q}\frac{q!}{i!}+\sum_{i=1}^{\infty}\frac{q!}{(q+i)!}=p(q-1)!\,.$$

(*Hint: multiply both sides of by* $q!$.)
b) Show that $\sum_{i=1}^{\infty}\frac{q!}{(q+i)!}<\sum_{i=1}^{\infty}\frac{1}{(q+1)^i}$. (*Hint: write out a few terms of the sum on the left.*)
c) Show that the sum on the left hand side in (b) cannot have an integer value.
d) Show that the other two terms in (a) have an integer value.
e) Conclude there is a contradiction *unless* the assumption that e is rational is false.

Exercise 1.14. Show that Liouville's theorem (Theorem 1.19) also holds for rational for rational numbers $\rho=\frac{r}{s}$ as long as $\frac{p}{q}\neq\frac{r}{s}$.

Exercise 1.15. a) Show that for all positive integers p and n, we have $p(n+1)n!\leq(n+p)!$.
b) Use (a) to show that

$$\sum_{k=n+1}^{\infty}10^{-k!}\leq\sum_{p=1}^{\infty}10^{-p(n+1)n!}=10^{-(n+1)n!}\left(1-10^{-(n+1)n!}\right)^{-1}.$$

c) Show that b) implies the affirmation after equation (1.1).

Exercise 1.16. a) Use a calculator to write down the decimal expansion of $\sqrt{2}$ in 10 decimal places.
b) How close to $\sqrt{2}$ is the decimal approximation $1414/1000$?
c) Compute $1393/985$ is 10 decimal places. How close is it to $\sqrt{2}$? (*Hint: compare with Theorem 1.15.*)

Exercise 1.17. Show that the inequality of Roth's theorem does not hold for all numbers. (*Hint: Let ρ be a Liouville number.*)

Definition 1.33. *Let A be a set. Its power set $P(A)$ is the set whose elements are the subsets of A. This always includes the empty set denoted by $\emptyset$.*

In the next two exercises, the aim is to show something that is obvious for finite sets, namely:

Theorem 1.34. *The cardinality of a power set is always (strictly) greater than that of the set itself.*

Exercise 1.18. a) Given a set A, show that there is an injection $f : A \to P(A)$. (*Hint: for every element $a \in A$ there is a set $\{a\}$.*)
b) Conclude that $|A| \leq |P(A)|$. (*Hint: see Definition 1.27.*)

Exercise 1.19. Let A be an arbitrary set. Assume that that there is a surjection $S : A \to P(A)$ and define

$$R = \{a \in A \,|\, a \notin S(a)\} . \tag{1.2}$$

a) Show that there is a $q \in A$ such that $S(q) = R$.
b) Show that if $q \in R$, then $q \notin R$. (*Hint: equation* (1.2).)
c) Show that if $q \notin R$, then $q \in R$. (*Hint: equation* (1.2).)
d) Use (b) and (c) and exercise 1.18, to establish that $|A| < |P(A)|$. (*Hint: see Definition 1.27.*)

In the next two exercises we show that the cardinality of $\mathbb{R}$ equals that of $P(\mathbb{N})$. This implies that that $|\mathbb{R}| > |\mathbb{N}|$, which also follows from Theorem 1.24.

Exercise 1.20. Let T be the set of sequences defined in the proof of Theorem 1.24. To a sequence $t \in T$, associate a set $S(t)$ in $P(\mathbb{N})$ as follows:

$$i \in S \text{ if } t(i) = 2 \quad \text{ and } \quad i \notin S \text{ if } t(i) = 0 .$$

a) Show that there is a bijection $S : T \to P(\mathbb{N})$.
b) Use the bijection f in the proof of Theorem 1.24 to show there is a bijection $K \to P(\mathbb{N})$.
c) Show that (a) and (b) imply that $|P(\mathbb{N})| = |K| = |T|$. (*Hint: see Definition 1.27.*)
d) Find an injection $K \to \mathbb{R}$ and conclude that $|P(\mathbb{N})| \leq |\mathbb{R}|$.

Exercise 1.21. a) Show that there is a bijection $\mathbb{R} \to (0,1)$.
b) Show that there is an injection $(0,1) \to T$. (*Hint: use usual binary (base 2) expansion of reals.*)
c) Use (a), (b), and exercise 1.20 (a), to show that $|\mathbb{R}| \leq |P(\mathbb{N})|$.
d) Use (c) and exercise 1.20 (d) to show that $|\mathbb{R}| = |P(\mathbb{N})|$.

Exercise 1.22. Show that having the same cardinality (see Definition 1.27) is an equivalence relation on sets.

Exercise 1.23. a) Fix some $n > 0$. Show that having the same remainder modulo n is an equivalence relation on $\mathbb{Z}$. (*Hint: for example, -8, 4, and 16 have remainder 4 modulo 12.*)
b) Show that addition respects this equivalence relation. (*Hint: If $a + b = c$, $a \sim a'$, and $b \sim b'$, then $a' + b' = c'$ with $c \sim c'$.*)
c) The same question for multiplication.

Exercise 1.24. a) Show that the set of algebraic numbers is countable. (*Hint: use Lemma 1.31.*)
b) Conclude that the transcendental numbers form an uncountable set.

Exercise 1.25. Base 60 number systems have a long history and are still used (think of the number of minute in an hour). Suppose you do not have a good theory of fractions. Why is base 60 convenient? (*Hint: what is the least common multiple of the numbers 1 through 6?*)

Exercise 1.26. a) Show that rectangular grid of n by m squares can be divided into squares of size d by d where d is a common divisor of n and m.
b) Show that in (a) the largest d equals $\gcd(n,m)$, see Figure 6.

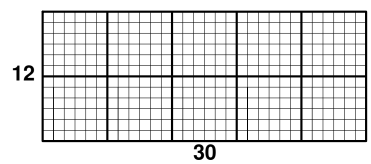


Figure 6. A rectangle of 30 by 12 squares can be subdivided into squares non larger than 6 by 6.

Exercise 1.27. Suppose two meshing gear wheels have n and m teeth, respectively. Each wheel has one marked tooth.
a) Show that the positions of the wheels after ℓ teeth are traversed is indicated by the projection of the point (ℓ,ℓ) on both in a rectangular coordinate system with n by m units. See Figure 7. (*Hint: each small square corresponds to the turn through one tooth on both wheels.* Show that the first time the marked teeth return exactly to their original position occurs when the first wheel has made $\operatorname{lcm}(n,m)/n = m/\gcd(n,m)$ complete turns and the second $\operatorname{lcm}(n,m)/n = n/\gcd(n,m)$.

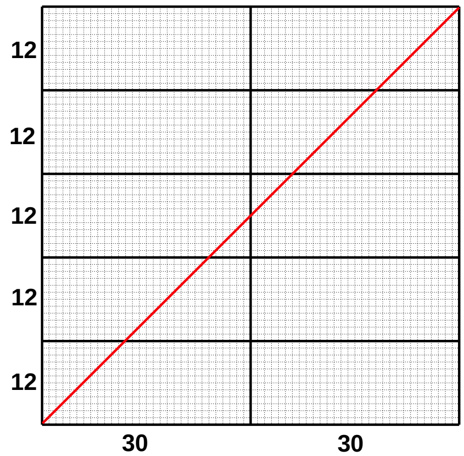


Figure 7. Two meshing gear wheels have 30, resp. 12 teeth. Each tiny square represents the turning of one tooth in each wheel. After precisely 5 turns of the first wheel and 2 of the second, both are back in the exact same position.

Chapter 2

The Fundamental Theorem of Arithmetic

Overview. We derive the Fundamental Theorem of Arithmetic. The most important part of that theorem says every integer can be uniquely written as a product of primes up to re-ordering of the factors, and up to factors -1. We discuss two of its most important consequences, namely the fact that the number of primes is infinite and the fact that non-integer roots are irrational.

On the way to proving the Fundamental Theorem of Arithmetic, we need Bézout's Lemma and Euclid's Lemma. The proofs of these well-known lemma's may appear abstract and devoid of intuition. To have some intuition, the student may *assume* the Fundamental Theorem of Arithmetic and derive from it each of these lemma's (see Exercise 2.9) and things will seem much more intuitive. The reason we do not do it that way in this book is of course that indirectly we use both results to establish the Fundamental Theorem of Arithmetic.

The principal difference between $\mathbb{Z}$ and $\mathbb{N}$ is that in $\mathbb{Z}$ addition has an inverse (subtraction). This makes $\mathbb{Z}$ a into a ring, a type of object we will encounter in Chapter 5. It will thus save us a lot of work and is not much more difficult to work in $\mathbb{Z}$ instead of in $\mathbb{N}$.

2.1. Bézout's Lemma

Definition 2.1. *The floor of a real number θ is defined as follows: $\lfloor \theta \rfloor$ is the greatest integer less than than or equal to θ. The fractional part $\{\theta\}$ of the number θ is defined as $\theta - \lfloor \theta \rfloor$. Similarly, the ceiling of θ, $\lceil \theta \rceil$, gives the smallest integer greater than or equal to θ.*

By the well-ordering principle, Corollary 1.10, the number $\lfloor \theta \rfloor$ and $\lceil \theta \rceil$ exist for any $\theta \in \mathbb{R}$. Given a number $\xi \in \mathbb{R}$, we denote its absolute value by $|\xi|$.

Lemma 2.2. *Given r_1 and r_2 with $r_2 > 0$, then there are q_2 and r_3 with $|r_3| < |r_2|$ such that $r_1 = r_2 q_2 + r_3$.*

Proof. Noting that $\frac{r_1}{r_2}$ is a rational number, we can choose the integer $q_2 = \lfloor \frac{r_1}{r_2} \rfloor$ so that

$$\frac{r_1}{r_2} = q_2 + e \, ,$$

where $e \in [0,1)$ (see Figure 8). The integer q_2 is called the *quotient*. Multiplying by r_2 gives the result. ■

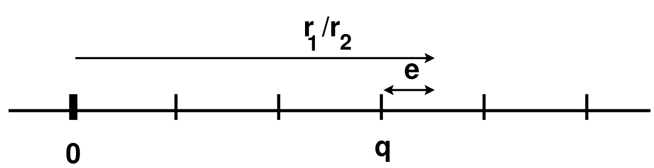


Figure 8. The division algorithm: for any two integers r_1 and r_2, we can find an integer q and a real $e \in [0,1)$ so that $r_1/r_2 = q_2 + e$.

Note that in this proof, in fact, $r_3 \in \{0, \cdots r_2 - 1\}$ and is unique. Thus among other things, this lemma implies that every integer has a unique residue (see Definition 1.8). More generally, we just require $|r_3| < |r_2|$, and there is more than one choice for q_2. This is typical in the more general context of rings (Chapter 8).

If $|r_1| < |r_2|$, then we can choose $q_2 = 0$. In this case, $\varepsilon = \frac{r_1}{r_2}$ and we learn nothing new. But if $|r_1| > |r_2|$, then $q_2 \neq 0$ and we have written r_1 as a multiple of r_2 plus a remainder r_3.

Definition 2.3. *Given r_1 and r_2 with $r_2 > 0$, the computation of q_2 and r_3 in Lemma 2.2 is called the division algorithm. Note that $r_3 = \mathrm{Res}_{r_2}(r_1)$ (see Definition 1.8).*

Remark 2.4. Lemma 2.2 is also called Euclid's division lemma. This is not to be confused with the Euclidean algorithm of Definition 3.3.

Lemma 2.5. (Bézout's Lemma) *Let a and b be such that $\gcd(a,b) = d$. Then $ax+by = c$ has integer solutions for x and y if and only if c is a multiple of d.*

Proof. Let S and $\nu(S)$ be the sets:

$$\begin{array}{rcl} S & = & \{ax+by : x,y \in \mathbb{Z},\, ax+by \neq 0\} \\ \nu(S) & = & \{|s| : s \in S\} \subseteq \mathbb{N} \cup \{0\} \end{array} .$$

Then $\nu(S) \neq \emptyset$ (it contains $|a|$ and $|b|$) and is bounded from below. Thus by the well-ordering principle of $\mathbb{N}$, it has a smallest element n. Then there is an element $d \in S$ that has that norm: $|d| = n$.

For that d, we use the division algorithm to establish that there are q and $r \geq 0$ such that

$$a = dq + r \;\text{ and }\; |r| < |d| \,. \tag{2.1}$$

Now substitute $d = ax+by$. A short computation shows that r can be rewritten as:

$$r = a(1-qx) + b(-qy) \,.$$

Suppose $r \neq 0$. Then this shows that $r \in S$. But we also know from (2.1) that $|r|$ is smaller than $|d|$. This is a contradiction because of the way d is defined. But $r = 0$ implies that d is a divisor of a. The same argument shows that d is also a divisor of b. Thus d is a common divisor of both a and b.

Now let e be *any* divisor of both a and b. Then $e \mid (ax+by)$, and so $e \mid d$. But if $e \mid d$, then $|e|$ must be smaller than or equal to $|d|$. Therefore, d is the *greatest* common divisor of both a and b.

By multiplying x and y by f, we achieve that for any multiple fd of d that

$$afx + bfy = fd \,.$$

On the other hand, let d be as defined above and suppose that x, y, and c are such that
$$ax+by=c\,.$$
Since d divides a and b, we must have that $d \mid c$, and thus c must be a multiple of d. ■

2.2. Corollaries of Bézout's Lemma

Lemma 2.6. (Euclid's Lemma) *Let a and b be such that $\gcd(a,b)=1$ and $a \mid bc$. Then $a \mid c$.*

Proof. By Bézout, there are x and y such that $ax+by=1$. Multiply by c to get:
$$acx+bcy=c\,.$$
Since $a \mid bc$, the left-hand side is divisible by a, and so is the right-hand side. ■

Euclid's lemma is so often used, that it will pay off to have a few of the standard consequences for future reference.

Theorem 2.7 (Cancellation Theorem). *Let $\gcd(a,b)=1$ and b positive. Then $ax =_b ay$ if and only if $x =_b y$.*

Proof. The statement is trivially true if $b=1$, because all integers are equal modulo 1.

If $ax =_b ay$, then $a(x-y) =_b 0$. The latter is equivalent to $b \mid a(x-y)$. The conclusion follows from Euclid's Lemma. Vice versa, if $x =_b y$, then $(x-y)$ is a multiple of b and so $a(x-y)$ is a multiple of b. ■

Used as we are to cancellations in calculations in $\mathbb{R}$, it is easy to underestimate the importance of this result. As an example, consider solving $21x =_{35} 21y$. It is tempting to say that this implies that $x =_{35} y$. But in fact, $\gcd(21,35)=7$ and the solution set is $x =_5 y$, as is easily checked. This example is in fact a special case of the following corollary.

Corollary 2.8. *Let $\gcd(a,b)=d$ and b positive. Then $ax =_b ay$ if and only if $x =_{b/d} y$.*

Proof. Divide by d to get $\frac{a}{d}x =_{\frac{b}{d}} \frac{a}{d}y$ and apply the cancellation theorem. ■

For the following results, recall the definition of primes in $\mathbb{Z}$ (Definition 1.4).

Corollary 2.9. *For any $n \geq 1$, p is prime and $p \mid \prod_{i=1}^{n} a_i$, if and only if there is $j \leq n$ such that $p \mid a_j$.*

Proof. If $p \mid a_j$, then $p \mid \prod_{i=1}^{n} a_i$. We prove the other direction by induction on n, the number of terms in the product. Let $S(n)$ be the statement of the corollary. $S(1)$ says: If p is prime and $p \mid a_1$, then $p \mid a_1$, which is trivially true.

For the induction step, suppose that for any $k > 1$, $S(k)$ is valid and let $p \mid \prod_{i=1}^{k+1} a_i$. Then

$$p \mid \left(\left(\prod_{i=1}^{k} a_i \right) a_{k+1} \right) .$$

Applying Euclid's Lemma, it follows that

$$p \mid \prod_{i=1}^{k} a_i \quad \text{ or, if not, then } \quad p \mid a_{k+1} .$$

In the former case $S(k+1)$ holds because $S(k)$ does. In the latter, we see that $S(k+1)$ also holds. ■

Corollary 2.10. *If p and q_i are prime and $p \mid \prod_{i=1}^{n} q_i$, then there is $j \leq n$ such that $p = q_j$.*

Proof. Corollary 2.9 says that if p and all q_i are primes, then there is $j \leq n$ such that $p \mid q_j$. Since q_j is prime, its only divisors are 1 and itself. Since $p \neq 1$ (by the definition of prime), $p = q_j$. ■

2.3. The Fundamental Theorem of Arithmetic

The last corollary of the previous section enables us to prove the most important result of this chapter.

Theorem 2.11 (The Fundamental Theorem of Arithmetic). *Every non-zero integer $n \in \mathbb{Z}$*
i) is a product of powers of primes (up to multiplication by units) and

ii) that product is unique (up to the order of multiplication and up to multiplication by the units).

Remark 2.12. The theorem is also called the *unique factorization theorem*. Its statement means that up to re-ordering of the p_i and factors ± 1, every integer n can be uniquely expressed as

$$n = \pm 1 \cdot \prod_{i=1}^{r} p_i^{\ell_i} ,$$

where the p_i are distinct primes.

Proof. First we prove (i). Define S to be the set of integers n that are not products of primes times a unit, and the set $\nu(S)$ their absolute values. If the set S is non-empty, then by the well-ordering principle (Theorem 1.9), $\nu(S)$ has a smallest element. Let a be one of the elements in S that minimize $\nu(S)$.

If a is prime, then it can be factored into primes, namely $a = a$, which contradicts the assumption. Thus a is a composite number, $a = bc$ and both b and c are non-units. Thus $|b|$ and $|c|$ are strictly smaller than $|a|$. By assumption, both b and c are products of primes. Then, of course, so is $a = bc$. But this contradicts the assumptions on a.

Next, we prove (ii). Let S be the set of integers that have more than one factorization and $\nu(S)$ the set of their absolute values. If the set S is non-empty, then, again by the well-ordering principle, $\nu(S)$ has a smallest element. Let a be one of the elements in S that minimize $\nu(S)$.

Thus we have

$$a = u \prod_{i=1}^{r} p_i = u' \prod_{i=1}^{s} p_i' ,$$

where at least some of the p_i and p_i' do not match up. Here, u and u' are units. Clearly, p_1 divides a. By Corollary 2.10, p_1 equals one of the p_i', say, p_1'. Since primes are not units, $\left|\frac{a}{p_1}\right|$ is strictly less than $|a|$. Therefore, by hypothesis, $\frac{a}{p_1}$ is uniquely factorizable. But then the primes in

$$\frac{a}{p_1} = u \prod_{i=2}^{r} p_i = u' \prod_{i=2}^{s} p_i' ,$$

all match up (up to units). ■

Remark 2.13. It is interesting to note that the proof of this theorem depends on *two* distinct characterizations of primes. In part (i), we use Definition 1.4, which essentially says that primes are numbers that cannot be factored into smaller numbers (the literal meaning of "irreducible"). But for part (ii), we essentially use the fact that if a prime p divides ab, then it divides a or b (or both). Now (through Corollary 2.10) we know both characterizations hold in $\mathbb{Z}$, but it will turn out that they are not equivalent in general (see Proposition 8.3).

If the reader investigates the arguments carefully, it will become clear that underneath it all lurks the division algorithm in $\mathbb{Z}$. To wit, we use Corollary 2.10 which Corollary 2.9 which uses Euclid's lemma which uses Bézout which finally uses the division algorithm. It is precisely this division algorithm that is not available in all rings, and which plays an important role in algebraic number theory, see Chapter 8).

Remark 2.14. The student might reflect on this and conclude that one *cannot* write 1 as a product of primes. So how come that in Theorem 2.11 we do not make an exception for the number 1 (or -1 for that matter). The answer is this: 1 is a unit times "the empty product" of primes, and this is unique. This piece of apparent *sophistry* actually turns out to be useful as we will see in Chapter 8 (corollary 8.14).

2.4. Corollaries of the Fundamental Theorem of Arithmetic

The unique factorization theorem is intuitive and easy to use. It is very effective in proving a great number of results. Some of these results can be proved with a little more effort without using the theorem (see exercise 2.6 for an example). We start with two somewhat technical results that we need for later reference.

Lemma 2.15. *We have*

$$\forall\, i \in \{1, \cdots n\} \;:\; \gcd(a_i, b) = 1 \quad \Longleftrightarrow \quad \gcd(\prod_{i=1}^{n} a_i, b) = 1\,.$$

Proof. The easiest way to see this uses prime power factorization. If $\gcd(\prod_{i=1}^{n} a_i, b) = d > 1$, then d contains a factor $p > 1$ that is a prime. Since p divides $\prod_{i=1}^{n} a_i$, at least one of the a_i must contain (by Corollary

2.9) a factor p. Since p also divides b, this contradicts the assumption that $\gcd(a_i, b) = 1$.

Vice versa, if $\gcd(a_i, b) = d > 1$ for some i, then also $\prod_{i=1}^{n} a_i$ is divisible by d. ■

Corollary 2.16. *For all a and b in $\mathbb{Z}$ not both equal to 0, we have that $\gcd(a,b) \cdot \operatorname{lcm}(a,b) = ab$ up to units.*

Proof. Given two numbers a and b, let $P = \{p_i\}_{i=1}^{k}$ be the list of all prime numbers occurring in the unique factorization of a or b. We then have:

$$a = u \prod_{i=1}^{s} p_i^{k_i} \quad \text{and} \quad b = u' \prod_{i=1}^{s} p_i^{\ell_i} ,$$

where u and u' are units and k_i and ℓ_i in $\mathbb{N} \cup \{0\}$. Now define:

$$m_i = \min(k_i, \ell_i) \quad \text{and} \quad M_i = \max(k_i, \ell_i) ,$$

and let the numbers m and M be given by

$$m = \prod_{i=1}^{s} p_i^{m_i} \quad \text{and} \quad M = \prod_{i=1}^{s} p_i^{M_i} .$$

Since $m_i + M_i = k_i + \ell_i$, it is clear that the multiplication $m \cdot M$ yields ab.

Now all we need to do, is showing that m equals $\gcd(a,b)$ and that M equals $\operatorname{lcm}(a,b)$. Clearly m divides both a and b. On the other hand, any integer greater than m has a unique factorization that *either* contains a prime not in the list P and therefore divides neither a nor b, *or*, if not, at least one of the primes in P in its factorization has a power greater than m_i. In the last case m is not a divisor of at least one of a and b. The proof that M equals $\operatorname{lcm}(a,b)$ is similar. ■

A question one might ask is: how many primes are there? In other words, how long can the list of primes in a factorization be? Euclid provided the answer around 300BC.

Theorem 2.17 (Infinitude of Primes). *There are infinitely many primes.*

Proof. Suppose the list P of all primes is finite, so that $P = \{p_i\}_{i=1}^{n}$. Define the integer d as the product of all primes (to the power 1):

$$d = \prod_{i=1}^{n} p_i .$$

If $d+1$ is a prime, we have a contradiction. So $d+1$ must be divisible by a prime p_i in P. But then we have

$$p_i \mid d \qquad \text{and} \qquad p_i \mid d+1 \; . \tag{2.2}$$

But since $(d+1)(1)+d(-1)=1$, Bézout's lemma implies that $\gcd(d,d+1)=1$, which contradicts equation (2.2). ■

One the best known consequences of the fundamental theorem of arithmetic is probably the theorem that follows below. A special case, namely $\sqrt{2}$ is irrational (see Theorem 1.14), was known to Pythagoras in the 6th century BC.

Theorem 2.18. *Let $n>0$ and $k>1$ be integers. Then $n^{\frac{1}{k}}$ is either an integer or irrational.*

Proof. Assume $n^{\frac{1}{k}}$ is rational. That is: suppose that there are integers a and b such that

$$n^{\frac{1}{k}} = \frac{a}{b} \qquad \Longrightarrow \qquad n \cdot b^k = a^k \; .$$

Divide out any common divisors of a and b, so that $\gcd(a,b)=1$. Then by the fundamental theorem of arithmetic, $b=\prod_{i=1}^{s} p_i^{m_i}$ and $a=\prod_{i=s+1}^{r} p_i^{\ell_i}$ (a and b share no prime factors) and so

$$n \prod_{i=1}^{s} p_i^{km_i} = \prod_{i=s+1}^{r} p_i^{k\ell_i} \; .$$

The primes p_i on the left and right side are distinct. This is only possible if $\prod_{i=1}^{s} p_i^{km_i}$ equals 1. But then n is the k-th power of an integer. ■

2.5. The Riemann Hypothesis

Analytic continuation will be discussed in more detail in Chapter 11. For now, we note that it is akin to replacing e^x where x is real by e^z where z is complex. A better example is the series $\sum_{j=0}^{\infty} z^j$. This series diverges for $|z|>1$. But as an analytic function, it can be replaced by $(1-z)^{-1}$ on all of $\mathbb{C}$ except at the pole $z=1$ where it diverges.

Analytic continuations are meaningful because they are unique. The reason this is true is roughly as follows (for details, see Theorem 11.22).

Analytic functions are functions that are differentiable, that is to say, wherever the derivative is non-zero, the derivative equals a scaling times a rotation. Equivalently, they are locally given by a convergent power series. If f and g are two analytic continuations to a region U of a function h given on a region $V \subset U$, then the difference $f-g$ is zero on V. One can then show that the power series of $f-g$ must be zero on the entire region U. Hence, analytic continuations f and g are unique.

Definition 2.19. *The Riemann zeta function* $\zeta(z)$ *is a complex function defined on* $\{z \in \mathbb{C} \,|\, \mathrm{Re}\, z > 1\}$ *by*

$$\zeta(z) = \sum_{n=1}^{\infty} n^{-z} .$$

On other values of $z \in \mathbb{C}$ *it is defined by the analytic continuation of this function (except at* $z = 1$ *where it has a simple pole).*

In analytic number theory, it is common to denote the argument of the zeta function by s, while in other branches of complex analysis z is the go-to complex variable. We will stick to the latter. Note that

$$n^{-z} = e^{-\ln n \,\mathrm{Re}\, z - i \ln n \,\mathrm{Im}\, z},$$

and so $|n^{-z}| = n^{-\mathrm{Re}\, z}$. Therefore for $\mathrm{Re}\, z > 1$ the series is absolutely convergent. More about this in Chapter 11. At this point, the student should remember – or look up in [**3**] – the fact that absolutely convergent series can be re-arranged arbitrarily without changing the sum. This leads to the following proposition.

Proposition 2.20 (Euler's Product Formula). *For* $\mathrm{Re}\, z > 1$ *we have*

$$\zeta(z) := \sum_{n=1}^{\infty} n^{-z} = \prod_{p \text{ prime}} (1 - p^{-z})^{-1} .$$

There are two common proofs of this formula. It is worth presenting both.

proof 1. The first proof uses the Fundamental Theorem of Arithmetic. First, we use the geometric series

$$(1 - p^{-z})^{-1} = \sum_{k=0}^{\infty} p^{-kz}$$

to rewrite the right-hand side of the Euler product. This gives

$$\prod_{p \text{ prime}} (1-p^{-z})^{-1} = \left(\sum_{k_1=0}^{\infty} p_1^{-k_1 z}\right)\left(\sum_{k_2=0}^{\infty} p_2^{-k_2 z}\right)\left(\sum_{k_3=0}^{\infty} p_3^{-k_3 z}\right)\cdots$$

Re-arranging terms yields

$$\cdots = \sum_{k_1,k_2,k_3,\cdots \geq 0} \left(p_1^{k_1} p_2^{k_2} p_3^{k_3} \cdots\right)^{-z} .$$

By the Fundamental Theorem of Arithmetic, the expression $\left(p_1^{k_1} p_2^{k_2} p_3^{k_3} \cdots\right)$ runs through all positive integers exactly once. Thus upon re-arranging again we obtain $\sum_{n=1}^{\infty} n^{-z}$. ■

proof 2. The second proof, the one that Euler used, employs a sieve method. This time, we start with the left-hand side of the Euler product. If we multiply ζ by 2^{-z}, we get back precisely the terms with n even. So

$$\left(1-2^{-z}\right)\zeta(z) = 1 + 3^{-z} + 5^{-z} + \cdots = \sum_{2 \nmid n} n^{-z} .$$

Subsequently we multiply this expression by $(1-3^{-z})$. This has the effect of removing the terms that remain where n is a multiple of 3. It follows that eventually

$$\left(1-p_\ell^{-z}\right)\cdots\left(1-p_1^{-z}\right)\zeta(z) = \sum_{p_1 \nmid n, \cdots p_\ell \nmid n} n^{-z} .$$

The argument used in Eratosthenes' sieve (Section 1.1) now serves to show that in the right-hand side of the last equation all terms other than 1 disappear as ℓ tends to infinity. Therefore, the left-hand side tends to 1, which implies the proposition. ■

The most important theorem concerning primes is probably the following. We will give a proof in Chapter 12.

Theorem 2.21 (Prime Number Theorem). *Let $\pi(x)$ denote the prime counting function, that is: the number of primes less than or equal to x with $x \geq 2$. Then*

$$1)\ \lim_{x\to\infty} \frac{\pi(x)}{(x/\ln x)} = 1 \quad \text{and} \quad 2)\ \lim_{x\to\infty} \frac{\pi(x)}{\int_2^x \ln t \, dt} = 1 ,$$

where ln *is the natural logarithm.*

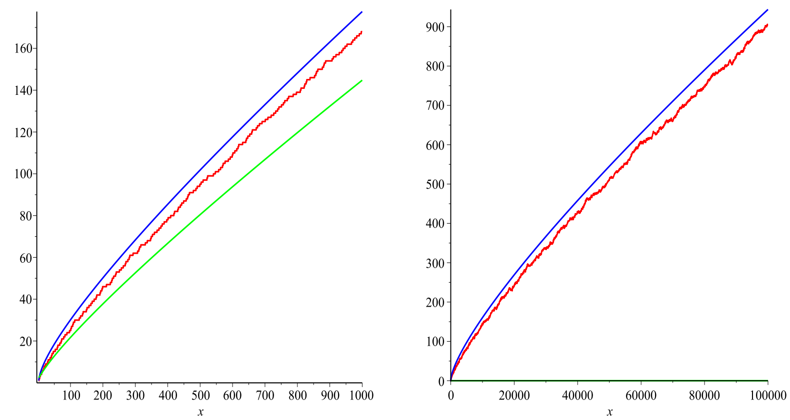


Figure 9. On the left, the function $\int_2^x \ln t\,dt$ in blue, $\pi(x)$ in red, and $x/\ln x$ in green. On the right, we have $\int_2^x \ln t\,dt - x/\ln x$ in blue, $\pi(x) - x/\ln x$ in red. Note the different scales.

The first estimate is the one we will prove directly in Chapter 12. It turns out the second is equivalent to it (exercise 12.10). However, it is this one that gives the better estimate of $\pi(x)$. In Figure 9 on the left, we plotted, for $x \in [2, 1000]$, from top to bottom the functions $\int_2^x \ln t\,dt$ in blue, $\pi(x)$ in red, and $x/\ln x$. In the right-hand figure, we augment the domain to $x \in [2, 10^5]$. and plot the *difference* of these functions with $x/\ln x$. It now becomes clear that $\int_2^x \ln t\,dt$ is indeed a much better approximation of $\pi(x)$. From this figure one may be tempted to conclude that $\int_2^x \ln t\,dt - \pi(x)$ is always greater than or equal to zero. This, however, is false. It is known that there are infinitely many n for which $\pi(n) > \int_2^n \ln t\,dt$. The first such n is called the *Skewes number*. Not much is known about this number[1], except that it is less than 10^{317}.

Perhaps the most important open problem in all of mathematics is the following. It concerns the analytic continuation of $\zeta(z)$ given above.

Conjecture 2.22 (**Riemann Hypothesis**). *All non-real zeros of $\zeta(z)$ lie on the line* $\mathrm{Re}\, z = \frac{1}{2}$.

In his only paper on number theory [**60**], Riemann realized that the hypothesis enabled him to describe detailed properties of the distribution

[1] In 2020.

of primes in terms of of the location of the non-real zero of $\zeta(z)$. This completely unexpected connection between so disparate fields — analytic functions and primes in $\mathbb{N}$ — spoke to the imagination and led to an enormous interest in the subject[2] In further research, it has been shown that the hypothesis is also related to other areas of mathematics, such as, for example, the spacings between eigenvalues of random Hermitian matrices [**52**], and even physics [**12**, **15**].

2.6. Exercises

Exercise 2.1. Apply the division algorithm to the following number pairs. (*Hint: replace negative numbers by positive ones.*)
a) 110 , 7.
b) 51 , -30.
c) -138 , 24.
d) 272 , 119.
e) 2378 , 1769.
f) 270 , 175560.

Exercise 2.2. In this exercise we will exhibit the division algorithm applied to polynomials $x+1$ and $3x^3+2x+1$ with coefficients in $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$.
a) Apply long division to divide 3021 by 11. (*Hint:* $3021 = 11\cdot 275 - 4$.)
b) Apply the exact same algorithm to divide $3x^3+2x+1$ by $x+1$. In this algorithm, x^k behaves as 10^k in (a). (*Hint: at every step, cancel the highest power of x.*)
c) Verify that you obtain $3x^3+2x+1 = (x+1)(3x^2-3x+5)-4$.
d) Show that in general, if p_1 and p_2 are polynomials such that the degree of p_1 is greater or equal to the degree of p_2, then

$$p_1 = q_2 p_2 + p_3 \ ,$$

where the degree of p_3 is less than the degree of p_2. (*Hint: perform long division as in (b). Stop when the degree of the remainder is less than that of p_2.*)
e) Why does this division not work for polynomials with coefficients in $\mathbb{Z}$? (*Hint: replace $x+1$ by $2x+1$.*)

Exercise 2.3. a) For a, b in $\mathbb{Z}$, let $\gcd(a,b)=1$. Show that if $a \mid c$ and $b \mid c$, then $ab \mid c$. (*Hint: observe that $a \mid by$ and use Euclid's lemma.*)
b) Show that $ax =_m c$ has a solution if and only $\gcd(a,m) \mid c$. (*Hint: note that $ax+my=c$ for some y and use Bézout.*)

[2]This area of research, complex analysis methods to investigate properties of primes, is now called *analytic number theory*. We take this up in Chapters 11 and 12.

Exercise 2.4. a) Compute by long division that $3021 = 11 \cdot 274 + 7$.
b) Conclude from exercise 2.2 that $3021 = 11(300 - 30 + 5) - 4$. (*Hint: let* $x = 10$.)
c) Conclude from exercise 2.2 (b) that

$$3 \cdot 16^3 + 2 \cdot 16 + 1 = 17(3 \cdot 16^2 - 3 \cdot 16 + 5) - 4 \,.$$

(*Hint: let* $x = 16$.)

Exercise 2.5. a) Use unique factorization to show that any composite number n must have a prime factor less than or equal to $\sqrt{n}$.
b) Use that fact to prove: If we apply Eratosthenes' sieve to $\{2, 3, \cdots n\}$, it is sufficient to sieve out numbers less than or equal to $\sqrt{n}$.

Exercise 2.6. We give an *elementary*[a] proof of Corollary 2.16.
a) Show that $a \cdot \frac{b}{\gcd(a,b)}$ is a multiple of a.
b) Show that $\frac{a}{\gcd(a,b)} \cdot b$ is a multiple of b.
c) Conclude that $\frac{ab}{\gcd(a,b)}$ is a multiple of both a and b and thus greater than or equal to $\operatorname{lcm}(a,b)$.
d) Show that $a / \left(\frac{ab}{\operatorname{lcm}(a,b)} \right) = \frac{\operatorname{lcm}(a,b)}{b}$ is an integer. Thus $\frac{ab}{\operatorname{lcm}(a,b)}$ is a divisor of a.
e) Similarly, show that $\frac{ab}{\operatorname{lcm}(a,b)}$ is a divisor of b.
f) Conclude that $\frac{ab}{\operatorname{lcm}(a,b)} \leq \gcd(a,b)$.
h) Finish the proof.

[a] The word elementary has a complicated meaning, namely a proof that does not use some at first glance unrelated results. In this case, we mean a proof that does not use unique factorization. It does not imply that the proof is easier. Indeed, the proof in the main text seems much easier once unique factorization is understood.

Exercise 2.7. It is possible to extend the definition of gcd and lcm to more than two integers (not all of which are zero). For example $\gcd(24, 27, 54) = 3$.
a) Compute $\gcd(6, 10, 15)$ and $\operatorname{lcm}(6, 10, 15)$.
b) Give an example of a triple whose gcd is one, but every pair of which has a gcd greater than one.
c) Show that there is no triple $\{a, b, c\}$ whose lcm equals abc, but every pair of which has lcm less than the product of that pair. (*Hint: consider* $\operatorname{lcm}(a,b) \cdot c$.)

Exercise 2.8. a) Give the prime factorization of the following numbers: 12, 392, 1043, 31, 128, 2160, 487.
b) Give the prime factorization of the following numbers: $12 \cdot 392$, $1043 \cdot 31$, $128 \cdot 2160$.
c) Give the prime factorization of: $1,250000$, 63^3, 720, and the product of the last three numbers.

Exercise 2.9. Use the Fundamental Theorem of Arithmetic to prove:
a) Bézout's Lemma.
b) Euclid's Lemma.

Exercise 2.10. For positive integers m and n, suppose that $m^\alpha = n$. Show that $\alpha = \frac{a}{b}$ with $\gcd(a,b) = 1$ if and only if

$$m = \prod_{i=1}^{s} p_i^{k_i} \quad \text{and} \quad n = \prod_{i=1}^{s} p_i^{\ell_i} \quad \text{with} \quad \forall\, i \,:\, ak_i = b\ell_i \,.$$

Exercise 2.11. Let E be the set of even numbers. Let a, c in E, then c is divisible by a if there is a $b \in E$ so that $ab = c$. Define a prime p in E as a number in E such that there are no a and b in E with $ab = p$.
a) List the first 30 primes in E.
b) Does Euclid's lemma hold in E? Explain.
c) Factor 60 into primes (in E) in two different ways.

Exercise 2.12. See exercise 2.11. Show that any number in E is a product of primes in E. (*Hint: follow the proof of Theorem 2.11, part (i).*)

Exercise 2.13. See exercise 2.11 which shows that unique factorization does not hold in $E = \{2,4,6,\cdots\}$. The proof of unique factorization uses Euclid's lemma. In turn, Euclid's lemma was a corollary of Bézout's lemma, which depends on the division algorithm. Where exactly does the chain break down in this case?

Exercise 2.14. Let $L = \{p_1, p_2, \cdots\}$ be the list of all (infinitely many) primes, ordered according ascending magnitude. Show that $p_{n+1} \leq \prod_{i=1}^{n} p_i$. (*Hint: consider* $d = \prod_{i=1}^{n} p_i$ *and let* p_{n+1} *be the smallest prime divisor of* $d-1$. *See the proof of Theorem 2.17.*)

A much stronger version of exercise 2.14 is the so-called Bertrand's Postulate. That theorem says that for every $n \geq 1$, there is a prime in $\{n+1,\cdots,2n\}$. It was proved by Chebyshev. Subsequently the proof was simplified by Ramanujan and Erdös [**1**].

Exercise 2.15. Let p and q be primes greater than 3.
a) Show that $\mathrm{Res}_{12}(p) = r$ with $r \in \{1,5,7,11\}$. (The same holds for q.)
b) Show that $24 \mid p^2 - q^2$. (*Hint: use (a) to show that* $p^2 = 24x + r^2$ *and check all cases.*)

Exercise 2.16. A *square full integer* is an integer n that has a prime factor and each prime factor occurs with a power at least 2. A *square free integer* is an integer n such that each prime factor occurs with a power at most 1.
a) If n is square full, show that there are positive integers a and b such that $n = a^2b^3$.
b) Show that every integer greater than one is the product of a square free number and a square full number.

Exercise 2.17. Let $L = \{p_1, p_2, \cdots\}$ be the list of all primes, ordered according ascending magnitude. The numbers $E_n = 1 + \prod_{i=1}^{n} p_i$ are called *Euclid numbers*.
a) Check the primality of E_1 through E_6.
b) Show that $E_n =_4 3$. (*Hint: $E_n - 1$ is twice an odd number.*)
c) Show that for $n \geq 3$ the decimal representation of E_n ends in a 1. (*Hint: look at the factors of E_n.*)

Exercise 2.18. *Twin primes* are a pair of primes of the form p and $p+2$.
a) Show that the product of two twin primes plus one is a square.
b) Show that $p > 3$, the sum of twin primes is divisible by 12. (*Hint: see exercise 2.15*)

Exercise 2.19. Show that there arbitrarily large gaps between successive primes. More precisely, show that every integer in $\{n!+2, n!+3, \cdots n!+n\}$ is composite for any $n \geq 2$.

The usual statement for the fundamental theorem of arithmetic includes only natural numbers $n \in \mathbb{N}$ (i.e. not $\mathbb{Z}$) and the common proof uses induction on n. We review that proof in the next two problems.

Exercise 2.20. a) Prove that 2 can be written as a product of primes.
b) Let $k > 2$. Suppose all numbers in $\{1, 2, \cdots k\}$ can be written as a product of primes (or 1). Show that $k+1$ is either prime or composite.
c) If in (b), $k+1$ is prime, then all numbers in $\{1, 2, \cdots k+1\}$ can be written as a product of primes (or 1).
d) If in (b), $k+1$ is composite, then there is a divisor $d \in \{2, \cdots k\}$ such that $k+1 = dd'$.
e) Show that the hypothesis in (b) implies also in this case, all numbers in $\{1, 2, \cdots k+1\}$ can be written as a product of primes (or 1).
f) Use the above to formulate the inductive proof that all elements of $\mathbb{N}$ can be written as a product of primes.

Exercise 2.21. The set-up of the proof is the same as in exercise 2.20. Use induction on n. We assume the result of that exercise.
a) Show that $n = 2$ has a unique factorization.
b) Suppose that if for $k > 2$, $\{2, \cdots k\}$ can be uniquely factored. Then there are primes p_i and q_i, not necessarily distinct, such that

$$k+1 = \prod_{i=1}^{s} p_i = \prod_{i=1}^{r} q_i \,.$$

c) Show that then p_1 divides $\prod_{i=1}^{r} q_i$ and so, Corollary 2.10 implies that there is a $j \leq r$ such that $p_1 = q_j$.
d) Relabel the q_i's, so that $p_1 = q_1$ and divide n by $p_1 = q_1$. Show that

$$\frac{k+1}{q_1} = \prod_{i=2}^{s} p_i = \prod_{i=2}^{r} q_i \,.$$

e) Show that the hypothesis in (b) implies that the remaining p_i equal the remaining q_i. (*Hint:* $\frac{k}{q_1} \leq k$.)
f) Use the above to formulate the inductive proof that all elements of $\mathbb{N}$ can be uniquely factored as a product of primes.

Here is a different characterization of gcd and lcm. We prove it as a corollary of the prime factorization theorem.

Corollary 2.23. *(1) A common divisor $d > 0$ of a and b equals* $\gcd(a,b)$ *if and only if every common divisor of a and b is a divisor of d.*
(2) Also, a common multiple $d > 0$ of a and b equals $\operatorname{lcm}(a,b)$ *if and only if every common multiple of a and b is a multiple of d.*

Exercise 2.22. Use the characterization of $\gcd(a,b)$ and $\operatorname{lcm}(a,b)$ given in the proof of Corollary 2.16 to prove Corollary 2.23.

Exercise 2.23. We develop the proof of Theorem 2.17 as it was given by Euler. We start by *assuming* that there is a finite list L of k primes. We will show in the following steps how that assumption leads to a contradiction. We order the list according to ascending order of magnitude of the primes. So $L = \{p_1, p_2, \cdots, p_k\}$ where $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, and so forth, up to the last prime p_k.
a) Show that $\prod_{i=1}^{k} \frac{p_i}{p_i - 1}$ is finite, say M.
b) Show that for $r > 0$,

$$\prod_{i=1}^{k} \frac{p_i}{p_i - 1} = \prod_{i=1}^{k} \frac{1}{1 - p_i^{-1}} > \prod_{i=1}^{k} \frac{1 - p_i^{-r-1}}{1 - p_i^{-1}} = \prod_{i=1}^{k} \left(\sum_{j=0}^{r} p_i^{-j} \right) .$$

c) Use the fundamental theorem of arithmetic to show that there is an $\alpha(r) > 0$ such that

$$\prod_{i=1}^{k} \left(\sum_{j=0}^{r} \frac{1}{p_i^j} \right) = \sum_{\ell=1}^{\alpha(r)} \frac{1}{\ell} + R ,$$

where R is a non-negative remainder.
d) Show that for all K there is an r such that $\alpha(r) > K$.
e) Thus for *any* K, there is an r such that

$$\prod_{i=1}^{k} \left(\sum_{j=0}^{r} \frac{1}{p_i^j} \right) \geq \sum_{\ell=1}^{K} \frac{1}{\ell} .$$

f) Conclude with a contradiction between a) and e). (*Hint: the harmonic series* $\sum \frac{1}{\ell}$ *diverges or see exercise 2.24 c).*)

Exercise 2.24. In this exercise we consider the Riemann zeta function for real values of z greater than 1.
a) Show that for all $x > -1$, we have $\ln(1 + x) \leq x$.
b) Use Proposition 2.20 and a) to show that

$$\ln \zeta(z) = \sum_{p \text{ prime}} \ln \left(1 + \frac{p^{-z}}{1 - p^{-z}} \right) \leq \sum_{p \text{ prime}} \frac{p^{-z}}{1 - p^{-z}} \leq \sum_{p \text{ prime}} \frac{p^{-z}}{1 - 2^{-z}} .$$

c) Use the following argument to show that $\lim_{z \searrow 1} \zeta(z) = \infty$.

$$\sum_{n=1}^{\infty} n^{-1} > \sum_{n=1}^{\infty} n^{-z} > \int_1^{\infty} x^{-z}\, dx .$$

(*Hint: for the last inequality, see Figure 10.*)
d) Show that b) and c) imply that $\sum_{p \text{ prime}} p^{-z}$ diverges as $z \searrow 1$.
e) Use (d) to show that — in some sense — primes are more frequent than squares in the natural numbers. (*Hint:* $\sum_{n=1}^{\infty} n^{-2}$ *converges.*)

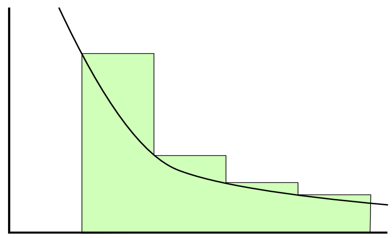


Figure 10. Proof that $\sum_{n=1}^{\infty} f(n)$ is greater than $\int_1^{\infty} f(x)\,dx$ if f is positive and (strictly) decreasing.

Exercise 2.25. a) Let p be a fixed prime. Show that the probability that two independently chosen integers in $\{1,\cdots,n\}$ are divisible by p tends to $1/p^2$ as $n \to \infty$. Equivalently, the probability that they are *not* divisible by p tends to $1 - 1/p^2$.
b) Make the necessary assumptions, and show that the probability that two two independently chosen integers in $\{1,\cdots,n\}$ are *not* divisible by *any* prime tends to $\prod_{p \text{ prime}} \left(1 - p^{-2}\right)$. (*Hint: you need to assume that the probabilities in (a) are independent and so they can be multiplied.*)
c) Show that from (b) and Euler's product formula, it follows that for 2 random (positive) integers a and b to have $\gcd(a,b) = 1$ has probability $1/\zeta(2) \approx 0.61$.
d) Show that for $d > 1$ and integers $\{a_1, a_2, \cdots a_d\}$ that probability equals $1/\zeta(d)$. (*Hint: the reasoning is the same as in (a), (b), and (c).*)
e) Show that for real $d > 1$:

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

For the middle inequality, see Figure 11.
f) Show that for large d, the probability that $\gcd(a_1, a_2, \cdots a_d) = 1$ tends to 1.

Exercise 2.26. This exercise in based on exercise 2.25.
a) In the $\{-4,\cdots,4\}^2 \backslash (0,0)$ grid in $\mathbb{Z}^2$, find out which proportion of the lattice points is visible from the origin, see Figure 12.
b) Use exercise 2.25 (c) to show that in a large grid, this proportion tends to $1/\zeta(2)$.
c) Use exercise 2.25 (d) to show that as the dimension increases to infinity, the proportion of the lattice points $\mathbb{Z}^d$ that are visible from the origin, increases to 1.

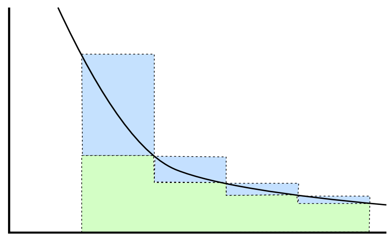


Figure 11. Proof that $\sum_{n=1}^{\infty} f(n)$ (shaded in blue and green) minus $f(1)$ (shaded in blue) is less than $\int_1^{\infty} f(x)\,dx$ if f is positive and (strictly) decreasing to 0.

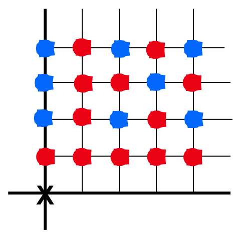


Figure 12. The origin is marked by "$\times$". The red dots are visible from $\times$; between any blue dot and $\times$ there is a red dot. The picture shows exactly one quarter of $\{-4,\cdots,4\}^2\backslash(0,0) \subset \mathbb{Z}^2$.

Exercise 2.27. We note here that $\zeta(2) = \frac{\pi^2}{6}$.
a) Show that the irrationality of π implies that $\zeta(2)$ is irrational.
b) Show that (a) and Proposition 2.20 yield another proof of the infinity of primes.

Chapter 3

Linear Diophantine Equations

Overview. A Diophantine equation is a polynomial equation in two or more unknowns and for which we seek to know what integer solutions it has. We determine the integer solutions of the simplest linear Diophantine equation $ax + by = c$. The central element this reasoning is the Euclidean algorithm. That algorithm has much wider applications. We discuss a few of those.

3.1. The Euclidean Algorithm

Lemma 3.1. *In the division algorithm of Lemma 2.2, we have* $\gcd(r_1, r_2) = \gcd(r_2, r_3)$.

Proof. On the one hand, we have $r_1 = r_2q_2 + r_3$, and so any common divisor of r_2 and r_3 must also be a divisor of r_1 (and of r_2). Vice versa, since $r_1 - r_2q_2 = r_3$, we have that any common divisor of r_1 and r_2 must also be a divisor of r_3 (and of r_2). ■

Thus by calculating r_3, the residue of r_1 modulo r_2, we have simplified the computation of $\gcd(r_1, r_2)$. This is because r_3 is strictly smaller (in absolute value) than both r_1 and r_2. In turn, the computation of $\gcd(r_2, r_3)$ can be simplified similarly, and so the process can be repeated. Since the r_i form a monotone decreasing sequence in $\mathbb{N}$, this process must end when $r_{n+1} = 0$ after a finite number of steps. We then have $\gcd(r_1, r_2) = \gcd(r_n, 0) = r_n$.

Corollary 3.2. *Given $r_1 > r_2 > 0$, apply the division algorithm until $r_n > r_{n+1} = 0$. Then $\gcd(r_1, r_2) = \gcd(r_n, 0) = r_n$. Since r_i is decreasing, the algorithm always ends.*

Definition 3.3. *The repeated application of the division algorithm to compute $\gcd(r_1, r_2)$ is called the <u>Euclidean</u> <u>algorithm</u>.*

We now give a framework to reduce the messiness of these repeated computations. Suppose we want to compute $\gcd(188, 158)$. We do the following computations:

$$\begin{array}{rcl} 188 & = & 158 \cdot 1 + 30 \\ 158 & = & 30 \cdot 5 + 8 \\ 30 & = & 8 \cdot 3 + 6 \\ 8 & = & 6 \cdot 1 + 2 \\ 6 & = & 2 \cdot 3 + 0 \end{array} \quad ,$$

We see that $\gcd(188, 158) = 2$. The numbers that multiply the r_i are the quotients of the division algorithm (see the proof of Lemma 2.2). If we call them q_i, the computation looks as follows:

$$\begin{array}{rcl} r_1 & = & r_2 \, q_2 + r_3 \\ r_2 & = & r_3 \, q_3 + r_4 \\ \vdots & \vdots & \vdots \\ r_{n-3} & = & r_{n-2} \, q_{n-2} + r_{n-1} \\ r_{n-2} & = & r_{n-1} \, q_{n-1} + r_n \\ r_{n-1} & = & r_n \, q_n + 0 \end{array} \quad , \tag{3.1}$$

where we use the convention that $r_{n+1} = 0$ while $r_n \neq 0$. Observe that with that convention, (3.1) consists of $n-1$ steps. A much more concise form (in part based on a suggestion of Katahdin [**37**]) to render this computation is as follows.

$$\begin{array}{c|c|c|c|c|c|c|} & q_n & q_{n-1} & \cdots & q_3 & q_2 & \\ \hline 0 & r_n & r_{n-1} & \cdots & r_3 & r_2 & r_1 \end{array} \tag{3.2}$$

Thus, each step $r_{i+1} \mid r_i \mid$ is similar to the usual long division, except that its quotient q_{i+1} is placed above r_{i+1} (and not above r_i), while its remainder

r_{i+2} is placed all the way to the left of of r_{i+1}. The example we worked out before, now looks like this:

$$\begin{array}{c|c|c|c|c|c|c|}
 & 3 & 1 & 3 & 5 & 1 & \\
\hline
0 & 2 & 6 & 8 & 30 & 158 & 188
\end{array} \tag{3.3}$$

There is a beautiful visualization of this process outlined in exercise 3.2.

3.2. A Particular Solution of $ax+by=c$

Another interesting way to encode the computations done in equations (3.1) and (3.2), is via matrices.

$$\begin{pmatrix} r_{i-1} \\ r_i \end{pmatrix} = \begin{pmatrix} q_i & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} r_i \\ r_{i+1} \end{pmatrix} . \tag{3.4}$$

Denote the matrix in this equation by Q_i. Its determinant equals -1, and so it is invertible. In fact,

$$Q_i = \begin{pmatrix} q_i & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad Q_i^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -q_i \end{pmatrix} .$$

These matrices Q_i are very interesting. We will use them again to study the theory of continued fractions in Chapter 6. For now, as we will see in Theorem 3.4, they give us an explicit algorithm to find a solution to the equation $r_1x+r_2y = r\gcd(r_1,r_2)$. Note that from Bézout's lemma (Lemma 2.5), we already know this has a solution. But the next result gives us a simple way to actually calculate a solution. In what follows X_{ij} means the (i,j) entry of the matrix X.

Theorem 3.4. *Give r_1 and r_2, a solution for x and y of $r_1x+r_2y = r\gcd(r_1,r_2)$ is given by*

$$x = r\left(Q_{n-1}^{-1}\cdots Q_2^{-1}\right)_{2,1} \quad \text{and} \quad y = r\left(Q_{n-1}^{-1}\cdots Q_2^{-1}\right)_{2,2} .$$

Proof. Let r_i, q_i, and Q_i be defined as above, and set $r_{n+1}=0$. From equation (3.4), we have

$$\begin{pmatrix} r_i \\ r_{i+1} \end{pmatrix} = Q_i^{-1} \begin{pmatrix} r_{i-1} \\ r_i \end{pmatrix} \quad \Longrightarrow \quad r \begin{pmatrix} r_{n-1} \\ r_n \end{pmatrix} = rQ_{n-1}^{-1} \cdots Q_2^{-1} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} .$$

Observe that $r_{n+1}=0$ and so $\gcd(r_1,r_2)=r_n$ and

$$\begin{pmatrix} r_{n-1} \\ r_n \end{pmatrix} = \begin{pmatrix} x_{n-1} & y_{n-1} \\ x_n & y_n \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} .$$

The theorem follows immediately by setting $x=x_n$ and $y=y_n$. ■

In practice, rather than multiplying all these matrices, it may be more convenient to solve equation (3.1) or (3.2) "backward", as the expression goes. This can be done as follows. Start with

$$\gcd(r_1,r_2) = r_n = r_{n-2} - r_{n-1}\, q_{n-1} \ ,$$

which follows from equation (3.1). The line above it in that same equation gives $r_{n-1} = r_{n-3} - r_{n-2}\, q_{n-2}$. Use this to eliminate r_{n-1} in favor of r_{n-2} and r_{n-3}. So,

$$\begin{aligned} \gcd(r_1,r_2) \quad = \quad r_n \quad &= \quad r_{n-2} - (r_{n-3} - r_{n-2}\, q_{n-2})\, q_{n-1} \\ &= \quad r_{n-2}\,(1+q_{n-1}q_{n-2}) + r_{n-3}(-q_{n-1}) \ . \end{aligned}$$

This computation can be done still more efficiently by employing the notation of equation (3.2) again.

	$+$	$-$	$+$	$-$	$+$	$\cdots$
	q_n	q_{n-1}	q_{n-2}	q_{n-3}	q_{n-4}	$\cdots$
0	r_n	r_{n-1}	r_{n-2}	r_{n-3}	r_{n-4}	$\cdots$
	1					
	0	$-q_{n-1}$	1			
			$q_{n-1}q_{n-2}$	$-q_{n-1}$		
				$-q_{n-3}(1+q_{n-1}q_{n-2})$	$1+q_{n-1}q_{n-2}$	$\cdots$

The algorithm proceeds as follows. Number the columns from right to left, so that r_i (in row 1) and q_i (in row 2) are in the ith column. (The signs in row "0" serve only to keep track of the signs of the coefficients in row 3 and below.) In the first two rows, the algorithm proceeds from right to left.

From r_{i-1} and r_i determine q_i and r_{i+1} by $r_{i-1} = r_i q_i + r_{i+1}$. The division guarantees that these exist, but they may not be unique (see exercise 7.17). In rows 3 and below, the algorithm proceeds from left to right. Each column has at most two non-zero entries. Start with column $n+1$ which has only zeroes and column n which has one 1. The bottom non-zero entry of column i equals the sum of column $i+1$ times q_i times (-1). The top non-zero entry of column i equals the sum of the entries in column $i+2$. Finally, we obtain that $r_n = r_2 x + r_1 y$, where x is the sum of the entries in the 2nd column (rows 3 and below) and y, the sum of the entries (row 3 and below) of the 1st column.

Applying this to the example gives

$$
\begin{array}{c|c|c|c|c|c|c}
 & + & - & + & - & + & - \\
\hline
 & 3 & 1 & 3 & 5 & 1 & 0 \\
\hline
0 & 2 & 6 & 8 & 30 & 158 & 188 \\
\hline
 & 1 & & & & & \\
 & & -1 & 1 & & & \\
 & & & 3 & -1 & & \\
 & & & & -20 & 4 & \\
 & & & & & 21 & -21
\end{array}
\tag{3.5}
$$

Adding the last two lines gives that $2 = 158(25) + 188(-21)$.

3.3. Solution of the Homogeneous equation $ax+by=0$

Proposition 3.5. *The general solution of the homogeneous equation $r_1 x + r_2 y = 0$ is given by*

$$x = k\,\frac{r_2}{\gcd(r_1,r_2)} \qquad \text{and} \qquad y = -k\,\frac{r_1}{\gcd(r_1,r_2)}\ ,$$

where $k \in \mathbb{Z}$.

Proof. On the one hand, by substitution the expressions for x and y into the homogeneous equation, one checks they are indeed solutions. On the other hand, x and y must satisfy

$$\frac{r_1}{\gcd(r_1,r_2)}\,x = -\frac{r_2}{\gcd(r_1,r_2)}\,y\ .$$

The integers $\frac{r_i}{\gcd(r_1,r_2)}$ (for i in $\{1,2\}$) have greatest common divisor equal to 1. Thus Euclid's lemma applies and therefore $\frac{r_1}{\gcd(r_1,r_2)}$ is a divisor of y while $\frac{r_2}{\gcd(r_1,r_2)}$ is a divisor of x. ■

A different proof of this lemma goes as follows. The set of all solution in $\mathbb{R}^2$ of $r_1x+r_2y=0$ is given by the line $\ell := \left\{ t \begin{pmatrix} r_2 \\ -r_1 \end{pmatrix} : t \in \mathbb{R} \right\}$ orthogonal to $\begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$. To obtain all its *lattice points* (i.e., points that are also in $\mathbb{Z}^2$), both tr_2 and $-tr_1$ must be integers. The smallest positive number t for which this is possible, is $t = \frac{1}{\gcd(r_1,r_2)}$.

3.4. The General Solution of $ax+by=c$

Definition 3.6. *Let r_1 and r_2 be given. The equation $r_1x+r_2y=0$ is called homogeneous*[1]*. The equation $r_1x+r_2y=c$ when $c \neq 0$ is called inhomogeneous. An arbitrary solution of the inhomogeneous equation is called a particular solution. By general solution, we mean the set of all possible solutions of the full (homogeneous or inhomogeneous) equation.*

It is useful to have some geometric intuition relevant to the equation $r_1x+r_2y=c$. In $\mathbb{R}^2$, we set $\vec{r} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$, $\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix}$, etcetera. The standard inner product is written as $(\cdot,\cdot)$. The set of points in $\mathbb{R}^2$ satisfying the above inhomogeneous equation thus lie on the line $m \subset \mathbb{R}^2$ given by $(\vec{r},\vec{x}) = c$. This line is orthogonal to the vector $\vec{r}$ and its distance to the origin (measured along the vector $\vec{r}$) equals $\frac{|c|}{\sqrt{(\vec{r},\vec{r})}}$. The situation is illustrated in Figure 13.

It is a standard result from linear algebra that the problem of finding all solutions of a inhomogeneous equation comes down to to finding one

[1]The word "homogeneous" in daily usage receives the emphasis often on its second syllable ("ho-MODGE-uhnus"). However, in mathematics, its emphasis is *always* on the third syllable ("ho-mo-GEE-nee-us"). A probable reason for the daily variation of the pronunciation appears to be conflation with the word "homogenous" (having the same genetic structure). For details, see wiktionary.

solution of the inhomogeneous equation, and finding the general solution of the homogeneous equation.

Lemma 3.7. *Let $(x^{(0)}, y^{(0)})$ be a particular solution of $r_1x + r_2y = c$. The general solution of the inhomogeneous equation is given by $(x^{(0)} + z_1, y^{(0)} + z_2)$ where (z_1, z_2) is the general solution of the homogeneous equation $r_1x + r_2y = 0$.*

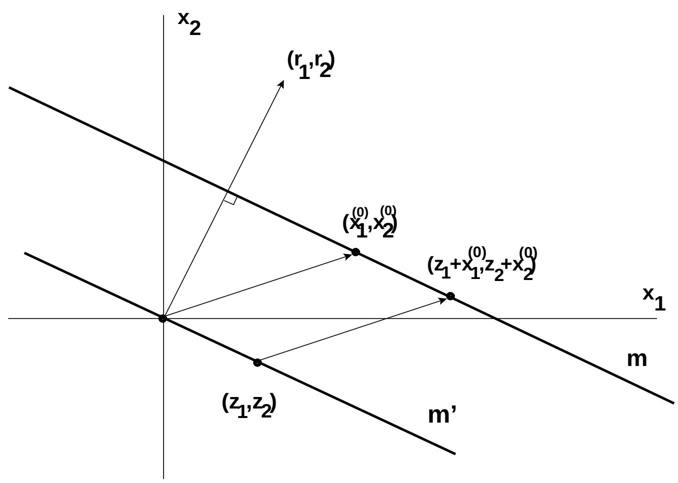


Figure 13. The general solution of the inhomogeneous equation $(\vec{r}, \vec{x}) = c$ in $\mathbb{R}^2$.

Proof. Let $\begin{pmatrix} x^{(0)} \\ y^{(0)} \end{pmatrix}$ be that particular solution. Let m be the line given by $(\vec{r}, \vec{x}) = c$. Translate m over the vector $\begin{pmatrix} -x^{(0)} \\ -y^{(0)} \end{pmatrix}$ to get the line m'. Then an integer point on the line m' is a solution $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ of the homogeneous equation if and only if $\begin{pmatrix} x^{(0)} + z_1 \\ y^{(0)} + z_2 \end{pmatrix}$ on m is also an integer point (see Figure 13). ∎

Bézout's Lemma says that $r_1x + r_2y = c$ has a solution if and only if $\gcd(r_1, r_2) \mid c$. Theorem 3.4 gives a particular solution of that equation (via the Euclidean algorithm). Putting those results and Proposition 3.5 together, gives our final result.

Corollary 3.8. *Given r_1, r_2, and c, the general solution of the equation $r_1x + r_2y = c$, where $\gcd(r_1, r_2) \mid c$, is the sum of a particular solution of Theorem 3.4 and the general* homogeneous *solution of $r_1x + r_2y = 0$ of Proposition 3.5.*

3.5. Recursive Solution of x and y in the Diophantine Equation

Theorem 3.4 has two interesting corollaries. The first is in fact stated in the proof of that theorem, and the second requires a very short proof. We will make extensive use of these two results in Chapter 6 when we discuss continued fractions.

Corollary 3.9. *Given r_1, r_2, and the successive quotients q_2 through q_n as in equation* (3.1)*. Then for $i \in \{3, \cdots, n\}$, the solution for (x_i, y_i) in $r_i = r_1x_i + r_2y_i$ is given by:*

$$\begin{pmatrix} r_i \\ r_{i+1} \end{pmatrix} = Q_i^{-1} \cdots Q_2^{-1} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} .$$

Corollary 3.10. *Given r_1, r_2, and the successive quotients q_2 through q_n as in equation* (3.1)*. Then x_i and y_i of Corollary 3.9 can be solved as follows:*

$$\begin{pmatrix} x_i & y_i \\ x_{i+1} & y_{i+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_i \end{pmatrix} \begin{pmatrix} x_{i-1} & y_{i-1} \\ x_i & y_i \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} .$$

Proof. The initial condition follows, because

$$\begin{array}{rcl} r_1 & = & r_1 \cdot 1 + r_2 \cdot 0 \\ r_2 & = & r_1 \cdot 0 + r_2 \cdot 1 \end{array} .$$

Notice that, by definition,

$$\begin{pmatrix} r_i \\ r_{i+1} \end{pmatrix} = \begin{pmatrix} r_1 x_i + r_2 y_i \\ r_1 x_{i+1} + r_2 y_{i+1} \end{pmatrix} = \begin{pmatrix} x_i & y_i \\ x_{i+1} & y_{i+1} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} .$$

From Corollary 3.9, we now have that

$$\begin{pmatrix} r_i \\ r_{i+1} \end{pmatrix} = Q_i^{-1} \begin{pmatrix} r_{i-1} \\ r_i \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} x_i & y_i \\ x_{i+1} & y_{i+1} \end{pmatrix} = Q_i^{-1} \begin{pmatrix} x_{i-1} & y_{i-1} \\ x_i & y_i \end{pmatrix} .$$

From this, one deduces the equations for x_{i+1} and y_{i+1}. ■

We remark that the recursion in Corollary 3.10 can also be expressed as

$$\begin{array}{rcl} x_{i+1} & = & -q_i x_i + x_{i-1} \\ y_{i+1} & = & -q_i y_i + y_{i-1} \end{array} .$$

3.6. The Chinese Remainder Theorem

We now present an important generalization of these ideas. First we need a small update of Definition 1.2.

Definition 3.11. *Let $\{b_i\}_{i=1}^k$ be non-zero integers. Their greatest common divisor, $\gcd(b_1, \cdots, b_k)$, is the maximum of the numbers that are divisors of every b_i; their least common multiple, $\operatorname{lcm}(b_1, \cdots, b_k)$, is the least of the positive numbers that are multiples of of every b_i.*

Surprisingly, for this more general definition, the generalization of Corollary 2.16 is false. For an example, see exercise 2.7. However, other important properties do generalize.

Lemma 3.12. *Let $\{b_i\}_{i=1}^k$ be non-zero integers.*
(i) If m is a common divisor of the b_i, then $m \mid \gcd(b_1, \cdots b_k)$.
(ii) If M is a common multiple of the b_i, then $\operatorname{lcm}(b_1, \cdots b_k) \mid M$.

Proof. The proof follows from unique factorization and is similar to that of Corollary 2.16. Suppose $b_j = \prod_{i=1}^s p_i^{k_{ij}}$, where $k_{ij} \geq 0$. Set

$$m_i = \min_j k_{ij} \qquad \text{and} \qquad M_i = \max_j k_{ij} ,$$

Then

$$\gcd(b_1,\cdots b_k)=\prod_{i=1}^{s} p_i^{m_i} \qquad \text{and} \qquad \operatorname{lcm}(b_1,\cdots b_k)=\prod_{i=1}^{s} p_i^{M_i}\,.$$

Any common divisor of the b_i must be equal to $\prod_{i=1}^{s} p_i^{\ell_i}$ with $\ell_i \leq m_i$ and similar for common multiples. ∎

Theorem 3.13 (Chinese Remainder Theorem). *Let $n=\prod_{i=1}^{k} b_i$, where b_i are positive integers such that* $\gcd(b_j,b_i)=1$ *for $i\neq j$. The set of solutions of*

$$\forall i \in \{1,\cdots,k\}\ :\ z =_{b_i} c_i$$

is given by

$$z =_n \sum_{j=1}^{k} \frac{n}{b_j} x_j c_j \quad \textit{where } x_i \textit{ satisfies} \quad \frac{n}{b_i} x_i =_{b_i} 1\,.$$

Proof. Note that $\gcd(n/b_i,b_i)=1$. So by Bézout, there are x_i and y_i (for $i\in\{1,\cdots,k\}$) so that

$$\frac{n}{b_i} x_i + b_i y_i = 1 \quad \Longleftrightarrow \quad \frac{n}{b_i} x_i =_{b_i} 1\,.$$

For these x_i, we have

$$\sum_{j=1}^{k} \frac{n}{b_j} x_j =_{b_i} 1\,.$$

Thus $z=\sum_{j=1}^{k} \frac{n}{b_j} x_j c_j$ is a particular solution. By Lemma 3.12, the homogeneous equation has solution $z =_n 0$. The proof is completed by observing that the general solution is the sum of a particular solution plus the solutions to the homogeneous equation. ∎

3.7. Polynomials

In this section, we illustrate that the division and Euclidean algorithms have much wider applications than just the integers, see also exercises 2.2 and 2.4.

Definition 3.14. *A polynomial f in $\mathbb{Q}[x]$ of positive degree is irreducible over Q if it cannot be written as a product of two polynomials in $\mathbb{Q}[x]$ with positive degree. Recall (Definition 1.18) that f is minimal polynomial in*

$\mathbb{Q}[x]$ for ρ if f is a non-zero polynomial in $\mathbb{Q}[x]$ of minimal degree such that $f(\rho) = 0$.

Definition 3.15. *Let f and g in $\mathbb{Q}[x]$. The greatest common divisor of f and g, or $\gcd(f,g)$, is a polynomial in $R[x]$ with maximal degree that is a factor of both f and g. The least common multiple of f and g, or $\operatorname{lcm}(f,g)$, is a polynomial in $\mathbb{Q}[x]$ with minimal degree that has both f and g as factors.*

Remark 3.16. If p is minimal for ρ, it must be irreducible, because if not, one of its factors with smaller degree would also have ρ as a root.

We mention without proof (but see exercise 2.2) that in $\mathbb{Q}[x]$ the division algorithm holds: given r_1 and r_2, then there are q_2 and r_3 such that

$$r_1 = r_2 q_2 + r_3 \quad \text{such that} \quad \text{degree}(r_3) < \text{degree}(r_1)\,.$$

Remark 3.17. To make this valid without exceptions, we adopt the convention that the degree of a non-zero constant equals 0, while the degree of 0 equals $-\infty$. For example, if $r_1 = r_2 = 1$, the inequality for r_3 still holds. The student is likely already familiar with these facts.

It is important to understand that for this to work, division of coefficients is essential. For example, with coefficients in $\mathbb{Z}$, we cannot express $2x^2+1$ as a multiple of $3x+1$ plus a remainder of smaller degree. However, in $\mathbb{Q}[x]$ we can divide coefficients and thus follow the reasoning of Section 2.1 and show the following. See also exercise 3.22).

The gcd of two polynomials can be computed in the same two ways we have seen before, and the proofs are the same. One is done by factoring both polynomials and multiplying together the common factors to the lowest power as in the proof of Corollary 2.23. Note though that factoring polynomials is hard. The other is applying the Euclidean Algorithm as in equation (3.1). An example is given in exercise 3.22. The relation between lcm and gcd of two polynomials is the same as in the proof of Corollary 2.23.

3.8. Exercises

Exercise 3.1. Let ℓ be the line in $\mathbb{R}^2$ given by $y = \rho x$, where $\rho \in \mathbb{R}$.
a) Show that ℓ intersects $\mathbb{Z}^2$ if and only if ρ is rational.
b) Given a rational $\rho > 0$, find the intersection of ℓ with $\mathbb{Z}^2$. (*Hint: set $\rho = \frac{r_1}{r_2}$ and use Proposition 3.5.*)

Exercise 3.2. This problem was taken (and reformulated) from [**30**].
a) Tile a 188 by 158 rectangle by squares using what is called a greedy algorithm [a]. The first square is 158 by 158. The remaining rectangle is 158 by 30. Now the optimal choice is five 30 by 30 squares. What remains is an 30 by 8 rectangle, and so on. Explain how this is a visualization of equation (3.3). See Figure 14.
b) Consider equation (3.1) or (3.2) and use a) to show that

$$r_1 r_2 = \sum_{i=2}^{n} q_i r_i^2 \,.$$

(*Hint: assume that* $r_1 > r_2 > 0$, $r_n \neq 0$, *and* $r_{n+1} = 0$.)

[a] By "greedy" we mean that at every step, you choose the biggest square possible and as many of them as possible. In general a greedy algorithm always makes a locally optimal choice.

Figure 14. A 'greedy' (or locally best) algorithm to tile the the 188×158 rectangle by squares. The 3 smallest — and barely visible — squares are 2×2. Note how the squares spiral inward as they get smaller. See exercise 3.13.

Exercise 3.3. In (3.1), assume that $r_1 > r_2 > 0$. What happens if you start the Euclidean algorithm with $r_2 = r_1 \cdot 0 + r_3$ instead of $r_1 = r_2 \cdot q_2 + r_3$?

Exercise 3.4. Apply the Euclidean algorithm to find the greatest common divisor of the following number pairs. (*Hint: replace negative numbers by positive ones. For the division algorithm applied to these pairs* (r_1, r_2)*, see exercise 2.1*)
a) 110 , 7.
b) 51 , -30.
c) -138 , 24.
d) 272 , 119.
e) 2378 , 1769.
f) 270 , 175560.

Exercise 3.5. Determine if the following Diophantine equations admit a solution for x and y. If yes, find a (particular) solution. (*Hint: Use one of the algorithms in Section 3.2.*)
a1) $110x+7y=13$.
a2) $110x+7y=5$.
b1) $51x-30y=6$.
b2) $51x-30y=7$.
c1) $-138x+24y=7$.
c2) $-138x+24y=6$.
d1) $272x+119y=54$.
d2) $272x+119y=17$.
e1) $2378x+1769y=300$.
e2) $2378x+1769y=57$.
f1) $270x+175560y=170$.
f2) $270x+175560y=150$.

Exercise 3.6. Find all solutions for x and y of the following (homogeneous) Diophantine equations. (*Hint: Use one of the algorithms in Section 3.2.*)
a) $110x+7y=0$.
b) $51x-30y=0$.
c) $-138x+24y=0$.
d) $272x+119y=0$.
e) $2378x+1769y=0$.
f) $270x+175560y=0$.

Exercise 3.7. Find the *general* solution for x and y in all problems of exercise 3.5 that admit a solution. (*Hint: use Corollary 3.8.*)

Exercise 3.8. Use Corollary 3.10 to express x_i and y_i in the successive remainders r_i in each of the items in exercise 3.4. (*Hint: you need to know the* q_i *for each item in exercise 3.4.*)

Exercise 3.9. Consider the line ℓ in $\mathbb{R}^3$ defined by $\ell(\xi) = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} \xi$, where $\xi \in \mathbb{R}$ and the r_i are integers.
a) Show that $\ell(\xi) \in \mathbb{Z}^3 \backslash \{\vec{0}\}$ if and only if $\xi = \frac{t}{\gcd(r_1,r_2,r_3)}$ and $t \in \mathbb{Z}$.
b) Show that this implies that if any of the r_i is irrational, then ℓ has no non-zero points in common with $\mathbb{Z}^3$.

Definition 3.18. *The sequence $\{F_i\}_{i=0}^{\infty}$ of Fibonacci numbers F_i is defined as follows*

$$F_0 = 0\,, \quad F_1 = 1\,, \quad \forall\, i > 1\,:\, F_{i+1} = F_i + F_{i-1}\,.$$

Exercise 3.10. Denote the golden mean , or $\frac{1+\sqrt{5}}{2} \approx 1.618$, by g.
a) Show that $g^2 = g+1$ and thus for $n \in \mathbb{Z}$: $g^{n+1} = g^n + g^{n-1}$.
b) Show that $F_3 \geq g^1$ and $F_2 \geq g^0$.
c) Use induction to show that $F_{n+2} \geq g^n$ for $n > 0$.
d) Use the fact that $5\log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \approx 1.045$, to show that $F_{5k+2} > 10^k$ for $k \geq 0$.

Exercise 3.11. Consider the equations in (3.1) and assume that $r_{n+2} = 0$ and $r_{n+1} > 0$.
a) Show that $r_{n+1} \geq F_2 = 1$ and $r_n \geq F_3 = 2$. (*Hint: $r(i)$ is strictly increasing.*)
b) Show that $r_1 \geq F_{n+2}$.
c) Suppose r_1 and r_2 in $\mathbb{N}$ and $\max\{r_1,r_2\} < F_{n+2}$. Show that the Euclidean Algorithm to calculate $\gcd(r_1,r_2)$ takes at most $n-1$ iterates of the division algorithm.

Exercise 3.12. Use exercises 3.10 and 3.11 to show that the Euclidean Algorithm to calculate $\gcd(r_1,r_2)$ takes at most $5k-1$ iterates where k is the number of decimal places of $\max\{r_1,r_2\}$. (*This is known as Lamé's theorem.*)

Exercise 3.13. Apply the greedy algorithm of exercise 3.2 (a) to the rectangle whose sides have length 1 and g (see exercise 3.10 (a)). At step 0, we start with the 1×1 square.
a) Use exercise 3.10 (a) to show at that step i, you get one $g^{-i} \times g^{-i}$ square (see Figure 15).
b) Use exercise 3.2 (b) to show that $g = \sum_{i=0}^{\infty} g^{-2i}$.
c) Use this construction, but now with a $F_{n+1} \times F_n$ Fibonacci rectangle, to show that $F_{n+1}F_n = \sum_{i=1}^{n} F_i^2$. For F_i, see Definition 3.18.
d) Show that in polar coordinates (r, θ) the red spiral connecting the corners of the squares in Figure 15 is given by $r = Cg^{2\theta/\pi}$ for some C.(*Note: this is called the golden spiral.*)

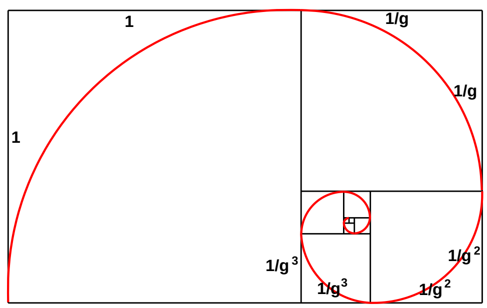


Figure 15. The greedy algorithm of exercise 3.2 (a) applied to the golden mean rectangle. The spiral connecting the corners of the square is known as the golden spiral. (In actual fact we used a 55 by 34 rectangle as an approximation. An approximation to a true spiral was created by fitting circular segments to the corners.)

Exercise 3.14. a) Write the numbers 287, 513, and 999 in base 2, 3, and 7, using the division algorithm. Do not use a calculating device. (*Hint: start with base 10. For example:*

$$\begin{array}{rcl} 287 & = & 28 \cdot 10 + \underline{\mathbf{7}} \\ 28 & = & 2 \cdot 10 + \underline{\mathbf{8}} \\ 2 & = & 0 \cdot 10 + \underline{\mathbf{2}} \end{array}$$

Hence the number in base 10 is $\underline{\mathbf{2}} \cdot 10^2 + \underline{\mathbf{8}} \cdot 10^1 + \underline{\mathbf{7}} \cdot 10^0$.)
b) Show that to write n in base b takes about $\log_b n$ divisions.

Exercise 3.15. Use Theorem 3.13 to solve:

$$\begin{aligned} z &=_2 1 \\ z &=_3 2 \\ z &=_5 3 \\ z &=_7 5\,. \end{aligned}$$

Exercise 3.16. The Fibonacci numbers F_n are defined in Definition 3.18.
a) Use the method of equation (3.1) to show that $\gcd(F_n, F_{n+1}) = 1$.
b) Determine the q_i in (a).
c) Use recursion to show that $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}^n = \begin{pmatrix} F_{n-1} & F_n \\ F_n & F_{n+1} \end{pmatrix}$.
d) Show that (c) implies that $F_{n+1}F_{n-1} - F_n^2 = (-1)^n$. (*Hint: in (c) take the determinant.*)

Exercise 3.17. Use Theorem 3.13 to solve:

$$\begin{aligned} z &=_{F_n} F_{n-1} \\ z &=_{F_{n+1}} F_n\,. \end{aligned}$$

where F_n are the Fibonacci numbers of Definition 3.18. (*Hint: you need to use exercise 3.16 (a) and (d).*)

Exercise 3.18. (*The Chinese remainder theorem generalized.*) Suppose $\{b_i\}_{i=1}^n$ are positive integers. We want to know all z that satisfy

$$z =_{b_i} c_i \qquad \text{for } i \in \{1, \cdots n\}\,.$$

a) Set $B = \operatorname{lcm}(b_1, b_2, \cdots b_n)$ and show that the homogeneous problem is solved by

$$z =_B 0\,.$$

b) Show that if there is a particular solution then

$$\forall i \neq j \,:\, c_i =_{\gcd(b_i, b_j)} c_j\,.$$

c) Formulate the general solution when the condition in (b) holds.

Exercise 3.19. Use exercise 3.18 to solve:

$$\begin{aligned} z &=_6 15 \\ z &=_{10} 6 \\ z &=_{15} 10\,. \end{aligned}$$

See also exercise 2.7.

Exercise 3.20. There is a reformulation of the Euclidean algorithm that will be very useful in Chapter 6.
a) Rewrite the example in Section 3.1 as follows.

$$\begin{aligned}
\frac{30}{158} &= \frac{188}{158} - 1 \\
\frac{8}{30} &= \frac{158}{30} - 5 \\
\frac{6}{8} &= \frac{30}{8} - 3 \\
\frac{2}{6} &= \frac{8}{6} - 1
\end{aligned} .$$

Note that the right hand side is a fraction *minus* its integer part.
b) Now rewrite this again as

$$\begin{aligned}
\frac{30}{158} &= \frac{1}{158/188} - 1 \\
\frac{8}{30} &= \frac{1}{30/158} - 5 \\
\frac{6}{8} &= \frac{1}{8/30} - 3 \\
\frac{2}{6} &= \frac{1}{6/8} - 1
\end{aligned} .$$

Exercise 3.21. a) Apply the Euclidean algorithm to $(r_1, r_2) = (14142, 10000)$. (*Hint: you should get* $(q_2, \cdots, q_{10}) = (1,2,2,2,2,2,1,1,29)$.)
b) Show that for $i \in \{2, \cdots, 8\}$:

$$\frac{r_{i+1}}{r_i} = \frac{1}{r_{i-1}/r_i} - \left\lfloor \frac{1}{r_{i-1}/r_i} \right\rfloor ,$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. (*Hint: see also exercise 3.20.*)

Exercise 3.22. For this exercise, read Section 3.7 carefully. All polynomials are in $\mathbb{Q}[x]$ (that is: with coefficients in $\mathbb{Q}$). Let $p_1(x) = x^7 - x^2 + 1$, $p_2(x) = x^3 + x^2$, and $e(x) = 2 - x$.
a) Use the Euclidean Algorithm to determine $\gcd(p_1, p_2)$. *Hint: We list the steps of the Euclidean algorithm:*

$$\begin{array}{ccllcl} (x^7 - x^2 + 1) & = & (x^3 + x^2) & (x^4 - x^3 + x^2 - x + 1) & + & (-2x^2 + 1) \\ (x^3 + x^2) & = & (-2x^2 + 1) & (-\frac{1}{2}x - \frac{1}{2}) & + & (\frac{1}{2}x + \frac{1}{2}) \\ (-2x^2 + 1) & = & (\frac{1}{2}x + \frac{1}{2}) & (-4x + 4) & + & (-1) \\ (\frac{1}{2}x + \frac{1}{2}) & = & (-1) & (-\frac{1}{2}x - \frac{1}{2}) & + & (0) \end{array},$$

b) Explain why there are polynomials g_p and h_p such that

$$p_1(x) g_p(x) + p_2(x) h_p(x) = e(x).$$

c) Use "backward solving" to find a *particular solution* of the equation in (b).
d) Find the general (homogeneous) solution of

$$p_1(x) g_0(x) + p_2(x) h_0(x) = 0.$$

e) Use (c) and (d) to give the general solution of the inhomogeneous equation (the one in (b)).

Exercise 3.23. All polynomials are in $\mathbb{Q}[x]$. Let $p(x)$ be a polynomial and $p'(x)$ its derivative.
a) Show that if $p(x)$ has a multiple root λ of order $k > 1$, then $p'(x)$ has that same root of order $k - 1$. (*Hint: Differentiate* $p(x) = h(x)(x - \lambda)^k$.)
b) Use exercise 3.22, to give an algorithm to find a polynomial $q(x)$ that has the same roots as $p(x)$, but all roots are simple (i.e. no multiple roots). (*Hint: you need to divide* p *by* $\gcd(p, p')$.)

Exercise 3.24. Assume that every polynomial f of degree $d \geq 1$ has at least 1 root, prove the fundamental theorem of algebra. (*Hint: let* ρ *be a root and use the division algorithm to write* $f(x) = (x - \rho)q(x) + r$ *where* r *has degree 0.*)

In Proposition 11.20, we will prove that every polynomial with complex coefficients has at least one zero in $\mathbb{C}$. Together with the result of exercise 3.24, this establishes the following important theorem.

Theorem 3.19 (Fundamental Theorem of Algebra). *A polynomial in* $\mathbb{C}[x]$ *(the set of polynomials with complex coefficients) of degree* $d \geq 1$ *has exactly* d *roots, counting multiplicity.*

Exercise 3.25. Let f and p be polynomials in $\mathbb{Q}[x]$ with root ρ and suppose that p is minimal (Definition 1.18). Show that $p \mid f$. (*Hint: use the division algorithm and 2.4 to write* $f(x) = p(x)q(x) + r(x)$ *where* r *has degree less than* g.)

Chapter 4

Number Theoretic Functions

Overview. We study *number theoretic functions*. These are functions defined on the positive integers with values in $\mathbb{C}$. In the context of number theory, the value typically depends on the arithmetic nature of its argument (i.e. whether it is a prime, and so forth), rather than just on the size of its argument. An example is $\tau(n)$ which equals the number of positive divisors of n.

4.1. Multiplicative Functions

Definition 4.1. *Number theoretic functions, arithmetic functions, or sequences are functions defined on the positive integers (i.e.* $\mathbb{N}$*) with values in* $\mathbb{C}$.

Note that outside number theory, the term *sequence* is the one that is most commonly used. We will use these terms interchangeably.

Definition 4.2. *A multiplicative function is a sequence such that* $\gcd(a,b) = 1$ *implies* $f(ab) = f(a)f(b)$. *A completely multiplicative function is one where the condition that* $\gcd(a,b) = 1$ *is not needed.*

Note that completely multiplicative implies multiplicative (but not vice versa). The reason this definition is interesting, is that it allows us to evaluate the value of a multiplicative function f on any integer as long as we can compute $f(p^k)$ for any prime p. Indeed, using the fundamental theorem of arithmetic,

$$\text{if} \quad n = \prod_{i=1}^{r} p_i^{\ell_i} \qquad \text{then} \qquad f(n) = \prod_{i=1}^{r} f(p_i^{\ell_i}) \,,$$

as follows immediately from Definition 4.2.

Proposition 4.3. *Let f be a multiplicative function on the integers. Then*

$$F(n) = \sum_{d|n} f(d)$$

is also multiplicative.

Proof. Let $n = \prod_{i=1}^{s} p_i^{\ell_i}$. The summation $\sum_{d|n} f(d)$ can be written out using the previous lemma and the fact that f is multiplicative:

$$\begin{aligned} F(n) &= \textstyle\sum_{a_1=0}^{\ell_1} \cdots \sum_{a_s=0}^{\ell_s} f(p_1^{a_1}) \cdots f(p_r^{a_r}) \\ &= \textstyle\prod_{i=1}^{s} \left(\sum_{a_i=0}^{\ell_i} f(p_i^{a_i}) \right) \;. \end{aligned}$$

Exercise 4.3 provides a visual explanation for the second equality.

Now let a and b two integers greater than 1 and such that $\gcd(a,b) = 1$ and $ab = n$. Then by the unique factorization theorem a and b can be written as:

$$a = \prod_{i=1}^{r} p_i^{\ell_i} \qquad \text{and} \qquad b = \prod_{i=r+1}^{s} p_i^{\ell_i}$$

Applying the previous computation to a and b yields that $f(a)f(b) = f(n)$. ■

Perhaps the simplest multiplicative functions are the ones where $f(n) = n^k$ for some fixed k. Indeed, $f(n)f(m) = n^k m^k = f(nm)$. In fact, this is a completely multiplicative function. Thus Proposition 4.3 implies that the functions σ_k defined below are multiplicative.

Definition 4.4. *Let $k \in \mathbb{R}$. The multiplicative function $\sigma_k : \mathbb{N} \to \mathbb{R}$ gives the sum of the k-th power of the* positive *divisors of n. Equivalently:*

$$\sigma_k(n) = \sum_{d|n} d^k \;.$$

Note that the multiplicativity of σ_k follows directly from Proposition 4.3. Special cases are when $k = 1$ and $k = 0$. In the first case, the function is simply the sum of the positive divisors and the subscript '1' is usually dropped. When $k = 0$, the function is usually called τ, and the function's value is the number of positive divisors of its argument.

Theorem 4.5. *Let $n = \prod_{i=1}^{r} p_i^{\ell_i}$ where the p_i are primes. Then for $k \neq 0$*

$$\sigma_k(n) = \prod_{i=1}^{r} \left(\frac{p_i^{k(\ell_i+1)} - 1}{p_i^k - 1} \right) ,$$

while for $k = 0$

$$\sigma_0(n) = \tau(n) = \prod_{i=1}^{r} (\ell_i + 1) \ .$$

Proof. By Proposition 4.3, $\sigma_k(n)$ is multiplicative, so it is sufficient to compute for some prime p:

$$\sigma_k(p^\ell) = \sum_{i=0}^{\ell} p^{ik} = \frac{p^{k(\ell+1)} - 1}{p^k - 1} .$$

Thus $\sigma_k(n)$ is indeed a product of these terms. ■

However, there are other interesting multiplicative functions beside the powers of the divisors. The Möbius function defined below is one of these, as we will see.

Definition 4.6. *The Möbius function $\mu : \mathbb{N} \to \mathbb{Z}$ is given by:*

$$\mu(n) = \begin{cases} 1 & \text{if} \quad n = 1 \\ 0 & \text{if} \quad \exists p > 1 \text{ prime} \quad \text{with} \quad p^2 \mid n \\ (-1)^r & \text{if} \quad n = p_1 \cdots p_r \quad \text{and} \quad p_i \text{ are distinct primes} \end{cases} .$$

Definition 4.7. *We say that n is square free if there is no prime p such that $p^2 \mid n$.*

Lemma 4.8. *The Möbius function μ is multiplicative.*

Proof. By unique factorization, we are allowed to assume that

$$n = ab \qquad \text{where} \qquad a = \prod_{i=1}^{r} p_i^{\ell_i} \qquad \text{and} \qquad b = \prod_{i=r+1}^{s} p_i^{\ell_i} \ .$$

If a equals 1, then $\mu(ab) = \mu(a)\mu(b) = 1\mu(b)$, and similar if $b = 1$. If either a or b is not square free, then neither is $n = ab$, and so in that case, we again have $\mu(ab) = \mu(a)\mu(b) = 0$. If *both* a and b are square free, then r (in the definition of μ) is *strictly additive* and so $(-1)^r$ is strictly multiplicative, hence multiplicative. ■

4.2. Additive Functions

Also important are the additive functions to which we will return in Chapter 12.

Definition 4.9. *An additive function is a sequence such that* $\gcd(a,b) = 1$ *implies* $f(ab) = f(a) + f(b)$*. A completely addititive function is one where the condition that* $\gcd(a,b) = 1$ *is not needed.*

Here are some examples.

Definition 4.10. *Let* $\omega(n)$ *denote the number of* distinct *prime divisors of n and let* $\Omega(n)$ *denote the* total *number of prime divisors of n. These functions are called the prime omega functions.*

So if $n = \prod_{i=1}^{s} p_i^{\ell_i}$, then

$$\omega(n) = s \qquad \text{and} \qquad \Omega(n) = \sum_{i=1}^{s} \ell_i \,.$$

The additivity of ω and the complete additivity of Ω should be clear. By way of example, since $72 = 2^3 \cdot 3^2$, $\omega(72) = 2$ while $\Omega(72) = 5$.

4.3. Möbius inversion

Lemma 4.11. *Define* $\varepsilon(n) \equiv \sum_{d|n} \mu(d)$*. Then* $\varepsilon(1) = 1$ *and for all* $n > 1$*,* $\varepsilon(n) = 0$*.*

Proof. Lemma 4.8 says that μ is multiplicative. Therefore, by Proposition 4.3, ε is also multiplicative. It follows that $\varepsilon(\prod_{i=1}^{r} p_i^{\ell_i})$ can be calculated by evaluating a product of terms like $\varepsilon(p^\ell)$ where p is prime. For example, when p is prime, we have

$$\begin{aligned} \varepsilon(p) &= \mu(1) + \mu(p) = 1 + (-1) = 0 \qquad \text{and} \\ \varepsilon(p^2) &= \mu(1) + \mu(p) + \mu(p^2) = 1 - 1 + 0 = 0\,. \end{aligned}$$

Thus one sees that $\varepsilon(p^\ell)$ is zero *unless* $\ell = 0$. ■

Lemma 4.12. *For $n \in \mathbb{N}$, define*

$$S_n \equiv \left\{(a,b) \in \mathbb{N}^2 : \exists d > 0 \text{ such that } d \mid n \text{ and } ab = d\right\} \text{ and}$$
$$T_n \equiv \left\{(a,b) \in \mathbb{N}^2 : b \mid n \text{ and } a \mid \frac{n}{b}\right\} .$$

Then $S_n = T_n$.

Proof. Suppose (a,b) is in S_n. Then $ab \mid n$ and so

$$\left.\begin{array}{r} ab = d \\ d \mid n \end{array}\right\} \implies b \mid n \quad \text{and} \quad a \mid \frac{n}{b} .$$

And so (a,b) is in T_n. Vice versa, if (a,b) is in T_n, then by setting $d \equiv ab$, we get

$$\left.\begin{array}{r} b \mid n \\ a \mid \dfrac{n}{b} \end{array}\right\} \implies d \mid n \quad \text{and} \quad ab = d .$$

And so (a,b) is in S_n. ■

Theorem 4.13. (Möbius inversion) *Let $F : \mathbb{N} \to \mathbb{C}$ be any number theoretic function and μ the Möbius function. Then the following equation holds*

$$F(n) = \sum_{d|n} f(d)$$

if and only if $f : \mathbb{N} \to \mathbb{C}$ satisfies

$$f(d) = \sum_{a|d} \mu(a) F\left(\frac{d}{a}\right) = \sum_{\{(a,b)\,:\,ab=d\}} \mu(a) F(b) .$$

Proof. $\Longleftarrow$: We show that substituting f gives F. Define H as

$$H(n) \equiv \sum_{d|n} f(d) = \sum_{d|n} \sum_{a|d} \mu(a) F\left(\frac{d}{a}\right) .$$

Then we need to prove that $H(n) = F(n)$. This proceeds in three steps. For the first step we write $ab = d$, so that now

$$H(n) \equiv \sum_{d|n} f(d) = \sum_{d|n} \sum_{ab=d} \mu(a) F(b) . \tag{4.1}$$

For the second step we apply Lemma 4.12 to the set over which the summation takes place. This gives:

$$H(n) = \sum_{b|n} \sum_{a|\frac{n}{b}} \mu(a) F(b) = \sum_{b|n} \left(\sum_{a|\frac{n}{b}} \mu(a) \right) F(b) . \tag{4.2}$$

Finally, Lemma 4.11 implies that the term in parentheses equals $\varepsilon\left(\frac{n}{b}\right)$. This equals 0, except when $b = n$ when it equals 1. The result follows.

$\Longrightarrow$: By the previous part, we already know one solution for f if we are given that $F(n) = \sum_{d|n} f(d)$. So suppose there are two solutions f and g. We have:

$$F(n) = \sum_{d|n} f(d) = \sum_{d|n} g(d) .$$

We show by induction on n that $f(n) = g(n)$.

Clearly $F(1) = f(1) = g(1)$. Now suppose that for $i \in \{1, \cdots k\}$, we have $f(i) = g(i)$. Then

$$F(k+1) = \left(\sum_{d|(k+1),\, d \le k} f(d) \right) + f(k+1) = \left(\sum_{d|(k+1),\, d \le k} g(d) \right) + g(k+1) .$$

The desired equality for $k+1$ follows from the induction hypothesis. ■

Remark 4.14. It is important that multiplicativity plays no role in this argument.

4.4. Euler's Phi or Totient Function

Definition 4.15. *Euler's phi function, also called the totient function is defined as follows: $\varphi(n)$ equals the number of integers in $\{1, \cdots n\}$ that are relative prime to n (see Figure 16).*

Lemma 4.16 (Gauss' Theorem). *For $n \in \mathbb{N}$: $n = \sum_{d|n} \varphi(d)$.*

Proof. Define $S(d,n)$ as the set of integers m between 1 and n such that $\gcd(m,n) = d$:

$$S(d,n) = \{m \in \mathbb{N} : m \le n \text{ and } \gcd(m,n) = d\} .$$

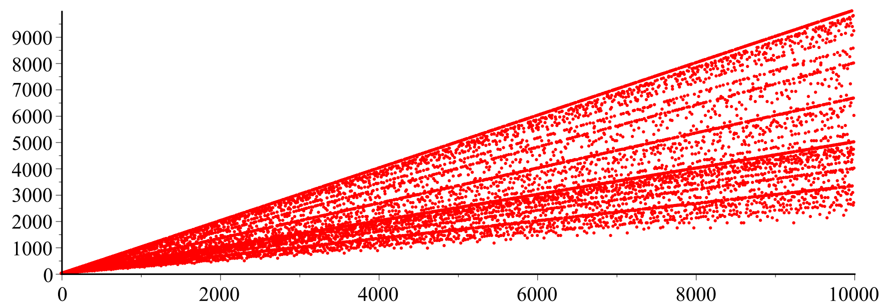


Figure 16. The totient function $\varphi(n)$ versus n. Its subtle structure is clearly visible, see also exercise 4.7.

Since every for natural number $m \leq n$ has a unique $\gcd(m,n)$ which is a divisor of n, we get

$$n = \sum_{d|n} |S(d,n)| \, .$$

Because the definition of S_n can be rewritten as

$$S(d,n) = \left\{ m \in \mathbb{N} : m \leq n \text{ and } \gcd\left(\frac{m}{d}, \frac{n}{d}\right) = 1 \right\} ,$$

the cardinality $|S(d,n)|$ of $S(d,n)$ is given by $\varphi\left(\frac{n}{d}\right)$. Thus we obtain:

$$n = \sum_{d|n} |S(d,n)| = \sum_{d|n} \varphi\left(\frac{n}{d}\right) \, .$$

As d runs through all divisors of n in the last sum, so does $\frac{n}{d}$. Therefore the last sum is equal to $\sum_{d|n} \varphi(d)$, which proves the lemma. ■

Theorem 4.17. *Let $\prod_{i=1}^{r} p_i^{\ell_i}$ be the prime power factorization of n. Then $\varphi(n) = n \prod_{i=1}^{r} \left(1 - \frac{1}{p_i}\right)$.*

Proof. [1] Apply Möbius inversion to Lemma 4.16:

$$\varphi(d) = \sum_{a|d} \mu(a) \frac{d}{a} = d \sum_{a|d} \frac{\mu(a)}{a} \, . \tag{4.3}$$

[1] There is a conceptually simpler — but in its details much more challenging — proof if you are familiar with the inclusion-exclusion principle. We review that proof in exercise 4.13.

The functions μ and $a \to \frac{1}{a}$ are multiplicative. It is easy to see that the product of two multiplicative functions is also multiplicative. Therefore φ is also multiplicative (Proposition 4.3). Thus for n as given,

$$\varphi(n) = \varphi\left(\prod_{i=1}^{r} p_i^{\ell_i}\right) = \prod_{i=1}^{r} \varphi\left(p_i^{\ell_i}\right) . \tag{4.4}$$

So it is sufficient to evaluate the function φ on prime powers. Noting that the divisors of the prime power p^ℓ are $\{1, p, \cdots p^\ell\}$, we get from equation (4.3)

$$\varphi(p^\ell) = p^\ell \sum_{j=0}^{\ell} \frac{\mu(p^j)}{p^j} = p^\ell \left(1 - \frac{1}{p}\right) .$$

Substituting this into equation (4.4) completes the proof. ■

From this proof we obtain the following corollary.

Corollary 4.18. *Euler's phi function is multiplicative.*

4.5. Dirichlet and Lambert Series

We will take a quick look at some interesting series without worrying too much about their convergence, because we are ultimately interested in the analytic continuations that underlie these series. For that, it is sufficient that there is convergence in any open non-empty region of the complex plane.

Definition 4.19. *Let f, g, and F be arithmetic functions (see Definition 4.1). Define the Dirichlet convolution of f and g, denoted by $f * g$, as*

$$(f * g)(n) \equiv \sum_{ab=n} f(a)g(b) .$$

This convolution is a very handy tool. Similar to the usual convolution of sequences, one can think of it as a sort of multiplication. It pays off to first define a few standard number theoretic functions.

Definition 4.20. *We use the following notation for certain standard sequences. The sequence $\varepsilon(n)$ is 1 if $n = 1$ and otherwise returns 0, $\mathbf{1}(n)$ always returns 1, and $I(n)$ returns n (so $I(n) = n$).*

The function ε acts as the identity of the convolution. Indeed,

$$(\varepsilon * g)(n) = \sum_{ab=n} \varepsilon(a)g(b) = g(n)\,.$$

Note that $I(n)$ is the identity *as a function*, but should not be confused with the identity *of the convolution* (ε). In other words, $I(n) = n$ but $I * f \neq f$.

We can now do some very *cool*[2] things of which we can unfortunately give but a few examples. As a first example, the Möbius inversion of Theorem 4.13

$$F(n) = \sum_{d|n} f(d) \quad \Longleftrightarrow \quad f(d) = \sum_{\{(a,b)\,:\,ab=d\}} \mu(a)\,F(b)\;,$$

can be more succinctly translated as follows:

$$F = \mathbf{1} * f \quad \Longleftrightarrow \quad f = \mu * F\,. \tag{4.5}$$

This leads to the next example. The first of the following equalities holds by Lemma 4.16, the second follows from Möbius inversion (4.5).

$$I = \mathbf{1} * \varphi \quad \Longleftrightarrow \quad \varphi = \mu * I\,. \tag{4.6}$$

And the best of these examples is gotten by substituting the identity ε for F in equation (4.5):

$$\varepsilon = \mathbf{1} * f \quad \Longleftrightarrow \quad f = \mu * \varepsilon = \mu\,. \tag{4.7}$$

Thus μ is the convolution inverse of the sequence $(1,1,1\cdots)$. This immediately leads to an unexpected[3] expression for $1/\zeta(z)$ of equation (4.8).

Definition 4.21. *Let $f(n)$ is an arithmetic function (or sequence). A* _Dirichlet series_ *is a series of the form $F(z) = \sum_{n=1}^{\infty} f(n)n^{-z}$. Similarly, a* _Lambert series_ *is a series of the form $F(x) = \sum_{n=1}^{\infty} f(n)\,\frac{x^n}{1-x^n}$.*

The prime example of a Dirichlet series is – of course – the Riemann zeta function of Definition 2.19, $\zeta(z) = \sum \mathbf{1}(n)\,n^{-z}$.

Lemma 4.22. *For the product of two Dirichlet series we have*

$$\left(\sum_{n=1} f(n)n^{-z}\right)\left(\sum_{n=1} g(n)n^{-z}\right) = \sum_{n=1}^{\infty} (f * g)(n)\,n^{-z}\,.$$

[2] A very unusual word in mathematics textbooks.

[3] The fact that this follows so easily, justifies the use of the word referred to in the previous footnote

Proof. This follows easily from re-arranging the terms in the product:

$$\sum_{a,b\geq 1}^{\infty} \frac{f(a)g(b)}{(ab)^z} = \sum_{n=1}^{\infty} \left(\sum_{ab=n} f(a)g(b) \right) n^{-z}.$$

We collected the terms with $ab = n$. ■

Can we find $f(n)$ such that $\frac{1}{\zeta(z)} = \sum f(n)(n)\, n^{-z}$? Yes! Because Lemma 4.22 translates $1 = \zeta(z) \cdot \frac{1}{\zeta(z)}$ as

$$\varepsilon = \mathbf{1} * f.$$

And equation (4.7) gives that $f = \mu$, or

$$\frac{1}{\zeta(z)} = \sum_{n\geq 1} \frac{\mu(n)}{n^z}. \tag{4.8}$$

Recall from Chapter 2 that one of the chief concerns of number theory is the location of the non-real zeros of ζ. At stake is Conjecture 2.22 which states that all its non-real zeros are on the line $\mathrm{Re}\, z = 1/2$. The original definition of the zeta function is as a series that is absolutely convergent for $\mathrm{Re}\, z > 1$ only. Equation (4.8) converges in that same region, and so establishes that at least in $\mathrm{Re}\, z > 1$ there are no zeroes. A (weak) partial result in the direction of the Riemann Hypothesis!

It is also important to establish that the analytic continuation of ζ is valid for *all* $z \neq 1$. The next result serves as a first indication that $\zeta(z)$ can indeed be continued for values $\mathrm{Re}\, z \leq 1$.

Corollary 4.23. *Let ζ be the Riemann zeta function and σ_k as in Definition 4.4, then*

$$\zeta(z-k)\zeta(z) = \sum_{n=1}^{\infty} \frac{\sigma_k(n)}{n^z}.$$

Proof.

$$\zeta(z-k)\zeta(z) = \sum_{a\geq 1} a^{-z} \sum_{b\geq 1} b^k b^{-z} = \sum_{n\geq 1} n^{-z} \sum_{b\,|\,n} b^k.$$

■

Lemma 4.24. *A Lambert series can re-summed as follows:*

$$\sum_{n=1}^{\infty} f(n) \frac{x^n}{1-x^n} = \sum_{n=1}^{\infty} (\mathbf{1} * f)(n)\, x^n.$$

Proof. First use that

$$\frac{x^b}{1-x^b} = \sum_{a=1}^{\infty} x^{ab}\,.$$

This gives that

$$\sum_{b=1} f(b)\,\frac{x^b}{1-x^b} = \sum_{b=1}^{\infty}\sum_{a=1}^{\infty} f(b)\,x^{ab}\,.$$

Now set $n = ab$ and collect terms. Noting that $(\mathbf{1}*f)(b) = \sum_{b\nmid n} f(b)$ yields the result. ■

Corollary 4.25. *The following equality holds*

$$\sum_{n\geq 1} \varphi(n)\,\frac{x^n}{1-x^n} = \frac{x}{(1-x)^2}\,.$$

Proof. We have

$$\sum_{n\geq 1} \varphi(n)\,\frac{x^n}{1-x^n} = \sum_{n\geq 1} (\mathbf{1}*\varphi)(n)\,x^n = \sum_{n\geq 1} I(n)\,x^n\,.$$

The first equality follows from Lemma 4.24 and the second from Lemma 4.16. The last sum can be computed as $x\,\frac{d}{dx}(1-x)^{-1}$ which gives the desired expression. ■

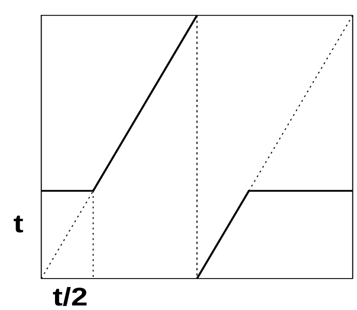


Figure 17. A one parameter family f_t of maps from the circle to itself. For every $t \in [0,1]$ the map f_t is constructed by truncating the map $x \to 2x \mod 1$ as indicated in this figure.

The last result is of importance in the study of dynamical systems. In figure 17, the map f_t is constructed by truncating the map $x \to 2x \mod 1$

for $t \in [0,1]$. Corollary 4.25 can be used to show that the set of t for which f_t does not have a periodic orbit has measure ("length") zero [**71**, **72**], even though that set is uncountable.

4.6. Exercises

Exercise 4.1. Decide which functions are not multiplicative, multiplicative, or completely multiplicative (see Definition 4.2).
a) $f(n) = 1$.
b) $f(n) = 2$.
c) $f(n) = \sum_{i=1}^{n} i$.
d) $f(n) = \prod_{i=1}^{n} i$.
e) $f(n) = n$.
f) $f(n) = n^k$.
g) $f(n) = \sum_{d|n} d$.
h) $f(n) = \prod_{d|n} d$.

Exercise 4.2. a) Let $h(n) = 0$ when n is even, and 1 when n is odd. Show that h is multiplicative.
b) Now let $H(n) = \sum_{d|n} h(d)$. Show without using Proposition 4.3 that H is multiplicative. (*Hint: write $a = 2^k \prod_{i=1}^{r} p_i^{\ell_i}$ by unique factorization, where the p_i are odd primes. Compute the number of odd divisors. Similarly for b.*)
c) What does Proposition 4.3 say?

Exercise 4.3. In Figure 18 a large volume in $\mathbb{R}^3$ with coordinates x, y, and z is chopped up into smaller rectangular boxes of dimensions x_i by y_j by z_k as indicated. See the proof of Proposition 4.3.
a) Show that the volume of the big box equals $\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} x_i y_j z_k$. (*Hint: add the volumes of the small boxes.*)
b) Show that the volume of the big box equals $\left(\sum_{i=1}^{n_1} x_i\right)\left(\sum_{j=1}^{n_2} y_j\right)\left(\sum_{k=1}^{n_3} z_k\right)$. (*Hint: compute the dimensions of the big box.*)

Exercise 4.4. a) Compute the numbers $\sigma_1(n) = \sigma(n)$ of Definition 4.4 for $n \in \{1, \cdots, 30\}$ without using Theorem 4.5.
b) What is the *only* value n for which $\sigma(n) = n$?
c) Show that $\sigma(p) = p + 1$ whenever p is prime.
d) Use (c) and multiplicativity of σ to check the list obtained in (a).
e) For what values of n in the list of (a) is $n \mid \sigma(n)$? (*Hint: 6 and 28.*)

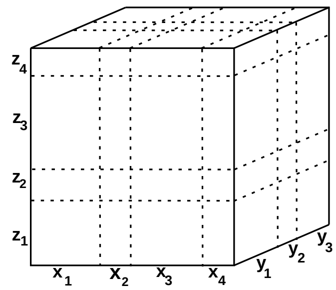


Figure 18. Two ways of computing the volume of a big box: add the volumes of the small boxes, or compute the dimensions of the big box.

Exercise 4.5. a) Compute the numbers $\sigma_0(n) = \tau(n)$ of Definition 4.4 for $n \in \{1, \cdots, 30\}$ without using Theorem 4.5.
b) What is the *only* value n for which $\tau(n) = 1$?
c) Show that $\tau(p) = 2$ whenever p is prime.
d) Use (c) and multiplicativity of τ to check the list obtained in (a).

Exercise 4.6. a) Compute the numbers $\varphi(n)$ of Definition 4.15 for $n \in \{1, \cdots, 30\}$ without using Theorem 4.17.
b) What is $\varphi(p)$ when p is a prime?
c) How many positive numbers less than pn are *not* divisible by p?
d) Use (c) and multiplicativity of φ to check the list obtained in (a).

Exercise 4.7. Consider Figure 19 and prove the following statements.
a) $\lim_{i\to\infty} \varphi(p_i)/p_i = 1$. (*Hint: use Theorem 4.17.*)
b) $\lim_{i\to\infty} \varphi(2p_i)/2p_i = 1/2$, where p_i are the odd primes.
c) If $n_i = \prod_1^k p_i$, then $\lim_{i\to\infty} \varphi(n_i)/n_i = 0$. (*Hint: use Theorem 4.17 and Proposition 2.20 plus the fact that $\sum n^{-1}$ diverges. Note that the n_i form a useful basis for number systems if you want to minimize division, see exercise 1.25*)

Exercise 4.8. a) Compute the numbers $\mu(n)$ of Definition 4.6 for $n \in \{1, \cdots, 30\}$.
b) What is $\mu(p)$ when p is a prime?
c) Use (c) and multiplicativity of μ to check the list obtained in (a).

Exercise 4.9. Let $\tau(n)$ be the number of distinct *positive* divisors of n. Answer the following question without using Theorem 4.5.
a) Show that τ is multiplicative.
b) If p is prime, show that $\tau(p^k) = k + 1$.
c) Use the unique factorization theorem, to find an expression for $\tau(n)$ for $n \in \mathbb{N}$.

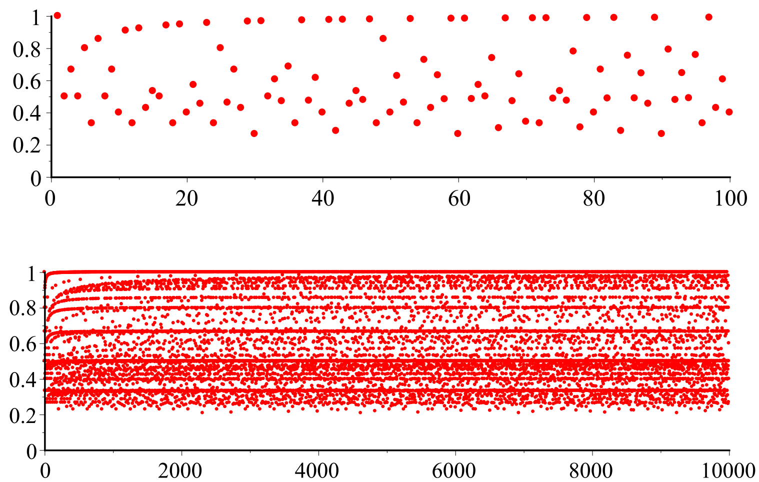


Figure 19. The fraction of numbers in $\{1, \cdots, n\}$ relatively prime to n, or $\varphi(n)/n$, versus n.

Exercise 4.10. Two positive integers a and b are called *amicable* if $\sigma(a) = \sigma(b) = a + b$. The smallest pair of amicable numbers is is formed by 220 and 284.
a) Use Theorem 4.5 to show that 220 and 284 are amicable.
b) The same for 1184 and 1210.

Exercise 4.11. A positive integer n is called *perfect* if $\sigma(n) = 2n$.
a) Show that n is perfect if and only if the sum of its positive divisors less than n equals n.
b) Show that if p and $2^p - 1$ are primes, then $n = 2^{p-1}(2^p - 1)$ is perfect. (*Hint: use Theorem 4.5 and exercise 4.4(c).*)
c) Use exercise 1.12 to show that if $2^p - 1$ is prime, then p is prime, and thus $n = 2^{p-1}(2^p - 1)$ is perfect.
d) Check that this is consistent with the list in exercise 4.4.

Exercise 4.12. Draw the following directed graph G: the set of vertices V represent 0 and the natural numbers between 1 and 50. For $a, b \in V$, a directed edge ab exists if $\sigma(a) - a = b$. Finally, add a loop at the vertex representing 0. Notice that every vertex has 1 outgoing edge, but may have more than 1 incoming edge.
a) Find the cycles of length 1 (loops). The non-zero of these represent perfect numbers.
b) Find the cycles of length 2 (if any). A pair of numbers a and b that form a cycle of length 2 are called *amicable numbers*. Thus for such a pair[a] , $\sigma(b) - b = a$ and $\sigma(a) - a = b$.
c) Find any longer cycles. Numbers represented by vertices in longer cycles are called *sociable numbers*.
d) Find numbers whose path ends in a cycle of length 1. These are called *aspiring numbers*.
e) Find numbers (if any) that have no incoming edge. These are called *untouchable numbers*.
f) Determine the paths starting at 2193 and at 562. (*Hint: both end in a cycle (or loop).*)

[a] As of 2017, about 10^9 amicable number pairs have been discovered.

A path through this graph is called an *aliquot sequence*. The so-called *Catalan-Dickson conjecture* says that every aliquot sequence ends in some finite cycle (or loop). However, even for a relatively small number such as 276, it is unknown (in 2017) whether its aliquot sequence ends in a cycle.

Exercise 4.13. In this exercise, we give a different proof of Theorem 4.17. It uses the principle of inclusion-exclusion [**61**]. We state it here for completeness. Let S be a finite set with subsets A_1, A_2, and so on through A_r. Then, if we denote the cardinality of a set A by $|A|$,

$$\left|S-\bigcup_{i=1}^{r} A_i\right| = |S|-|S_1|+|S_2|-\cdots+(-1)^r|S_r| \,, \qquad (4.9)$$

where $|S_\ell|$ is the *sum of the sizes of all intersections of* ℓ *members* of $\{A_1,\cdots A_r\}$.
Now, in the following we keep to these conventions. Using prime factorization, write

$$n = \textstyle\prod_{i=1}^r p_i^{k_i} \,,$$
$$A_i = \{z\in S\,|\,p_i \text{ divides } z\} \,.$$
$$S=\{1,2\cdots n\} \quad \text{and} \quad R=\{1,2\cdots r\} \,,$$
$$I_\ell \subseteq R \quad \text{such that} \quad |I_\ell| = \ell \,.$$

a) Show that $\varphi(n) = \left|S-\bigcup_{i=1}^r A_i\right|$. (*Hint: any number that is not co-prime with n is a multiple of at least one of the* p_i.)
b) Show that $\left|A_i\right| = \frac{n}{p_i}$.
c) Show that $\left|\bigcap_{i\in I_\ell} A_i\right| = n\prod_{i\in I_\ell}\frac{1}{p_i}$. (*Hint: use Lemma 3.12.*)
d) Show that $\left|S_\ell\right| = n\sum_{I_\ell\subseteq R}\prod_{i\in I_\ell}\frac{1}{p_i}$.
e) Show that the principle of inclusion-exclusion implies that $\left|S-\bigcup_{i=1}^r A_i\right| = n+n\sum_{\ell=1}^r(-1)^\ell\sum_{I_\ell\subseteq R}\prod_{i\in I_\ell}\frac{1}{p_i}$.
f) Show that $n+n\sum_{\ell=1}^r(-1)^\ell\sum_{I_\ell\subseteq R}\prod_{i\in I_\ell}\frac{1}{p_i} = n\prod_{i=1}^r(1-\frac{1}{p_i})$. Notice that this implies Theorem 4.17. (*Hint: write out the product* $\prod_{i=1}^r(1-\frac{1}{p_i})$.)

Exercise 4.14. Let $F(n) = n = \sum_{d|n} f(n)$. Use the Möbius inversion formula (or $f(n) = \sum_{d|n}\mu(d)F(\frac{n}{d})$) to find $f(n)$. (*Hint: substitute the Möbius function of Definition 4.6 and use multiplicativity where needed.*)

Exercise 4.15. a) Compute the sets S_n and T_n of Lemma 4.12 *explicitly* for $n=4$ and $n=12$.
b) Perform the resummation done in equations 4.1 and 4.2 explicitly for $n=4$ and $n=12$.

Exercise 4.16. Recall the definition of Dirichlet convolution $f * g$ of the arithmetic functions f and g (Definition 4.19).
a) Show that the set A of arithmetic functions with addition forms an Abelian group (see Definition 5.19).
b) Show that Dirichlet convolution is *associative*[a], that is:

$$(f * g) * h = f * (g * h)\ .$$

c) Show that Dirichlet convolution is *distributive* over addition, that is:

$$f * (g + h) = f * g + f * h\ .$$

d) The binary operation Dirichlet convolution has an identity ε (Definition 4.20), defined by

$$f * \varepsilon = \varepsilon * f = f\ .$$

Show that the function ε of Lemma 4.11 is the identity of the convolution.
e) Show that Dirichlet convolution is *commutative*, that is:

$$f * g = g * f\ .$$

(*Note: In this exercise we proved that the set of arithmetic functions with addition and convolution is a commutative ring, see Definitions 5.20 and 5.26. This ring is sometimes called the* *Dirichlet ring* .)

[a] Associativity is a property whose importance is sometimes hush-hushed a bit. We chose to elaborate it, see exercise 5.23

Exercise 4.17. Use exercise 4.16 to prove the following:
a) Show that the Dirichlet convolution of two multiplicative functions is multiplicative.
b) Show that the sum of two multiplicative functions is not necessarily multiplicative. (*Hint:* $\varepsilon + \varepsilon$.)

Exercise 4.18. See Definition 4.10. Define $f(n) \equiv \tau(n^2)$ and $g(n) \equiv 2^{\omega(n)}$.
a) Compute $\omega(n)$, $f(n)$, and $g(n)$ for n equals 10^n and $6!$.
b) For p prime, show that $\tau(p^{2k}) = \sum_{d|p^k} 2^{\omega(d)} = 2k+1$. (*Hint: use Theorem 4.5.*)
c) Show that f is multiplicative. (*Hint: use that τ is multiplicative.*)
d) Use (d) to show that g is multiplicative.
e) Show that

$$\tau(n^2) = \sum_{d|n} 2^{\omega(d)}\ .$$

Exercise 4.19. Let $S(n)$ denote the number of square free divisors of n with $S(1) = 1$ and $\omega(n)$ the number of distinct prime divisors of n. See also Definition 4.10.
a) Show that $S(n) = \sum_{d|n} |\mu(d)|$. (*Hint: use Definition 4.6.*)
b) Show that $S(n) = 2^{\omega(n)}$. (*Hint: let W be the set of prime divisors of n. Then every square free divisor corresponds to a subset — product — of those primes. How many subsets of primes are there in W?*)
c) Conclude that

$$\sum_{d|n} |\mu(d)| = 2^{\omega(n)} .$$

Exercise 4.20. Define the *Liouville λ-function* by $\lambda(1) = 1$ and $\lambda(n) = (-1)^{\Omega(n)}$.
a) Compute $\lambda(10^n)$ and $\lambda(6!)$.
b) Show that λ is multiplicative. (*Hint: $\Omega(n)$ is completely additive.*)
c) Use Proposition 4.3 to show that $F(n) = \sum_{d|n} \lambda(d)$ is multiplicative.
d) For p prime, show that

$$\sum_{d|p^k} \lambda(d) = \sum_{i=0}^{k} (-1)^i$$

which equals 1 if k is even and 0 if k is odd.
e) Use (c) and (d) to conclude that

$$F(n) = \sum_{d|n} \lambda(d) = \begin{cases} 1 & \text{if } n = m^2 \\ 0 & \text{else} \end{cases} .$$

Exercise 4.21. Let f be a multiplicative function.
Define $q(n) \equiv \sum_{d|n} \mu(d)f(d)$, where μ is the Möbius function.
a) Show that $f(1) = 1$.
b) Show that $f\mu$ (their product) is multiplicative.
c) Use Proposition 4.3 to show that $q(n)$ is multiplicative.
d) Show that if p is prime, then $q(p^k) = f(1) - f(p) = 1 - f(p)$.
e) Use (c) and (d) to show that

$$q(n) = \sum_{d|n} \mu(d)f(d) = \prod_{p \text{ prime},\ p|n} (1 - f(p)) .$$

Exercise 4.22. Use exercise 4.21 (e) and the definition of ω in exercise 4.18 and λ in exercise 4.20 to show that

$$\sum_{d|n} \mu(d)\lambda(d) = 2^{\omega(n)} .$$

Exercise 4.23. a) Show that for all $n \in \mathbb{N}$, $\mu(n)\mu(n+1)\mu(n+2)\mu(n+3) = 0$. (*Hint: divisibility by 4.*)
b) Show that for any integer $n \geq 3$, $\sum_{k=1}^{n} \mu(k!) = 1$. (*Hint: use (a).*)

Exercise 4.24. a) Use Euler's product formula and the sequence μ of Definition 4.6 to show that

$$\frac{1}{\zeta(z)} = \prod_{p \text{ prime}} \left(1 - p^{-z}\right) = \prod_{p \text{ prime}} \left(\sum_{i \geq 0} \mu(p^i) p^{-iz}\right).$$

b) Without using equation (4.7), prove that the expression in (a) equals $\sum_{n \geq 1} \mu(n) n^{-z}$. (*Hint: since μ is multiplicative, you can write a proof re-arranging terms as in the first proof of Euler's product formula.*)

Exercise 4.25. a) Use equation (4.8) to show that

$$\frac{\zeta(z-1)}{\zeta(z)} = \sum_{a \geq 1} \frac{a}{a^z} \sum_{b \geq 1} \frac{\mu(b)}{b^z}.$$

b) Show that $I * \mu = \varphi$.
c) Use Lemma 4.22, (a), and (b) to show that

$$\frac{\zeta(z-1)}{\zeta(z)} = \sum_{n \geq 1} \frac{\varphi(n)}{n^z}.$$

Exercise 4.26. a) Use Corollary 4.23 to show that

$$\zeta(z-k) = \sum_{a \geq 1} \frac{\sigma_k(a)}{a^z} \sum_{b \geq 1} \frac{\mu(b)}{b^z}.$$

b) Show that

$$\zeta(z-k) = \sum_{n \geq 1} (\sigma_k * \mu)(n) n^{-z},$$

where $*$ means the Dirichlet convolution (Definition 4.19).

Exercise 4.27. Show that $\zeta(z)$ has no zeroes and no poles in the region $\Re(z) > 1$. (*Hint: use that $\zeta(z)$ converges for $\Re(z) > 1$ and* (4.8).)

Chapter 5

Modular Arithmetic and Primes

Overview. We return to the study of primes in $\mathbb{N}$. This is related to the study of modular arithmetic (the properties of addition and multiplication in $\mathbb{Z}_b$), because $a \in \mathbb{N}$ is a prime if and only if there are no non-trivial divisors or, expressed differently, there is no $0 < b < a$ so that $a =_b 0$. Modular arithmetic concerns itself with computations involving addition and multiplication in $\mathbb{Z}$ modulo b, denoted by $\mathbb{Z}_b$, i.e. calculations with residues modulo b (see Definition 1.8). One common way of looking at this is to consider integers x and y that differ by a multiple of b as *equivalent* (see exercise 5.1). We write $x \sim y$. One then proves that the usual addition and multiplication is well-defined for these equivalence classes. This is done in exercise 5.2.

5.1. Euler's Theorem and Primitive Roots

The *order* of an element g is the smallest positive integer k such that $g * g * \cdots * g$, repeated k times and usually written as g^k, equals e. One can show that the elements $\{e, g, g^2, \cdots, g^{k-1}\}$ also form a *group* (Definition 5.19). More details can be found in [**28**], [**54**], or [**34**]. In the case at hand, $\mathbb{Z}_b$, we have a structure with two operations, namely addition with identity element 0 and multiplication with identity element 1. We could therefore define the order of an element in $\mathbb{Z}_b$ with respect to addition and with respect to multiplication. As an example, we consider the element 3 in $\mathbb{Z}_7$ (see Figure 20):

$$3+3+3+3+3+3+3 =_7 0 \quad \text{and} \quad 3\cdot 3\cdot 3\cdot 3\cdot 3\cdot 3 =_7 1 \,.$$

The first gives 7 as the <u>additive</u> <u>order</u> of 3, and the second gives 6 for the <u>multiplicative</u> <u>order</u>. For our current purposes, however, it is sufficient to work only with the multiplicative version.

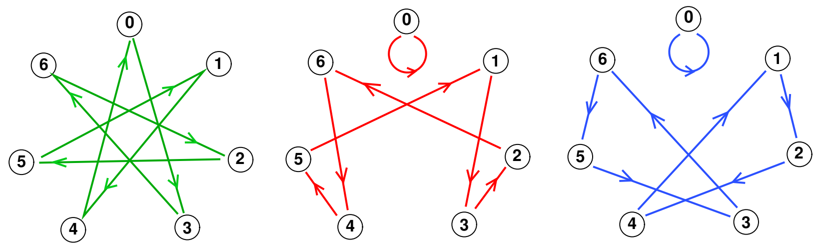


Figure 20. Left, the orbits in $\mathbb{Z}_7$ under addition of 3; middle, the orbits under multiplication by 3; and right, the orbit under multiplication by 2. Observe that the multiplicative graphs are not connected and have less symmetries.

Definition 5.1. *The (multiplicative) <u>order</u> of <u>a modulo b</u>, written as* $\mathrm{Ord}_b^\times(a)$, *is the smallest positive number k such that* $a^k =_b 1$. *(If there is no such k, the order is* ∞.*)*

Recall that φ denotes Euler's phi or totient function (Definition 4.15).

Definition 5.2. *i) A <u>complete set of residues</u> modulo b is a set of b integers in* $\mathbb{Z}$ *that has exactly one integer in each congruence class (modulo b).*
ii) A <u>reduced set of residues</u> modulo b is the subset of (i) of elements that are relatively prime to b ($\gcd(a,b)=1$*).*

As an example, the set $\{0,1,2,\cdots,11\}$ is a complete set of residues modulo 12, while $\{1,5,7,11\}$ is a reduced set of residues modulo 12.

Lemma 5.3. *Suppose* $\gcd(a,b)=1$. *If the numbers* $\{x_i\}$ *form a complete set of residues modulo b (reduced set of residues modulo b), then* $\{ax_i\}$ *is a complete set of residues modulo b (reduced set of residues modulo b).*

Proof. Let $\{x_i\}$ be a *complete* set of residues modulo b. Then the b numbers $\{ax_i\}$ form complete set of residues *unless* two of them are congruent. But that is impossible by Theorem 2.7.

Let $\{x_i\}$ be a *reduced* set of residues modulo b. Then, as above, no two of the $\varphi(b)$ numbers $\{ax_i\}$ are congruent modulo b. Furthermore, Lemma 2.15 implies that if $\gcd(a,b) = 1$ and $\gcd(x_i,b) = 1$, then $\gcd(ax_i,b) = 1$. Thus the set $\{ax_i\}$ is a reduced set of residues modulo b. ■

Theorem 5.4 (Euler). *Let* $a,b > 1$ *and* $\gcd(a,b) = 1$. *Then* $a^{\varphi(b)} =_b 1$.

Proof. Let $\{x_i\}_{i=1}^{\varphi(b)}$ be a reduced set of residues modulo b. Then by Lemma 5.3, $\{ax_i\}_{i=1}^{\varphi(b)}$ is a reduced set of residues modulo b. Because multiplication is commutative, we get

$$\prod_{i=1}^{\varphi(b)} x_i =_b \prod_{i=1}^{\varphi(b)} ax_i =_b a^{\varphi(b)} \prod_{i=1}^{\varphi(b)} x_i$$

Since $\gcd(x_i,a) = 1$, Lemma 2.15 implies that $\gcd\left(\prod_{i=1}^{\varphi(b)} x_i, a\right) = 1$. The cancelation theorem applied to the equality between the first and third terms proves the result. ■

Euler's theorem says that $\varphi(b)$ is a multiple of $\mathrm{Ord}_b^\times(a)$. But it does not say *what* multiple. In fact, in practice, that question is difficult to decide. It is of theoretical importance to decide when the two are equal.

Definition 5.5. *Let a and b positive integers with* $\gcd(a,b) = 1$. *If* $\mathrm{Ord}_b^\times(a) = \varphi(b)$, *then a is called a* <u>*primitive*</u> <u>*root*</u> *modulo b.*

For example, the smallest integer k for which $3^k =_7 1$ is 6 (see Figure 20). Since $\varphi(7) = 6$, we see that 3 is a primitive root of 7. Since multiplication is well-defined in $\mathbb{Z}_7$, it follows that $(3+7k)^6 =_7 3^6 =_7 1$. Thus $\{\cdots -4, 3, 10, \cdots\}$ are all primitive roots of 7. The only other non-congruent primitive root of 7 is 5. Not all numbers have primitive roots. For instance, 8 has none.

The importance of the notion of primitive root is perhaps more easily remembered via the next lemma.

Lemma 5.6. *a is a primitive root modulo b if and only if the orbit* $\{a^i \mod b\}_{i=1}^{\varphi(b)}$ *contains all reduced residues modulo b.*

Proof. If a is a primitive root, then all values of $\{a^i \mod b\}_{i=1}^{\varphi(b)}$ must be distinct, because if $a^i = a^j$ for some $i > j$ in $\{1, \cdots, \varphi(b)\}$, then $a^{i-j} =_b 1$, contradicting that a is a primitive root.

We prove the contrapositive[1] of the other direction. If $a^i =_b 1$ for some positive i less than $\varphi(b)$, then $a^{i+1} =_b a$ and the numbers start repeating so that $\{a^i \mod b\}_{i=1}^{\varphi(b)}$ cannot contain all reduced residues modulo b. ■

The salient fact about prime roots is that we know exactly when they occur. An accessible proof of Theorem 5.7 (i) can be found in [**18**]chapter 8 and part (ii) in [**4**]chapter 10.

Theorem 5.7. *i) An integer n has a primitive root if and only if n equals 1, 2, 4, p^k, or $2p^k$, where p is an odd prime and $k \geq 1$.*
ii) If n has a primitive root g, then it has $\varphi(\varphi(n))$ primitive roots given by g^i for every i such that $\gcd(i, \varphi(n)) = 1$.

The primitive root also has interesting connections with day-to-day arithmetic, namely the expression of rational numbers in any base. We use base 10 as an example.

Proposition 5.8. *Let a and n greater than 0 and* $\gcd(a,n) = \gcd(10,n) = 1$. *The expansion of a/n in base 10 is non-terminating and eventually periodic with period p, where (i) $p = \mathrm{Ord}_n^\times(10)$ and (ii) $p \mid \varphi(n)$.*

Proof. The proof proceeds in steps, each of which uses the division algorithm. Start by reducing a modulo n and call the result r_0.

$$a = nq_0 + r_0\,,$$

where $r_0 \in \{0, \cdots n-1\}$. Lemma 3.1 implies that $\gcd(a,n) = \gcd(r_0,n) = 1$. So in particular, $r_0 \neq 0$. The integer part of a/n is q_0. The next step is:

$$\frac{r_1}{n} := \left\{\frac{10r_0}{n}\right\} \qquad \text{or} \qquad 10r_0 = nq_1 + r_1\,,$$

where again $r_1 \in \{0, \cdots n-1\}$.

Note that $0 \leq 10r_0 < 10n$ and so $q_1 \in \{0, \cdots 9\}$. We now record the first digit "after the decimal point" of the decimal expansion: q_1. By Lemma 3.1,

[1]The contrapositive of $(P \Rightarrow Q)$ is $(\neg Q \Rightarrow \neg P)$ (or: not Q implies not P) and holds if and only if the former holds.

we have $\gcd(10r_0,n) = \gcd(r_1,n)$. In turn, this implies via Lemma 2.15 that $\gcd(r_0,n) = \gcd(r_1,n)$. And again, we see that $r_1 \neq 0$.

The process now repeats itself.

$$\frac{r_2}{n} := \left\{\frac{10r_1}{n}\right\} \qquad \text{or} \qquad 10\underbrace{(10r_0 - nq_1)}_{r_1} = nq_2 + r_2\,,$$

and we record the second digit after the decimal dot, $q_2 \in \{0,\cdots 9\}$. By the same reasoning, $\gcd(r_2,n) = 1$ and so $r_2 \neq 0$. One continues and proves by induction that $\gcd(r_i,n) = 1$. In particular, $r_i \neq 0$, so the expansion does not terminate.

Since the remainders r_i are in $\{1,\cdots n-1\}$, the sequence must be *eventually* periodic with (least positive) period p. At that point, we have

$$10^{k+p}r_0 =_n 10^k r_0\,.$$

By Theorem 2.7, we can cancel the common factors 10^k and r_0, and we obtain that $10^p =_n 1$. Since p is the least such (positive) number, we have proved (i). Item (ii) follows directly from Euler's Theorem. ■

Of course, this proposition easily generalizes to computations in any other base b. As an en example, we mention that if $\gcd(a,n) = 1$ and b is a primitive root of n, then the expansion of a/b has period $\varphi(n)$.

The next result follows by setting $y = x + k\varphi(b)$ in a^y and applying Euler's theorem. It has important applications in cryptography.

Corollary 5.9. *Let a and b be coprime with* $b > 1$.

$$x =_{\varphi(b)} y \quad\Longrightarrow\quad a^x =_b a^y\,.$$

5.2. Fermat's Little Theorem and Primality Testing

Euler's theorem has many other important consequences. It implies what is known as Fermat's little theorem, although it was not proved by Fermat himself, since, as he writes in the letter in which he stated the result, he feared "its being too long" [**18**][Section 5.2]. Not an isolated case, it would appear!

Corollary 5.10 (Fermat's little theorem). *If p is prime and* $\gcd(a,p) = 1$*, then* $a^{p-1} =_p 1$.

This follows from Euler's Theorem by noticing that for a prime p, $\varphi(p) = p - 1$. There is an equivalent formulation which allows p to be a divisor of a. Namely, if p is prime, then $a^p =_p a$. Notice that if $p \mid a$, then both sides are congruent to 0.

Primes are of great theoretical and practical value (think of encryption, for example). Algorithms for primality testing are therefore very useful. The simplest test to find out if some large number n is prime, consists of course of applying some version of Eratosthenes' sieve to the positive integers less than or equal to $\sqrt{n}$. To carry this out, we will have to perform on the order of $\sqrt{n}$ divisions.

Another possibility is to use the converse of Fermat's little theorem (Corollary 5.10). If n and p are distinct primes, we know that $p^{n-1} =_n 1$. The Fermat primality test for n consists of testing, for example, whether $2^{n-1} =_n 1$. If that fails, we know that n is not prime. However, the converse of Fermat's little theorem is not true! So even if $2^{n-1} =_n 1$, it could be that n is not prime; we will discuss this possibility at the end of this section. As it turns out, primality testing via Fermat's little theorem can be done much faster than the naive method, provided one uses fast *modular exponentiation* algorithms. We briefly illustrate this technique by computing 11^{340} modulo 341.

Start by expanding 340 in base 2 as done in exercise 3.14, where it was shown that this takes on the order of $\log_2 340$ (long) divisions.

$$\begin{array}{rcl}
340 & = & 170\cdot 2+\mathbf{0}\\
170 & = & 85\cdot 2+\mathbf{0}\\
85 & = & 42\cdot 2+\mathbf{1}\\
42 & = & 21\cdot 2+\mathbf{0}\\
21 & = & 10\cdot 2+\mathbf{1}\\
10 & = & 5\cdot 2+\mathbf{0}\\
5 & = & 2\cdot 2+\mathbf{1}\\
2 & = & 1\cdot 2+\mathbf{0}\\
1 & = & 0\cdot 2+\mathbf{1}
\end{array}$$

And so

$$340 = 101010100 \quad \text{in base } 2\ .$$

Next, compute a table of powers 11^{2^i} modulo 341, as done below. This can be done using very few computations. For instance, once $11^8 =_{341} 143$

has been established, the next up is found by computing 143^2 modulo 341, which gives 330, and so on. So this takes about $\log_2 340$ multiplications.

0	11^1	$=_{341}$	11
0	11^2	$=_{341}$	121
1	11^4	$=_{341}$	319
0	11^8	$=_{341}$	143
1	11^{16}	$=_{341}$	330
0	11^{32}	$=_{341}$	121
1	11^{64}	$=_{341}$	319
0	11^{128}	$=_{341}$	143
1	11^{256}	$=_{341}$	330

The first column in the table thus obtained now tells us which coefficients in the second we need to compute the result.

$$11^{340} =_{341} 330 \cdot 319 \cdot 330 \cdot 319 =_{341} 132\,.$$

Again, this takes no more than $\log_2 340$ multiplications. Thus altogether, for a number n and a computation in base b, this takes on the order of $2\log_b n$ multiplications plus $\log_b n$ divisions[2]. For large numbers, this is much more efficient than the $\sqrt{n}$ of the naive method.

As mentioned, the drawback is that we can get *false positives*. While there are partial converses to Fermat's little theorem, they do not yield computationally efficient improvements (see exercise 5.20).

Definition 5.11. *The number $n \in \mathbb{N}$ is called a* <u>*pseudoprime*</u> *to the base b if* $\gcd(b,n) = 1$ *and* $b^{n-1} =_n 1$ *but nonetheless n is composite. (When the base is 2, the clause* to the base 2 *is often dropped.)*

Some numbers pass all tests to every base and are still composite. These are called Carmichael numbers. The smallest Carmichael number is 561. It has been proved [**57**] that there are infinitely many of them.

Definition 5.12. *The number $n \in \mathbb{N}$ is called a* <u>*Carmichael*</u> <u>*number*</u> *if it is composite and it is a pseudoprime to every base.*

---

[2] Divisions take more computations than multiplications. We do not pursue this here.

The smallest pseudoprime is 341, because $2^{340} =_{341} 1$ while $341 = 11 \cdot 31$. In this case, one can still show that 341 is not a prime by using a different base: $3^{340} =_{341} 56$. Thus by Fermat's little theorem, 341 cannot be prime.

The reason that the method sketched here is still useful is that pseudoprimes are very much rarer than primes. The numbers below $2.5 \cdot 10^{10}$ contain on the order of 10^9 primes. At the same time, this set contains only 21853 pseudoprimes to the base 2. There are only 1770 integers below $2.5 \cdot 10^{10}$ that are pseudoprime to the bases 2, 3, 5, and 7. Thus if a number passes these four tests, it is overwhelmingly likely that it is a prime.

5.3. Fermat and Mersenne Primes

Through the ages, back to early antiquity, people have been fascinated by numbers, such as 6, that are the sum of their positive divisors other than itself, to wit: 6=1+2+3. Mersenne and Fermat primes, primes of the form $2^k \pm 1$, have also attracted centuries of attention. Note that if p is a prime other than 2, then $p^k \pm 1$ is divisible by 2 and therefore not a prime.

Definition 5.13. *(i) The <u>Mersenne</u> <u>numbers</u> are $M_k = 2^k - 1$. <u>Mersenne</u> <u>prime</u> is a Mersenne number that is also prime.*
(ii) The <u>Fermat</u> <u>numbers</u> are $F_k = 2^{2^k} + 1$. A <u>Fermat</u> <u>prime</u> is a Fermat number that is also prime.
(iii) The number $n \in \mathbb{N}$ is called a <u>perfect</u>, if $\sigma(n) = 2n$.

Lemma 5.14. *(i) If $ab = k$, then $(2^b - 1) \mid (2^k - 1)$.*
(ii) If $ab = k$ and a is odd, then $(2^b + 1) \mid (2^k + 1)$.

Proof. We only prove (ii); (i) can be proved similarly. So suppose that a is odd, then

$$2^b =_{2^b+1} -1 \implies 2^{ab} =_{2^b+1} (-1)^a =_{2^b+1} -1 \implies 2^{ab} + 1 =_{2^b+1} 0$$

which proves the statement. Notice that this includes the case where $b = 1$. In that case, we have $3 \mid (2^a + 1)$ (whenever a odd). ■

A proof using geometric series can be found in exercise 1.12. This lemma immediately implies the following.

Corollary 5.15. *i) If $2^k - 1$ is prime, then k is prime.*
ii) If $2^k + 1$ is prime, then $k = 2^r$.

So candidates for Mersenne primes are the numbers $2^p - 1$ where p is prime. This works for $p \in \{2,3,5,7\}$, but $2^{11} - 1 = 2047$ is the monkey-wrench. It is equal to $23 \cdot 89$ and thus is composite. After that, the Mersenne primes become increasingly sparse. For example, 8 of the first 11 Mersenne numbers are prime (M_{11}, M_{23}, M_{29} are not prime). However, among the first approximately 2.3 million Mersenne numbers, only 45 give Mersenne primes. As of this writing (2021), it is not known whether there are infinitely many Mersenne primes. In 2020, a very large Mersenne prime was discovered: $2^{82,589,933} - 1$. Mersenne primes are used in pseudo-random number generators.

Turning to primes of the form $2^k + 1$, the only candidates are $F_r = 2^{2^r} + 1$. Fermat himself noted that F_r is prime for $0 \leq r \leq 4\}$, and he conjectured that all these numbers were primes. Again, Fermat did not quite get it right! It turns out that the 5-th Fermat number, $2^{32} + 1$, is divisible by 641 (see exercise 5.11). In fact, as of this writing in 2017, there are *no other known Fermat primes* among the first 297 Fermat numbers! Fermat primes are also used in pseudorandom number generators.

Lemma 5.16. *If $2^k - 1$ is prime, then $k > 1$ and $2^{k-1}(2^k - 1)$ is perfect.*

Proof. If $2^k - 1$ is prime, then it must be at least 2, and so $k > 1$. Let $n = 2^{k-1}(2^k - 1)$. Since σ is multiplicative and $2^k - 1$ is prime, we can compute (using Theorem 4.5):

$$\sigma(n) = \sigma(2^{k-1})\sigma(2^k - 1) = \left(\sum_{i=0}^{k-1} 2^i\right) 2^k = (2^k - 1)\, 2^k = 2n$$

which proves the lemma. ■

Theorem 5.17 (Euler's Theorem). *Suppose $n > 0$ is even. Then n is of the form $2^{k-1}(2^k - 1)$ where $2^k - 1$ is prime if and only if n is perfect.*

Proof. One direction follows from the previous lemma. Thus we only need to prove that if an even number n is perfect, then it is of the form stipulated.

Since n is even, we may assume $n = q\,2^{k-1}$ where $k \geq 2$ and q is odd. Using multiplicativity of σ and the fact that $\sigma(n) = 2n$:

$$\sigma(n) = \sigma(q)\,(2^k - 1) = 2n = q\,2^k \,.$$

Thus

$$(2^k - 1)\,\sigma(q) - 2^k\,q = 0\,. \tag{5.1}$$

Since $2^k - (2^k - 1) = 1$, we know by Bézout that $\gcd((2^k - 1), 2^k) = 1$. Thus Proposition 3.5 implies that the general solution of the above equation is:

$$q = (2^k - 1)\,t \qquad \text{and} \qquad \sigma(q) = 2^k\,t\,, \tag{5.2}$$

where $t > 0$, because we know that $q > 0$.

Assume first that $t > 1$. The form of q, namely $q = (2^k - 1)\,t$, allows us to identify at least four distinct divisors of q. This gives that

$$\sigma(q) \geq 1 + t + (2^k - 1) + (2^k - 1)\,t = 2^k\,(t+1)\,.$$

This contradicts equation (5.2), and so $t = 1$.

Now use equation (5.2) again (with $t = 1$) to get that $n = q\,2^{k-1} = (2^k - 1)\,2^{k-1}$ has the required form. Furthermore, the same equation says that $\sigma(q) = \sigma(2^k - 1) = 2^k$ which proves that $2^k - 1$ is prime. ■

It is unknown at the date of this writing (2021) whether any odd perfect numbers exist.

5.4. A Divisive Issue: Rings and Fields

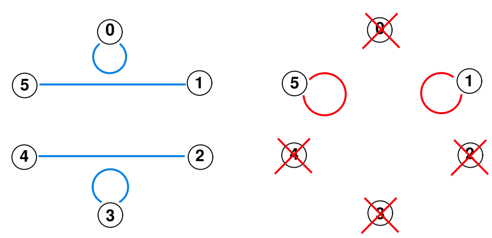


Figure 21. Left: the relation a is an additive inverse mod 6 of b. Right: the relation a is an multiplicative inverse mod 6 of b.

The next result is a game changer! It tells us that there is a unique element a^{-1} such that $aa^{-1} =_b 1$ if and only if a is in the reduced set of residues (modulo b). Thus *division* is only well-defined in the reduced set of residues modulo b. So, for example, the reduced set of residues modulo 15 equals $\{1, 2, 4, 7, 8, 11, 13, 14\}$. In this group, we can multiply and divide all we want. For example, the inverse of 8 in $\mathbb{Z}_{15}$ is 2 because $8 \cdot 2 =_{15} 1$.

In fact, this set forms a nice Abelian group (defined below) under multiplication. Another illustration is given in Figure 21 for $\mathbb{Z}_6$. On the left, we see that every element has an *additive* inverse. So, for example $4+2 =_6 0$, and so 4 and 2 are additive inverses. Notice that the relation is symmetric, so the edges in the graph have no arrows on them. However, on the right, we see that only 5 and 1 have a *multiplicative* inverse. So $5^2 =_6 1$ and similar for 1. Since 7 is a prime, all elements of $\mathbb{Z}_7$ except 0 have multiplicative inverses modulo 7 (see Figure 22).

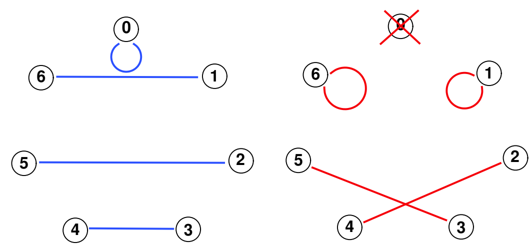


Figure 22. Left: the relation a is an additive inverse mod 7 of b. Right: the relation a is an multiplicative inverse mod 7 of b.

Proposition 5.18. *Let R be a reduced set of residues modulo b. Then*
i) for every $a \in R$, there is a unique a' in R such that $a'a =_b aa' =_b 1$,
ii) for every $a \notin R$, there exists no $x \in \mathbb{Z}_b$ such that $ax =_b 1$,
iii) let $R = \{x_i\}_{i=1}^{\varphi(b)}$, then also $R = \{x_i^{-1}\}_{i=1}^{\varphi(b)}$.

Proof. Statement (i): Since $\gcd(a,b) = 1$, the existence of a solution follows immediately from Bézout's Lemma. Namely a' solves for x in $ax + by = 1$. This solution must be in R, because a, in turn, is the solution of $a'x + by = 1$ and thus Bézout's Lemma implies that $\gcd(a',b) = 1$. Suppose we have two solutions $ax =_b 1$ and $ay =_b 1$, then uniqueness follows from applying the cancelation Theorem 2.7 to the difference of these equations.

Statement (ii): By hypothesis, $\gcd(a,b) > 1$. We have that $ax =_b 1$ is equivalent to $ax + by = 1$, which contradicts Bézout's lemma.

Statement (iii): This is similar to Lemma 5.3. By (1), we know that all inverses are in R. So if the statement is false, there must be two elements of R with the same inverse: $ax =_b cx$. This is impossible by cancellation (Theorem 2.7). ■

What this means is that in structures like $\mathbb{Z}_b$ addition and multiplication have a complicated relationship. Under addition, they form a group.

Definition 5.19. *A group is defined as a set G with an operation * satisfying:*
i) G is closed under the operation, or all a, b in G, $a*b \in G$.
ii) The operation is associative or $(a*b)*c = a*(b*c)$.
iii) R has an identity element e and for all a in G, $a*e = e*a = a$.
iv) Each $a \in G$ *has an inverse* a^{-1} *such that* $a*a^{-1} = a^{-1}*a = e$.
The group is called Abelian group if the operation is commutative or $a*b = b*a$*).*

It is important to realize that not all groups are commutative. The smallest non-commutative group is the group of symmetries of an equilateral triangle, S_3, with the composition as the group operation (see Figure 23). This group is isomorphic to the group of permutations of the symbols $\{1,2,3\}$. Thus the refection in the line through 1 in the triangle corresponds to swapping 2 and 3, while one possible rotation corresponds to $1 \to 2$, $2 \to 3$, and $3 \to 1$. The figure that the order in which we carry out the operations affects the outcome.

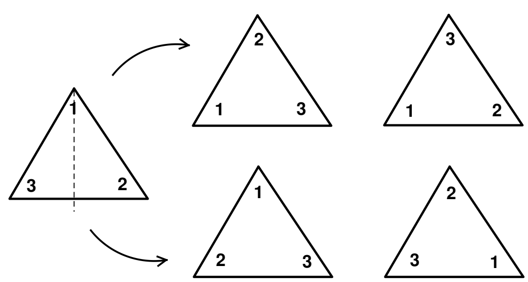


Figure 23. The group of symmetries of an equilateral triangle, S_3, is not commutative. The group operation is composition. In the top row, we perform first a rotation given by $1 \to 2$, $2 \to 3$, and $3 \to 1$ and then a refection in the dotted line through 1. In the bottom row, the reverse.

The *additive* group $\mathbb{Z}_b$ is generated by the element 1, because repeated addition of 1 gives the entire group. This also makes it clear that we can not leave any elements out and still obtain an additive group. But under multiplication, the story is more complicated. There is no multiplicative inverse of 0. But even if we exclude 0, then according to Proposition 5.18, we *only* get a *multiplicative* group if b is prime. Indeed, in general we only get a multiplicative group if we further restrict to the reduced set of residues modulo b. Let us illustrate the point by showing the tables for multiplication in $\mathbb{Z}_5$ and $\mathbb{Z}_6$. In the latter case, the only multiplicative group consists of the elements 1 and 5.

$\mathbb{Z}_5(\times)$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

$\mathbb{Z}_6(\times)$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

The optimistic reader might be inclined to think that maybe not all is lost, as long as things work for the most important number system, $\mathbb{Z}$ itself. Alas, a moment's thought reveals that multiplication in $\mathbb{Z}$, like multiplication in $\mathbb{Z}_b$ for b non-prime, does not have an inverse. Thus our hand is forced, and we define a structure where addition has all the nice properties — in particular, it has an inverse — and where we are a bit more prudent in assigning the characteristics of multiplication.

Definition 5.20. *A ring is defined as a set R which is closed under two operations, usually called addition and multiplication, and has the following properties:*
i) R with addition is an Abelian group (with additive identity 0).
ii) Multiplication in R is associative (see exercise 5.23).
iii) Multiplication is distributive over addition (that is: $a(b+c) = ab+bc$ *and* $(b+c)a = ba+ca$*).*
iv) R has a (multiplicative) identity denoted by 1 and $0 \neq 1$.
A commutative ring is a ring in which multiplication is commutative.

Remark 5.21. Note that $\mathbb{N}$ is not a ring, because addition is not invertible. We will from here on out consider the primes as a subset of $\mathbb{Z}$.

Remark 5.22. We will assume rings to be commutative and drop the adjective "commutative" for brevity, unless needed for clarity.

Remark 5.23. The requirement that $0 \neq 1$ only excludes the 0 ring ($R = \{0\}$).

Remark 5.24. An important example of an "almost ring" are the multiples $n\mathbb{Z}$ in $\mathbb{Z}$ for $n > 1$. Indeed, that set satisfies all the requirements of a ring *except* that it does not have a multiplicative identity. This is sometimes called a rng.

Definition 5.25. *A unit in a ring is an element that has a multiplicative inverse in the ring. This is also called an invertible element.*

On the other hand, other important sets, such as $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$, *do* have a well-defined multiplicative inverse (again excepting 0) much like $\mathbb{Z}_p$ for p prime. Thus we also need to define a structure where multiplication is treated on more equal footing with addition — it has an inverse.

Definition 5.26. *A field is a commutative ring for which multiplication by a non-zero number has an inverse. Equivalently, considered as a ring, all non-zero elements are units.*

But in generally, the words *division* and *multiplicative inverse* have to be used carefully in a ring.

Definition 5.27. *Let a, b, and x in a ring. We say that b is a divisor of a and write $b \mid a$ if there is a solution x of $bx = a$.*

The sets $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{Z}_b$ are all examples of rings, but of these only $\mathbb{Q}$ and $\mathbb{Z}_p$ with p prime are fields, because all elements are invertible as we saw in Proposition 5.18. The field of the integers modulo a prime p will be from now be denoted by $\mathbb{F}_p$, where p is understood to be a prime.

Rings and fields occur in all kinds of other situations and applications. We already looked at one interesting example of a ring, namely the arithmetic functions with addition and convolution as operations (exercise 4.16). Here are some other examples of rings that are not fields. Real numbers of the form $a + b\sqrt{3}$ where a and b in $\mathbb{Z}$, complex numbers of the form $a + ib$ or those of the form $a + ib\sqrt{6}$ where a and b in $\mathbb{Z}$. Other examples are the n by n matrices ($n \geq 2$). We have already seen the polynomials with rational coefficients exercise 3.22. They also form a ring. All of these rings have different properties. For instance, the ring of n by n matrices is not commutative. We will see later that not all rings (that are not fields) have primes.

It is useful to reflect a moment on how the absence of division influences how we think about such sets. It is precisely that curious absence that brings us to the study of primes, integers that have no non-trivial divisors at all. The situation in fields like $\mathbb{Z}_p$ (for prime p) or $\mathbb{R}$ is very different! Here multiplication *does* have an inverse, and thus given a and b not equal to 0, we can always write a as a non-trivial product as follows:

$$a =_p (ab)b^{-1} .$$

Here is another interesting observation. If we extend the integers to the rationals $\mathbb{Q}$, we obtain a field. Thus the problem of where the primes are goes away: in $\mathbb{Q}$ (or $\mathbb{R}$) we can always divide (except by 0), and there are no primes. Of course, since, even in mathematics, nothing is perfect, in the rationals we have other problems. If we allow the integers to be arbitrarily divided by other integers, we obtain the field of the *rational numbers*. It was a source of surprise and mystery to the ancients, that within the rational numbers we still cannot solve for x in $x^2 = 2$, although we can get arbitrarily good approximations. Those 'gaps' in the rational numbers, are the *irrational numbers*. We are then left with the thorny question of whether the reals containing both the rational and the irrational numbers still have gaps. How can we approximate irrational numbers using rational numbers? How can we calculate with the reals? Well, among other things you have to learn how to take limits, which is a whole other *can of worms*.

5.5. Wilson's Theorem

We end this chapter with one important application of division in $\mathbb{Z}_p$.

Lemma 5.28. *Let p be prime. Then $a^2 =_p 1$ if and only if $a =_p \pm 1$. Equivalently, $a \in \mathbb{Z}_p$ is its own multiplicative inverse if and only if $a =_p \pm 1$.*

Proof. We have

$$a^2 =_p 1 \iff a^2 - 1 =_p (a+1)(a-1) =_p 0 \iff p \mid (a+1)(a-1) .$$

Because p is prime, Corollary 2.9 says that the last statement holds if and only if *either* $p \mid a+1$ (and so $a =_p -1$) *or* $p \mid a-1$ (and so $a =_p +1$). ■

Perhaps surprisingly, this last lemma is false if p is not prime. For example, $4^2 =_{15} 1$, but $4 \neq_{15} \pm 1$.

Theorem 5.29 (Wilson's theorem). *If p prime in $\mathbb{Z}$, then $(p-1)! =_p -1$. If b is composite, then $(b-1)! \neq_b \pm 1$.*

Proof. This is true for p is 2 and 3. If $p > 3$, then Proposition 5.18 (3) and Lemma 5.28 imply that every factor a_i in the product $(p-1)!$ other than -1 or 1 has a unique inverse a_i' different from itself. The factors a_i' run through all factors 2 through $p-2$ exactly once. Thus in the product, we can pair each a_i different from ± 1 with an inverse a_i' distinct from itself. This gives

$$(p-1)! =_p (+1)(-1)\prod a_i a_i' =_p -1 \ .$$

The second part is easier. If b is composite, there are least residues a and d greater than 1 so that $ad =_b 0$. Now either we can choose a and d distinct and then $(b-1)!$ contains the product ad, and thus it equals zero mod b. Or else this is impossible and there exists a such that $a^2 =_b 0$. But then still $\gcd((b-1)!, b)$ is a multiple of a. Then, by Bézout, $(b-1)!$ mod b cannot be equal to ± 1. ∎

Wilson's theorem could be used to test primality of a number n. However, this takes n multiplications, which in practice is more expensive than trying to divide n by all numbers less than $\sqrt{n}$. Note, however, that if you want to compute a list of *all* primes between 1 and N, Wilson's theorem can be used much more efficiently. After computing $(k-1)!$ mod k to determine whether k is prime, it takes only 1 multiplication and 1 division to determine whether $k+1$ is prime.

5.6. Exercises

Exercise 5.1. a) Let $m > 0$. Show that $a =_m b$ is an equivalence relation on $\mathbb{Z}$. (*Use Definitions 1.7 and 1.28.*)
b) Describe the equivalence classes of $\mathbb{Z}$ modulo 6. (*Which numbers in $\mathbb{Z}$ are equivalent to 0? Which are equivalent to 1? Et cetera.)*
c) Show that the equivalence classes are identified by their residue, that is: $a \sim b$ if and only if $\mathrm{Res}_m(a) = \mathrm{Res}_m(b)$.

Note: If we pick one element of each equivalence class, such an element is called a representative of that class. The smallest non-negative representative of a residue class in $\mathbb{Z}_m$, is called the *least residue* (see Definition 1.8). The collection consisting of the smallest non-negative representative of each residue class is called a *complete set of least residues*.

Exercise 5.2. This exercise relies on exercise 5.1. Denote the set of equivalence classes of $\mathbb{Z}$ modulo m by $\mathbb{Z}_m$ (see Definition 1.7). Prove that addition and multiplication are well-defined in $\mathbb{Z}_m$, using the following steps.
a) If $a =_m a'$ and $b =_m b'$, then $\mathrm{Res}_m(a) + \mathrm{Res}_m(b) =_m \mathrm{Res}_m(a') + \mathrm{Res}_m(b')$. (*Hint: show that $a+b=c$ if and only if $a+b=_m c$. In other words: the sum modulo m only depend on* $\mathrm{Res}_m(a)$ *and* $\mathrm{Res}_m(b)$ *and not on which representative in the class (see exercise 5.1) you started with.*)
b) Do the same for multiplication.

Exercise 5.3. Let $n = \sum_{i=1}^{k} a_i 10^i$ where $a_i \in \{0,1,2,\cdots,9\}$.
a) Show that $10^k =_3 1$ for all $k \geq 0$. (*Hint: use exercise 5.2.*)
b) Show that $n =_3 \sum_{i=1}^{k} a_i$.
c) Show that this implies that n is divisible by 3 if and only the sum of its digits is divisible by 3.

Exercise 5.4. Let $n = \sum_{i=1}^{k} a_i 10^i$ where $a_i \in \{0,1,2,\cdots,9\}$. Follow the strategy in exercise 5.3 to prove the following facts.
a) Show that n is divisible by 5 if and only if a_0 is. (*Hint: Show that* $n =_5 a_0$.)
b) Show that n is divisible by 2 if and only if a_0 is.
c) Show that n is divisible by 9 if and only if $\sum_{i=1}^{k} a_i$ is.
d) Show that n is divisible by 11 if and only if $\sum_{i=1}^{k} (-1)^i a_i$ is.
e) Find the criterion for divisibility by 4.
f) Find the criterion for divisibility by 7. (*Hint: this is a more complicated criterion!*)

Exercise 5.5. a) Determine the period of the decimal expansion of the following numbers: 100/13, 13/77, and 1/17 through long division.
b) Use Proposition 5.8 to determine the period.
c) Check that this period equals a divisor of $\varphi(n)$.
d) The same questions for expansions in base 2 instead of base 10.

Exercise 5.6. a) Compute $2^{n-1} \mod n$ for n odd in $\{3\cdots 40\}$.
b) Are there any pseudo-primes in the list?

Exercise 5.7. Assume that n is a pseudoprime to the base 2.
a) Show that $2^n - 2 =_n 0$.
b) Show from (a) that $n \mid M_n - 1$. (*See Definition 5.13.*)
c) Use Lemma 5.14 to show that (b) implies that $M_n \mid 2^{M_n - 1} - 1$.
d) Conclude from (c) that if n is a pseudoprime in base 2, so is M_n.

Exercise 5.8. a) List $(n-1)! \mod n$ for $n \in \{2, \cdots, 16\}$.
b) Where does the proof of the first part of Wilson's theorem fail in the case of $n = 16$?
c) Does Wilson's theorem hold for $p = 2$? Explain!
d) Characterize the set of $n \geq 2$ for which $(n-1)! \bmod n$ is not in $\{0, -1\}$.

Exercise 5.9. a) Compute $7^{72} \mod 13$, using modular exponentiation.
b) Similarly for $484^{187} \mod 1189$.
c) Find $100! + 102! \mod 101$. (*Hint: Wilson.*)
d) Show that $1381! =_{1382} 0$. (*Hint: Wilson.*)

Exercise 5.10. a) For i in $\{1, 2, \cdots 11\}$ and j in $\{2, 3, \cdots 11\}$, make a table of $\mathrm{Ord}_j^{\times}(i)$, i varying horizontally. After the jth column, write $\varphi(j)$.
b) List the primitive roots i modulo j for i and j as in (a). (*Hint: the smallest primitive roots modulo j are:* $\{1, 2, 3, 2, 5, 3, \emptyset, 2, 3, 2\}$.)

Exercise 5.11. We show that the 5-th Fermat number, $2^{32} + 1$, is a composite number.
a) Show that $2^4 =_{641} -5^4$.(*Hint: add* 2^4 *and* 5^4.)
b) Show that $2^7 5 =_{641} -1$.
c) Show that $2^{32} + 1 = (2^7)^4 2^4 + 1 =_{641} 0$.
d) Conclude that F_5 is divisible by 641.

Exercise 5.12. a) Compute $\varphi(100)$. (*Hint: use Theorem 4.17.*)
b) Show that $179^{121} =_{100} 79^{121}$.
c) Show that $79^{121} =_{100} 79^1$. (*Hint: use Theorem 5.4*)
d) What are the last 2 digits of 179^{121}?

The following 5 exercises on basic cryptography are based on [**70**]. First some language. The original readable message is called the *plain text*. Encoding the message is called *encryption*. And the encoded message is often called the *encrypted* message or *code*. To revert the process, that is: to turn the encrypted message back into plain text, you often need a *key*. Below we will encode the letters by 0 through 25 (in alphabetical order). We encrypt by using a *multiplicative cipher*. This means that we will encrypt our text by multiplying each number by the cipher *modulo* 26, and then return the corresponding letter. For example, if we use the cipher 3 to encrypt the plain text `bob`, we obtain the encrypted text as follows $1.14.1 \to 3.42.3 \to 3.16.3$.

Exercise 5.13. a) Use the multiplicative cipher 3 to decode `DHIM`.
b) Show that an easy way to decode is multiplying by 9 (modulo 26). The corresponding algorithm at the number level is called division by 3 modulo 26.
c) Suppose instead that our multiplicative cipher was 4. Encode `bob` again.
d) Can we invert *this* encryption by using multiplication modulo 26? Explain why.

Exercise 5.14. Suppose we have an alphabet of q letters and we encrypt using the multiplicative cipher $p \in \{0, \cdots q-1\}$. Use modular arithmetic to show that the encryption can be inverted if and only $\gcd(p,q) = 1$. (*Hint: Assume the encryption of j_1 and j_2 are equal. Then look up and use the Unique Factorization theorem in Chapter 2.*)

Exercise 5.15. Assume the setting of exercise 5.14. Assume p and q are such that the encryption is invertible. What is the decryption algorithm? Prove it. (*Hint Find $r \in \{0, \cdots q-1\}$ such that $rp =_q 1$. Then multiply the encryption by r.*)

Exercise 5.16. Work out the last two problems if we encrypt using an *affine cipher* (a, p) . That is, the encryption on the alphabet $\{0, \cdots q-1\}$ is done as follows:

$$i \to a + pi \mod q$$

Work out when this can be inverted, and what the algorithm for the inverse is.

Exercise 5.17. Decrypt the code `V'ir Tbg n Frperg`.

Theorem 5.30 (Binomial Theorem). *If n is a positive integer, then*

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i} \text{ where } \binom{n}{i} = \frac{n!}{i!(n-i)!} .$$

Exercise 5.18. a) If p is prime, show that $\binom{p}{i}$ mod p equals 0 if $1 \le i \le p-1$ and equals 1 if $i = 0$ or $i = p$.
b) Evaluate $\binom{4}{i} \mod 4$ and $\binom{6}{i} \mod 6$. So where in (a) did you use the fact that p is prime?
c) Use (a) and the binomial theorem to show that if p is prime, then we have $(a+b)^p =_p a^p + b^p$.

Exercise 5.19. Let p be prime.
a) Show that $1^p =_p 1$.
b) Use exercise 5.18 (c) to show that for $k > 0$, if $k^p =_p k$, then $(k+1)^p =_p k+1$.
c) Conclude from (b) that for for all $n \in \mathbb{N}$, $n^p =_p n$. (*Hint: use induction.*)
d) Prove that for for all $n \in \mathbb{Z}$, $n^p =_p n$. (*Hint:* $(-n)^p =_p (-1)^p n^p$ *and assume p odd. Prove separately for* $p = 2$.)
e) Use (d) to prove Fermat's little theorem. (*Hint: use cancellation.*)

There are partial converses to Fermat's little theorem. But if our aim is testing for primality, these do not yield *computationally* efficient improvements. We give the simplest of these results here.

Lemma 5.31. *Suppose a and n in* $\mathbb{N}$ *such that* $a^{n-1} =_n 1$ *and that for all primes that divide* $n-1$ *we have* $a^{(n-1)/p} \neq_n 1$. *Then n is a prime.*

Exercise 5.20. In this exercise, we prove Lemma 5.31. For this purpose, abbreviate $\mathrm{Ord}_n^\times(a)$ by o and assume the condition of the lemma.
a) Show that $n-1 = oj$ for some $j \in \mathbb{N}$.
b) Show that if $j > 1$ in (a), there is a prime p dividing j such that

$$a^{(n-1)/p} =_n a^{o(j/p)} =_n 1\,.$$

c) Show that $j = 1$ and so $o = \mathrm{Ord}_n^\times(a) = n-1$.
d) Show that (c) implies the lemma. (*Hint: use Euler.*)
e) Use the lemma to show that 997 is prime. (*Hint: 996 has prime divisors 2, 3, and 83.*)

Theorem 3.13 and exercise 3.18 show how to solve linear congruences generally. Quadratic congruences are much more complicated. As an example, we look at the equation $x^2 =_p \pm 1$ in the following exercise.

Exercise 5.21. a) Show that Fermat's little theorem gives a solution of $x^2 - 1 =_p 0$ whenever p is an odd prime. (*Hint: consider* $x^{\frac{p-1}{2}}$.)
b) Use Lemma 5.28 to show that $x^{\frac{p-1}{2}} =_p \pm 1$.
c) Show that Wilson's theorem implies that for odd primes p

$$(-1)^{\frac{p-1}{2}} \left[\left(\frac{p-1}{2}\right)!\right]^2 =_p -1\,.$$

(*Hint: the left-hand side gives all reduced residues modulo p.*)
d) Use (c) to show that if $p =_4 1$ (examples are 13, 17, 29, etc), then $\left[\left(\frac{p-1}{2}\right)!\right]$ satisfies the quadratic congruence $x^2 + 1 =_p 0$.
e) Show that if $p =_4 3$ (examples are 7, 11, 19, etc), then the quadratic congruence $x^2 + 1 =_p 0$ has no solutions. (*Hint: we have* $x^4 =_p 1$ *and by Euler* $x^{\varphi(p)} =_p 1$*; derive a contradiction if* $p =_4 3$.)

Exercise 5.22. Given $b > 2$, let $R \subseteq \mathbb{Z}_b$ be the reduced set of residues and let $S \subseteq \mathbb{Z}_b$ be the set of solutions in $\mathbb{Z}_b$ of $x^2 =_b 1$ (or self inverses).
a) Show that $S \subseteq R$. (*Hint:Bézout.*)
b) Show that

$$\prod_{x \in R} x =_b \prod_{x \in S} x \quad (=_b 1 \text{ if } S \text{ is empty}).$$

c) Show that if S contains a, then it contains $-a$.
d) Show that if $a =_b -a$, then a and $-a$ are not in S.
e) Show that

$$\prod_{x \in R} x =_b (-1)^m \ \text{ some } m.$$

f) Show that

$$\prod_{x \in R} x =_b (-1)^{|S|/2}.$$

g) Compute $\prod_{x \in R} x$ in a few cases ($b = 6$, 8), and verify that (f) holds.

Definition 5.32. *The nth* <u>*Catalan*</u> <u>*number*</u> C_n *equals* $\frac{1}{n+1}\binom{2n}{n} = \binom{2n}{n} - \binom{2n}{n+1}$.

Exercise 5.23. Many common operations in $\mathbb{R}$ are not associative.
a) Compute 2^{3^4}, $4-3-2$, $4/3/2$. (*Hint: depending on how you place the parentheses, you get different answers.*) In the last two cases, the problem disappears if we recast the computation in terms of the (associative) operators $+$ and $\times$: compute $4+(-3)+(-2)$ and $4 \times \frac{1}{3} \times \frac{1}{2}$.
b) Show that the number of monotone lattice paths from $(0,0)$ to (a,b) where $a,b > 0$ equals $\binom{a+b}{a}$. (*Hint: place $a+b$ edges of which a are horizontal and b are vertical in any order.*)
c) For notational ease, indicate the non-associative operation by $*$. Show that the number of ways $*_{i=1}^{n+1} a_i$ can be interpreted equals the number of "good paths", that is: monotone lattice paths in $\mathbb{R}^2$ from $(0,0)$ to (n,n) that do not go above the diagonal. (*Hint: write the expression so that it has n opening parentheses "(" in it; there are n operations to be performed; reading from left to right, each (corresponds to a "right" move, each * to an "up" move.*)
d) Show that there is a bijection from the set of "bad paths", that is: monotone lattice paths in $\mathbb{R}^2$ from $(0,0)$ to (n,n) that touch the line $\ell : y = x+1$, to the set of monotone paths in $\mathbb{R}^2$ from $(0,0)$ to $(n-1,n+1)$. (*Hint: reflect the bad path in ℓ as indicated in Figure 24 and show this is invertible.*)
e) Use (c) and (d) to show that the number of good paths equals the number of monotone paths from $(0,0)$ to (n,n) minus the number of monotone paths from $(0,0)$ to $(n-1,n+1)$.
f) Use (e) to show that the number of interpretations in (c) equals C_n of Definition 5.32.

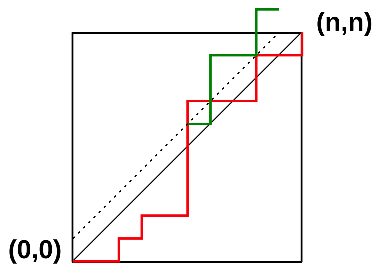


Figure 24. The part to the right of the intersection with $\ell : y = x+1$ (dashed) of a bad path (in red) is reflected. The reflected part in indicated in green. The path becomes a monotone path from $(0,0)$ to $(n-1,n+1)$.

Exercise 5.24. Show that the following sets with the usual additive and multiplicative operations are not fields:
a) The numbers $a+b\sqrt{3}$ where a and b in $\mathbb{Z}$.
b) The numbers of the form $a+ib\sqrt{6}$ where a and b in $\mathbb{Z}$.
c) $\mathbb{Z}_6$.
d) The 2 by 2 real matrices.
e) The polynomials with rational coefficients.
f) The Gaussian integers, i.e. the numbers $a+bi$ where a and b in $\mathbb{Z}$.
(*Hint: in each case, exhibit at least one element that does not have a multiplicative inverse.*)

Exercise 5.25. We revisit the Dirichlet ring of exercise 4.16.
a) Show that given an arithmetic function f, we have that if $f(1) \neq 0$

$$g * f = \varepsilon \quad \Longleftrightarrow \quad \begin{cases} g(1) = \frac{1}{f(1)} & \text{if } n = 1 \\ g(n) = \frac{-1}{f(1)} \sum_{d|n,d<n} f\left(\frac{n}{d}\right) g(d) & \text{if } n > 1 \end{cases}$$

(*Note: g is called the* <u>*Dirichlet*</u> <u>*inverse*</u> *of* f.)
b) Show that f is a unit if and only if $f(1) \neq 0$.
c) Compute the first 12 terms of the Dirichlet inverse of the Fibonacci sequence (Definition 3.18). (*Hint:* $(1,-1,-2,-2,-5,-4,-13,-16,-30,-45,-89,-122)$.)
d) Show $g(n) = -f(n)$ if n is prime.
e) What is the Dirichlet inverse of the (non-zero) constant function? (*Hint: Equation* (4.7).)

Part 2

Currents in Number Theory: Algebraic, Probabilistic, and Analytic

Chapter 6

Continued Fractions

Overview. The algorithm for continued fractions is really a reformulation of the Euclidean algorithm. However, the reformulated algorithm has had such a spectacular impact on mathematics that it deserves its own name and a separate treatment. One of the best introductions to this subject is the classic [**39**]. A generalization of these ideas can be found in [**11**].

6.1. The Gauss Map

Definition 6.1. *The Gauss map (see Figure 25) is the transformation* $T : [0,1] \to [0,1)$ *defined by*

$$T(\xi) = \frac{1}{\xi} - \left\lfloor \frac{1}{\xi} \right\rfloor = \left\{ \frac{1}{\xi} \right\} \qquad \text{and} \qquad T(0) = 0\,,$$

where we have used the notation of Definition 2.1.

Lemma 6.2. *Set* $q_i = \left\lfloor \frac{r_{i-i}}{r_i} \right\rfloor$ *as in equation* (3.1)*. Then the sequence* $\{r_i\}$ *defined by the Euclidean algorithm of Definition 3.3 satisfies:*

$$\left\{ \begin{array}{ll} \dfrac{r_{i+1}}{r_i} = \dfrac{1}{r_i/r_{i-1}} - q_i = T\left(\dfrac{r_i}{r_{i-1}}\right) & \text{and} \\ \dfrac{r_i}{r_{i-1}} = \dfrac{1}{q_i + \dfrac{r_{i+1}}{r_i}} & \end{array} \right. .$$

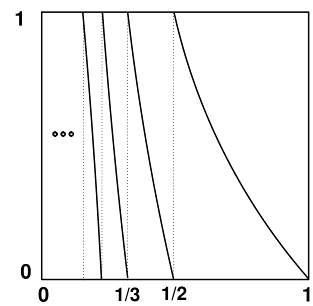


Figure 25. Four branches of the Gauss map.

Proof. From equation (3.1) or (3.4), we recall that that $\{r_i\}$ is a decreasing sequence and that $r_{i-1} = r_i q_i + r_{i+1}$ Upon dividing by r_i, we get

$$\frac{r_{i-1}}{r_i} = q_i + \frac{r_{i+1}}{r_i} \qquad \text{where} \qquad q_i = \left\lfloor \frac{r_{i-1}}{r_i} \right\rfloor ,$$

The first equation of the lemma is obtained by subtracting q_i from both sides, and the second equation is obtained by taking the reciprocal of both sides of the above equation. ■

In the exercises 3.20 and 3.21, we indicated by example how the Gauss map is related to the Euclidean algorithm. In particular, the lemma implies that

$$\frac{r_{i+1}}{r_i} = \frac{1}{r_i/r_{i-1}} - \left\lfloor \frac{1}{r_i/r_{i-1}} \right\rfloor . \tag{6.1}$$

6.2. Continued Fractions

The beauty of the relation in Lemma 6.2 is that, having sacrificed the value of $\gcd(r_1, r_2)$ — whose value we therefore may as well set at 1, we have a procedure that applies to rational numbers! There is no reason why this recursive procedure should be restricted to rational numbers. Indeed, very interesting things happen when we extend the procedure to also allow irrational starting values.

Definition 6.3. *Here is a modified version of the Euclidean algorthm. In the second equation of Lemma 6.2, write* $\omega_i = \frac{r_{i+1}}{r_i}$ *and* $a_i = \left\lfloor \frac{1}{\omega_i} \right\rfloor$ *(or,equivalently,* $a_i = \ell$ *if* $\omega_i \in \left(\frac{1}{\ell+1}, \frac{1}{\ell}\right]$*). Extend* ω *to allow for all values in* $[0, 1)$.

It is important to note that, in effect, we have set a_i equal to q_{i+1}. This very unfortunate bit of redefining is done so that the q_i mesh well with the Euclidean algorithm (see equation (3.2)) while making sure that the sequence of the a_i in Definition 6.4 below starts with a_1.

At any rate, with these conventions, the equations of Lemma 6.2 become (see also (6.1)):

$$\begin{cases} \omega_{i+1} = \dfrac{1}{\omega_i} - a_i = T(\omega_i) \;, \quad a_i = \lfloor 1/\omega_i \rfloor \;\text{ and} \\ \omega_i = \dfrac{1}{a_i + \omega_{i+1}} \end{cases} . \tag{6.2}$$

The way one thinks of this is as follows. The first equation defines a dynamical system[1]. Namely, given an initial value $\omega_1 \in [0,1)$, the repeated application of T gives a string of positive integers $\{a_1, a_2, \cdots\}$. The string ends if and only if after n steps $\omega_n = \frac{1}{\ell}$, and so $\omega_{n+1} = 0$. We show in Theorem 6.5 that this happens if and only if ω_1 is rational. The ℓth branch of T, depicted in Figure 25, has $I_\ell = (\frac{1}{\ell+1}, \frac{1}{\ell}]$ as its domain. It is easy to see that $a_i = \ell$ precisely if $\omega_i \in I_\ell$.

If, on the other hand, the $\{a_i\}$ are given, then we can use the second equation to *formally*[2] derive a possibly infinite quotient that characterizes ω_1. For, in that case, we have

$$\omega_1 \;=\; \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}} \;\; . \tag{6.3}$$

The expression stops after n steps, if $\omega_{n+1} = 0$. Else the expression continues forever, and we can only hope that converges to a limit. We now give some definitions.

Definition 6.4. *Let* $\omega_1 \in [0,1]$. *The expression*

$$\omega_1 \;=\; \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}} \;\stackrel{\text{def}}{\equiv}\; [a_1, a_2, a_3, \cdots] \;\; .$$

[1]A dynamical system is basically a rule that describes short term changes. Usually the purpose of studying such a system is to derive long term behavior, such as, in this case, deciding whether the sequence $\{a_i\}$ is finite, periodic, or neither.

[2]Here, "formally" means that we have an expression for ω_1, but (1) we don't yet know if the actual computation of that expression converges to that number, and on the other hand (2) we "secretly" do know that it converges, or we would not bother with it.

is called the <u>continued fraction expansion</u> *of* ω_1*. The finite truncations*

$$\frac{p_n}{q_n} = \frac{1}{a_1 + \frac{1}{a_2 + \cdots \frac{1}{a_n}}} \stackrel{\text{def}}{\equiv} [a_1, a_2, \cdots, a_n] \ .$$

are called the <u>continued fraction convergents</u> *(or* <u>continued fraction approximants</u>*) of* ω_1*. The coefficients* a_i *are called the* <u>continued fraction coefficients</u> *.*

Let us illustrate this definition with a few examples of continued fraction expansions:

$$\pi - 3 = [7, 15, 1, 292, 1, 1, 1, 2, \cdots] \,,$$
$$e - 2 = [1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, \cdots] \,,$$
$$\theta \equiv \sqrt{2} - 1 = [2, 2, 2, 2, 2, \cdots] \,,$$
$$g \equiv \frac{\sqrt{5} - 1}{2} = [1, 1, 1, 1, 1, \cdots] \,.$$

For example, $\pi - 3$ the sequence of continued fraction convergents starts out as: $\frac{1}{7}$, $\frac{15}{106}$, $\frac{16}{113}$, $\frac{4687}{33102}$, $\frac{4703}{33215}$, $\cdots$. The number g is also well-known. It is usually called the *golden mean*. Its continued fraction convergents are formed by the *Fibonacci numbers* defined in Definition 3.18 and given by $\{1, 1, 2, 3, 5, 8, 13, 21, \cdots\}$. Namely, the convergents are $\frac{1}{1}$, $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{5}$, $\frac{5}{8}$, and so forth.

We have defined continued fraction expansion *only* for numbers in ω in $[0, 1)$. This can be easily be remedied by adding a "zeroth" digit a_0 — signifying the floor of ω — to it. Thus the expansion of π would then become $[3; 7, 15, 1, 292, 1, \cdots]$. We do not pursue this further.

Theorem 6.5. *The continued fraction expansion of* $\omega \in [0, 1)$ *is finite if and only if* ω *is rational.*

Proof. If ω is rational, then by Lemma 6.2 and Corollary 3.2, the algorithm ends. On the other hand, if the expansion is finite, namely $[a_1, a_2, \cdots, a_n]$, then, from equation (6.3), we see that ω is rational. ■

Theorem 6.6. *For the continued fraction convergents, we have*

$$\begin{array}{rcl} p_n &=& a_n p_{n-1} + p_{n-2} \\ q_n &=& a_n q_{n-1} + q_{n-2} \end{array} \qquad \text{with} \qquad \begin{array}{ccc} q_0 = 1 & , & p_0 = 0 \\ q_{-1} = 0 & , & p_{-1} = 1 \end{array} \ ,$$

or, in matrix notation,

$$\begin{pmatrix} q_n & p_n \\ q_{n-1} & p_{n-1} \end{pmatrix} = A_n \begin{pmatrix} q_{n-1} & p_{n-1} \\ q_{n-2} & p_{n-2} \end{pmatrix} = A_n \cdots A_2 A_1 \,,$$

where

$$A_i = \begin{pmatrix} a_i & 1 \\ 1 & 0 \end{pmatrix} \qquad \text{and} \qquad A_i^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -a_i \end{pmatrix} \,.$$

Remark. We encountered A_i in Chapter 3 where it was called Q_{i+1}. We changed the name so we have convenient subscript that agrees with the standard notation. Note that the variables q_i are not the same as the q_i of Chapter 3.

Proof. From Definition 6.4, we have that $q_1 = a_1$ and $p_1 = 1$ and thus

$$\begin{pmatrix} q_1 & p_1 \\ q_0 & p_0 \end{pmatrix} = \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} = A_1 \,.$$

We proceed by induction. Suppose that the recursion holds for all $n \leq k$, then

$$\begin{array}{rcl} p_k & = & a_k p_{k-1} + p_{k-2} \\ q_k & = & a_k q_{k-1} + q_{k-2} \end{array} \,. \tag{6.4}$$

The definition of the convergents gives:

$$\frac{p_k}{q_k} = \frac{1}{a_1 + \cdots \frac{1}{a_k}} \qquad \text{and} \qquad \frac{p_{k+1}}{q_{k+1}} = \frac{1}{a_1 + \cdots \frac{1}{a_k + \frac{1}{a_{k+1}}}} \,.$$

Thus $\frac{p_{k+1}}{q_{k+1}}$ is obtained from $\frac{p_k}{q_k}$ by replacing a_k by $a_k + \frac{1}{a_{k+1}}$ or

$$\begin{array}{rcl} p_{k+1} & = & \left(a_k + \frac{1}{a_{k+1}}\right) p_{k-1} + p_{k-2} \\ q_{k+1} & = & \left(a_k + \frac{1}{a_{k+1}}\right) q_{k-1} + q_{k-2} \end{array} \,.$$

Using equation (6.4) gives

$$\begin{array}{rcl} p_{k+1} & = & p_k + \frac{1}{a_{k+1}} p_{k-1} \\ q_{k+1} & = & q_k + \frac{1}{a_{k+1}} q_{k-1} \end{array} \,.$$

The quotient $\frac{p_{k+1}}{q_{k+1}}$ does not change if if we multiply only the right-hand side of these equations by a_{k+1} to insure that both p_{k+1} and q_{k+1} are integers. This gives the result. ■

Corollary 6.7. *We have*

$$\begin{array}{ll} (i) & q_{n+1}p_n - q_n p_{n+1} = (-1)^{n+1} \\ (ii) & \dfrac{p_n}{q_n} - \dfrac{p_{n+1}}{q_{n+1}} = \dfrac{(-1)^{n+1}}{q_n q_{n+1}} \end{array} .$$

Proof. The left-hand side of the expression in (i) equals the determinant of $\begin{pmatrix} q_{n+1} & p_{n+1} \\ q_n & p_n \end{pmatrix}$, which, by Theorem 6.6, must equal the determinant of $A_{n+1}\cdots A_2 A_1$. Finally, each A_i has determinant -1. To get the second equation, divide the first by $q_{n+1}q_n$. ■

Corollary 6.8. *We have*

$$\begin{array}{ll} (i) & p_n \geq 2^{\frac{n-1}{2}} \;\text{ and }\; q_n \geq 2^{\frac{n-1}{2}} \\ (ii) & \gcd(p_n, q_n) = 1 \end{array} .$$

Proof. i) Iterating the recursion in Theorem 6.6 twice, we conclude that

$$p_{n+1} = (a_n a_{n-1} + 1)p_{n-1} + a_n p_{n-2} \geq 2p_{n-1} + p_{n-2}\,,$$

while $p_1 = 1$ and $p_2 \geq 2$. The same holds for q_n.
ii) By Corollary 6.7 (i) and Bézout. ■

Theorem 6.9. *For irrational ω, the limit $\lim_{n\to\infty} \frac{p_n}{q_n}$ exists and equals ω. In fact, $|\omega - p_n/q_n| < 1/(q_n q_{n+1})$.*

Remark 6.10. The way the convergents approxmate ω is illustrated in Figure 26.

Proof. If we replace n by $n-1$ in the equality of Corollary 6.7(ii), we get another equality. Adding it to the original equality gives:

$$\frac{p_{n-1}}{q_{n-1}} - \frac{p_{n+1}}{q_{n+1}} = \frac{(-1)^n}{q_{n-1}q_n} + \frac{(-1)^{n+1}}{q_n q_{n+1}} \quad \text{or}$$

$$\frac{p_{n+1}}{q_{n+1}} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n+1}}{q_n}\left(\frac{1}{q_{n-1}} - \frac{1}{q_{n+1}}\right).$$

By Theorem 6.6, the q_i are positive and strictly increasing, and so the right-hand side of the last equality is positive if n is even, and negative if n is odd. Thus the sequence $\{\frac{p_n}{q_n}\}_{n \text{ even}}$ is *increasing* while the sequence $\{\frac{p_n}{q_n}\}_{n \text{ odd}}$ is *decreasing*.

In addition, by substituting $2k$ for n in Corollary 6.7(ii), we see that the *decreasing* sequence (n odd) is bounded from below by the *increasing* sequence, and vice versa. Since a bounded monotone sequence of real numbers has a limit[3], the decreasing sequence has a limit ω_-. Similarly, the increasing sequence must have a limit ω_+. Now we use Corollary 6.7(ii) again to see that for all n, the difference between the two cannot exceed $\frac{1}{q_{n-1}q_n}$. So $\omega_+ = \omega_- = \omega$.

Clearly, ω lies between $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$, and so the final estimate of the theorem follows from Corollary 6.7(ii). ∎

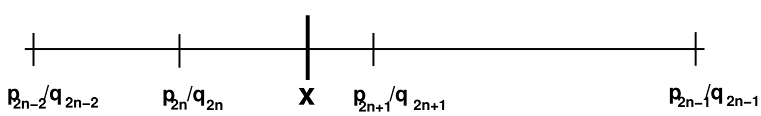


Figure 26. This figure illustrates how the convergents p_n/q_n approach their limit x.

Corollary 6.11. *Suppose ω is irrational. For every $n > 0$, we have $\frac{p_{2n}}{q_{2n}} < \omega < \frac{p_{2n+1}}{q_{2n+1}}$. If ω is rational, the same happens, until we obtain equality of ω and the last convergent.*

[3]This is the monotone convergence theorem, see for example [**53**]

6.3. Computing with Continued Fractions

Suppose we have a positive real ω_0 and want to know its continued fraction coefficients a_i. By the remark just before Theorem 6.5, we start by setting

$$a_0 = \lfloor \omega_0 \rfloor \qquad \text{and} \qquad \omega_1 = \omega_0 - a_0\,.$$

After that, we use Lemma 6.2, and get

$$a_i = \left\lfloor \frac{1}{\omega_i} \right\rfloor \qquad \text{and} \qquad \omega_{i+1} = \frac{1}{\omega_i} - a_i\,.$$

For example, we want to compute the a_i for

$$\omega_1 = \frac{1+\sqrt{6}}{5} \approx 0.6898979\cdots\,. \tag{6.5}$$

If you do this numerically, bear in mind that to compute all the a_i you need to know the number with infinite precision. This is akin to computing, say, the binary representation of ω_1: if we want infinitely many binary digits, we need to know all its decimal digits. To circumvent this issue, we keep the exact form of ω_1. This involves some careful manipulations with the square root. Here are the details. Since $\omega_1 \in (1/2, 1)$, we have $a_1 = 1$. Thus

$$\omega_2 = \frac{5}{1+\sqrt{6}} - 1 = \frac{4-\sqrt{6}}{1+\sqrt{6}}\,.$$

To get rid of the square root in the denominator, we multiply both sides by the "conjugate" $1-\sqrt{6}$ of the denominator. Note that $(1+\sqrt{6})(-1+\sqrt{6})$ gives $-1+6=5$. So we obtain

$$\omega_2 = \frac{4-\sqrt{6}}{1+\sqrt{6}} \cdot \frac{-1+\sqrt{6}}{-1+\sqrt{6}} = -2+\sqrt{6} \approx 0.45 \in \left(\frac{1}{3}, \frac{1}{2}\right] \quad \Longrightarrow \quad a_2 = 2\,.$$

Subsequently, we repeat the same steps to get

$$\omega_3 = \frac{1}{-2+\sqrt{6}} - 2 = \cdots = \frac{-2+\sqrt{6}}{2} \approx 0.225 \in \left(\frac{1}{5}, \frac{1}{4}\right] \quad \Longrightarrow \quad a_3 = 4\,.$$

This is beginning to look desperate, but rescue is on the way:

$$\omega_4 = \frac{2}{-2+\sqrt{6}} - 4 = -2+\sqrt{6} = \omega_2\,.$$

Now everything repeats, and thus we know the complete representation of ω_1 in terms of its continued fraction coefficients:

$$\omega_1 = \frac{1+\sqrt{6}}{5} = [1,2,4,2,4,2,4\cdots] = [1,\overline{2,4}]\,.$$

The reverse problem is also interesting. Suppose we just know the continued fraction coefficients $\{a_i\}_{i=1}^{\infty}$ of ω_1. We can compute the continued fraction convergents by using Theorem 6.6

$$\begin{pmatrix} q_n & p_n \\ q_{n-1} & p_{n-1} \end{pmatrix} = A_n \cdots A_2 A_1 \qquad \text{where} \qquad A_i = \begin{pmatrix} a_i & 1 \\ 1 & 0 \end{pmatrix} .$$

Theorem 6.9 assures us that the limit of the convergents $\{\frac{p_n}{q_n}\}_{i=1}^{\infty}$ indeed equals $\omega_1 = [a_1, a_2, \cdots]$. If also $a_0 > 0$, add a_0 to ω_1 in order to obtain ω_0. So in our example $\omega_1 = [1, \overline{2,4}]$, this is easy enough to do:

i :	0	1	2	3	4	5	$\cdots$
a_i :	-	1	2	4	2	4	$\cdots$
p_i :	0	1	2	9	20	89	$\cdots$
q_i :	1	1	3	13	29	129	$\cdots$

But, because the a_i are eventually periodic, we can also opt for a more explicit representation of ω_1. The periodic tail can be easily analyzed. Indeed, let

$$x = \cfrac{1}{2 + \cfrac{1}{4 + \cfrac{1}{2 + \cdots}}} \qquad \Longrightarrow \qquad x = \cfrac{1}{2 + \cfrac{1}{4 + x}}$$

After some manipulation, this simplifies to a quadratic equation for x with one root in $[0,1)$.

$$x^2 + 4x - 2 = 0 \quad \Longrightarrow \quad x = -2 \pm \sqrt{6} .$$

Select the root in $[0,1)$ as answer. Now we compute ω_1 as follows.

$$\omega_1 = \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{4 + \cdots}}} = \frac{1}{1+x} = \frac{1}{-1+\sqrt{6}} = \frac{1+\sqrt{6}}{5} ,$$

which agrees with our earlier choice of ω_1 in equation (6.5).

6.4. The Geometric Theory of Continued Fractions

We now give a brief description of the geometric theory of continued fractions. This description allows us to prove one of the most remarkable characteristics of the continued fraction convergents (Theorem 6.14). Another geometric description can be found in exercise 6.11.

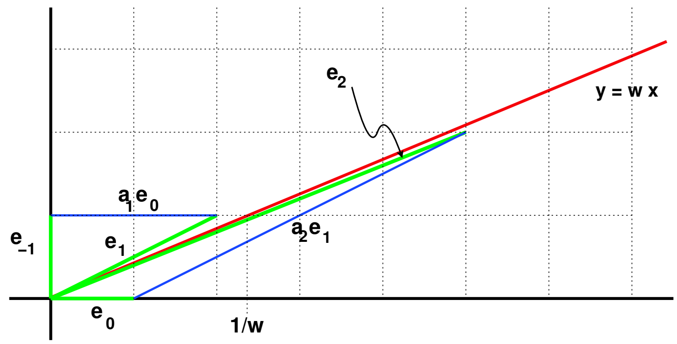


Figure 27. Successive approximations to the line $y = wx$. In green, the vectors e_i. In blue, $a_i e_{i-1}$ for $i = 1$ and 2. Note that $a_1 = a_2 = 2$ in this figure. Note that $a_1 = \lfloor 1/w \rfloor$.

The theory consists of constructing successive line segments that approximate the line $y = \omega_1 x$ in the Cartesian plane. The construction is inductive. Here is the first step.

Start with

$$e_{-1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad e_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} . \tag{6.6}$$

Although at first sight a little odd, it is the convention that e_{-1} is the basis vector along the y-axis and e_0 the one along the x-axis. To get the first new approximation, define

$$e_1 = a_1 e_0 + e_{-1} = \begin{pmatrix} a_1 \\ 1 \end{pmatrix} , \tag{6.7}$$

where we choose a_1 to be the largest integer so that e_1 and e_{-1} lie on the same side of $y = \omega_1 x$ (see Figure 27). From the figure, it is easy to see that in particular $e_1 = \begin{pmatrix} a_1 \\ 1 \end{pmatrix}$ and $a_1 = \lfloor 1/\omega_1 \rfloor$, the same as in Definition 6.3. This way, we can construct e_n as segments whose slopes are p_n/q_n. In that sense, e_n cam be considered the approximants of ω_1.

There is another way of seeing this. Note that ω_1 lies between the slopes of e_0 and e_1. Now define the two by two matrix A_1 as the matrix corresponding to the coordinate change T_1 such that $T_1(e_{-1}) = e_0$ and $T_1(e_0) = e_1$. Thus from equations (6.6) and (6.7), one concludes that the matrix A_1 satisfies

$$A_1 \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} = x_1 e_0 + x_2 e_1 \quad \text{and} \quad \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} = A_1^{-1}(x_1 e_0 + x_2 e_1) \; .$$

The first equation implies that, indeed, A_1 is the matrix we defined earlier (in Theorem 6.6). The second equation says that A_1^{-1} is the coordinate transform that gives the coordinates of a point in terms of the new basis e_0 and e_1. The new coordinates of the line $\begin{pmatrix} x \\ \omega_1 x \end{pmatrix}$ become

$$A_1^{-1} \begin{pmatrix} x \\ \omega_1 x \end{pmatrix} = \begin{pmatrix} \omega_1 x \\ x - a_1 \omega_1 x \end{pmatrix} = t \begin{pmatrix} 1 \\ \omega_1^{-1} - a_1 \end{pmatrix} \; ,$$

upon reparametrizing $t = \omega_1 x$. Thus the slope of that line in the new coordinates, ω_2, is the one given by equation (6.2). Since a_1 was chosen the greatest integer so that the new slope is non-negative, we obtain that ω_1 is contained in $[0,1)$.

Since $\omega_2 > 0$, the construction now repeats itself, so that we get

$$e_{n+1} = a_{n+1} e_n + e_{n-1} \; ,$$

as long as $\omega_n > 0$. By construction, ω_1 always lies between $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$. We obtain the same relations for the components of e_n as those in Theorem 6.6, and so

$$e_n = \begin{pmatrix} q_n \\ p_n \end{pmatrix} \; .$$

Consider the parallelogram $p(e_n, e_{n-1})$ spanned by e_n and e_{n-1}. The oriented area of $p(e_n, e_{n-1})$ is exactly the determinant of the matrix $\begin{pmatrix} q_n & p_n \\ q_{n-1} & p_{n-1} \end{pmatrix}$. One now obtains Corollary 6.7 again[4].

6.5. Closest Returns

Consideration of the line ωx in the plane gives us another insight, see Figure 28. The successive intersections with the vertical unit edges are in fact the iterates of the rotation $R_\omega : x \to x + \omega \mod 1$ on the circle starting with initial condition 0. A natural question that arises is: when do these iterates return close to their starting point?

Considering Theorem 6.9, we see that $R_\omega^{q_n}(0) = \omega q_n - p_n$ is very small. In this section, we prove something much better (Theorem 6.14), namely that these iterates are the *closest* returns.

Definition 6.12 (**Closest Returns**). *R_ω^q is a closest return if $R_\omega^q(0)$ is closer to 0 (on the circle) than $R_\omega^n(0)$ for any $0 < n < q$.*

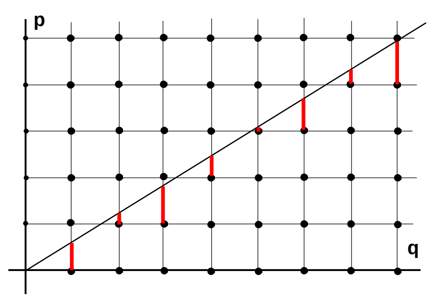


Figure 28. The line $y = \omega x$ and (in red) successive iterates of the rotation R_ω. Closest returns in this figure are q in $\{2, 3, 5, 8\}$.

[4]Geometrically, the proof of that corollary is most easily expressed in the language of exterior or wedge products. These follow the rules of determinant computations. Thus the relevant induction step is the following computation.

$$e_n \wedge e_{n-1} = (a_n e_{n-1} + e_{n-2}) \wedge e_{n-1} = -e_{n-1} \wedge e_{n-2} .$$

Lemma 6.13. *Define $d_n \equiv \omega q_n - p_n$. Then the sequence $\{d_n\}$ is alternating and its absolute value decreases monotonically. In fact,* $|d_{n+1}| < \frac{1}{1+a_{n+1}}|d_{n-1}|$.

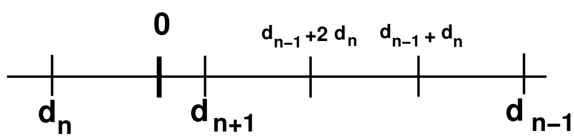


Figure 29. The geometry of successive closest returns. In this figure, $a_{n+1} = 3$. Note that $|d_{n-1}| > 4d_{n+1}$.

Proof. The sequence $\{\omega - \frac{p_n}{q_n}\}$ alternates in sign by construction (see Figure 26). Therefore, so does $\{d_n\}$. Recall that a_{n+1} is the largest integer such that

$$\begin{aligned}\omega q_{n+1} - p_{n+1} &= \omega(a_{n+1}q_n + q_{n-1}) - (a_{n+1}p_n + p_{n-1}) \\ &= (\omega q_{n-1} - p_{n-1}) + a_{n+1}(\omega q_n - p_n)\,,\end{aligned}$$

has the same sign as $\omega q_{n-1} - p_{n-1}$. This says that

$$d_{n+1} = d_{n-1} + a_{n+1}d_n\,.$$

Together with the fact that the d_n alternate, this implies that d_n is decreasing. So from Figure 29, one concludes that $(1+a_{n+1})|d_{n+1}| < |d_{n-1}|$. ■

Theorem 6.14 (The closest return property). $\frac{p}{q}$ *is a continued fraction convergent of ω if and only if $q = 1$ or*

$$|\omega q - p| < |\omega q' - p'|$$

for all q' such that $0 < q' < q$ and for all p'.

Proof. Note that if $a_1 > 1$, then $q = 1$ is not a continued fraction denominator, but it still (trivially) satisfies the above inequality.

We will first show by induction that the parallelogram $p(e_{n+1}, e_n)$ spanned by e_{n+1} and e_n contains no integer lattice points except on its four vertices. Clearly, this is the case for $p(e_{-1}, e_0)$. Suppose $p(e_n, e_{n-1})$ has the same property. The next parallelogram $p(e_{n+1}, e_n)$ is contained in a union of $a_{n+1}+1$ integer translates of the previous and careful inspection of Figure 30 shows that it inherits this property.

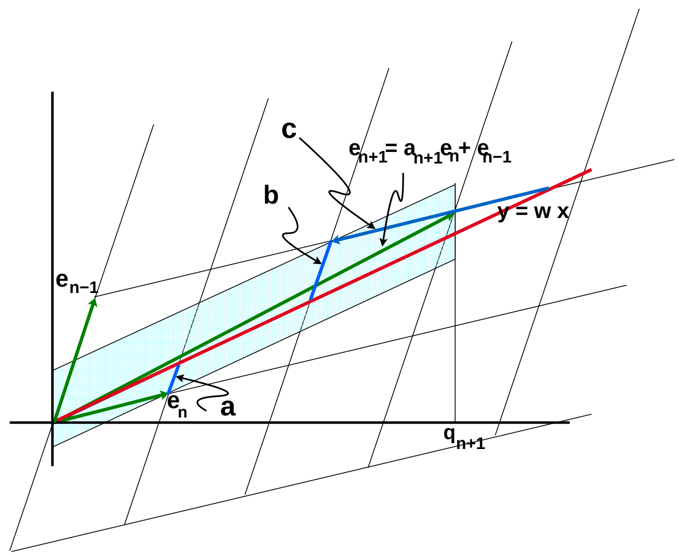


Figure 30. Drawing of the shaded parallelogram P, the line $y = \omega x$, and a few successive approximants. The green arrows correspond to e_{n-1}, e_n, and e_{n+1}. Here, a_{n+1} is taken to be 3.

Next we show, again by induction, that the $R_\omega^{q_n}$ are closest returns, and that there are no others. It is trivial that $R_\omega^{a_1}$ is a closest return, because by definition, $a_1\omega < 1$ and $(a_1+1)\omega_1 > 1$ (see (6.2)). Now suppose that up to $q = q_n$ the only closest returns are e_i, $i \leq n$. We have to prove that the next closest return is e_{n+1}. By Lemma 6.13, $d_{n+1} < d_n$. Now we only need to prove that there are no closest returns for q in $\{q_n+1, q_n+2, \cdots, q_{n+1}-1\}$.

To that purpose we consider Figure 30. Consider the shaded parallelogram P bounded by the vertical lines $x = 0$, $x = q_{n+1}$, and two lines with slope ω, one through e_n and one through $e_{n+1} - e_n$. From the first paragraph it follows that the only lattice points in P are the origin and the endpoints of e_n, and $e_{n+1} - e_n$, e_{n+1}. Since the segment c is parallel to and larger than e_n, we also have that $b > a$. Thus the distance of $e_{n+1} - e_n$ to ωx is greater than the distance of e_n and ωx. And so there is a band of width d_n around $y = \omega x$ in P that contain no points in $\mathbb{Z}^2$ except the origin, e_n, and e_{n+1}. ∎

Corollary 6.15. *If $\frac{p_n}{q_n}$ is a continued fraction convergent of ω, then*

$$|\omega q_n - p_n| \leq |\omega q' - p'|$$

for all q' such that $0 < q' < q_{n+1}$ and all p', with equality only if $q' = q_n$ and $p' = p_n$.

Proof. This follows from the statement of the previous theorem and the fact that there is no closest return for q' (strictly) between q_n and q_{n+1}. ■

6.6. Another Interpretation of the Convergents

Given a number $x_1 \in [0,1)$, we easily see that the first convergent $1/a_1$ maps to zero under the Gauss map T, that is: $T(p_1/q_1) = 0$. Furthermore, since

$$x_1 = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}} = \frac{1}{a_1 + x_2},$$

and $x_2 \in [0,1)$, we can conclude that x_1 lies in the domain $(\frac{1}{a_1+1}, \frac{1}{a_1}]$ of the a_1-branch of T, see Figure 25. More precisely, if $b_1 : I_1 \to [0,1)$ is the branch of T such that $x \in I_1$, then the end point of I_1 that maps to zero under T is the first convergent. It is this statement we wish to generalize.

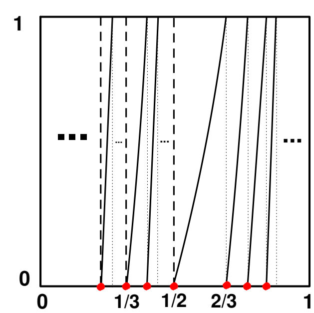


Figure 31. A few branches of the twice iterated Gauss map T^2. The points of $T^{-2}(\{0\})$ (the complete inverse image of 0 under T^2) are marked in red. The reader should compare this plot to Figure 25.

To get an idea what iterates of T look like, let's have a look at T^2 in Figure 31. T has a countable collection of branches with negative slope, each one mapping onto $[0,1]$. Thus T^2 has countably many branches (with positive slope) for every single branch $b : I \to [0,1]$ of T. In turn, each of the branches of T^2 also maps onto $[0,1]$. And so forth.

Proposition 6.16. *Given $x \in (0,1)$. Let $b_k : I_k \to [0,1]$ be the branch of T^k such that $x \in I_k$, then the kth convergent p_k/q_k of x is the (unique) end point of I_k that maps to zero under T^k.*

Proof. From the expression given in Definition 6.4 for $\frac{p_n}{q_n} = [a_1, a_2, \cdots, a_n]$, we see that $T([a_1, a_2, \cdots, a_n]) = [a_2, \cdots, a_n]$. Continuing by induction, we find

$$T^n([a_1, a_2, \cdots, a_n]) = T^{n-1}([a_2, \cdots, a_n]) = \cdots = T([a_n]) = 0\,.$$

So the nth convergent is indeed an nth pre-image of 0 under T.

Similarly, (6.2) implies that

$$x = [a_1, a_2, \cdots] = [a_1, a_2, \cdots, a_n, a_{n+1}, \cdots] = [a_1, a_2, \cdots, (a_n + x_{n+1})]\,.$$

Since $x_{n+1} \in [0,1)$, this is a single branch whose domain contains x. ■

By way of example, we look briefly at the golden mean $g = [1, 1, \cdots] \approx 0.61803\cdots$ in this context. The first convergent is $1/1 = 1$. We immediately remark something perhaps a little unexpected: while this convergent is pre-image of 0 that belongs to the same branch as g, it is *not* that element of $T^{-1}(0)$ that is closest to 0 under T, The next convergent of g is $1/2$. The same thing happens: again, the element of $T^{-2}(0)$ closest to g is in fact $2/3$.

This characterization of convergents is in fact very familiar. Indeed in the usual decimal expansion, based on the map $T : [0,1) \to [0,1)$ given by $T(x) = \{10x\}$, the third *convergent* of the golden mean mentioned above is $p_3/q_3 = [6,1,8]$, more commonly written as 0.618. Note that $T^3(p_3/q_3) = 0$ and that g lies in the domain of the 618 branch of T^3.

Another interesting observation is that the fact that all slopes of T are negative, means that the signs of the slopes of T^k equal $(-1)^k$. So, for odd k the convergents (the zeroes of the branches) are on the right of the interval of definition of the branch they belong to, and for the k even they are on the left side. This is convenient, because it implies that x is always 'sandwiched' between two successive convergents.

6.7. Exercises

Exercise 6.1. Give the continued fraction expansion of $\frac{13}{31}$, $\frac{21}{34}$, $\frac{34}{21}$, $\frac{n-1}{n}$ for $n > 1$, $\frac{n-1}{n^2}$ for $n > 1$ by following the steps in Section 6.3.

Exercise 6.2. a) Find the continued fraction expansion of the fixed points (i.e. solutions of $T(x) = x$ for T in Definition 6.1) of the Gauss map.
b) Use the continued fractions in (a) to find quadratic equations for the fixed points in (a).
c) Derive the same equations from $T(x) = x$.
d) Give the positive solutions of the quadratic equations in (b) and (c).

Exercise 6.3. Compute the continued fraction expansion for $\sqrt{n}$ for n between 1 and 15.

Exercise 6.4. Given the following continued fraction expansions, deduce a quadratic equation for x. (*Hint: see Section 6.3.*)
a) $x = [\overline{8}] = [8,8,8,8,8,\cdots]$.
b) $x = [3,\overline{6}] = [3,6,6,6,6,\cdots]$.
c) $x = [\overline{1,2,3}] = [1,2,3,1,2,3,\cdots]$.
d) $x = [4,5,\overline{1,2,3}] = [4,5,1,2,3,1,2,3,\cdots]$.

Exercise 6.5. In exercise 6.4:
a) solve the quadratic equations (leaving roots intact).
b) give approximate decimal expressions for x.
c) give the first 4 continued fraction convergents.

Exercise 6.6. Derive a quadratic equation for the number with continued fraction expansion: $[\overline{n}]$, $[m,\overline{n}]$, $[\overline{n,m}]$, $[a,b,\overline{n,m}]$.

Exercise 6.7. From the expressions given in Section 6.2, use Theorem 6.6 to compute the first 6 convergents of $\pi - 3$, $e - 2$, θ, and g.

Exercise 6.8. a) In exercise 6.7, numerically check how close the nth convergent of ω is to the actual value of ω.
b) Compare your answer to (a) with the decimal expansion approximation using i digits.
c) In exercise 6.7, check that the increasing/decreasing patterns of the approximants satisfies the one described in the proof of Theorem 6.9.

Exercise 6.9. What does the matrix in Theorem 6.6 correspond to in terms of the Euclidean algorithm of Chapter 3?

Exercise 6.10. Use Lemma 6.13 to show that

$$\left|\omega - \frac{p_{2n+1}}{q_{2n+1}}\right| < \frac{1}{q_{2n+1}} \prod_{i=1}^{n} \frac{1}{1+a_{2i+1}} .$$

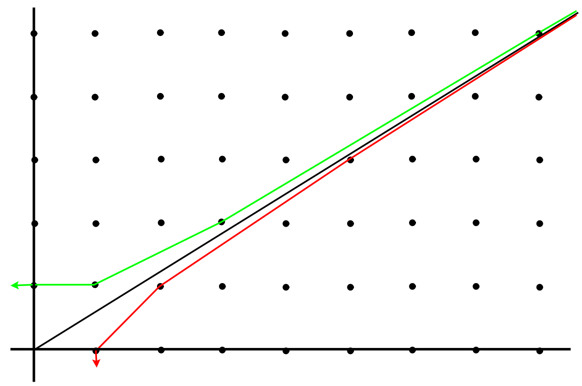


Figure 32. Black: thread from origin with golden mean slope; red: pulling the thread down from the origin; green: pulling the thread up from the origin.

Exercise 6.11. (Adapted from [**5**]) Consider the line ℓ given by $y = \omega x$ with $\omega \in (0,1)$ an irrational number. Visualize a thread lying on the line ℓ fastened at infinity on one end and at the origin at the other. An infinitely thin pin is placed at every lattice point in the positive quadrant. Since the slope of the thread is irrational, the thread touches none of the pins (except the one at the origin). Now remove the pin at the origin and pull the free end of the thread downward towards e_0 (as defined in the text). The thread will touch the pin at e_0 and certain other pins with slopes less than ω. Mark the nth of those pins as f_{2n} for $n \in \mathbb{N}$. Denote the points of the positive quadrant now lying *on or below* the thread by A. Next, pull the thread up towards e_{-1}. Mark the pins the thread touches, starting with e_{-1} as f_{2n-1} for $n \in \mathbb{N}$. Denote the points of the positive quadrant now lying *or or above* the thread by B. See Figure 32.
a) Show that $A \cup B$ contain all the lattice points of the positive quadrant.
b) Show that for all $n \in \mathbb{N}$, $f_n = (q_n, p_n)$ where (q_n, p_n) are as defined in the text.
c) Compute the slopes of the upper boundary of the region A. The same for the lower boundary of the region B.
d) Show that A and B are convex sets.

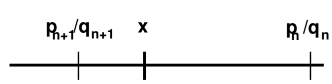


Figure 33. The placement of x between its convergents p_n/q_n and p_{n+1}/q_{n+1}.

Exercise 6.12. Assume x is irrational.
a) Use Corollary 6.7(ii) and Corollary 6.11 to show that

$$\left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n q_{n+1}} \,.$$

b) Use Lemma 6.13 to show that

$$\left| x - \frac{p_{n+1}}{q_{n+1}} \right| < \left| x - \frac{p_n}{q_n} \right| \,.$$

c) Use (a), (b), and Figure 33 to show that

$$\frac{1}{2q_n q_{n+1}} < \left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n q_{n+1}} \,.$$

(*Hint: note that x is closer to $\frac{p_{n+1}}{q_{n+1}}$ than to $\frac{p_n}{q_n}$.*)

Exercise 6.13. Use exercise 6.12 to generate bounds for the errors computed in exercise 6.8. Compare your answers.

Exercise 6.14. Use exercise 6.12 (a) to prove Theorem 1.15.

Exercise 6.15. a) Let $x \in [0,1)$ have periodic coefficients a_i. Show that x satisfies $x = \frac{ax+b}{cx+d}$ where a, b, c, and d are integers. (*Hint: see Section 6.3.*)
b) Show that x in (a) is an algebraic number of degree 2 (See Definition 1.17).
c) Show that if $x \in [0,1)$ has eventually periodic coefficients a_i, then x is an algebraic number of degree 2.

This is one direction of the following Theorem.

Theorem 6.17. *The continued fraction coefficients $\{a_i\}$ of a number x are eventually periodic if and only if x is an algebraic number of degree 2.*

It is not known (in 2022) whether the continued fraction coefficients of algebraic numbers of degree 3 exhibit a recognizable pattern.

Exercise 6.16. A natural question that arises is whether you can formulate continued fractions for polynomials in $\mathbb{Q}[x]$ or $\mathbb{R}[x]$ (suggested to us by [**20**]). We try this for the rational function $f(x) = \frac{x^3+x^2}{x^7-x^2+1}$. Referring to exercise 3.22 and the definition of a_i in the remark after Definition 6.3, we see that

$$\begin{array}{rcl} a_1 & = & (x^4 - x^3 + x^2 - x + 1) \\ a_2 & = & (-\frac{1}{2}x - \frac{1}{2}) \\ a_3 & = & (-4x+4) \\ \text{and } a_4 & = & (-\frac{1}{2}x - \frac{1}{2}) \end{array}$$

a) Compute the continued fraction convergents $\frac{p_n}{q_n}$ for $n \in \{1, \cdots 4\}$ of $f(x)$. (*Hint: perform the computations as given in Theorem 6.6.*)
b) In (a), you obtained the polynomials of exercise 3.22 up to a factor -1. Why? (*Hint: The* gcd *we computed in that exercise is actually -1. As stated in that exercise, we neglect constants when using the algorithm for polynomials. At any rate, in the quotient, the constant cancels.*)
c) Is there a theorem like the one in exercise 6.14?
d) Solve for y: $y = [\overline{x}]$. (*Hint: check exercise 6.6*)
e) Any ideas for other non-rational functions? (*Hint: check the web for Padé approximants.*)

Exercise 6.17. What is the mistake in the following reasoning?
We "prove" that countable = uncountable. First we show that a countably infinite product of countably infinite sets is countable.
$n = \prod_{i=1}^r p_i^{\ell_i}$ and there are infinitely many primes. Thus we can encode the natural numbers as infinite sequences $(\ell_1, \ell_2, \ell_3, \cdots)$ of natural numbers. That gives a bijection of infinite product of $\mathbb{N}$'s to $\mathbb{N}$. Therefore an infinite product of $\mathbb{N}$ is countable.
On the other hand, an infinite number of natural numbers $[q_1, q_2, \cdots]$ can be used to give the real numbers in $(0,1)$ in terms of their continued fraction expansion. This gives of bijection on to the interval. Therefore the infinite product of $\mathbb{N}$ is uncountable.

Exercise 6.18. Consider Figure 34. The first plot contains the points $\{(n,n)\}_{n=1}^{50}$ in standard polar coordinates, the first coordinate denoting the radius and the second, the angle with the positive x-axis in radians. The next plots are the same, but now for n ranging from 1 to 180, 330, and 2000, respectively.
a) Determine the first 4 continued fraction convergents of 2π.
b) Use (a) to explain why we appear to see 6, 19, 25, and 44 spiral arms.
c) Why does the curvature of the individual spiral arms appear to (a) alternate in sign and (b) decrease?

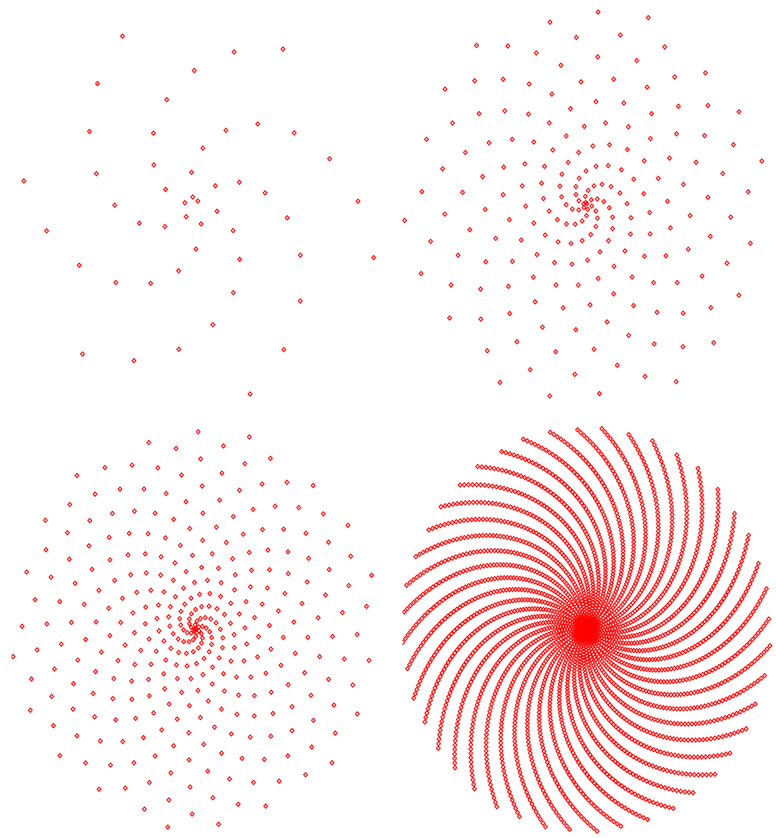


Figure 34. Plots of the points (n,n) in polar coordinates, for n ranging from 1 to 50, 180, 330, and 3000, respectively.

Exercise 6.19. The exercise depends on exercise 6.18. Suppose we restrict the points plotted in that exercise to primes (in $\mathbb{N}$) only. Consider the last plot (with 44 spiral arms) of Figure 34.
a) Show that each spiral arm corresponds to a residue class i modulo 44.
b) Show that if $\gcd(i,44) > 1$, that arm contain no primes (except possibly i itself), see the left plot of Figure 35.
c) Use Theorem 6.18 below to show that the primes tend (as $\max p \to \infty$) to be equally distributed over the co-prime arms.
d) Use Theorem 4.17 to determine the number of co-prime arms. Confirm this in the left plot of Figure 35.
e) Explain the new phenomenon occurring in the right plot of Figure 35.

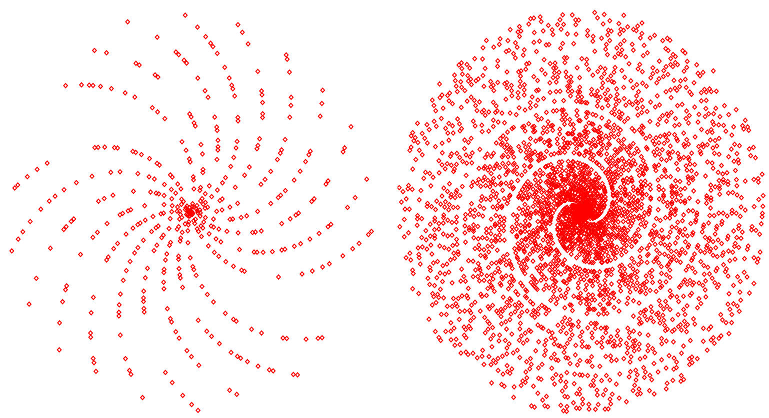


Figure 35. Plots of the prime points (p, p) (p prime) in polar coordinates with p ranging between 2 and 3000, and between 2 and 30000, respectively.

The following result will be proved in Chapter 13.

Theorem 6.18 (Prime Number Theorem for Arithmetic Progressions). *For given n, denote by r any of its reduced residues. Let $\pi(x;n,r)$ stand for the number of primes p less than or equal to x such that $Res_n(p) = r$. Then*

$$\lim_{x\to\infty} \frac{\pi(x;n,r)}{\pi(x)} = \frac{1}{\varphi(n)} .$$

Exercise 6.20. a) Visualize the continued fraction expansion of another irrational number $\rho \in (0,1)$ by plotting a polar plot of the numbers $(n, \frac{2\pi}{\rho} n)$ for various ranges of n as in exercise 6.18.
b) Visualize Theorem 6.18 as in exercise 6.19 (e).

Call $\frac{p}{q}$ is a closest approximation of ω if

$$|\omega - \frac{p}{q}| < |\omega - \frac{p'}{q'}|$$

for all q' such that $0 < q' < q$ and for all p'. In the following two exercises, we show that it is *not* true that p/q is a closest approximation if and only if it is a continued fraction. See Theorem 6.14.

Exercise 6.21. Set $\omega = e - 2 \approx 0.71828$.
a) Compute a_1 through a_4 numerically.
b) From (a), compute the convergents p_i/q_i for $i \in \{1,2,3\}$.
c) Show that $1/2$ (which is not a convergent) is a closest approximant. d) Show that $1/2$ is not a closest return in the sense of Theorem 6.14.

Exercise 6.22. Set $\omega = [\overline{3}] = -\frac{1}{2}(-3+\sqrt{13}) \approx 0.302775638$.
a) Check that $\omega =$ and that

$$\frac{p_0}{q_0} = \frac{0}{1}\,, \quad \frac{p_1}{q_1} = \frac{1}{3}\,, \quad \frac{p_2}{q_2} = \frac{3}{10}\,, \quad \frac{p_3}{q_3} = \frac{10}{33}\,, \quad \frac{p_4}{q_4} = \frac{33}{109}\,.$$

b) For $a \in \{1,2\}$ and $i \in \{1,2,3\}$, (numerically) check which ones of the numbers $\frac{ap_{i+1}+p_i}{aq_{i+1}+q_i}$ approximate ω better than $\frac{p_i}{q_i}$.
c) Use (b) to show that the continued fraction convergents are not the only closest approximants.

Exercise 6.23. Let $x^2 - bx - c = 0$, where b and c are positive integers.
a) Show that $x = \sqrt{c+bx}$ gives the unique positive solution.
b) Use (a) to show that *formally*

$$x = \sqrt{c + b\sqrt{c + b\sqrt{c + \cdots}}}\,.$$

c) Set $f(x) := \sqrt{c+bx}$. Show that the sequence $x_1 = \sqrt{c}$, $x_2 = \sqrt{c+b\sqrt{c}}$, et cetera, corresponds to $x_1 = f(0)$, $x_2 = f(f(0))$, and so on.
d) Show that these repeated images of 0 under f converge to the positive solution of (a) by studying figure 36.

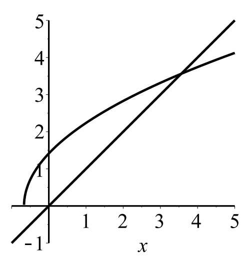


Figure 36. Left, the functions $f(x) = \sqrt{2+3x}$ and x. The fixed point x^* is the solution of $f(x) = x$ is the value sought in exercise 6.23. Show that for positive x, $\frac{x^*-f(x)}{x^*-x}$, and that that implies that $x_n \to x^*$.

Exercise 6.24. Let $x^2 - bx - c = 0$, where b and c are positive integers. See exercise 6.23.
a) Show that $x = b + c/x$ for $x > 0$, gives the unique positive solution.
b) Use (a) to show that *formally*

$$x = b + \frac{c}{b + \frac{c}{b + \cdots}}\,.$$

(*Note: this is an example of what is known as a generalized continued fraction . In turn, these can be generalized to arbitrary (finite) dimension* [**11**].)
c) Set $f(x) := b + c/x$. Show that the sequence $x_1 = b$, $x_2 = b + c/b$, $x_3 = b + \frac{c}{b + \frac{c}{b}}$ et cetera, corresponds to $x_1 = b$, $x_2 = f(b)$, $x_3 = f(f(b))$, and so on.
d) Show that these repeated images of 0 under f converge to the positive solution of (a) by studying

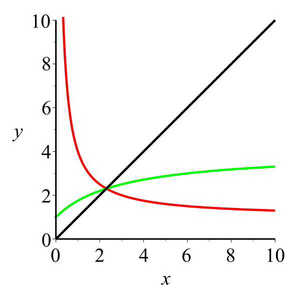


Figure 37. The functions $f(x) = 1 + 3/x$ (red), $f(f(x)$ (green), and x (black). Since $f(f(x))$ is increasing slowly, the sequence $\{0, f(0), f(f(0)), \cdots\}$ must converge to the fixed point.

Chapter 7

Fields, Rings, and Ideals

Overview. The characteristics of $\mathbb{Z}$ are so familiar to us, that it is hard to break through that familiarity to understand what makes things like unique factorization tick. Algebraic number theory and with it large swaths of algebra were developed to deal with more general number systems in order to overcome this problem. So in this chapter, we initially move away from numbers a little to study concepts of abstract algebra. This discipline of mathematics seems to start with a daunting barrage of definitions or *nomenclature*[1]. Here, we look at some of these and relate them as much as possible to their origins in number theory. An excellent introduction to abstract algebra is [**54**], while [**34**] is a standard among the more advanced texts.

7.1. Rings of Polynomials

Since one of our aims is to study factorization properties in certain sets of algebraic integers — which are defined through polynomials — we need to start by studying sets of polynomials. Broadly speaking, there are two important cases. The coefficients of the polynomials belong either to a *ring* such as $\mathbb{Z}$ or — an important special case — they belong to a *field* such as $\mathbb{Q}$. In what follows we denote a ring by R and a field by F.

[1] From Latin *nomen* or 'name' and *calare* or 'to call'. So — taken quite literally — *name-calling*.

Definition 7.1. *A ring $R[x]$ of polynomials is the set of polynomials with coefficients in a (commutative) ring R without zero divisors*[2] *(unless otherwise mentioned).*

Without the extra requirements, the resulting ring would have very strange properties indeed. For example, if R consists of the integers modulo 6, then, indeed, very strange factorizations can happen:

$$(2x-3)(3x+2) =_6 6x^2 - 5x - 6 =_6 x\,.$$

So, in particular, the degree of the product is not equal to the sum of the degrees of the factors. Dropping commutativity would lead to another strange problem. Given $f \in R[x]$, we may want to evaluate f at $c \in R$ by substituting the value c for x. Suppose for example that R is the non-commutative ring of 2 by 2 matrices. Set for some $a \in R$,

$$f(x) = (x-a)(x+a) = x^2 - a^2\,.$$

But if we substitute another 2 by 2 matrix c for x such that the matrices a and c do not commute, then the above equality does not hold anymore. However, if R satisfies the two requirements, one can prove that the resulting polynomial ring has no zero divisors, evaluations are safe, and that the degree of a product is additive (see [**34**][sections 8.5 and 8.6] for details).

Definition 7.2. *Recall (Definition 1.18) that f is minimal polynomial in $R[x]$ for ρ if f is a non-zero polynomial in $R[x]$ of minimal degree such that $f(\rho) = 0$. A polynomial f in $R[x]$ of positive degree is irreducible over R if it cannot be written as a product of two polynomials in $R[x]$ with positive degree. A polynomial f in $R[x]$ is prime over R if if whenever f divides gh (g and h in $R[x]$), it must divide g or h.*

Definition 7.3. *Let f and g in $R[x]$. The greatest common divisor of f and g, or $\gcd(f,g)$, is a polynomial in $R[x]$ with maximal degree that is a factor of both f and g. The least common multiple of f and g, or $\operatorname{lcm}(f,g)$, is a polynomial in $R[x]$ with minimal degree that has both f and g as factors.*

Remark 7.4. If p is minimal for ρ, it must be irreducible, because if not, one of its factors with smaller degree would also have ρ as a root.

[2] This means that if for a, b in R, we have that $ab = 0$, then $a = 0$ or $b = 0$, see Definition 8.4.

It turns out that in the special case where the coefficients of the polynomials are taken from a *field* F, the result is a ring $F[x]$ that is very reminiscent of the trusty old ring $\mathbb{Z}$. The underlying reason for this similarity is that in $F[x]$, the division algorithm works (see exercise 7.1): given r_1 and r_2, then there are q_2 and r_3 such that[3]

$$r_1 = r_2 q_2 + r_3 \quad \text{such that} \quad deg(r_3) < \deg(r_1)\,.$$

Recall that the gcd of two polynomials in $F[x]$ can be computed by factoring both polynomials and multiplying together the common factors to the lowest power as in the proof of Corollary 2.23. Since factoring polynomials is hard, it is often easier to just use the Euclidean algorithm. An example is given in exercise 3.22. The relation between lcm and gcd of two polynomials is the same as in the proof of Corollary 2.23. The minimal polynomials of $F[x]$ are "like" the primes in $\mathbb{Z}$. We will see later that this implies unique factorization, and that primes and irreducibles are the same[4]. We give a few properties that will be immediately useful[5]

Proposition 7.5. *Given $\rho \in \mathbb{C}$ and $p \in F[x]$ so that $p(\rho) = 0$ and $p' \neq 0$.*
i) p is minimal for ρ if and only if p is irreducible.
ii) If p is minimal, it has no repeated roots.

Proof. If p is minimal, see Remark 7.4. On the other hand, if f is irreducible and p is minimal for ρ, then the division algorithm tells us that there are polynomials q and r such that

$$f = pq + r\,,$$

where r has degree strictly less than p. Since $p(\rho) = f(\rho) = 0$, we have $r(\rho) = 0$. But since p is minimal, we must have $r(x) = 0$. Thus $p \mid f$. But f is irreducible, so q must be a constant and f is also minimal. This proves (i).

To prove (ii), suppose that p has a repeated root α. Since $p \in F[x]$, we have that also p' (its derivative) in $F[x]$. But if

$$p(x) = (x-\alpha)^2 r(x) \quad \text{then} \quad p'(x) = 2(x-\alpha)r(x) + (x-\alpha)^2 r'(x)\,.$$

[3] Since remark 3.17, we adopt the convention that the degree of a non-zero constant equals 0, while the degree of 0 equals $-\infty$.

[4] In fact, the fact that the division algorithm works, makes this ring a Euclidean domain (Definition 8.11).

[5] But in Corollary 8.13 we will get much more: irreducibles equal primes and unique factorization.

The latter is non-zero and of lower degree and still has a root α. This contradicts the minimality of p. ■

An even simpler argument gives the following result.

Lemma 7.6. *Given $\rho \in \mathbb{C}$ and p minimal for ρ in $F[x]$. If $f \in F[x]$ has a root ρ, then $p \mid f$.*

Proof. We use again the division algorithm to establish that

$$f = pq + r\,,$$

where r has degree less than p. Since $f(\rho) = p(\rho) = 0$, also $r(\rho)$ must be zero, contradicting the minimality of p *unless* $r(x) = 0$. The lemma follows. ■

Theorem 7.7. *Given $a(x)$ and $b(x)$ in $F[x]$, there are g and h in $F[x]$ satisfying*

$$a(x)\,g(x) + b(x)\,h(x) = c(x)$$

if and only if c is a multiple of $\gcd(a,b)$.

Proof. We paraphrase the proof of Lemma 2.5 with "degree" replacing "absolute value". Let S and $\nu(S)$ be the sets:

$$\begin{array}{rcl} S & = & \{a(x)g(x) + b(x)h(x) \,:\, a(x)g(x) + b(x)h(x) \neq 0\} \\ \nu(S) & = & \{\deg(s) : s \in S\} \subseteq \mathbb{N} \cup \{-\infty, 0\}\,. \end{array}$$

Again $\nu(S)$ is non-empty, and so by well-ordering, it must have a smallest element, say δ, the degree of a polynomial $d(x)$. If $\delta = 0$, then $d(x)$ is a constant $\gamma \in F$. After dividing by γ, we see that $\gcd(a,b) = 1$ since no common factor can have degree less than 0.

If $\delta > 0$, we use the division algorithm exactly as in the proof of Lemma 2.5 and conclude that $d(x)$ is a divisor (or factor) of both $a(x)$ and $b(x)$.

Suppose e is a factor of both a and b. Since $d(x) = a(x)g(x) + b(x)h(x)$, we see that e must also be a factor of d. And thus d is the greatest common divisor.

The proof is finished by repeating the last paragraph of the proof of Lemma 2.5 to show that $a(x)\,g(x) + b(x)\,h(x) = c(x)$ has a solution if and only if c is a multiple of d. ■

Next, we present a result that holds for more general rings of the form $R[x]$ (though not for all). For simplicity, however, we give the result for $\mathbb{Z}[x]$. It says that if we can factor a polynomial in $\mathbb{Z}[x]$ as a product of polynomials with rational coefficients, then, in fact, those coefficients are integers.

Lemma 7.8 (**Gauss' Lemma**). *Let $A_\ell \in \mathbb{Z}$, and $b_i, c_j \in \mathbb{Q}$. If*

$$\sum_{\ell=0}^{m+n} A_\ell x^\ell = \left(\sum_{i=0}^{m} b_i x^i\right)\left(\sum_{j=0}^{n} c_j x^j\right),$$

then $b_i, c_j \in \mathbb{Z}$.

Proof. Let $A := \gcd(\{A_\ell\})$ and set $a_\ell = A_\ell/A$. In addition, we fix integers B and C such that Bb_i and Cc_j are integers and $\gcd(\{Bb_i\}) = \gcd(\{Cc_j\}) = 1$. We then get

$$\sum_{\ell=0}^{m+n} ABCa_\ell x^\ell = \left(\sum_{i=0}^{m} Bb_i x^i\right)\left(\sum_{j=0}^{n} Cc_j x^j\right).$$

We now show that $ABC = \pm 1$ and so all three are ± 1. Given any prime p in $\mathbb{Z}$, let r be the minimum of the index i such that $p \nmid Bb_i$, and s the minimum of the index j such that $p \nmid Cc_j$. From the way the coefficient $ABCa_{r+s}$ is computed, see Figure 38, it immediately follows that $p \nmid ABCa_{r+s}$. Since we can find such r and s for every prime p, the result follows. ■

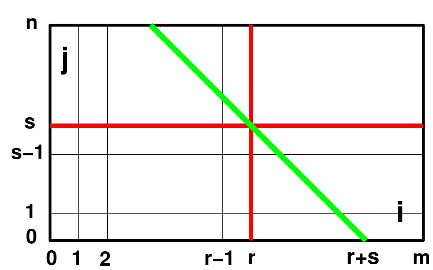


Figure 38. $ABCa_{r+s}$ is the sum of the Bb_iCc_j along the green line in the $i-j$ diagram. The red lines indicate where $p \nmid Bb_i$ and $p \nmid Cc_j$. So all contributions except Bb_rCc_s are divisible by p. Thus $p \nmid ABCa_{r+s}$.

A similar argument gives another useful result.

Lemma 7.9 (Eisenstein's Criterion). *Let $a(x) = \sum_{i=0}^{d} a_i x^i$ be a polynomial in $\mathbb{Z}[x]$ of degree $d \geq 2$. If there is a prime p such that*

$$p \mid a_i \ \text{ if } \ i < d \quad \text{and} \quad p \nmid a_d \quad \text{and} \quad p^2 \nmid a_0\,,$$

then a is irreducible over $\mathbb{Q}$.

Proof. Suppose a is reducible over $\mathbb{Q}$, then by Lemma 7.8, it is reducible over $\mathbb{Z}$. So let $a(x) = b(x)c(x)$ where b has degree $m > 0$ and c has degree $n > 0$ and $d = m+n$. Since $a_d = b_m c_n$ and $a_0 = b_0 c_0$, the requirements on the a_i give, without loss of generality,

$$p \nmid b_m \quad \text{and} \quad p \nmid c_n \quad \text{and} \quad p \nmid b_0\,.$$

There is a smallest integer $s \leq n < d$ such that $p \nmid c_s$. We compute a_s:

$$a_s = \sum_{i=0}^{\min\{s,m\}} b_i c_{s-i}\,.$$

All terms on both sides are divisible by p except $b_0 c_s$, which is impossible. ∎

We end this section with a note on some notation that can be confusing. We can "adjoin" x to a ring R in two ways. If we use square brackets $[\cdot]$, we take $R[x]$ to be the minimal (smallest) *ring* that contains both R and x. On the other hand, parentheses $(\cdot)$ are used to indicate the minimal (smallest) *field* that contains both R and x. On the other hand, A little reflection leads to the following definition.

$$\begin{aligned} R[x] \quad &:= \quad \{f(x) \,:\, f \text{ is a polynomial over } R\}\,, \\ R(x) \quad &:= \quad \left\{ \frac{f(x)}{g(x)} \,:\, f,g \text{ are polynomials over } R \right\}. \end{aligned} \tag{7.1}$$

Here, x can be a place holder or an actual number. In the former case, $R(x)$ denotes the rational functions in x, and $R[x]$ are the polynomials.

The ring of power series (not just polynomials of finite degree) is indicated by $R[[x]]$. For a *field* F, the field of quotients or fractions $F[[x]]$ is written as $F((x))$. This field consists of the quotients of power series. Consider $f(x) = \sum_{i=0}^{\infty} a_i x^i$ and $g(x) = \sum_{i=0}^{\infty} b_i x^i$. Then if $b_0 \neq 0$, the quotient f/g can be formally reduced to a power series:

$$\frac{f(x)}{g(x)} := c_0 + c_1 x + c_2 x^2 + \cdots = \frac{a_0}{b_0} + \left(\frac{a_1}{b_0} - \frac{a_0 b_1}{b_0^2} \right) x + \cdots\,. \tag{7.2}$$

If $b_0 = 0$ and $b_1 \neq 0$, then employing the same method (exercise 7.2) we get

$$\frac{f(x)}{g(x)} = \frac{1}{x}\left\{\frac{a_0}{b_1} + \left(\frac{a_1}{b_1} - \frac{a_0 b_2}{b_1^2}\right)x + \cdots\right\}. \tag{7.3}$$

Continuing this way, we see that the $F((x))$ is the set of formal Laurent series (which is how it is usually defined):

$$F((x)) = \left\{\sum_{i=n}^{\infty} c_i x^i \ : \ n \in \mathbb{Z} \ \text{and} \ c_i \in F\right\}.$$

For a *ring* R, the notation $R((x))$ is best avoided because it is ambiguous: in this case the field of quotients is not the same as the set of Laurent series over R.

7.2. Ideals

Definition 7.10. *A non-empty subset I of a ring R is called an ideal*[6] *if*
i) For all i and j in I, $i \pm j$ is in I (closed under addition and negatives).
ii) For all x in R and i in I, xi and[7] *ix are in I (it "absorbs" products).*
The smallest ideal containing the elements i and j will be indicated[8] *by $\langle i, j\rangle$.*
A principal ideal is an ideal that is generated by a single element, that is: it is of the form Ri. An ideal I is a maximal ideal if there is no other ideal L so that $I \subsetneq L \subsetneq R$.

To guide our considerations, we look at $\mathbb{Z}$ first. In $\mathbb{Z}$ it is clear that for any $j \in \mathbb{Z}$, the corresponding ideal $\langle j\rangle$ is given by the set $j\mathbb{Z}$ of integer multiples of j. The relation $3 \mid 15$ can now be replaced by $\langle 3\rangle \supseteq \langle 15\rangle$.

Addition of ideals is defined as in the following example

$$\langle 6\rangle + \langle 15\rangle := \{n + m : n \in \langle 6\rangle, m \in \langle 15\rangle\} = \langle \gcd(6, 15)\rangle.$$

Notice that the last equality is not trivial. It in fact encodes Bézout's lemma (Lemma 2.5). In turn, this says that $\langle 6\rangle + \langle 15\rangle$ is the smallest ideal containing both $\langle 6\rangle$ and $\langle 15\rangle$. We also say that $\langle 6\rangle + \langle 15\rangle$ is the ideal *generated by*

[6] Usually "fraktur" letters ($\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{c}$ ...) are used for ideals. On a blackboard or whiteboard, these are hard to distinguish from normal letters. So instead we will use capital letters to indicate ideals.

[7] One of the two is sufficient if R is commutative.

[8] In most texts parentheses $(\cdot)$ are used. We want to avoid ambiguity with the notation for an n tuple $(i, j, \cdots)$.

6 and 15. This is more conveniently written as $\langle 6, 15\rangle$. More generally, for ideals A and B, we have that

$$A+B := \{a+b \,:\, a \in A, b \in B\} \ . \tag{7.4}$$

This example also illustrates the fact that $\langle 6\rangle + \langle 15\rangle$ is a *principal* ideal. In fact, in $\mathbb{Z}$, it is easy to see that every ideal I is principal. One can use Bźout to show that I is generated by its least positive element. Another non-trivial example of a principal ideal is the set of polynomials q satisfying $q(\rho) = 0$ in the ring of polynomials over a field F. Indeed, we need to refer to Lemma 7.6 to establish that this is the case (work out the details in exercise 7.8).

Next, we look at multiplication of ideals . If ideals are to behave like numbers, then the product of two ideals should also be an ideal. At first glance, one would think the collection of products of one element in $\langle 6\rangle$ and one in $\langle 15\rangle$ would do the trick. This is indeed the case in $\mathbb{Z}$ (exercise 7.4). However, in general this construct is *not* closed under addition (exercise 7.5). Thus we define AB as the smallest ideal containing the products of one element in A and one in B, or

$$AB := \left\{ \sum_{i=1}^{k} a_i b_i \,:\, a_i \in A, b_i \in B, k \in \mathbb{N} \right\} . \tag{7.5}$$

The relation between ring and ideal is very similar to the one between group and normal subgroup (Definition 7.30). In fact, since a ring R is an Abelian group with respect to addition, any ideal in R is a normal subgroup. There is one interesting difference: a normal subgroup is also a group. In contrast an ideal (like the even numbers) may not have a multiplicative identity and so it is not a ring (see Remark 5.24). The remainder of this section spells out the relation between rings and their ideals.

Definition 7.11. *Given two rings I and J, a ring homomorphism is a map $f : I \to J$ that preserves addition and multiplication and their respective identities 0 and 1. The kernel of a ring homomorphism is the pre-image of the additive identity 0. A ring isomorphism is a ring homomorphism that is a bijection. The word "'ring" is often omitted.*

Proposition 7.12. *i) The quotient R/K of a ring R by an ideal K is a ring. ii) The kernel K of a ring homomorphism $f : R \to H$ is an ideal.*

Proof. K is an ideal and thus a normal subgroup of the Abelian additive group R. Thus R/K is a group under addition (exercise 7.6). We have to show that multiplication is well-defined and is associative, distributive, and has an identity (Definition 5.20).

Multiplication in R/K is well-defined if for all a, a', b, and b' in R such that $a-a'$ and $b-b'$ are in K, we have

$$(a+K)(b+K)-(a'+K)(b'+K) \subseteq K\,.$$

The left hand side can be expanded as

$$ab-a'b'+(a-a')K+K(b-b')+K\cdot K=$$

$$(a-a')b+a'(b-b')+(a-a')K+K(b-b')+K\cdot K\,.$$

The *absorption* property of the product does the rest.

Associativity and distributivity now follow easily. For example, since $[ab]c=a[bc]$ in R and multiplication is well-defined, we must have

$$[(a+K)(b+K)]\,(c+K)=(a+K)\,[(b+K)(c+K)]\,.$$

Similarly for distributivity. Again, by absorption, $(a+K)(1+K)\subseteq(a+K)$ and so $1+K$ is the multiplicative identity. This proves (i).

The proof of (ii) is rather trivial. Just use Definitions 7.10 and 7.11. Choose x and y in the kernel of f and conclude that $f(x\pm y)=0$ and that for any $r\in G$, $f(rx)$ also equals 0. ■

Theorem 7.13 (Fundamental Homomorphism Theorem). *If $f:R\to H$ is a surjective ring homomorphism with kernel K, then H is (ring) isomorphic to R/K.*

Proof. Define the map $\varphi:R/K\to H$ by

$$\varphi(K+x):=f(x)\,.$$

We need to prove that (a) φ is a bijection, that (b) it preserves addition and that (c) it preserves multiplication.

To prove (a), note that φ is onto because f is. So next suppose that $\varphi(K+x)=\varphi(K+y)$. Because f preserves addition, we get $f(x-y)=0$ and therefore $x-y\in K$. Injectivity follows: because $K+x=(K+(x-y))+y$ and $K+K=K$, we get $K+x=K+y$.

The proofs of (b) and (c) are almost identical. We prove only (c).

$$\varphi(K+x)\varphi(K+y) = f(x)f(y) = f(xy) = \varphi(K+xy).$$

But by the absorbing property of ideals, $(K+x)(K+y) = K+xy$. ■

The idea that quotients of certain structures are isomorphic to structures they map onto, is important not only in algebra (groups, modules) but also in topology and analysis (quotient spaces). For instance, $\mathbb{R}/\mathbb{Z}$ with the right topology is homeomorphic to the standard (unit) circle. See Figure 39

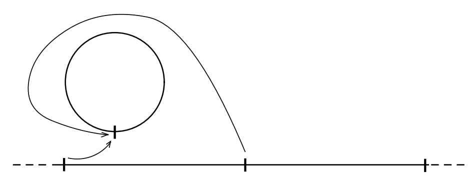


Figure 39. Intuitively we wrap $\mathbb{R}$ around a circle of length 1, so that points that differ by an integer land on the same point.

Theorem 7.13 has the surprising consequence, for example, that there are no non-trivial (ring) homomorphisms $\mathbb{C} \to \mathbb{R}$ (see exercise 7.7).

Corollary 7.14. *A ring homomorphism $f : F \to R$ where F is a field is either trivial (zero) or injective.*

Proof. If f is not injective, it has a non-trivial kernel, which by Theorem 7.13, is an ideal I in the field F. So I contains a a non-zero element i. Now pick any $x \in F$. Then by Definition 7.10 (ii), $xi^{-1} \cdot i = x$ is in I. Thus $I = F$, and hence $f(F) = 0$. ■

In many common cases, the conclusion if the fundamental homomorphism theorem is intuitively obvious. For example, we did not need it to prove that $\mathbb{Z}/5\mathbb{Z}$ is isomorphic to $\mathbb{Z}_5$. However, in Theorem 7.16 below the conclusion is not self-evident and we make essential use of it.

7.3. Fields and Extensions

As a first example, let us consider the field $\mathbb{Q}$ and adjoin the number π (or any other transcendental number). We denote the smallest field containing both by $\mathbb{Q}(\pi)$. The pair of fields $(\mathbb{Q}(\pi), \mathbb{Q})$ in this example is called a *field* *extension*. $\mathbb{Q}(\pi)$ is the *extension field* of $\mathbb{Q}$. By equation (7.1), it consists of all quotients of polynomials. Since π is transcendental, there exists *no* polynomial p with rational coefficients so that $p(\pi) = 0$. Thus none of these expressions simplify. Therefore this set is isomorphic to $\mathbb{Q}(x)$. An extension of this nature is also called a transcendental extension .

In order to get something both new *and* manageable, we should adjoin a number α to the field $\mathbb{Q}$ that requires us to take only *finitely* many powers of α into account. This is done by taking α to be an algebraic number. Such as extension is called *finite* or *algebraic* .

A simple example tells the whole story. Let us take $\alpha = \sqrt{2}$ and study $\mathbb{Q}(\sqrt{2})$. Clearly, $\alpha^{2+i} = 2\alpha^i$, so any polynomial over $\mathbb{Q}$ in α can be rewritten as $a + b\sqrt{2}$ with a and b in $\mathbb{Q}$. Any quotient of polynomials in α can therefore be written as

$$\frac{a+b\sqrt{2}}{c+d\sqrt{2}} = \frac{(a+b\sqrt{2})(c-d\sqrt{2})}{(c-d\sqrt{2})(c-d\sqrt{2})} = \frac{ac-2bd}{c^2-2d^2} + \frac{(bc-ad)\sqrt{2}}{c^2-2d^2}\,.$$

Since 2 is square free, the denominator is not zero and hence every element in the field $\mathbb{Q}(\sqrt{2})$ can be written as $e + f\sqrt{2}$ with e and f in $\mathbb{Q}$. This holds is general as the next result shows.

Proposition 7.15. *Let $F(\rho)$ a finite extension of a field F. Suppose p is a minimal polynomial for ρ and has degree d. Then, as sets,*

$$F(\rho) = \left\{ \sum_{i=0}^{d-1} a_i \rho^i \;:\; a_i \in F \right\}.$$

Proof. Clearly, $\{1, \rho, \cdots, \rho^{d-1}\}$ are independent over F (otherwise the minimal polynomial would have degree less than d) and since a field is closed under addition, subtraction, and multiplication, and so $F(\rho)$ must contain all expressions $\sum_{i=0}^{d-1} a_i \rho^i$.

We only need to check that $F(\rho)$ is closed under (multiplicative) inversion. So choose $b_i \in F$ such that $f(x) := \sum_{i=0}^{d-1} b_i x^i$ is not 0. The minimal polynomial p for ρ is irreducible (Proposition 7.5); it can have only trivial factors in common with f. Thus by Theorem 7.7, there are polynomials s and t such that

$$f(x)s(x) + p(x)t(x) = 1\,.$$

Using the minimal polynomial, s can be reduced to have degree less than d. Substitute ρ for x in this equation to obtain (since $p(\rho)=0$ and $f(\rho)\neq 0$)

$$s(\rho) = 1/f(\rho)\,.$$

Thus $F(\rho)$ is indeed closed under (multiplicative) inversion. ■

All we are doing in this last proof, really, is taking an arbitrary quotient f/g of polynomials f and g in ρ and reducing it using the minimal polynomial. That insight leads to a sharper result.

Theorem 7.16. *Let $F(\rho)$ a finite extension of a field F. Suppose p is a minimal polynomial for ρ. Then $F(\rho)$ is ring isomorphic to $F[x]/\langle p(x)\rangle$.*

Proof. Define a map $\sigma_\rho : F[x] \to F(\rho)$ as follows. Given a polynomial f,

$$\sigma_\rho(f) = f(\rho)\,.$$

Clearly, σ_ρ is a ring homomorphism, because

$$\sigma_\rho(f\cdot g) = \sigma_\rho(f)\sigma_\rho(g) \quad \text{and} \quad \sigma_\rho(f+g) = \sigma_\rho(f)+\sigma_\rho(g)\,.$$

Since

$$\sigma_\rho\left(\sum_{i=0}^{d-1} a_i x^i\right) = \sum_{i=0}^{d-1} a_i \rho^i\,,$$

Proposition 7.15 shows that σ_ρ is onto. By Proposition 7.12, the kernel of σ_ρ is an ideal, and by Lemma 7.6 it is the ideal $\langle p(x)\rangle$ generated by $p(x)$. Thus by the fundamental homomorphism theorem, $F(\rho)$ is isomorphic to $F[x]/\langle p(x)\rangle$. ■

Remark 7.17. The map σ_ρ is called an evaluation map .

This is all very well, but what if we adjoin another algebraic element, β, to $\mathbb{Q}(\alpha)$? What does $\mathbb{Q}(\alpha,\beta)$ look like? Are the results we just proved still useful? The answer, miraculously, is yes. And the reason is the primitive element theorem below (Theorem 7.19).

Let us look at an example again. Adjoin $\beta=\sqrt{3}$ to the previous example $\mathbb{Q}(\alpha)=\mathbb{Q}(\sqrt{2})$, and consider $\mathbb{Q}(\alpha,\beta)$. Since the squares of α and β are integers, it is clear that every element of $\mathbb{Q}(\sqrt{2},\sqrt{3})$ can be written as

$$a+b\sqrt{2}+c\sqrt{3}+d\sqrt{6} \qquad \text{where} \qquad a,b,c,d\in\mathbb{Q}\,.$$

What is *not* immediately obvious is that 1, $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{6}$ are linearly independent over the rationals, but let us assume that for now (see Lemma 7.33 in the exercises).

Remark 7.18. We obtain a 4 dimensional vector space over $\mathbb{Q}$ with a basis formed by the vectors $\{1, \sqrt{2}, \sqrt{3}, \sqrt{6}\}$.

Now we make the "inspired guess[9]" that in this example $\mathbb{Q}(\alpha+\beta)$ is identical to $\mathbb{Q}(\alpha,\beta)$! To verify that, set $\gamma = \sqrt{2}+\sqrt{3}$. Clearly, $\gamma \in \mathbb{Q}(\alpha,\beta)$ and so

$$\mathbb{Q}(\gamma) \subseteq \mathbb{Q}(\alpha,\beta)\,.$$

A simple computation indeed yields

$$\gamma^2 = 5+2\sqrt{6}\,, \quad \gamma^3 = 11\sqrt{2}+9\sqrt{3}\,, \quad \gamma^4 = 49+20\sqrt{6}\,. \tag{7.6}$$

And so $\gamma^3 - 9\gamma$ generates $\sqrt{2}$, $\gamma^3 - 11\gamma$ generates $\sqrt{3}$, while $\gamma^2 - 5$ generates $\sqrt{6}$. Thus

$$\mathbb{Q}(\alpha,\beta) \subseteq \mathbb{Q}(\gamma)\,.$$

We have established that $\mathbb{Q}(\gamma) = \mathbb{Q}(\alpha,\beta)$. That we can do this in general, is the content of the primitive element theorem.

Theorem 7.19 (Primitive Element Theorem). *Every finite extension of an infinite field is a simple extension.*

Remark 7.20. Any single element that generates the finite extension is called a primitive element.

Proof. Let F be an infinite field and $K := F(\alpha,\beta,\gamma,\cdots,\delta)$ a finite (algebraic) extension. Suppose we can find a single generator φ for α and β. We can then repeat the argument to find a generator θ for φ and γ, and so forth. Thus it is sufficient to prove this result for $F(\alpha,\beta)$.

Let p and q be minimal polynomials in $F[x]$ for α and β, respectively. The roots of p are $\{\alpha_i\}_{i=1}^m$ with $\alpha_1 \equiv \alpha$ and those of q are $\{\beta_i\}_{i=1}^n$ with $\beta_1 \equiv \beta$. Now define for $c \neq 0$ in F

$$r(x) := p(\alpha + c\beta - cx)\,.$$

[9] "Inspired guess" is pretentious way of saying that I do not want to say where I got this (but see the proof of Theorem 7.19).

This polynomial has several intriguing properties. First, it is a member of the field $F(\alpha + c\beta)[x]$, for it has coefficients in $F(\alpha + c\beta)$. Furthermore, its roots are given by

$$\alpha_1 + c\beta - cx = \alpha_i \quad \Longleftrightarrow \quad x_i = \frac{\alpha_1 - \alpha_i}{c} + \beta_1 \,.$$

For $i = 1$, we of course get $\beta = \beta_1$ as a root. But now, since F is infinite, we fix a value of c^* of c such that none of the other roots equals β_i for $i > 1$.

Since both $q \in F[x] \subseteq F(\alpha + c^*\beta)[x]$ and $r \in F(\alpha + c^*\beta)[x]$ and both have β as a root, Lemma 7.6 implies that the minimal polynomial d for β in $F(\alpha + c^*\beta)[x]$ must be a divisor of both q and r. But these two share only one root, and therefore $d \in F(\alpha + c^*\beta)[x]$ has degree one:

$$s(x) = a_1 x + a_0 = a_1(x - \beta)\,.$$

Clearly, the a_i are in $F(\alpha + c^*\beta)$, but then so does $\beta = a_0/a_1$, and the same holds for $\alpha = (\alpha + c^*\beta) - c^*\beta$. Thus $\alpha + c^*\beta$ generates $F(\alpha, \beta)$. ■

Thus a primitive element generates the whole field extension through addition and multiplication (and their inverses). In contrast, a primitive *root* (Definition 5.5) is an element of $\mathbb{F}_p$ (the elements of $\mathbb{Z}_p$ with addition and multiplication as operations) whose *powers* generate $\mathbb{F}_p$.

As mentioned in our last example, $\mathbb{Q}(\gamma)$ is in fact a vector space over $\mathbb{Q}$. From (7.6), it is clear that $\gamma^4 - 10\gamma^2 + 1 = 0$. Therefore $\mathbb{Q}(\gamma)$ has four basis vectors, like $\mathbb{Q}(\alpha, \beta)$, namely $\{1, \gamma, \gamma^2, \gamma^3\}$ span the space $\mathbb{Q}(\gamma)$. The scalars are elements of $\mathbb{Q}$. As such, it is somewhat confusingly denoted by $\mathbb{Q}(\gamma)/\mathbb{Q}$ in the literature, though this is not to be interpreted as a quotient. The ***dimension*** of the vector space is denoted by $[\mathbb{Q}(\gamma) : \mathbb{Q}]$ and is also commonly called the ***degree of the extension***. Notice that

$$[\mathbb{Q}(\alpha, \beta) : \mathbb{Q}] = [\mathbb{Q}(\alpha, \beta) : \mathbb{Q}(\alpha)] \cdot [\mathbb{Q}(\alpha) : \mathbb{Q}]\,.$$

This holds much more generally (see [**68**] or [**35**]).

7.4. The Algebraic Integers

We look at the ring of all algebraic integers and show that it unsuitable for the study factorization into primes or irreducibles for it has neither primes nor irreducibles.

Theorem 7.21. *The set $\mathscr{A}$ of algebraic integers forms a ring with no zero divisors.*

We can take advantage of the fact that algebraic integers are complex numbers, which in turn form a commutative field (and thus a ring) without zero divisors. Many of the properties mentioned in Definition 5.20 *as well as* the absence of zero divisors are thus automatically satisfied. To make a long story short, we only need to prove that $\mathscr{A}$ is closed under additive inversion, under addition, and under multiplication. The first is easy. Suppose that $\theta \in \mathscr{A}$ is a root of $x^d + a_{d-1}x^{d-1} + \cdots + a_0$, where the a_i are in $\mathbb{Z}$. Then, of course, $-\theta$ is a root of the same polynomial with the odd a_i replaced by $-a_i$. The remaining two criteria have a very interesting constructive proof. To understand it, we need to define the Kronecker product.

Definition 7.22. *Given two matrices A and B, their Kronecker product $A \otimes B$ is given by*

$$A \otimes B := \begin{pmatrix} A_{11}B & A_{12}B & A_{13}B & \cdots \\ A_{21}B & A_{22}B & \cdots & \cdots \\ \vdots & \vdots & & \end{pmatrix} .$$

Lemma 7.23. *Suppose that A and B be square matrices of dimension a and b, respectively and denote by I_a and I_b the identity matrices of the appropriate dimension. If A has eigenpair[10] (α,x) and B, (β,y). Then*
i) $A \otimes B$ has eigenpair $(\alpha\beta, x \otimes y)$, and
ii) $A \otimes I_b + I_a \otimes B$ has eigenpair $(\alpha+\beta, x \otimes y)$.

Proof. We have that

$$\begin{pmatrix} A_{11}B & A_{12}B & A_{13}B & \cdots \\ A_{21}b & A_{22}B & \cdots & \cdots \\ \vdots & \vdots & & \end{pmatrix} \begin{pmatrix} x_1 y \\ x_2 y \\ \vdots \end{pmatrix} = \begin{pmatrix} A_{11}x_1By + A_{12}x_2By + \cdots \\ A_{21}x_1By + A_{22}x_2By + \cdots \\ \vdots \end{pmatrix} ,$$

which equals $Ax \otimes By$ or $\alpha x \otimes \beta y$. Using Definition 7.22 again, it is easy to check that this in turn equals $\alpha\beta x \otimes y$. This proves item (i).

[10] This means that $Ax = \alpha x$.

By (i), $A \otimes I_b$ has eigenpair $(\alpha, x \otimes y)$, and $I_a \otimes B$ has eigenpair $(\beta, x \otimes y)$. Adding the two gives item (ii). ■

As an example, consider the matrices

$$A = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \quad \text{and } B = \begin{pmatrix} 0 & 3 \\ 1 & 0 \end{pmatrix},$$

with eigenvalues $\pm\sqrt{3}$ and $\pm\sqrt{2}$. We obtain:

$$A \otimes B = \begin{pmatrix} 0 & 0 & 0 & 6 \\ 0 & 0 & 2 & 0 \\ 0 & 3 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{and } A \otimes I_b + I_a \otimes B = \begin{pmatrix} 0 & 3 & 2 & 0 \\ 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 3 \\ 0 & 1 & 1 & 0 \end{pmatrix},$$

with eigenvalues $\pm\sqrt{6}$ (of multiplicity 2) and $\pm\sqrt{3} \pm \sqrt{2}$, respectively. The characteristic polynomials are $(x^2-6)^2$ and $x^4 - 10x^2 + 1$, respectively. The polynomial was obtained earlier from equation (7.6).

Proof of Theorem 7.21. We only need to prove that $\mathscr{A}$ is closed under addition and under multiplication. So let α and β be in $\mathscr{A}$. Then α is a root of a monic[11] polynomial $p_A(x)$ of degree a and the same for β and $p_B(x)$ of degree b. Suppose $p_A(x) = x^a + \sum_{i=0}^{a-1} a_i x^i$. The so-called *companion matrix* , that is: the $a \times a$ matrix whose characteristic polynomial equals p_A, is

$$A = \begin{pmatrix} 0 & 0 & 0 & \cdots & -a_0 \\ 1 & 0 & 0 & \cdots & -a_1 \\ 0 & 1 & 0 & \cdots & -a_2 \\ & \vdots & & \vdots & \\ 0 & \cdots & 0 & 1 & -a_{a-1} \end{pmatrix},$$

and similarly a matrix B can be defined for p_B. Now form the matrices mentioned in Lemma 7.23 (i) and (ii). Then $\alpha\beta$ and $\alpha+\beta$ are roots of these polynomials or of factors of these polynomials. Since a factor (over

[11] A *monic* polynomial is a polynomial whose leading term has coefficient 1.

$\mathbb{Z}$) of a monic polynomial is monic, we see that both $\alpha\beta$ and $\alpha+\beta$ are algebraic integers. ■

While, like $\mathbb{Z}$, the algebraic integers $\mathscr{A}$ form a ring, that ring does not "look" like $\mathbb{Z}$ *at all*! We will take this up later when we prove that $\mathscr{A}$ is dense in the complex numbers and has no irreducibles and no primes (Theorem 8.6). So to study factorization, we must look at more restricted collections of algebraic integers.

Examples of more restricted rings of integers are $\mathbb{Z}(\gamma)$, the ring consisting of numbers of the form $\sum_{i=0}^{d-1} c_i\gamma^i$ with $c_i \in \mathbb{Z}$, where γ is algebraic of degree d. To see that $\mathbb{Z}(\gamma)$ is a ring is trivial, since we do not have to worry about multiplicative inverses, which was the only complication in Proposition 7.15.

We end this section with a slightly confusing definition and a warning in the form of a Lemma.

Definition 7.24. *Consider the field $\mathbb{Q}(\gamma)$. The integers of $\mathbb{Q}(\gamma)$ are those elements in $\mathbb{Q}(\gamma)$ that are algebraic integers.*

This is *not* necessarily the same as the set $\mathbb{Z}(\gamma)$! As an example we will prove the lemma below in exercise 7.20.

Lemma 7.25. *Let j be square free. The integers of $\mathbb{Q}(\sqrt{j})$ are precisely the elements of the ring $\mathbb{Z}(\frac{1}{2}(1+\sqrt{j}))$ if $j =_4 1$, and $\mathbb{Z}(\sqrt{j})$ else.*

7.5. Rings of Quadratic Numbers and Modules

Let j be a non-zero square free integer $j \in \mathbb{Z}$ (see exercise 2.16) not equal to 0 or 1. Then $\sqrt{j}$ is a *algebraic integer* of degree 2. If j is negative, we can think of $\mathbb{Z}[\sqrt{j}]$ and $\mathbb{Q}[\sqrt{j}]$ as subsets of the complex plane. If j is positive, then they are subset of the real line. In both cases $\mathbb{Z}[\sqrt{j}]$ and $\mathbb{Q}[\sqrt{j}]$ are countable (see Theorem 1.26). All elements of $\mathbb{Z}[\sqrt{j}]$ are algebraic integers of degree 2, because they are roots of

$$(x-a-b\sqrt{j})(x-a+b\sqrt{j}) = x^2-2ax+a^2-b^2j = 0\,, \tag{7.7}$$

and that degree 2 polynomial cannot be factored over the integers.

We can look at $\mathbb{Z}[\sqrt{j}]$ as having two basis vectors

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 \quad \text{and} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \sqrt{j}\,.$$

The elements of $\mathbb{Z}[\sqrt{j}]$ are precisely the linear combinations $a \cdot 1 + b \cdot \sqrt{j}$. Just like a the vector space of remark 7.18! The only difference is that the "scalars" now belong to a ring and not a field. The resulting construction is called a module.

Definition 7.26. *A* <u>*module*</u> *M (or* <u>*left*</u> <u>*module*</u> *) is a set with the same structure as a finite-dimensional vector space, except that its scalars form a commutative ring R (and not a field as in a vector space). Scalars multiply the elements of M from the left. (If in a non-Abelian ring, scalars multiply from the right, the result is called a* <u>*right module*</u>*.)*

Next, we interpret multiplication by $\alpha = a + b\sqrt{j}$ in $\mathbb{Z}[\sqrt{j}]$ when $\sqrt{j}$ is an algebraic integer of degree 2. Clearly, it is linear, because

$$\alpha(c + d\sqrt{j}) = c\alpha 1 + d\alpha\sqrt{j}\,.$$

Therefore, α can be seen as a matrix. Identify 1 with $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\sqrt{j}$ with $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Then the equations $(a+b\sqrt{j})1 = a + b\sqrt{j}$ and $(a+b\sqrt{j})\sqrt{j} = bj + a\sqrt{j}$ can be rewritten as

$$A\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} \quad \text{and} \quad A\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} bj \\ a \end{pmatrix}\,.$$

Thus we can use elementary linear algebra to see that

$$A = \begin{pmatrix} a & bj \\ b & a \end{pmatrix}\,. \tag{7.8}$$

What is interesting here, is that the determinant of A

$$\det A = a^2 - b^2 j \tag{7.9}$$

is clearly an integer and cannot be zero, because if $jb^2 - a^2 = 0$, because j is square free. The beauty of this is that this allows us to study factorization in complicated rings like $\mathbb{Z}[\sqrt{j}]$ using the tools of a simpler ring, namely $\mathbb{Z}$. All we have to do is to phrase factorization in $\mathbb{Z}[\sqrt{j}]$ in terms of the *determinant of A*. In number theory, this is known as the *norm of* α.

Definition 7.27. *The field norm, or simply norm*[12]*, of an element* α *of* $\mathbb{Z}[\sqrt{j}]$ *or* $\mathbb{Q}[\sqrt{j}]$ *is the determinant of the matrix that represents multiplication by* α*. It will be denoted by* $N(\alpha)$*. The trace of that matrix will be call the trace of* α *and is denoted by* $T(\alpha)$*.*

A fundamental result about determinants from linear algebra ($\det AB = \det A \det B$) gives a handy rule.

Corollary 7.28. *The norm of a ring of quadratic integers is a completely multiplicative function:* $N(\alpha\beta) = N(\alpha)N(\beta)$*. (See Definition 4.2.)*

Remark 7.29. Suppose $\alpha = a + b\sqrt{j}$ in $\mathbb{Z}[\sqrt{j}]$. From equation (7.9), we also get $N(\alpha) = \alpha\overline{\alpha}$ where $\overline{\alpha} = a - b\sqrt{j}$. $\overline{\alpha}$ is called the conjugate of α. Note that if j is negative, the conjugate $\overline{\alpha}$ corresponds to the usual complex conjugate of α and so the norm $N(\alpha)$ corresponds to the usual absolute value squared $|\alpha|^2$.

All this can be seamlessly generalized to $\mathbb{Z}[\beta]$ where β is some algebraic number of degree $d > 2$. We then get a d-dimensional module.

[12]This is another case of assigning a name that gives rise to confusion: the "norm" as defined here can be negative! Nonetheless, this seems to be the most common name for this notion, and so we'll adhere to it.

7.6. Exercises

Exercise 7.1. The reader might want to review exercises 3.22 to 3.25 first. Let f and g in $F[x]$. We will show that there are polynomials q and r in $F[x]$ such that

$$f = gq + r \quad \text{and} \quad \deg(r) < \deg(g)\,. \tag{7.10}$$

a) Show that this holds if $\deg(g) > \deg(f)$.
b) Now let $n = \deg(f) \geq \deg(g) = m$ and $f(x) = \sum_{i=0}^{n} a_i x^i$ and $g(x) = \sum_{i=0}^{m} b_i x^i$. Define

$$f_j(x) = f(x) - \frac{a_n}{b_m} x^{n-m} g(x)\,,$$

where f_j has degree j. Show that $j \leq n-1$. (*Hint: by assumption, a_n and b_m are not zero.*)
c) Show that the computation in (b) can be repeated with f replaced by f_j as long as $j \geq m$. (*Hint: we are just formalizing long division here.*)
d) Show that $r(x) = f_i(x)$, where f_i is the first of the f_j to have degree less than m.
e) Show that the leading term of $q(x)$ in (7.10) is $\frac{a_n}{b_m} x^{n-m}$.

Exercise 7.2. We perform long division to divide $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots$ by $g(x) = b_0 + b_1 x + b_2 x^2 + \cdots$. In contrast to exercise 7.1, now consider the *constant term* as the *leading term*, the next leading term is the one linear in x, and so on.
a) Assume $b_0 \neq 0$, then $f - \frac{a_0}{b_0} g$ cancels the constant term. So the first term of the quotient equals $\frac{a_0}{b_0}$. Find the next two terms. (*Hint: see equation 7.2.*)
b) Assume $b_0 = 0$ and $b_1 \neq 0$. Divide f by $xg(x) = b_1 + b_2 x + \cdots$ using the method in (a). Find the first three terms of the quotient. (*Hint: see equation 7.3.*)

Exercise 7.3. We prove that $\mathbb{Z}(i) = \mathbb{Q}[i]$.
a) Show that $\mathbb{Z}[i]$ is the set $\{a + bi \,:\, a, b \in \mathbb{Z}\}$. (*Hint: $i^2 \in \mathbb{Z}$.*)
b) Show that $\mathbb{Z}(i)$ equals $\{(a+bi)/(c+di) \,:\, a, b, c, d \in \mathbb{Z}\}$.
c) From (b), rewrite $\mathbb{Z}(i)$ as $\{r + si \,:\, r, s \in \mathbb{Q}\}$.
d) Show that $\mathbb{R}[i] = \mathbb{C}$.

Exercise 7.4. a) Given two ideals $\langle a \rangle$ and $\langle b \rangle$ in $\mathbb{Z}$. Show that

$$\langle a \rangle \cdot \langle b \rangle = \left\{ \sum_{i=1}^{k} n_i m_i ab \,:\, n_i, m_i \in \mathbb{Z}, k \in \mathbb{N} \right\}.$$

b) Use (a) to prove that in $\mathbb{Z}$

$$\langle a \rangle \cdot \langle b \rangle = \langle ab \rangle\,.$$

Exercise 7.5. Consider the ideals $I = \langle 2, x\rangle$ and $J = \langle 3, x\rangle$ in $\mathbb{Z}[x]$.
a) Show that

$$I = g_1(x)2 + g_2(x)x \quad \text{and} \quad J = h_1(x)3 + h_2(x)x\,,$$

where g_i and h_i are arbitrary elements of $Z[x]$.
b) Show that $3x$ and $-2x$ *can* be written as $(g_1(x)2 + g_2(x)x)(h_1(x)3 + h_2(x)x)$.
c) Use (b) to show that x must be in the ideal IJ.
d) Show that x *cannot* be written as $(g_1(x)2 + g_2(x)x)(h_1(x)3 + h_2(x)x)$.
e) Use (d) to show that IJ is *not* equal to the "naive" definition of the product of ideals, $\{ab : a \in I, b \in J\}$.

Definition 7.30. *Let G be a group and N a group contained in G. Then N is a normal subgroup of G if for every $n \in N$ and every $x \in G$, also $x^{-1}nx \in N$. In other words, if $n \in N$, then every conjugate $x^{-1}nx$ of n is also in N.*

Exercise 7.6. a) Show that for a not necessarily Abelian additive group G with a subgroup I, we have

$$(a+I)+(b+I) = (a+b-b+I)+(b+I) = (a+b)+(-b+I+b)+I\,.$$

b) Show that (a) implies that addition of cosets is well-defined if I is normal.
c) Let I be a normal subgroup of a group R, then R/I is a group. Where in the proof do you need normality. (*Hint: check the items in Definition 5.20 (1).*)
d) Let $h : R \to H$ be a homomorphism of groups. Show that $\ker h$ is a normal subgroup. (*Hint: write $h(x^{-1}nx) = h(x^{-1})h(n)h(x)$. What is $h(n)$?*)

We have proved the fundamental homomorphism for groups. This is really a slightly weaker version of Theorem 7.13. So we will record it here as a corollary.

Corollary 7.31 (Fundamental Homomorphism Theorem for Groups). *If $f : R \to H$ is a surjective group homomorphism with kernel K, then H is group isomorphic to R/K and K is a normal subgroup.*

Exercise 7.7. Show that there is no non-trivial (ring) homomorphism $\mathbb{C} \to \mathbb{R}$. (*Hint: use Corollary 7.14 to show that the kernel of f is $\{0\}$. Use $i^2 = -1$ to see $f(i)$ is undefined.*)

Exercise 7.8. Let ρ be an algebraic number with minimal polynomial p.
a) Show that the set of polynomials q in $\mathbb{Q}[x]$ such that $q(\rho) = 0$ form an ideal. (*Hint: use only Definition 7.10.*)
b) Show that this is a principal ideal. (*Hint: Lemma 7.6.*)

Exercise 7.9. a) Solve the polynomial $\gamma^4 - 10\gamma^2 + 1 = 0$ using the standard quadratic formula and then taking a square root again. Show that

$$\gamma = \pm\sqrt{5 \pm 2\sqrt{6}}\,.$$

b) Show that the root with the two '+' signs equals $\sqrt{2} + \sqrt{3}$.

Exercise 7.10. a) Show that $-\frac{1}{2} + \frac{i}{2}\sqrt{3}$ is an algebraic integer. (*Hint: compute* $(x + \frac{1}{2} - \frac{i}{2}\sqrt{3})(x + \frac{1}{2} + \frac{i}{2}\sqrt{3})$.)
b) Use a computation similar to (a) to show that $-\frac{1}{2} + \frac{1}{2}\sqrt{3}$ satisfies $x^2 + x - \frac{1}{2} = 0$.
c) Show that (b) implies that $-\frac{1}{2} + \frac{1}{2}\sqrt{3}$ is *not* a algebraic integer. (*Hint: what if that number also satisfied* $x^2 + bx + c = 0$ *with* b *and* c *in* $\mathbb{Z}$*?*)

Exercise 7.11. Consider primes p and q (in $\mathbb{Z}$). Use Lemma 7.23 to find minimal polynomials for $\sqrt{p}\sqrt{q}$ and $\sqrt{p} + \sqrt{q}$.

Exercise 7.12. Let ρ be algebraic integer with minimal polynomial $p(x) = x^d + \sum_{i=0}^{d-1} c_i x^i$ ($c_i \in \mathbb{Z}$).
a) Use Lemma 7.23 to show that for all a and b in $\mathbb{Z}$, $a + b\rho$ is also an algebraic integer of degree at most d. (*Hint: Let* C *be the companion matrix for the minimal polynomial for* ρ*; the lemma leads to considering the characteristic polynomial of* $aI + bC$.)
b) Show that $q(a + b\rho) = 0$ if q is the polynomial given by

$$q(x) = (x-a)^d + \sum_{i=0}^{d-1} c_i b^{d-i} (x-a)^i\,.$$

c) Show that if $b \neq 0$, then if $q(x)$ can be factored over the integers by $f(x)g(x)$, then $p(x)$ can be factored by $b^{-d} f(bx+a)g(bx+a)$.
d) Conclude that q is the minimal polynomial for $a + b\rho$ ($b \neq 0$).

Theorem 7.21 and the next two exercises imply the following. Theorem 8.6 provides more information.

Proposition 7.32. *The set* $\mathscr{A}$ *forms a integral domain but not a field and* $\mathscr{A}$ *is dense in* $\mathbb{C}$.

Exercise 7.13. a) Show that the algebraic numbers are closed under multiplicative inversion. (*Hint: let d be the degree of the polynomial p and consider the polynomial $q(x) := x^d p(x^{-1})$.*)
b) Show that if the degree d polynomial $p \in \mathbb{Z}[x]$ is irreducible, then so is $q(x) := x^d p(x^{-1})$. (*Hint: $q(x) = f(x)g(x)$ implies $p(x^{-1}) = f(x^{-1})g(x^{-1})$.*)
c) Use (b) to prove the following. An algebraic integer α is a unit (is invertible) if and only if α has minimal polynomial $p(x) = x^d + \sum_{i=0}^{d-1} a_i x^i$ with $a_0 = \pm 1$. (*Hint: in a minimal polynomial, a_0 cannot be zero.*)
d) Conclude that the algebraic integers do not form a field.

Exercise 7.14. a) For any real $\alpha > 1$, and any $n \in \mathbb{N}$, we can choose $k = \lfloor \alpha^n \rfloor$. Show that

$$k^{\frac{1}{n}} \leq \alpha < (k+1)^{\frac{1}{n}} .$$

b) Use (a) to show that

$$(k+1)^{\frac{1}{n}} - k^{\frac{1}{n}} < \alpha \left(2^{\frac{1}{n}} - 1\right) .$$

c) Show that the algebraic integers are dense in $\{x \in \mathbb{R} : x \geq 1\}$. (*Hint: $k^{1/n}$ is an algebraic integer.*)
d) Extend the conclusion in (c) to all of $\mathbb{R}$ by using exercise 7.12 (a).
e) Use (d) and Lemma 7.23 to prove that $\mathscr{A}$ is dense in $\mathbb{C}$.

Exercise 7.15. a) Use the method of Section 7.3 to find the minimal polynomial in $\mathbb{Z}[x]$ for $\sqrt{2} + \sqrt{3} + \sqrt{5}$. (*Hint:* $x^8 - 40x^6 + 352x^4 - 960x^2 + 576$.)

Exercise 7.16. Suppose $\rho \in \mathscr{A}$ is not a unit and has minimal (monic) polynomial p in $\mathbb{Z}[x]$.
a) Show that $q(x) = p(x^2)$ has root $\sqrt{\rho}$.
b) Show that any factor in $\mathbb{Z}[x]$ of q is monic.
c) Show that $\sqrt{\rho}$ is not a unit. (*Hint: if it is, then its square must be too.*)
d) Conclude that ρ is not irreducible.

Exercise 7.17. We apply the Euclidean algorithm in $\mathbb{Z}[\sqrt{-1}]$ to $17+15i$ and $7+5i$. Compare with the computations in Section 3.2 and exercise 3.22.
a) Check all computations in the following diagram.

	$+$	$-$	$+$	$-$
	$2+2i$	$-2-i$	3	0
0	$-1+i$	-4	$7+5i$	$17+15i$
	1			
		$2+i$	1	
			$-6-3i$	$-2-i$

b) Check all computations in the following diagram.

	$+$	$-$	$+$	$-$	$+$
	$1+2i$	$1-i$	$1-i$	2	0
0	$1+i$	$-1+3i$	$3+5i$	$7+5i$	$17+15i$
	1				
		$-1+i$	1		
			$-2i$	$-1+i$	
				$-2+4i$	$1-2i$

c) From the diagram in (a), compute values for x and y in $\mathbb{Z}[\sqrt{-1}]$ such that

$$-1+i=(7+5i)x+(17+15i)y\,.$$

(*Hint: follow instructions in Section 3.2.*)
d) From the diagram in (b), compute values for x and y in $\mathbb{Z}[\sqrt{-1}]$ such that

$$1+i=(7+5i)x+(17+15i)y\,.$$

e) Compute $\gcd(17+15i, 7+5i)$ (up to invertible elements).
f) Compute $\operatorname{lcm}(17+15i, 7+5i)$ (up to invertible elements). (*Hint: see Corollary 2.16.*)

Exercise 7.18. Find a greatest common divisor and a least common multiple for each of the following pairs of Gaussian integers. (*Hint: see exercise 7.17.*)
a) $7+5i$ and $3-5i$.
b) $8+38i$ and $9+59i$.
c) $-9+19i$ and $52+68i$.

Exercise 7.19. a) Show that the arithmetic functions (Definition 4.1) with the operations addition and Dirichlet convolution (Definition 4.19 form a commutative ring. (*Hint: see exercise 4.16*).
b) Show that the same does not hold for the multiplicative (Definition 4.1) arithmetic functions. (*Hint: see exercise 4.17*).
c) Show that the functions $f : \mathbb{R} \to \mathbb{R}$ together with the operations addition and multiplication form a commutative ring.
d) Is the ring in (c) a domain?
e) Show that the square integrable functions $f : [0,\infty) \to [0,\infty)$ together with the operations addition and convolution form almost a commutative ring. (*Hint: only the multiplicative identity is missing.*)
f) Look up Titchmarsh's convolution theorem and show that it implies that the ring in (e) (with the "Dirac delta function" added) is a domain.

Exercise 7.20. a) Show that all elements of $\mathbb{Q}[\sqrt{j}]$, $j \in \mathbb{Z}$, are algebraic numbers. (*Hint: see equation* (7.7).)
b) Now let j be square free and show that if $a+b\sqrt{j}$ is an integer of $\mathbb{Q}[\sqrt{j}]$, then

$$2a \in \mathbb{Z} \quad \text{and} \quad a^2 - b^2 j \in \mathbb{Z}\,.$$

c) Show (b) implies that if $a \in \mathbb{Z}$, then $b \in \mathbb{Z}$. (*Hint: set* $b = \frac{p}{q}$ *where* $\gcd(p,q) = 1$.)
d) Show that (b) implies that if $a \in \mathbb{Z} + \frac{1}{2}$, then $4b^2 j \in 4\mathbb{Z}+1$.
e) Show that in (d) we obtain that $b \in \mathbb{Z}+\frac{1}{2}$ and $j =_4 1$. (*Hint: set* $b = \frac{p}{q}$ *where* $\gcd(p,q) = 1$ *and conclude that* $q = 2$. *Then show that* $p^2 j = 2n+1$ *implies that* $j =_4 1$.)
f) Use (c) and (e) to show that if $j =_4 1$, the integers of $\mathbb{Q}[\sqrt{j}]$ are given by

$$I = \left\{a+b\sqrt{j} \,:\, a,b \in \mathbb{Z}\right\} \cup \left\{a+\frac{1}{2}+\left(b+\frac{1}{2}\right)\sqrt{j} \,:\, a,b \in \mathbb{Z}\right\}.$$

g) Use (f) to prove Lemma 7.25.

In the remainder of these exercises, we use the following notation. We let $\{m_1, \cdots, m_n\}$ denote a collection of distinct, *square free* (non-zero) integers in $\mathbb{Z}$. In the interest of brevity of notation, we write $\varepsilon_1 q_1 + \cdots + \varepsilon_n q_n$ as $(\varepsilon \cdot q)_{(n)}$ in exercise 7.21 and $\varepsilon_1 a_1 \sqrt{m_1} + \cdots + \varepsilon_n a_n \sqrt{m_n}$ as $(\varepsilon \cdot a\sqrt{m})_{(n)}$ in exercise 7.22. (If $m < 0$, take the root in the upper half plane of $\mathbb{C}$.) We also set $S_n = \{-1,+1\}^n$, $n > 0$. Our ultimate aim is proving Proposition 7.34. Along the way, though, we will find some interesting gems. Our approach is inspired by [**16**]; more general results can be found in [**64**].

Exercise 7.21. Define the following polynomial of degree 2^n in x with parameters q_1 through q_n :

$$P_n(x) := \prod_{\varepsilon \in S_n} \Big(x + (\varepsilon \cdot q)_{(n)} \Big) \,,$$

a) Show that for $n > 1$, $P_n(x)$ is equal to

$$P_n(x) = \sum_{\ell, j_1, \cdots, j_n} \kappa_{\ell, j_1, \cdots, j_n} x^\ell q_1^{j_1} \cdots q_n^{j_n}$$

where the κ's are integers.
b) Show from the definition of P_n that replacing q_i by $-q_i$ leaves P_n invariant.
c) Show that (b) implies that the powers j_i are *even*.

Exercise 7.22. a) Define the polynomial of degree 2^n

$$P_n(x) := \prod_{\varepsilon \in S_n} \Big(x + \big(\varepsilon \cdot a\sqrt{m}\big)_{(n)} \Big) \,,$$

and show that for $n \geq 1$, $P_n(x)$ satisfies

$$P_n(x) = P_{n-1}(x + a_n\sqrt{m_n}) \cdot P_{n-1}(x - a_n\sqrt{m_n}) \,.$$

b) Use exercise 7.21 (c) to show that $P_n \in \mathbb{Z}[x]$.
c) Use that $P_{n-1} \in \mathbb{Z}[x]$ to show that

$$\begin{aligned} P_{n-1}\,(x \pm a_n\sqrt{m_n}) &= \sum_{i=0}^{n-1} b_i (x \pm a_n\sqrt{m_n})^i \\ &= \pm\sqrt{m_n}\, O_{n-1}(x) + E_{n-1}(x) \,, \end{aligned}$$

where O_{n-1} and E_{n-1} are in $\mathbb{Z}[x]$.
d) Show that $E_{n-1}(x)$ is not the zero polynomial. (*Hint:* x^{2^n}.). Then show that if $O_{n-1}(x)$ is zero, we obtain a contradiction as follows.

$$P_{n-1}\,(x + a_n\sqrt{m_n}) = P_{n-1}\,(x - a_n\sqrt{m_n})$$

implies that the roots of those two polynomials are equal. But

$$\bigcup_{\varepsilon \in S_{n-1}} \Big\{ \big(\varepsilon \cdot a\sqrt{m}\big)_{(n-1)} + a_n\sqrt{m_n} \Big\} = \bigcup_{\varepsilon \in S_{n-1}} \Big\{ \big(\varepsilon \cdot a\sqrt{m}\big)_{(n-1)} - a_n\sqrt{m_n} \Big\} \,.$$

The following lemma (proved in Exercise 7.23) is interesting in its own right, but, as we will see, it also has important applications.

Lemma 7.33. *Let* $\{m_1, \cdots, m_\ell\}$ *be a collection of distinct, square free integers in* $\mathbb{Z}$*. Then for* $a_i \in \mathbb{Z}$*,* $\sum_{i=1}^{\ell} a_i\sqrt{m_i} = 0$ *if and only if* $a_i = 0$ *for all i.*

Exercise 7.23. This exercise relies heavily on exercise 7.22. Given ℓ distinct square free integers in $\mathbb{Z}$, $\{m_1, m_2, \cdots, m_\ell\}$. Assume the lemma is false. Thus let $r > 1$ be the smallest integer such that upon re-ordering the m_i, the first r terms are linearly dependent over $\mathbb{Z}$:

$$\sum_{i=1}^{r} a_i \sqrt{m_i} = 0 \quad \text{and} \quad \forall i \;\; a_i \neq 0\,. \tag{7.11}$$

Define the polynomials P_j, O_j, and E_j as in exercise 7.22.
a) Show that $r > 2$. (*Hint: this is trivial if one of the m_i is negative and one is positive; if they have the same sign, square the relation $a_1\sqrt{m_1} = -a_2\sqrt{m_2}$.*)
b) Show that (7.11) implies that $O_{r-1}(0) = E_{r-1}(0) = 0$. (*Hint: show that 0 is a root of P_r, so one of its factors $P_{r-1}(x \pm a\sqrt{m_r})$ has a root 0.*)
c) Show that (b) implies that there is an $\varepsilon \in S_{r-1}$ so that

$$\begin{aligned} \left(\varepsilon \cdot a\sqrt{m}\right)_{(r-1)} + a_r\sqrt{m_r} &= 0 \quad \text{and} \\ \left(\varepsilon \cdot a\sqrt{m}\right)_{(r-1)} - a_r\sqrt{m_r} &= 0\,. \end{aligned}$$

d) Show that this proves Lemma 7.33. (*Hint: adding the equalities in (c) contradicts minimality of r.*)

The next exercise is another one of the promised ‘gems’. We find the minimal polynomial for $\sum_{i=1}^{n} \sqrt{p_i}$ where the p_i are distinct primes in $\mathbb{N}$.

Exercise 7.24. Let $\{p_1, \cdots, p_n\}$ be a collection of positive primes in $\mathbb{Z}$, and for $\varepsilon \in S_n$, set

$$\gamma_\varepsilon := (\varepsilon \cdot \sqrt{p})_{(n)}$$

Furthermore, denote by $\{m_i\}_{i=1}^{2^n}$ the 2^n distinct[a] values of $\prod_{i=1}^n p_i^{\delta_i}$ for $\delta_i \in \{0,1\}$ including $1 = \prod p^0$.

a) Show that

$$\gamma_\varepsilon^\ell = \sum_{i=1}^{2^n} b_{\ell,i}\eta_i\sqrt{m_i}\,.$$

where $b_{\ell,i} \in \mathbb{Z}$ and $\eta_i = \pm 1$ is determined exclusively by ε and m_i. (*Hint:if $m_i = p_2p_5p_9$, say, then $\eta_i = \varepsilon_2\varepsilon_5\varepsilon_9$.*)

b) Let $P = \sum_{i=1}^k a_i x^i$ in $\mathbb{Z}[x]$. Show that

$$P(\gamma_\varepsilon) = \sum_{i=1}^{2^n} c_{\ell,i}\eta_i\sqrt{m_i}\,.$$

where $b_{\ell,i} \in \mathbb{Z}$ and η_i as in (a).

c) Use Lemma 7.33 to show that (b) implies:

$$\exists\, \varepsilon \in S_n \text{ such that } P(\gamma_\varepsilon) = 0 \implies \forall\, \varepsilon \in S_n \text{ such that } P(\gamma_\varepsilon) = 0\,.$$

d) Define the polynomial of degree 2^n

$$P_n(x) := \prod_{\varepsilon \in S_n}\left(x + (\varepsilon \cdot \sqrt{p})_{(n)}\right)\,.$$

Show that P_n defined in (b) is the minimal polynomial for any of the γ_ε. (*Hint: by (c) the minimal polynomial must have degree at least 2^n, and by exercise 7.22 (c), $P_n \in \mathbb{Z}[x]$.*)

[a] Here we use positivity of the primes, because, for instance, $(-2)\cdot(-3) = 2\cdot 3$. However, if the collection of primes contains no pair $\{-p, p\}$, this difficulty goes away.

Now we come to the main result of these exercises.

Proposition 7.34. *Let $\{p_1, \cdots, p_n\}$ denote any collection of* distinct *primes in $\mathbb{N}$.*
i) Let $\gamma_n := \gamma_{\varepsilon=\mathbf{1}} = \sum_{i=1}^n \sqrt{p_i}$. Then $\mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_n}) = \mathbb{Q}(\gamma_n)$
ii) Denote $F_n = \mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_n})$. The degree $[F_n : F_{n-1}]$ equals 2 and so $[F_n : \mathbb{Q}] = 2^n$.
iii) Items (i) and (ii) also hold for $\{\varepsilon_1 p_1, \cdots, \varepsilon_n p_n\}$ where $\varepsilon \in S_n$.

The first part of this proposition actually says that γ_n is a primitive element (see Theorem 7.19) for F_n. Part (ii) says that the fields F_n form an infinite tower of fields (see definition 7.35), each having degree 2 with respect to the previous field.

Definition 7.35. *A tower of fields is a (finite or infinite) sequence of successive field extensions $F_1 \subseteq F_2 \subseteq \cdots \subseteq F_n \subseteq \cdots$.*

Exercise 7.25. a) Show that $[F_n : \mathbb{Q}] \leq 2^n$. (*Hint: see preamble of exercise 7.24.*)
b) Show that $[F_n : \mathbb{Q}] = 2^n$. (*Hint: use Lemma 7.33.*)

The next exercise proves part (i) of Proposition 7.34. Note that by definition of γ_n, $\mathbb{Q}(\gamma_n) \subseteq F_n$. So all we need to show is the reverse inclusion.

Exercise 7.26. Define P_n as in exercise 7.24 (d) and γ_n as in item (i) of the proposition.
a) Proceed as in exercise 7.22, but with $a_i\sqrt{m_i}$ replaced by $\sqrt{p_i}$, to show that

$$P_{n-1}\,(x \pm \sqrt{p_n}) = E_{n-1}(x) \pm \sqrt{p_n}\,O_{n-1}(x)\,,$$

where O_{n-1} and E_{n-1} are non-zero polynomials in $\mathbb{Z}[x]$.
b) Use (a) and exercise 7.24 (d) to show that $O_{n-1}(\gamma_n) \neq 0$.
c) Show that $P_{n-1}(\gamma_n - \sqrt{p_n}) = 0$. (*Hint: from the definition of P_n.*)
d) Show that

$$E_{n-1}(\gamma_n) \pm \sqrt{p_n}\,O_{n-1}(\gamma_n) = 0\,,$$

and thus $\mathbb{Q}(\gamma_n)$ contains $\sqrt{p_n}$.
f) Use that fact that the order of the primes is arbitrary to show that $\mathbb{Q}(\gamma_n)$ contains $\sqrt{p_i}$ for any $i \in \{1, \cdots, n\}$.

Finally, we carefully checking the proofs and making sure that adjoining negative primes does not cause any problems. See for example the footnote to exercise 7.24. Note that in this case, we generally get a tower in $\mathbb{C}$, not in $\mathbb{R}$.

Exercise 7.27. Check the proofs in exercises 7.22 through 7.26 to make sure that adjoining negative p_i does not cause problems.

Chapter 8

Factorization in Rings

Overview. We now get back to factorization. It is instructive to go back to the discussion of the proof of unique factorization in $\mathbb{Z}$ (Section 2.3) at this point. Our familiarity with $\mathbb{Z}$ may hide underlying structures from us. To circumvent this familiarity, we study factorization in rings. Perhaps unexpectedly, at this level of generality, pretty much anything can happen, as we show in the first section below. We then add various ingredients to rings in an effort to end up with an abstract structure that guarantees unique factorization. Unless mentioned otherwise, we restrict to commutative rings.

8.1. So, How Bad Does It Get?

Recall that even in $\mathbb{Z}$, we have unique factorization up to factor -1 (see remark 2.12). So the best we can reasonably hope for in a general ring is to have unique factorization up to multiplication by units (Definition 5.25) and up to re-ordering. In this section, we dash those hopes. Let us start by revising our basic notions to this more general context.

Definition 8.1. *Given a ring R and an element r that is not zero or a unit. Then r is* <u>reducible</u> *if it is a product of two non-units (or non-invertible elements). If r is not equal to a product of two non-invertible elements it is called* <u>irreducible</u> *(or not reducible). If whenever $r \mid ab$, then $r \mid a$ or $r \mid b$ (or both), then it is called a* <u>prime</u>.

The important observation here is that the two characteristics of primes in $\mathbb{Z}$ that mentioned in remark 2.13 have been separated, because they do not coincide in general rings: irreducibles and primes become two different things[1].

Next we must realize that in general rings, we cannot necessarily order divisors according to their absolute value as we do in $\mathbb{Z}$ (see Definition 1.2). Instead, in the new definition we order divisors according to the partial order given by the division relation.

Definition 8.2. *Let R be a integral domain and α and β non-zero elements. A greatest common divisor $g = \gcd(\alpha, \beta)$ is a common divisor of both α and β such that for any common divisor γ we have $\gamma \mid g$. A least common multiple $\ell = \operatorname{lcm}(\alpha, \beta)$ is a common multiple of both α and β such that for any common multiple γ we have $\ell \mid \gamma$.*

So given a general ring, pick an arbitrary element, what different identities can it have? Well, it can be irreducible, reducible, a unit, or 0. These categories are mutually exclusive. In addition, every non-zero, non-unit element can also be prime or non-prime. But the primes and irreducibles are not necessarily the same. The next result gives a sample of the truly bizarre behaviors of factorizations in general (commutative) rings.

Proposition 8.3. *i) In a ring that is also a field, there are no primes or irreducibles.*
ii) The set of algebraic integers $\mathscr{A}$ form a proper ring (i.e. not a field) that has no irreducibles and no primes.
iii) In the ring $\mathbb{Z}_6$, the element 2 is prime, but not irreducible.
iv) In the ring $\mathbb{Z}[\sqrt{-5}]$, the element 3 is irreducible, but not prime.
v) In $\mathbb{Z}[\sqrt{-3}]$, the $\gcd$ of 4 and $2+2\sqrt{-3}$ does not exist.
vi) In $\mathbb{Z}_6[x]$, $2x(1+3h(x))^n$ divides $4x^2$ for any polynomial h and any $n \geq 0$.

Proof. (i) Recall that in a field, every non-zero element is a unit, and so there are no primes or irreducibles.

(ii) Pick any non-zero, non-unit $\rho \in \mathscr{A}$. According to exercise 7.13 (c), ρ has minimal polynomial

$$p(x) = \sum_{i=0}^{d} a_i x^i \quad \text{ with } \quad a_d = 1 \text{ and } a_0 \neq -1, 0, +1\,.$$

[1] And the meaning of "prime" has changed to confuse non-algebraists. But we're not falling for it!

Set $a = b = \sqrt{\rho}$. Then a is a root of $q(x) = p(x^2)$ by exercise 7.13 (b). Now $q \in \mathbb{Z}[x]$ is monic and any polynomial factor of q must also be monic. Therefore is in $\mathscr{A}$. Since $\rho = ab$, ρ is reducible. Clearly, we also have $\rho \mid ab$. But if ρ divides a (or b), then a/ρ is in $\mathscr{A}$. Since $\mathscr{A}$ is closed under multiplication, its square, which equals ρ^{-1} would then also be in $\mathscr{A}$. This contradicts our initial choice of ρ. Hence ρ cannot divide a or b, and so ρ is not prime.

(iii) Suppose $2 \mid ab$ in $\mathbb{Z}_6$. Then in $\mathbb{Z}$, 2 divides $ab+6m$. But that means that ab is even and thus a (or b) has a factor 2. But then in $\mathbb{Z}_6$, 2 divides a (or b). Therefore 2 is prime in $\mathbb{Z}_6$. On the other hand, $2 \cdot 4 =_6 2$. Since both 2 and 4 are non-invertible, 2 is reducible.

(iv) Suppose the number 3 equals the product xy, where x and y in $\mathbb{Z}[\sqrt{-5}]$. Clearly, x and y cannot both be real, because 3 is prime and irreducible in $\mathbb{Z}$. If both are non-real, then $b \neq 0$ and each has absolute value at least $\sqrt{5}$, and $|xy| \geq 5$, a contradiction. If one of them is non-real, then so is their product, another contradiction. Therefore, one of x or y must be a unit. This proves that 3 is irreducible in $\mathbb{Z}[\sqrt{-5}]$. But on the other hand,

$$(2+i\sqrt{5})(2-i\sqrt{5}) = 9 \quad \Longrightarrow \quad 3 \mid (2+i\sqrt{5})(2-i\sqrt{5})\,.$$

But since $\frac{(2\pm i\sqrt{5})}{3} \notin \mathbb{Z}[\sqrt{-5}]$, 3 does not divide either of these factors.

(v) Since

$$4 = 2 \cdot 2 = (1+\sqrt{-3})(1-\sqrt{-3})\,,$$

both 2 and $(1+\sqrt{-3})$ are divisors of 4. They are also divisors of $(2+2\sqrt{-3})$. however, it is a simple check to see that 2 and $(1+\sqrt{-3})$ do not divide each other. In other words, there is no *mximal* common divisor in this case.

(vi) Using the binomial theorem, we see that modulo 6

$$2x \cdot 2x \cdot (1+3h(x))^n =_6 4x^2 \sum_{i=0}^{n} \binom{n}{i} 3^i h(x)^i =_6 4x^2\,,$$

because $4 \cdot 3^i =_6 0$ for $i > 0$. ∎

8.2. Integral Domains

In order to "tame" factorizations, the first thing to do is to require the absence of zero divisors.

Definition 8.4. *An* *integral domain* *or* *domain* *is a commutative ring* R *with no zero divisors (i.e. if* $a \neq 0$ *and* $b \neq 0$*, then* $ab \neq 0$*).*

Thus, in an integral domain, if we have $ab = 0$, then we can conclude that either $a = 0$ or $b = 0$ or both. This applies to the situation where we have $a(x-y) = 0$. If $a \neq 0$, we must have $x = y$. This immediately implies (see Theorem 2.7) the following.

Theorem 8.5 (**Cancellation Theorem**)**.** *In an integral domain, if* $a \neq 0$*, then* $ax = ay$ *if and only if* $x = y$*. (See also Theorem 2.7.)*

Polynomials whose coefficients form an integral domain are themselves an integral domain (see Section 3.7 and Definition 7.1). Other examples are the fields $\mathbb{F}_p$ of the integers modulo a prime p. In this context, Lagrange's theorem (Theorem8.32) is interesting: it says that an degree n polynomial over a *field* has at most n roots. So,

$$x^2 + 5x + 6 =_{11} 0 \quad \Longrightarrow \quad (x+2)(x+3) =_{11} 0 \,.$$

And this implies that $x =_{11} -2 =_{11} 9$ or $x =_{11} -3 =_{11} 8$. If we work modulo 12, factoring does not solve the problem. For example, $x^2 + 5x + 6$ modulo 12 has roots $\{1, 6, 9, 10\}$.

A (non-Abelian) ring that *does* have zero divisors are the 2 by 2 matrices with coefficients in $\mathbb{Z}$. In fact, if $N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, then $N^2 = 0$.

Note that $\mathbb{C}$ does not have zero divisors. Therefore, the same holds for any subset of $\mathbb{C}$, such as the set $\mathscr{A}$ of all algebraic integers. Propositions 7.32 and 8.3 (ii) imply the following remarkable facts.

Theorem 8.6. *The set* $\mathscr{A}$ *forms a integral domain but not a field and*
i) $\mathscr{A}$ *is dense in* $\mathbb{C}$*, ii)* $\mathscr{A}$ *has no irreducibles, iii)* $\mathscr{A}$ *has no primes.*

It might seem that we have not done much to tame the factorization process. However, the following result indicates that we on the right track.

Theorem 8.7. *Any prime* p *in an integral domain* R *is irreducible.*

Proof. Suppose that the prime p satisfies $p = ab$. We need to show that a or b is a unit. Certainly $p \neq 0$ divides ab, and so, from Definition 8.1, $p \mid a$ or $p \mid b$. Assume the former. So there is a c such that $pc = abc = a$. We then get The cancellation theorem gives, of course, that $bc = 1$, and so b has an inverse and therefore is a unit. Similar if we assume $p \mid b$. ■

Like any ring in $\mathbb{C}$, $\mathbb{Z}[\sqrt{-5}]$ is an integral domain. So Proposition 8.3 (iv) shows that the converse is false. However, here is an interesting lemma that implies (once again) that $\mathbb{F}_p$ is a field (see Proposition 5.18). The proof is essentially the same as that of Lemma 5.3.

Lemma 8.8. *A finite integral domain is a field.*

Proof. Fix some $a \neq 0$ in the integral domain R. Consider the (finitely many) elements $\{ax\}_{x \in R}$. Either all these elements are all distinct, or two are the same. But if $ax = ay$, the cancellation theorem gives a contradiction. If they are all distinct, then there is an x such that $ax = 1 \in R$. Thus a has a multiplicative inverse. ■

Theorem 8.9. *Let R be an integral domain in which every element has a factorization into irreducibles. Every irreducible is a prime if and only if factorization into irreducibles is always unique.*

Proof. First, suppose that every irreducible is a prime and assume that the following are two factorizations of $x \in \mathbb{R}$ into irreducibles.

$$x = up_1 \cdots p_k = u' q_1 \cdots q_\ell \,.$$

Now if p_1 is a prime, upon relabeling the q_i, it must divide q_1. Since q_1 is irreducible, we must have $p_1 = q_1$ up to units. Doing finitely many steps, one proves that the factorization is unique.

Next, suppose that q is irreducible and that there are non-zero a and b such that $q \mid ab$. This implies $qc = ab$. We factor both sides of this last equation into irreducibles.

$$uq(p_1 \cdots p_k) = u'(q_1 \cdots q_\ell)(q_{\ell+1} \cdots q_m) \,.$$

By unique factorization, q must equal to q_1 (upon relabeling and up to units) and thus it divides a or b. ■

Definition 8.10. *An integral domain R is a unique factorization domain[2] if every element admits a unique factorization into irreducibles. This is often abbreviated to UFD.*

[2] The word “domain” serves as a reminder that R must be an integral domain.

By Theorems 8.7 and 8.9, in a UFD, "prime" and "irreducible" are synonymous. In a UFD, the notions of greatest common divisor and least common multiple are well-defined. The reason these notions are well-defined can be found in the proof of Corollary 2.16. To repeat that argument, suppose that

$$\alpha = u\prod_{i=1}^{s} p_i^{k_i} \quad \text{and} \quad \beta = u'\prod_{i=1}^{s} p_i^{\ell_i} ,$$

where u and u' are units and k_i and ℓ_i in $\mathbb{N} \cup \{0\}$. Now define:

$$m_i = \min(k_i, \ell_i) \quad \text{and} \quad M_i = \max(k_i, \ell_i) .$$

Then, of course, we have

$$\gcd(\alpha, \beta) = \prod_{i=1}^{s} p_i^{m_i} \quad \text{and} \quad \operatorname{lcm}(\alpha, \beta) = \prod_{i=1}^{s} p_i^{M_i} .$$

The p_i are unique up to a unit. And so are the gcd and lcm, since the product of units is a unit.

We still need to be slightly cautious. For instance, in $\mathbb{Z}[i]$, which is a UFD, the units are ± 1 and $\pm i$. The gcd of $2i$ and -4 is 2 up to units, that is: ± 2 or $\pm 2i$.

8.3. Euclidean Domains

The next step in the taming process, is to make sure there is a division algorithm.

Definition 8.11. *A Euclidean function on a ring R is a function $E: R\backslash\{0\} \to \mathbb{N} \cup \{0\}$ that satisfies:*
i) For all ρ_1 and ρ_2 in R, there are κ and ρ_3 in R such that $\rho_1 = \kappa\rho_2 + \rho_3$ and $E(\rho_3) < E(\rho_2)$ and
ii) For all α and γ in $R\backslash\{0\}$, we have $E(\alpha\gamma) \geq E(\alpha)$.
A Euclidean ring or Euclidean domain is an integral domain R for which there is a Euclidean function.

In a Euclidean domain, we can perform the division algorithm of Lemma 2.2[3]. All statements and proofs in Chapter 2 from Bézout's Lemma (Lemma 2.5) on, up to and including Corollary 2.16, hold with minor modifications. For example, we need to use $E(n)$ instead of the norm of n. Theorem 2.7

[3]The name "Euclidean domain" derives from the alternative name of that algorithm, see remark 2.4.

needs the reformulation given by Theorem 8.5. Corollary 2.8 would need to be reformulated (which we omit). Among other things, the unique factorization, and the Euclidean algorithm of Chapter 3, which in turn led us to continued fractions, follow from these. So the consequences of having a Euclidean function are indeed *staggering*! Exercise 8.12 investigates the relation between the two chapters.

In Euclidean domains the notions of prime and irreducible are again happily reunited.

Proposition 8.12. *Let R be a Euclidean domain. If $p \in R$ is irreducible, then p is prime.*

Proof. Suppose p is irreducible and $p \mid ab$ and let g be a $gcd(a,p)$. Then there are h and k such that $p = gh$ and $a = gk$. Since p is irreducible, either g or h is a unit. Suppose first that h is a unit. Then $a = ghh^{-1}k = ph^{-1}k$ and so $p \mid a$. If, on the other hand, g is a unit, then g divides 1 (the multiplicative identity). Of course, 1 is a common divisor of a and p, and thus we also have $\gcd(a,p) = 1$. Euclid's lemma (Lemma 2.6) gives that $p \mid b$. ■

Corollary 8.13. *Let R be a Euclidean domain. Then*
i) $p \in R$ is prime if and only if p is irreducible.
ii) Every element admits a unique factorization into powers of primes up to re-ordering and products of units.

Proof. Item (i) follows from the previous proposition together with Theorem 8.7. Theorem 8.9 implies item(ii). ■

Polynomial rings over a *field*, such as $\mathbb{Q}[x]$ or $\mathbb{R}[x]$, are a great examples of Euclidean domains. We already saw in Section 7.1, that the degree is a Euclidean function in these rings.

We finally come to the reason to introduce *empty products* in Remark 2.14.

Corollary 8.14. *A field F is a Euclidean domain and therefore has unique factorization. Namely, every non-zero $x \in F$ is a unit times the <u>empty product</u> of primes. In particular, there are no primes and no irreducible numbers in F.*

Proof. We take x and y in F and write $x = yq + 0$, where $q = y^{-1}x$. So every remainder maps to zero[4]. ■

Thus the results in Chapter 2 starting with Theorem 2.17 (the infinitude of primes) do *not* generalize to all Euclidean domains. The problem in the proof of Theorem 2.17 is that it crucially depends on adding "1" to some number in order to get a "bigger" number. The rest of that Chapter depends on the embedding of the integers in the real numbers (or even $\mathbb{C}$).

The last result, together with Definitions 8.11, 8.4, and 5.20, immediately implies the following.

Corollary 8.15. *We have the following inclusions:*
fields $\subsetneq$ Eucl. domains $\subsetneq$ UFDs $\subsetneq$ domains $\subsetneq$ comm. rings $\subsetneq$ rings.

8.4. Example and Counter-Example

As an example we consider the elements of the set $\mathbb{Z}[\sqrt{-1}]$. These are usually called the Gaussian integers (see Figure 40). From equations 7.8 and 7.9, we can infer that $\alpha = a + bi$ can be represented in matrix form as:

$$\alpha = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \qquad \text{with} \qquad N(\alpha) = a^2 + b^2 \ .$$

It is easy to check that multiplication of these matrices is commutative — after all, multiplication of the underlying complex numbers is commutative.

Proposition 8.16. *The Gaussian integers form a Euclidean domain with the norm as Euclidean function.*

Proof. For j a square free integer, $N(\alpha)$ is the square of the absolute value of α, and so it is a positive integer. So the second requirement of Definition 8.11 follows immediately from Corollary 7.28. It remains to prove that the first requirement is satisfied.

Given any ρ_1 and ρ_2 in $\mathbb{Z}[i]$, we can certainly choose κ and ρ_3 so that

$$\rho_1 = \kappa\rho_2 + \rho_3 \ .$$

[4] This is one of reasons we added 0 to the image of E in Definition 8.11

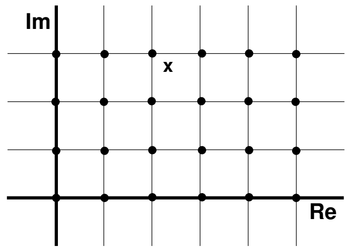


Figure 40. The Gaussian integers are the lattice points in the complex plane; both real and imaginary parts are integers. For an arbitrary point $z \in \mathbb{C}$ — marked by x in the figure, a nearby integer is $k_1 + ik_2$ where k_1 is the closest integer to $\mathrm{Re}(z)$ and k_2 the closest integer to $\mathrm{Im}(z)$. In this case that is $2+3i$.

(For example, $\kappa = 0$ and $\rho_3 = \rho_1$.) Dividing by ρ_2 gives

$$\rho_1\rho_2^{-1} = \kappa + \rho_3\rho_2^{-1}\,. \tag{8.1}$$

We choose κ to be the closest[5] Gaussian integer to $\rho_1\rho_2^{-1}$ (indicated by "x" in Figure 40). Recalling that in this case, the norm corresponds to the usual absolute value squared, we immediately see from the figure that we can choose κ so that $N(\rho_3\rho_2^{-1}) \leq 1/2$. And thus with that choice, using Corollary 7.28,

$$N(\rho_3) = N(\rho_3\rho_2^{-1})N(\rho_2) \leq \frac{1}{2}N(\rho_2) \tag{8.2}$$

which proves the first requirement. ■

The computation that leads from equation (8.1) to equation (8.2) can also be done explicitly. Let $\rho_1 = a+bi$ and $\rho_2 = c+di$. It is an easy computation to see that

$$\rho_1\rho_2^{-1} = \frac{ac+bd}{c^2+d^2} + i\,\frac{-ad+bc}{c^2+d^2}\,.$$

We want to express this as a Gaussian integer $\kappa = k_1 + ik_2$ plus a remainder $\rho_3\rho_2^{-1} = \varepsilon_1 + i\varepsilon_2$ whose norm is less than 1. We choose k_1 to be the integer closest (or one of the integers closest) to $\frac{ac+bd}{c^2+d^2}$, and k_2, the integer closest to $\frac{-ad+bc}{c^2+d^2}$. With those choices, the remainders

$$\varepsilon_1 = \frac{ac+bd}{c^2+d^2} - k_1 \quad \text{and} \quad \varepsilon_2 = \frac{-ad+bc}{c^2+d^2} - k_2$$

[5]If there is more than one closest Gaussian integer, pick any one of them.

are each not greater than $\frac{1}{2}$ in absolute value. Thus

$$\rho_3 = (\varepsilon_1 + i\varepsilon_2)(c + id) ,$$

with norm $(\varepsilon_1^2 + \varepsilon_2^2)(c^2 + d^2)$ by Corollary 7.28. Since the ε_i are no greater than $\frac{1}{2}$, (8.2) follows.

The computation in the foregoing proof will be important, and so it is useful to summarize it even more succinctly.

Definition 8.17. *A fundamental domain of $\mathbb{Z}[i]$ is a simply connected region in $\mathbb{C}$ such that it contains exactly one representative of every set $z + \mathbb{Z}[i]$. Usually one takes the unit square as a fundamental domain for $\mathbb{Z}[i]$.*

Remark 8.18. For j negative and square free, N is a Euclidean function on $\mathbb{Z}[\sqrt{j}]$ if and only if in a fundamental domain, the distance to the nearest algebraic integer is strictly less than 1.

Note that in $\mathbb{Z}$, to get a small remainder we simply choose the *floor* of $\rho_1\rho_2^{-1}$ for the equivalent of κ (see the proof of Lemma 2.2). But in the above proof — working the Gaussian integers — it is clear that in general there is no obvious natural choice for $\kappa = k_1 + ik_2$ that makes $N(\varepsilon)$ less than 1. In exercise 8.3, we look in some more detail at the possible choices for k_1 and k_2. So the Euclidean algorithm applied to, say, $17+15i$ and $7+5i$ may lead to different computations. We gave an example of this in exercise 7.17.

Proposition 8.19. *The ring $\mathbb{Z}[\sqrt{-6}]$ does not have the unique factorization (into irreducibles) property. Therefore this ring is not a Euclidean domain.*

Proof. $\mathbb{Z}[\sqrt{-6}]$ (see Figure 41) is an integral domain, because it is a subring of $\mathbb{C}$. We show that $\mathbb{Z}[\sqrt{-6}]$ does not have unique factorization in two steps. The first step is to observe that

$$10 = 2 \cdot 5 = (2 + i\sqrt{6})(2 - i\sqrt{6}) .$$

We are done if we show that 2, 5, and $2 \pm i\sqrt{6}$ are irreducible. Assume $2 = \alpha\gamma$, both non-units. Taking the norm[6] (always using Corollary 7.28), we get

$$4 = N(\alpha)N(\gamma) .$$

[6]This part of this proof illustrates how to use the norm to reduce the question whether a number in a Euclidean domain R is irreducible to the same question in $\mathbb{Z}$.

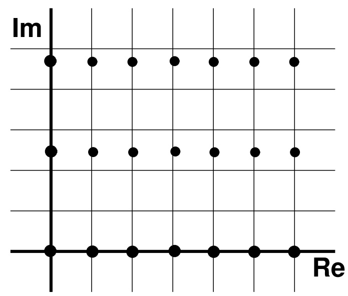


Figure 41. A depiction of $\mathbb{Z}[\sqrt{-6}]$ in the complex plane; real parts are integers and imaginary parts are multiples of $\sqrt{6}$.

Thus each of the norms equals 2. But $2 = a^2 + 6b^2$ has no integer solutions, hence 2 is irreducible. The exact same argument applied to 5 gives that

$$25 = N(\alpha)N(\gamma) .$$

Each of the norms now must equal 5. But again $5 = a^2 + 6b^2$ has no integer solutions. If we apply the argument to $2 \pm i\sqrt{6}$, we obtain

$$10 = N(\alpha)N(\gamma) .$$

Thus either α must have norm 2 and β must have norm 5, or vice versa. But the previous arguments show that both are impossible. ■

8.5. Ideal Numbers

In this section, we explain how ideals arise from the study of factorization into primes in rings of algebraic integers. We base this description loosely on the historical record as described in chapter 21 of the excellent book [**69**]. For the definition and basic properties of ideals, we refer to Section 7.2. We start by reformulating gcd and lcm in the language of ideals.

Definition 8.20. *Let A and B ideals. The <u>greatest common divisor</u> of A and B is the smallest ideal that* contains *both of these. It is denoted by* $\gcd(A,B)$. *The <u>least common multiple</u> of A and B is the largest ideal that* is contained *in both A and B. It is denoted by* $\operatorname{lcm}(A,B)$.

Recall that an ideal $\langle j\rangle$ in $\mathbb{Z}$ is maximal if and only if j is prime. For if j is not prime, the ideal generated by a divisor of j contains $\langle j\rangle$. On the other hand, consider the ideal $\langle p, j\rangle$. The fact that it is generated by $\gcd(p, j)$ is non-trivial: it follows from Bézout (Lemma 2.5).

Now let us see how this pans out in some examples of ideals in rings of algebraic integers. Start by considering the ring $\mathbb{Z}[\sqrt{-3}]$ of algebraic integers (see equation (7.7)) displayed in the left of Figure 42. We start by showing that this ring does not have the unique factorization property. Knowing that

$$4 = 2\cdot 2 = (1+i\sqrt{3})(1-i\sqrt{3})\,, \tag{8.3}$$

the proof of that statement is almost verbatim that of Proposition 8.19 (see exercise 8.21. This exercise goes on to show that 4 admits no factorization *at all* into primes!).

What is interesting here is that the numbers 2 and $(1\pm i\sqrt{3})$ belong to the same maximal ideal.

Lemma 8.21. *$I = \langle 2, 1+i\sqrt{3}\rangle$ is a maximal ideal in $R = \mathbb{Z}[\sqrt{-3}]$.*

Proof. I is depicted in red in the left of Figure 42. It clearly contains both 2 and $1+i\sqrt{3}$. It clearly forms a lattice and so is closed under addition. Next we check the absorption property of the ideal. Denote the two generators by x and y for brevity. For any elements α, β, and γ of R, we must have

$$\alpha(\beta x + \gamma y) = \delta x + \varepsilon y \in I\,.$$

It is an easy but tedious exercise to check that for any integers a, b, c, and d

$$(a+ib\sqrt{3})\cdot 2 + (c+id\sqrt{3})\cdot(1+i\sqrt{3}) = (a-b-2d)\cdot 2 + (c+d+2b)\cdot(1+i\sqrt{3})\,.$$

And so all these elements lie in the lattice I.

If we add I any element *not* in I, then the resulting set contains the differences 1 and $i\sqrt{3}$ (see Figure 42). Taking the closure under addition, it immediately follows that we obtain all of $\mathbb{Z}[\sqrt{-3}]$. Thus I is maximal. ∎

The upshot is that we are tempted (or, rather, Kummer was [**69**]) to think of the set I as the set of multiples of some hidden or "ideal"[7] prime Q. Then both 2 and $(1\pm i\sqrt{3})$ are multiples of this "ideal" number Q (up to units at least). This way, lo and behold, unique factorization into irreducibles or primes is restored!

[7] Hence the name "ideal".

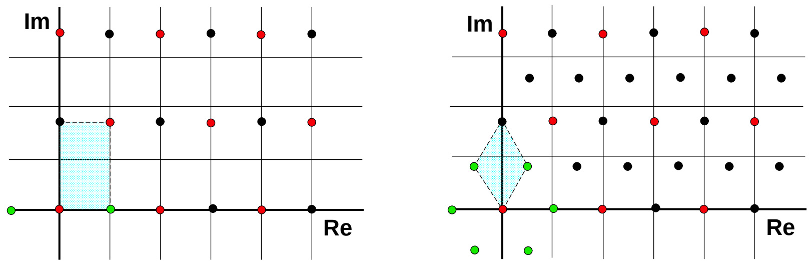


Figure 42. Left, the elements of the ring $\mathbb{Z}[\sqrt{-3}]$. Right, the ring $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-3})]$. The units of each ring are indicated in green and the ideals $\langle 2, 1+\sqrt{-3}\rangle$ on the left and $\langle 2\rangle$ on the left are indicated in red. Fundamental domains (Definition 8.17) are shaded in blue.

There is more than a grain of truth in this. Recall that the ring $R' = \mathbb{Z}[\frac{1}{2}(1+i\sqrt{3})]$ is the ring of integers in $\mathbb{Q}(\sqrt{-3})$ (Lemma 7.25). This ring, depicted on the right of Figure 42, contains the *units* (drawn in green) $\frac{1\pm i\sqrt{3}}{2}$. Clearly, 2 and $1+i\sqrt{3}$ are now the same up to a unit. Therefore, this time around 2 generates I. In other words, R' contains R, and has the same set I as an ideal, only now it is a *principal* ideal. Indeed, in R', equation (8.3) does not represent *distinct* factorizations of 4, precisely because in this ring, 2 and $1+i\sqrt{3}$ differ by a unit.

Finally, we finish this section by checking that indeed the norm is not a Euclidean function for $\mathbb{Z}[\sqrt{-3}]$, while it is for $\mathbb{Z}[\frac{1}{2}(1+i\sqrt{3})]$. Thus this ring is a Euclidean domain and so, by Corollary 8.13, primes and irreducibles are the same, and factorization is unique. This ring is an important example and has its own name; its elements are called the Eisenstein integers .

Proposition 8.22. *i) The norm in $\mathbb{Z}[\sqrt{-3}]$ is not a Euclidean function.*
ii) The norm in $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-3})]$ is a Euclidean function.

Proof. According to Remark 8.18, the norm — which in these two cases is positive — is a Euclidean function if and only if it is less than 1 in a fundamental domain. In both cases, the norm of a number is simply the square of the usual absolute value of that number. The fundamental domains are shaded in Figure 42.

Proof of (i). The fundamental domain D is given by a rectangle of height $|h| = \sqrt{3}$ and width 1 (see Figure 43). The diagonals in D have length $\sqrt{1+3} = 2$ and so we have that the distance to the nearest algebraic integer is between 0 and 1. It equals 1 at the intersection of the diagonals. Thus N fails to be a Euclidean function.

Proof of (ii). The fundamental domain consists of two isosceles triangles, one of which is depicted on the right of Figure 43. Its height d is $\frac{1}{2}\sqrt{3}$ and its base has length 1. We are looking for the point that maximizes the distance to the nearest point of the triangle. That point lies at height y on the bisector of the top-angle and its its distance $d-y$ to the three points of the triangle is the same. Thus we compute

$$\frac{1}{2^2} + y^2 = (d-y)^2 \quad \Longrightarrow \quad y = \frac{4d^2-1}{8d} \quad \Longrightarrow \quad d-y = \frac{4d^2+1}{8d}\,.$$

This evaluates to $d - y = \frac{\sqrt{3}}{3}$ which is less than 1. ■

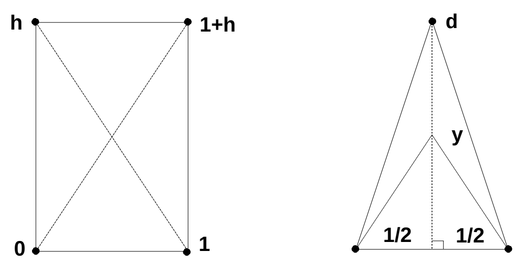


Figure 43. Left, the fundamental domain of $\mathbb{Z}[\sqrt{-3}]$. Here, $h = i\sqrt{3}$. Right, one of the 2 isosceles triangles that constitute the fundamental domain of $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-3})]$. Its height d equals $\frac{1}{2}\sqrt{3}$. The point that maximizes the distance to the closest of the 3 corner points lies on the bisector of the top angle at height y.

It is surprising that in the first part of the proof, the criterion of Euclidean fails *at only 1 point* in the fundamental domain. An analyst might suspect that somehow we can get around the exception because it has measure zero. Note, however, that (8.3) shows that $\mathbb{Z}[\sqrt{-3}]$ does not have have unique factorization and thus there is no Euclidean function (Proposition 8.13).

8.6. Principal Ideal Domains

Definition 8.23. *A principal ideal domain is an integral domain in which every ideal is a principal ideal. This is usually abbreviated to PID.*

We now complete the containments given in Corollary 8.15.

Theorem 8.24. *We have the following inclusions:*
fields $\subsetneq$ ED's $\subsetneq$ PID's $\subsetneq$ UFD's $\subsetneq$ domains $\subsetneq$ comm. rings $\subsetneq$ rings.

Proof. In view of Corollary 8.15, we only need to prove (i) that a Euclidean domain is a PID, (ii) that a PID is a UFD, and (iii) that the three categories are not equal. We leave (iii) for the next section.

i) In a Euclidean domain, the trivial ideal $\{0\}$ is of course a principal ideal (as it has only one element). Let E be the Euclidean function in D. Fix a non-trivial ideal I and pick $x \in I$ that minimizes E on $I\backslash\{0\}$. Pick any other $y \in I$. Then by the division algorithm

$$y = xq + r \quad \text{and} \quad E(r) < E(x)\,.$$

But since $y - xq \in I$, r is in I, and so $E(r)$ must be zero by the minimality of x. Hence x generates y.

ii) Suppose x_0 is an element of a principal ideal domain D that cannot be written as a a product of irreducibles. Then, clearly, there are non-zero non-units x_1 and y_1 so that $x_0 = x_1 y_1$. But by definition of x_0, at least one of x_1 and y_1 cannot be written as a product of irreducibles. Suppose that is x_1. Now x_1 divides x_0, and we get $\langle x_0\rangle \subsetneq \langle x_1\rangle$. We can apply the same arguments to x_1, and so on. Thus we get what is called an (infinite) ascending chain of ideals:

$$\langle x_0\rangle \subsetneq \langle x_1\rangle \cdots \subsetneq \langle x_n\rangle \cdots\,.$$

We define $I = \cup_{i=0}^{\infty} \langle x_i\rangle$. It is easy to see that I is an ideal (Definition 7.10). But because D is a PID, I must have a single generator p. The element p must reside in $\langle x_n\rangle$ for some n. Since p generates $\langle x_n\rangle$ it must in fact be equal to x_n. Thus the ascending chain must end, contradicting the hypothesis on x_0, which implies that every element in D can be written as a product of irreducibles.

It is then sufficient by Theorem 8.9 to show that every irreducible p is also prime. Let element a not in $\langle p\rangle$ and consider the ideal $\langle p,a\rangle$. Because

D is a PID, there is a q that generates this ideal: $\langle q\rangle = \langle p,a\rangle$. But then we must have

$$\langle q\rangle = D\,,$$

because if not, p has a non-trivial divisor q. In particular, we get that there are x and y so that

$$1 = px + ay \quad \Longrightarrow \quad \forall b\in D:\; b = pxb + ayb$$

But this implies that if $p \mid ab$ and $p\nmid a$, then we must have $p \mid b$. Thus p is prime. ■

Common PID's are $\mathbb{Z}$ and $F[x]$, but these are also Euclidean domains.

8.7. ED, PID, and UFD are Different

PID's that are *not* Euclidean domains are a not so easy to come by. Here we show, following **[74]**, that $\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$ is an example of this. Recall that by Lemma 7.25, $\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$ is the set of integers of $\mathbb{Q}[\sqrt{-19}]$.

Lemma 8.25. *In* $\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$*, the units are* ±1*, while 2 and 3 are irreducible.*

Proof. For brevity, we set $\theta = \frac{1+\sqrt{-19}}{2}$ and denote $R=\mathbb{Z}[\theta]$. The norm of $a+b\theta$ satisfies (see, for example, remark 7.29)

$$N(a+b\theta) = \left(a+\frac{b}{2}\right)^2 + \frac{19b^2}{4} = a^2+ab+5b^2 \in \mathbb{N}\cup\{0\}\,.$$

We have that the norm of units must be ±1, so

$$\left(a+\frac{b}{2}\right)^2 + \frac{19b^2}{4} = 1\,.$$

Clearly, the only solutions are $a=\pm1$ and $b=0$.

By the multiplicative property of the norm, if 2 is reducible we have

$$2 = xy \quad \Longrightarrow \quad N(2) = 4 = N(x)N(y)\,.$$

$N(x)$ and $N(y)$ are natural numbers and not equal to 0 or ±1. The only solution is $N(x)=N(y)=2$ which is easily seen to be impossible. Hence 2 is irreducible. The same reasoning works for 3. ■

Proposition 8.26. $\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$ *is not a Euclidean domain.*

Proof. We start by assuming that R is a Euclidean domain (ED) and derive a contradiction. So let E denote the Euclidean function[8] of Definition 8.11 and let m be an element of R that minimizes E over the set of non-zero, non-unit elements. In a ED, we are allowed to use the division algorithm, so

$$2 = mq + r \qquad \text{with} \qquad E(r) < E(m)\,.$$

From the inequality and the assumption on m, we see that r must be zero or a unit. So by Lemma 8.25, $r \in \{0, \pm 1\}$. Now if $r = 1$, then $mq = 2 - r = 1$ and so m is invertible, contradicting the assumption on m.

If r equals 0 or -1, we need to do one more step. In this case, mq equals 2 or 3. By Lemma 8.25 these numbers are irreducible, and thus q must be a unit (since m is not), whence $m \in \{\pm 2, \pm 3\}$. We apply the division algorithm to θ:

$$\theta = mq' + r' \qquad \text{with} \qquad E(r') < E(m)\,.$$

So $\theta - r'$ is divisible by m where $r' \in \{0, \pm 1\}$, that is to say: by 2 or 3. But it is easy to see that any of these numbers divided by 2 or 3 are not in R. ∎

Theorem 8.27. *The set R of the integers of $\mathbb{Q}[\sqrt{-19}]$ is a PID but not a Euclidean domain.*

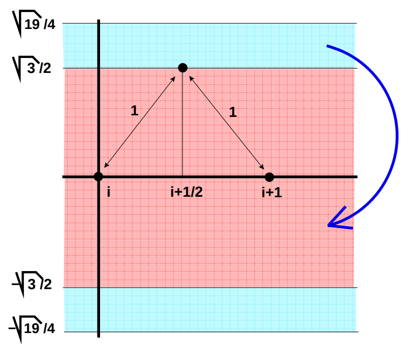


Figure 44. Points in the area red shaded are a distance less than from an integer in $\mathbb{Z}$. The blue area maps into the red under $x \to 2x - \sqrt{19}/4$ indicated by the arrow. We note that $\sqrt{19}/4 \approx 1.09$ and $\sqrt{3}/2 \approx 0.87$.

[8] Some unknown function, not necessarily the norm.

Proof. Of course, the second part is settled by the previous result. So here we just prove that R is a PID. So consider any non-zero ideal I in R and pick a b in I which minimizes the norm $N(b)$ on the non-zero elements of I. Now *assume* I is *not* principal, then certainly bR will not equal I. In this case, we choose any $a \in I \backslash bR$ and investigate what happens. By the absorption property, we have that

$$\forall p, q \in R \;:\; ap - bq \in I\,.$$

We will show, however, that

$$\exists p, q \in R \;:\; ap - bq \neq 0 \quad \text{and} \quad N(ap - bq) < N(b)\,, \tag{8.4}$$

which contradicts our choice of b, and therefore disproves the assumption that I is not principal. By the multiplicativity of norms, (8.4) will be proved if $N(ap/b - q) < 1$. By remark 7.29 then, (8.4) is equivalent to

$$\exists p, q \in R \;:\; ap - bq \neq 0 \quad \text{and} \quad \left|\frac{a}{b}\,p - q\right| < 1\,. \tag{8.5}$$

Clearly, we can choose q so that the *real part* of $ap/b - q$ is not zero. Then add a multiple of $i\sqrt{19}/2$ to q so that the *imaginary part* of $ap/b - q$ is in $(-\sqrt{19}/4, \sqrt{19}/4]$. Note that $ap/b - q \neq 0$. If in fact the imaginary part is in $(-\sqrt{3}/2, \sqrt{3}/2)$ (shaded red in Figure 44), then by subtracting an integer (in $\mathbb{Z}$) from q we are done. If, however, the imaginary part of $ap/b - q$ lands in $[\sqrt{3}/2, \sqrt{19}/4]$, then we multiply both p and q by 2 and subtract $i\sqrt{19}/2$. One can check (see exercise 8.24) that the complex map $g : z \to 2z - i\sqrt{19}/4$ maps the top blue shaded area in Figure 44 into the area shaded in red. The argument for the lower blue area is identical. ■

Theorem 8.28. *The set* $\mathbb{Z}[x]$ *is a UFD but not a PID.*

Proof. We start by showing that $I = \langle 2, x\rangle$ is not a principal ideal in $\mathbb{Z}[x]$. Let $p \in I$. Then $p(x) = 2f(x) + xg(x)$ and so $p(0) = 2n$ for some $n \in \mathbb{Z}$. If p generates the ideal I, then we must also have

$$2 = p(x)a(x) \qquad \text{and} \qquad x = p(x)b(x)\,.$$

The first equality implies that p has degree 0 and $p(x) = 2n$, while the second then yields that $x = 2nb(x)$ which is impossible. So I is not principal.

Given $f \in \mathbb{Z}[x]$. It is not surprising that any factorization of f in $\mathbb{Z}[x]$ is also a factorization in $\mathbb{Q}[x]$. However, the reverse is also true by Gauss' lemma (Lemma 7.8). Now f as an element of $\mathbb{Q}[x]$ has a unique factorization by Corollary 8.13 and the fact that the degree is a Euclidean function in $\mathbb{Q}[x]$. Thus the same holds in $\mathbb{Z}[x]$. ∎

Many results about factorization of rings of quadratic integers are known. We mention a few without proof.

Proposition 8.29. [**32**] *For d square free, the norm in* $\mathbb{Q}(\sqrt{d})$ *is a Euclidean functions if and only if d is an element of*

$$\{-11,-7,-3,-2,-1,2,3,5,6,7,11,13,17,19,21,29,33,37,41,57,73\}.$$

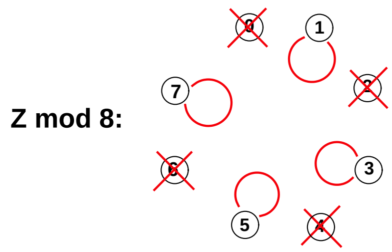
There are exist quadratic fields, such as $\mathbb{Q}[\sqrt{69}]$*, that are Euclidean but whose norm is not a Euclidean function* [**21**].
Furthermore (Baker-Heegner-Stark Theorem , see [**66**]*), for negative square free d, the integers of* $\mathbb{Q}(\sqrt{d})$ *form a PID and not a Euclidean domain if and only if*

$$d \in \{-163,-67,-43,-19\}.$$

Is has been conjectured that for positive (square free) d, there are infinitely many values for which the integers of $\mathbb{Q}(\sqrt{d})$ *have unique factorization.*
For d square free, if the integers of $\mathbb{Q}(\sqrt{d})$ *are a UFD, then they are also a PID.*

8.8. Exercises

Exercise 8.1. Use Definition 8.1 (see also Proposition 5.18) to find the units, irreducibles (see Proposition 8.3 (iii), and primes in:
a) $\mathbb{Z}_6$,
b) $\mathbb{Z}_7$, and
c) $\mathbb{Z}_8$.
(*Hint: the multiplicative inverses and multiplication tables of* $\mathbb{Z}_6$ *and* $\mathbb{Z}_7$ *are given in Figures 21 and 22 and the tables after Definition 5.19. For* $\mathbb{Z}_8$*, the information is in Figure 45.*)



	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7
2	0	2	4	6	0	2	4	6
3	0	3	6	1	4	7	2	5
4	0	4	0	4	0	4	0	4
5	0	5	2	7	4	1	6	3
6	0	6	4	2	0	6	4	2
7	0	7	6	5	4	3	2	1

Figure 45. Left: the relation a is a multiplicative inverse mod 8 of b. Right: the multiplication table of $\mathbb{Z}_8$.

Exercise 8.2. Let R be an integral domain. Consider the set

$$R \times \{R \backslash \{0\}\} = \{(a,b) : a,b \in R, b \neq 0\}.$$

Define an equivalence relation $\sim$ as follows.

$$(a,b) \sim (c,d) \quad \text{if} \quad ad = bc.$$

Frac(R) is the collection of equivalence classes with addition and multiplication:

$$(a,b)+(c,d) = (ad+bc,bd) \qquad \text{and} \qquad (a,b)\cdot(c,d) = (ac,bd).$$

It is not hard (but tedious) to show [**28**][Chapter 8] that $\sim$ is indeed an equivalence and that Frac(R) is the minimal field containing R. Frac(R) is called the <u>field of fractions</u> or <u>field of quotients</u> of R.

a) Show that addition and multiplication are well-defined in Frac(R).

b) What is the field of fractions of $\mathbb{Z}$?

c) The identity is not used in the definition of Frac(R). What is the "field of fractions" of the "rng" (see remark 5.24) $m\mathbb{Z}$ where $m > 1$ in $\mathbb{N}$?

d) Why is it necessary to require that R has no zero divisors?

Exercise 8.3. We apply the Euclidean algorithm in $\mathbb{Z}[\sqrt{-1}]$ as in Section 8.4. For the notation, see the proof of Proposition 8.16. Suppose $\rho\gamma^{-1}$ falls in the unit square depicted in Figure 46. We have drawn four quarter circles of radius 1 in the unit square, denoted by a, b, c, and d.
a) Show that we cannot always choose $\kappa = \kappa_1 + i\kappa_2$ where κ_1 is the *floor* of the real part of $\kappa + \rho\gamma^{-1}$ and κ_2 the *floor* of the imaginary part. (*Hint: Consider the region "northeast" of the quarter circle a.*)
b) Compute the coordinates of the points A, B, C, and D indicated in the figure. (*Hint: Because of the symmetries of the figure, the x coordinate of A equals 1/2. et cetera.*)
c) Show that if $\rho\gamma^{-1}$ falls in the interior of the convex shape $FACE$, then there are four possible choices for κ so that $N(\rho) < N(\gamma)$.
d) Estimate the area of the convex shape $FACE$. (*Hint: It is contained in a square with sides of length BD and it contains a square with sides of length AC.*)
e) Is it possible that there is only one value for κ so that $N(\rho) < N(\gamma)$?

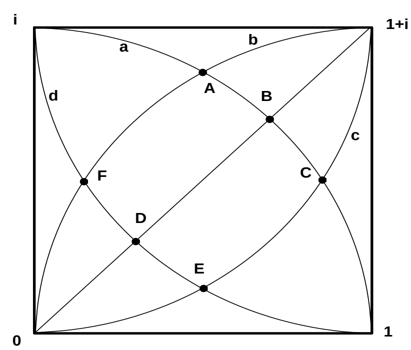


Figure 46. Possible values of $\rho\gamma^{-1}$ in the proof of Proposition 8.16.

In the following proposition and in exercises 8.4, 8.5, and 8.6, we study the primes in $\mathbb{Z}[\sqrt{-1}]$ — called Gaussian primes. Recall that the Gaussian integers from a Euclidean domain (Proposition 8.16), and so we have unique factorization and primes and irreducibles are the same (Corollary 8.13). We use the following notation. C (for "cross") denotes the set $\mathbb{Z} \cup i\mathbb{Z}$ minus the origin. Recall that the units in $\mathbb{Z}[\sqrt{-1}]$ are $\{\pm 1, \pm i\}$ and those in $\mathbb{Z}$ are $\{\pm 1\}$. The notation π means a prime in $\mathbb{Z}[\sqrt{-1}]$, whereas p means a positive prime in $\mathbb{Z}$.

Proposition 8.30 (Gaussian Primes). *A number $\pi \in \mathbb{Z}[\sqrt{-1}]$ is prime if:*
i) $\pi \in C$ and $|\pi|$ equals a prime p in $\mathbb{Z}$ with $p =_4 3$,
ii) $\pi \notin C$ and $|\pi|^2$ equals a prime p in $\mathbb{Z}$ with $p = 2$ or $p =_4 1$.
iii) Furthermore, if π is reducible then (i) and (ii) cannot hold. (So "if" can be replaced by "if and only if".)

For an illustration of the Gaussian primes, see Figure 47.

Exercise 8.4. a) Show that
$$N(\pi) = \pi\overline{\pi} = \prod_i p_i^{k_i}.$$
b) Show that $N(\pi)$ equals p or p^2 (up to units). (*Hint: π must divide one of the primes, say p, in (a).*)
c) Use (b) to show that if $\pi \in C$, then $N(\pi) = p^2$ and so $|\pi| = p$.
d) Use unique factorization in $\mathbb{Z}[\sqrt{-1}]$ to show that if $\pi \notin C$, then $N(\pi) = p$. (*Hint: can $p \cdot p = \pi \cdot \overline{\pi}$?*)

Exercise 8.5. a) Use exercise 5.21 (c) to show that if $p =_4 1$ and p prime in $\mathbb{Z}$, then there is m such that $p \mid m^2 + 1$.
b) Show that if $p =_4 1$, then p is *not* a prime in $\mathbb{Z}[\sqrt{-1}]$. (*Hint: use that $p \mid (m+i)(m-i)$.*) Also show that 2 is *not* a prime in $\mathbb{Z}[\sqrt{-1}]$.
c) Show that $a^2 + b^2 \neq_4 3$. (*Hint: compute modulo 4.*)
d) Show that if a prime p in $\mathbb{Z}$ does not have residue 1 or 3 modulo 4, then $p = 2$.
e) Use exercise 8.4 (c) and (b) of this exercise to prove Proposition 8.30 (i).
f) Then use exercise 8.4 (d) and (c) and (d) of this exercise to prove part (ii).

Exercise 8.6. a) Show that for a reducible γ in $\mathbb{Z}[\sqrt{-1}]$, $N(\gamma)$ is not prime in $\mathbb{Z}$. (*Hint: use Corollary 7.28.*)
b) Use (a) to show that a reducible γ cannot satisfy Proposition 8.30 (ii).
c) Assume γ in C and $\gamma = \alpha\beta$ up to units. Show that if α and β are in C, then $|\gamma|$ is not prime in $\mathbb{Z}$.
d) Assume γ in $\mathbb{N}$ and $\gamma = \alpha\beta$ up to units and that α and β are <u>not</u> in C. Show that if $\gamma = p$, then $|\alpha| = |\beta|$, and therefore are conjugates (*Hint: use Corollary 7.28.*). Show that this implies that $N(\gamma)$ has the form $a^2 + b^2$.
e) Show that (c) and (d) and exercise 8.5 (c) imply that γ cannot satisfy Proposition 8.30 (i).
f) Extend the reasoning in (d) and (e) to all of C.

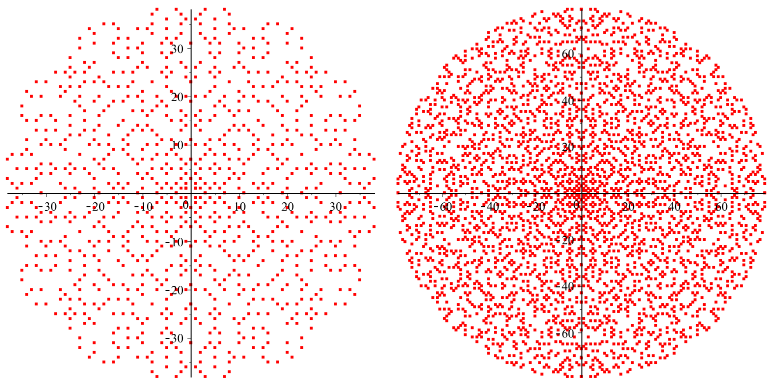


Figure 47. The Gaussian primes described in Proposition 8.30. There are approximately 950 within a radius 40 of the origin (left figure) and about 3300 within a radius 80 (right figure).

Exercise 8.7. Again, we consider numbers in the ring $R = \mathbb{Z}[\sqrt{-1}]$.
a) Show that if $b^n - 1$ is prime in $\mathbb{Z}[\sqrt{-1}]$, then $b-1$ is a unit.
b) Use (a) to show that b must be 2 or $1 \pm i$.
c) Use Proposition 8.30 (i) to show that if $b = 2$, we obtain the usual Mersenne primes (Definition 5.13) as primes in $\mathbb{Z}[\sqrt{-1}]$.
d) Show that if n is not prime, then $b^n - 1$ is not prime. (*Hint: as in exercise 1.12 (i).*)
e) Show that

$$N((1 \pm i)^n - 1) = 2^n - 2^{1+\frac{n}{2}} \cos \frac{n\pi}{4} + 1 \,.$$

(*Hint:* $(1 \pm i) = 2^{1/2} e^{\pm i\pi/4}$ *and* $e^{i\varphi} + e^{-i\varphi} = 2\cos\varphi$.)
f) Show that $(1 \pm i)^n - 1$ is prime if and only if its norm is prime and n is odd. (*Hint: use (d) to show that n must be odd, and then Proposition 8.30.*)

The primes in exercise 8.7 are a generalization of the Mersenne primes of Definition 5.13. These primes in $\mathbb{Z}[\sqrt{-1}]$ of the form $(1 \pm i)^n - 1$ are called Gaussian Mersenne primes . A similar construction works also in the Eisenstein integers $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-3})]$; the resulting primes are called Eisenstein Mersenne primes . For more details, see **[14]**.

Exercise 8.8. Given the ring $R = \mathbb{Z}[\sqrt{-5}]$.
a) Show that 2 is irreducible. (*Hint: suppose* $2 = \beta\gamma$*, where* β *and* γ *are non-units. Use Corollary 7.28 to see that* $N(\beta) = N(\gamma) = 2$*. Solve for the coefficients of* β *and* γ*.*)
b) Show that 3 is irreducible. (*Hint: as (a).*)
c) Use (a) and (b) to show that $1 \pm i\sqrt{5}$ are irreducible.
d) Show that $\mathbb{Z}[\sqrt{-5}]$ is a not Euclidean domain. (*Hint: Show it does not have unique factorization.*)

Exercise 8.9. Given the ring $R = \mathbb{Z}[\sqrt{2}]$.
a) Show that R has no zero divisors. (*Hint: If* $\alpha\beta = 0$*, then one of the norms must be zero by Corollary 7.28. Solve for the coefficients.*)
b) Suppose

$$\rho_1 = \kappa\rho_2 + \rho_3 \ ,$$

where $\rho_1 = a + b\sqrt{2}$, $\rho_2 = c + d\sqrt{2}$, $\kappa = k_1 + k_2\sqrt{2}$, and $\rho_3 = \varepsilon_1 + \varepsilon_2\sqrt{2}$. Show that

$$\rho_1\rho_2^{-1} = \frac{ac - 2bd}{c^2 - 2d^2} + \frac{-ad + bc}{c^2 - 2d^2}\sqrt{2} \ .$$

c) Choose k_1 to be the integer closest to $\frac{ac-2bd}{c^2-2d^2}$ and k_2 the one closest to $\frac{-ad+bc}{c^2-2d^2}$. Show that the remainder has norm with absolute value less than 1. (*Hint: recall that the norm is* $a^2 - 2b^2$*!*)
d) Show that the ring $\mathbb{Z}[\sqrt{2}]$ is a Euclidean domain (*Hint: use Corollary 7.28.*)

Exercise 8.10. a) Show that in $\mathbb{Z}[\sqrt{-5}]$, $(2 + i\sqrt{5})$ is irreducible but not prime. (*Hint: follow the proof of Proposition 8.3 (iv), except now start with* $3 \cdot 3 = 9$ *to prove non-primality.*)
b) Show that in $\mathbb{Z}_6$, 3 is prime but not irreducible. (*Hint: follow the proof of Proposition 8.3 (iii).*)
c) Find other counterexamples.

Exercise 8.11. a) Solve $3x =_b 6x$ where b is 11, 12, 13, 14, 15.
b) If b is such that $\mathbb{Z}_b$ is an integral domain, solve by factoring. c) Use a result in Chapter 5 to show that $\mathbb{Z}_b$ is an integral domain and hence a field if and only if p is prime.
d) Give a direct proof that a field is an integral domain. (*Hint: if a and b are non-zero elements of F, then* $abb^{-1}a^{-1} = 1$*.*)

Exercise 8.12. a) Prove Lemmas 2.5 and 2.6 for a Euclidean domain.
b) Theorem 8.5 follows immediately from the absence of zero divisors (Definition 8.4). In Chapter 2, we take the absence of zero divisors in $\mathbb{Z}$ for granted. Why do we need Euclid's Lemma (Lemma 2.6) — whose proof uses that division algorithm — to prove Theorem 2.7? (*Hint: does the cancellation take place in $\mathbb{Z}$?)*

In the next exercise, we prove:

Lemma 8.31. *Let $d \in \mathbb{Z}$ be square free. $\alpha \in \mathbb{Z}[\sqrt{d}]$ is a unit if and only if $N(\alpha) = \pm 1$.*

Exercise 8.13. a) Show that if α is a unit, $N(\alpha^{-1}) = \frac{1}{N(\alpha)}$.
b) Use (a) to show that the norm of a unit must be ± 1.
c) Vice versa, show that if $N(\alpha) = \pm 1$, then α is invertible. (*Hint: a matrix with determinant ± 1 is invertible. Show that the inverse matrix corresponds to an element of $\mathbb{Z}[\sqrt{d}]$.*)

Exercise 8.14. Consider $\mathbb{Z}[\sqrt{-6}]$ and define $a_\pm = 2 \pm \sqrt{-6}$.
a) Show that $a_- a_+ = 10 = 2 \cdot 5$.
b) Show that $a_\pm$, 2, and 5 are irreducible in $\mathbb{Z}[\sqrt{-6}]$. (*Hint: if $a_+ = \alpha\beta$ is reducible, then $N(a_+) = N(\alpha)N(\beta)$. By Lemma 8.31, we may assume $N(\alpha) = 2$. Solve that equation. And so forth.)*
c) Show that $a_\pm$, 2, and 5 are not primes. (*Hint: for $a_\pm$, use (a)).*
d) Show that unique factorization does not hold. (*Hint: see (a)).*
e) Show that Euclid's lemma 2.6 does not hold here. (*Hint: use Definition 8.2.*)

Exercise 8.15. a) Which ones of the sets in exercise 5.24 are integral domains?
b) Euclidean domains?

Exercise 8.16. a) Show that ± 1 and $\pm 1 \pm \sqrt{2})$ are units of $\mathbb{Z}[\sqrt{2}]$. (*Hint: see Lemma 8.31.*)
b) Show if α is a unit, then for all $n \in \mathbb{Z}$, α^n is a unit.
c) Show that $\mathbb{Z}[\sqrt{2}]$ has infinitely many units.
d) Find solutions of the quadratic equation $a^2 - 2b^2 = \pm 1$. (*Note: an equation of the form $a^2 - db^2 = 0$ where d is square free, is called Pell's equation* .)

One can show that the set of units of $\mathbb{Z}[\sqrt{2}]$ is $\{\pm(1+\sqrt{2})^n : n \in \mathbb{Z}\}$.

Exercise 8.17. Given the ring $R = \mathbb{Z}[\sqrt{10}]$.
a) Show that there is no $\alpha \in R$ with $N(\alpha) = \pm 2$. (*Hint: write* $\alpha = a + b\sqrt{10}$ *and try to solve for the coefficients of* α *in* $\mathbb{Z}_{10}$.)
b) Show that there is no $\alpha \in R$ with $N(\alpha) = \pm 5$. (*Hint: write* $\alpha = a + b\sqrt{10}$. *Then in* $\mathbb{Z}_5$, *show that* $a =_5 0$. *It follows that* $25k^2 - 10b^2 = \pm 5$. *Divide by 5 and solve in* $\mathbb{Z}_5$.)
c) Use (a) and (b) to show that 2 and 5 are irreducible. (*Hint: assume that* $2 = \alpha\beta$, *show that then* $N(\alpha) = \pm 2$, *et cetera.)*
d) Use (a) and (b) to show that $\sqrt{10}$ is irreducible.
e) Show that $\mathbb{Z}[\sqrt{10}]$ is a not Euclidean domain. (*Hint: Show that 10 does not have unique factorization.*)

Exercise 8.18. Given a *field* F, we form the ring $F[x]$ of polynomials. For this exercise, read Section 3.7 again.
a) Use exercise 7.1 to show that the ring $F[x]$ is a Euclidean domain with the degree d (of the polynomial) as a Euclidean function.
b) What goes wrong in (a) if $F = \mathbb{Z}$? (*Hint: give a counter-example.)*
c) What are the "primes" in $F[x]$. (*Hint: see Proposition 7.5 and Corollary 8.13.*)
d) $p_1(x) = x^2 + 1$ is reducible over $\mathbb{C}$, $\mathbb{R}$, or $\mathbb{Q}$? What about $p_2(x) = x^2 - 2$?
e) Show that the degree in $R[x]$ is an *additive* function if R is a domain.

Exercise 8.19. Given a field F.
a) Show that for any $\alpha \in F$ and p in $F[x]$, there are q and r in $F[x]$ such that $p(x) = (x - \alpha)q(x) + r(x)$, where $r(x)$ is a constant. (*Hint: the degree is a Euclidean function.*)
b) Show that in (a), $p(\alpha) = 0$ if and only if $r = 0$. (*Hint: Substitute* $x = \alpha$.)
c) Use (b) to show that if $p_n \in F[x]$ of degree n has a root, then $p_n(x) = p_{n-1}(x)(x - \alpha)$ where p_{n-1} has degree $n - 1$.
d) Use (c) to show that a degree n polynomial in $F[x]$ has at most n roots. (*Compare with exercises 3.22 and 7.17*)

We state the last result of exercise 8.19. It has important applications.

Theorem 8.32 (Lagrange's Theorem). *If f is a degree n polynomial with coefficients in a field F, then $f(x) = 0$ has at most n solutions.*

Exercise 8.20. Define the product of ideals A and B as the smallest ideal containing $\{a_i b_i : a_i \in A, b_i \in B\}$.
a) Show that AB must contain $\left\{\sum_{i=1}^{k} a_i b_i : a_i \in A, b_i \in B \; k \in \mathbb{N}\right\}$.
b) Show that the set in (a) is an ideal.
c) Suppose A is generated by $\{x_i\}$ and B by $\{y_j\}$. Show that AB is the ideal generated by $\{a_i y_j\}$.
d) Use (c) to show that for I and J as in exercise 7.5, $IJ = \langle 6, x\rangle$. (*Hint:* x^2 *is in* $\langle x\rangle$, *and so forth.*)

Exercise 8.21. a) Show that 2 and $1 \pm i\sqrt{3}$ are irreducible in $\mathbb{Z}[\sqrt{-3}]$. (*Hint: follow the proof of Proposition 8.19.*)
b) Use (a) to show that up to units, there are two factorizations in $\mathbb{Z}[\sqrt{-3}]$ of 4 (see equation (8.3)).
c) Use equation (8.3) to show that 4 is not prime.
d) Show that 2 and $(1 \pm i\sqrt{-3})$ are not prime. (*Hint: see Proposition 8.3.*)
e) Conclude that 4 does not admit *any* factorization into primes in $\mathbb{Z}[\sqrt{-3}]$.
f) Show that 2 and $(1 \pm i\sqrt{-3})$ *are* prime in $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-3})]$.

Exercise 8.22. a) Modify the first part of the proof of Proposition 8.22 to show that the norm is a Euclidean function for $\mathbb{Z}[\sqrt{-1}]$ and $\mathbb{Z}[\sqrt{-2}]$ but not for $\mathbb{Z}[\sqrt{-n}]$ for $n \geq 3$.
b) Modify the second part of the proof of Proposition 8.22 to show that the distance to the nearest lattice point of $\mathbb{Z}[\frac{1}{2}(1+\sqrt{j})]$ is less than 1 if $j \in \{-11, \cdots, -1\}$. (*Hint: the height y of the equidistant point in triangle on the left of Figure 43 must be such that $d - y < 1$ where $d = \frac{1}{2}\sqrt{|j|}$.*)
c) Show that with Lemma 7.25, this implies that the norm is a Euclidean function for the integers of $\mathbb{Q}[\sqrt{j}]$ where $j \in \{-11, -7, -3, -2, -1\}$.

Exercise 8.23. Use Definition 7.10 to show that I in part (ii) of the proof of Theorem 8.24 is an ideal.

Exercise 8.24. Consider the map $g : \mathbb{C} \to \mathbb{C}$, defined in the proof of part (ii) of Theorem 8.27. a) Show that $g\left(\frac{\sqrt{19}}{4}\right) = 0$.
b) Show that $-\frac{\sqrt{3}}{2} < g\left(\frac{\sqrt{3}}{2}\right) < 0$.
c) Show that (a) and (b) imply that g maps the blue region in Figure 44 into the red region.

Exercise 8.25. Consider the ring $\mathbb{Z}[x]$.
a) Show that the ideal $I := \langle 3, 5x \rangle$ is not principal. (*Hint: see proof of Theorem 8.28.*)
b) Show by direct computation that I does not generate $\mathbb{Z}[x]$. (*Hint: solve $1 = 3f(x) + 5xg(x)$.*)
c) Show that (b) also follows *directly* from (a). (*Hint: 1 generates all of $\mathbb{Z}[x] \supseteq I$.*)
d) Find $\gcd(3, 5x)$ and $\operatorname{lcm}(3, 5x)$.
e) Show that Bézout does not hold in this ring.

Chapter 9

Ergodic Theory

Overview. This time we venture seemingly very distant from number theory. The reason is that we wish to investigate what properties "typical" real numbers have. By "typical" we mean "almost all"; and to define "almost all", we have to delve fairly deeply into measure theory, one of the backbones of abstract analysis. In this chapter, we will point to the technical problems that need to be addressed, and then quickly state the most important result (the Birkhoff ergodic theorem). In Chapter 10 we will then move to the implications for number theory. The proof of the Birkhoff ergodic theorem will be postponed to Chapter 14.

The ergodic theorem (there are various versions) is arguably one of the most important theorems in mathematics. In essence, it offers a means to replace the study of long-term behavior of complex systems by much simpler statistical reasoning, allowing quantitative predictions for the long term behavior of such a system. Considerations of this nature gave arose out of, and contributed to, an important branch of theoretical physics, namely statistical physics [**31**, **49**]. It is also widely applied in number theory, probability theory, functional analysis, and other fields of study. The discussion whether or not 'physical' systems tend to be ergodic has had a profound impact on science, in particular physics [**31**, **49**]. The use of probabilistic methods to study number theory is often referred to as probabilistic number theory.

9.1. The Trouble with Measure Theory

In analysis we can distinguish short intervals from long ones by looking at their "length" even though both have the same cardinality (see Definition 1.27). The notion of length works perfectly well for simple sets such as intervals. But if we want to consider more general sets – such as Cantor sets — it is definitely very useful to have a more general notion of length, which we denote by *measure*. However, there is a difficulty in formulating a rigorous mathematical theory of measure for arbitrary sets. The source of the difficulty is that there are, in a sense, too many sets. Recall that the real line is uncountable (see Theorem 1.24). The collection of subsets of the line is in fact the same as the power set (Definition 1.33) $P(\mathbb{R})$ of the the real line. And thus the cardinality of the collection of subsets is strictly larger than that of the real numbers (Theorem 1.34), making it a truly very big set.

A reasonable theory of measure for arbitrary subsets of $\mathbb{R}$ should have some basic properties that are consistent with with intuitive notions of "length". If we denote the measure of a set A by $\mu(A)$, then we would like μ to have the following properties.

1) $\mu : P(\mathbb{R}) \to [0,\infty]$.
2) For any interval I: $\mu(I)$ equals the length of I.
3) μ is translation invariant.
4) For a countable collection of disjoint sets A_i: $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

The problem is that no such function exists. Among all the possible sets, we can construct an — admittedly pretty weird — set for which the last three properties cannot simultaneously hold.

To explain this more easily, let us replace $\mathbb{R}$ by the circle $S = \mathbb{R}/\mathbb{Z}$. Now define an equivalence relation (Definition 1.28) in S as follows: $a \sim b$ if $a - b$ is rational. Each element of S clearly belongs to some equivalence class (it is equivalent to itself), and cannot belong to two distinct equivalence classes, because if $a \sim b$ and $a \sim c$, then also the difference between b and c is rational, and hence they belong to the same class. Note that each equivalence class is countable, and so (see exercise 1.8) there are uncountably many equivalence classes.

For every one of these equivalence classes, we pick exactly one representative. The union of these representatives forms a set V. A set constructed this way is known as a Vitali set . Now by requirement (1), any set, no matter how exotic its construction, should have a measure that is a real number. We choose V as our set and let its "measure" be equal to some $\varepsilon \in [0,\infty]$. Let $r : \mathbb{N} \to \mathbb{Q}$ be a bijection between $\mathbb{N}$ and the rationals in S. Consider the union of the translates

$$\cup_{i=1}^{\infty} (V + r_i) .$$

By definition of V, this union covers the entire circle without any overlaps. So by requirement (2) above, the measure of the union is 1. By requirement (3), each of the translates of V must have the same measure, ε. Since the translates of V are disjoint, requirement (4) implies that

$$1 = \sum_{i=1}^{\infty} \varepsilon ,$$

which is clearly impossible!

The construction of the set V just outlined is admittedly a little vague. It is not clear at all *how* exactly we could choose an individual representative, much less how we could achieve that feat for each of the uncountably many equivalence classes. If we wanted to draw a picture of the set V, we'd get nowhere[1]. Does this construction V really exist as an honest set? It turns out that one needs to invoke the axiom of choice[2] to make sure that V exists.

The consensus in current mathematics (2020) is to accept the axiom of choice (though the consensus is not unanimous [**44**]). One consequence of that is that if we want to define a measure, then at least one of those four requirements above needs to be dropped or weakened. The measure theoretic answer to this quandary is to restrict the collection sets for which we can determine a measure. This means, that of the properties (1) through (4), we restrict property (1) to hold only for certain sets. These are called the Lebesgue measurable sets . More generally, not all measures are "length-like", and so we may drop the second requirement. In that case, we speak of measurable sets .

[1] I tried.

[2] The axiom of choice states that for any set A, there exists a function $f : P(A) \to A$ that assigns to each non-empty subset of A assigns an element of that subset. For more details, see [**36**].

9.2. Measure and Integration

To surmount the difficulty sketched in the previous section so that we can define measure and integration unambiguously turns to be technically very involved. This section serves just to give an idea of that complication and its resolution. The interested student should consult the literature, such as the excellent introduction [**8**]. In chapter 14, we provide some more details.

Recall that a set $O \subseteq \mathbb{R}$ is an open set usually[3] means that for all $x \in O$ there is a positive ε so that $(x-\varepsilon, x+\varepsilon)$ is also contained in O. Closed sets are defined as sets whose complement is an open set. Vice versa, the complement of a closed set is open.

Definition 9.1. *Consider the smallest collection of sets closed under complementation, countable intersection, and countable union that contains the open sets. These are called the Borel sets.*

This is simply a way of saying that the Borel sets are the open sets plus all sets that can be obtained from these by complementations, countable intersections, and countable unions.

Definition 9.2. *The outer measure of a set S is*

$$\mu_{out}(S) = \inf \sum_k \ell(I_k)\,.$$

where the infimum is over the countable covers of S by disjoint open intervals I_k.

It turns out that the outer measure is a measure on the Borel sets. This takes some effort to prove and we refer to the literature ([**8**] and [**40**] for a slightly different formulation of essentially those same ideas). To give an idea, one ingredient is that every open sets in $\mathbb{R}$ is a countable union of disjoint open intervals[4] (see exercise 9.4), so the outer measure of an open sets can be calculated easily. After establishing that the outer measure is a measure on the Borel sets, the theory then proceeds by augmenting the Borel sets by arbitrary sets of outer measure zero.

[3]This is called the standard topology on $\mathbb{R}$. It is possible to have different conventions for what the open sets in $\mathbb{R}$ are.

[4]Open sets in $\mathbb{R}^2$ are countable disjoint unions of open rectangles, and so forth in $\mathbb{R}^n$, $n > 2$.

Definition 9.3. *A set S is called Lebesgue measurable if it contains a Borel set B whose outer measure equals $\mu_{out}(S)$. That outer measure is the Lebesgue measure of S. (Thus the measure of an interval equals its length.)*

More informally, Lebesgue measurable sets are Borel sets modulo sets of outer measure zero. One can work out that the collection of Lebesgue measurable sets is also closed under complementation, countable intersection, and countable union. As a consequence of these facts, we have the following result.

Proposition 9.4. *i) A set $S \subset \mathbb{R}$ is Lebesgue measurable if and only if there exist closed sets $C_i \subseteq S$ such that*

$$\mu_{out}(S \backslash \cup_{i=1}^{\infty} C_i) = 0\,.$$

ii) A set $S \subset \mathbb{R}$ is Lebesgue measurable if and only if there exist open sets $O_i \supseteq S$ such that

$$\mu_{out}(\cap_{i=1}^{\infty} O_i \backslash S) = 0\,.$$

Proof. For the first part of this proof, see also exercise 9.1. Observe that every closed set $C_i \subseteq S$ is the complement of an open set $O_i \supseteq S$ and vice versa. So if $\cup_{i=1}^{n} C_i$ contains nearly all of S, then its complement $\cap_{i=1}^{\infty} O_i$ very little more than the complement S^c of S and vice versa. Since complementation preserves the Lebesgue measurable sets (by definition 9.3), (i) and (ii) are equivalent.

A countable union of closed sets is Borel. Since the outer measure is a measure on Borel sets, (i) says that S is a Borel set plus something of outer measure zero. This implies Definition 9.3.

Vice versa, Definition 9.3 says that a Lebesgue measurable set consists of a Borel set contained in a countable collection of disjoint open intervals I_i (by Definition 9.2) *plus* possibly a set Z of outer measure zero. The latter set can be covered by a collection of intervals of arbitrarily small outer measure. ■

Finally, we can define a measure more generally — i.e. not "length-like" or Lebesgue — as follows and show that it satisfies the above characteristics, if one limits the definition to measurable sets.

Definition 9.5. *A measure μ is a non-negative function from a collection*[5] *of μ-measurable sets to $[0,\infty]$ such that $\mu(\emptyset) = 0$ and for every countable sequence of disjoint (measurable) sets S_i:*

$$\mu(\cup_{i=1}^{\infty} S_i) = \sum_{i=1}^{\infty} \mu(S_i)\,.$$

A couple of remarks are in order. The first is the observation that the measurable sets for some arbitrary measure are not constructed in this theory (except for the Lebesgue measure). Rather, they are part of the definition of measure. Which sets are measurable? The sets on which μ is defined.

We remark further that this definition implies that in general sub-additivity holds:

$$\mu(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)\,. \tag{9.1}$$

The reason is that the measure of the union equals the sum of the measures of the disjoint "new" parts A_i' of A_i, i.e. A_i *minus* the intersection of A_i with the A_j where $j < i$. Since $A_i' \cup (A_i \backslash A_i') = A_i$ and this is a disjoint union, we have $\mu(A_i') \leq \mu(A_i)$. Hence the sub-additivity.

Thus a Lebesgue measure μ is a sub-additive function from the (Lebesgue) measurable sets to the positive reals (including infinity) and the measurable sets are constructed so that properties (2), (3), and (4) in Section 9.1 hold, while a more general measure does not have to satisfy (2) and (3). We summarize this as follows.

Corollary 9.6. *The Lebesgue measure μ on $\mathbb{R}$ or $\mathbb{R}/\mathbb{Z}$ satisfies the following properties*
1) μ : measurable sets $\to [0,\infty]$ and $\mu(\emptyset) = 0$.
2) For any interval I: $\mu(I)$ equals the length of I.
3) μ is translation invariant.
4) For a countable collection of disjoint sets A_i: $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.
More generally, a (non Lebesgue) measure satisfies (1) and (3).

We need some more technical terms.

[5] The collection of measurable sets must be closed under complementation and countable unions and intersections

Definition 9.7. *If we have a space X and a collection Σ of measurable sets, then the pair (X,Σ) is called a measurable space. A function $T : X \to X$ is called measurable if the inverse image under T of any measurable set is measurable. A triple (X,Σ,μ) is called a measure space. A probability measure is a measure that assigns a measure 1 to the entire space.*

The Lebesgue integral of a measurable function $f : X \to \mathbb{R}$ with respect to the measure μ is written as

$$I = \int f\, d\mu\,.$$

Assume $f(x)$ is non-negative. To approximate the Lebesgue integral I, one

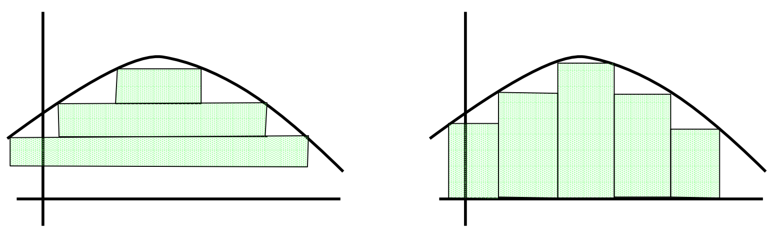


Figure 48. A comparison between approximating the Lebesgue integral (left) and the Riemann integral (right).

partitions the *range* of f into small pieces $[y_i, y_{i+1}]$. For each such layer, the contribution is the measure of the inverse image $f^{-1}(\{y : y \geq y_{i+1}\})$ times $y_{i+1} - y_i$. Sets of measure zero are neglected. Summing all contributions, one obtains an approximation of the Lebesgue integral (see Figures 48 and 86). The Lebesgue integral itself is defined as the limit (if it exists) of these. The Lebesgue integral of a not necessarily non-negative function f is computed by splitting up f into its non-negative part f^+ and its negative part f^-, so that $f = f^+ + f^-$. The integral of f is then defined as

$$I = \int f^+\, d\mu - \int (-f^-)\, d\mu\,.$$

We'll see in Section 14.2 that the domains of f^+ and f^- are measurable so that this operation is well-defined. A function f is called integrable , or μ-integrable for clarity, if $\int |f|\, d\mu$ exists and is finite. It turns out that the Lebesgue integral generalizes the Riemann integral[6] we know from calculus (see exercise 9.6).

[6] Recall that the Riemann integral is approximated by partitioning the *domain* of f, see Figure 48.

This level of technical sophistication means that the fundamental theorems in measure theory require a substantial mastery of the formalism. Since pursuing all the technicalities would take a considerable effort and would lead us well and far away from number theory, we have suppressed some details in this chapter.

9.3. The Birkhoff Ergodic Theorem

The context here is that we have a measurable transformation T from a measure space (X,Σ,μ) to itself. The situation is quite general. The measure μ is not necessarily the Lebesgue measure, but we will assume that it is a probability measure, that is: $\int_X d\mu = \mu(X) = 1$.

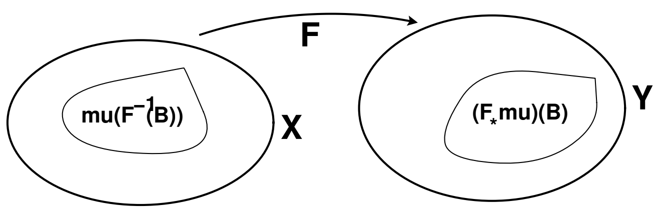


Figure 49. The pushforward of a measure μ.

Definition 9.8. *Let $F : X \to Y$ be a measurable transformation and μ a measure on X. The pushforward $F_*\mu$ of the measure μ is a measure on Y defined as*

$$(F_*\mu)(B) := \mu\left(F^{-1}(B)\right),$$

for every measurable set B in Y (see Figure 49).

Definition 9.9. *Let $T : X \to X$ be measurable. We say that T preserves the (probability) measure μ, or, equivalently, that μ is an invariant measure , if $T_*\mu = \mu$. That is to say, if for every measurable set B, $\mu\left(T^{-1}(B)\right) = \mu(B)$.*

Theorem 9.10 (Birkhoff or Pointwise Ergodic Theorem). *Let $T : X \to X$ be a transformation that preserves the probability measure μ. If $f : X \to \mathbb{R}$ is an integrable function, the limit of the time average*

$$\langle f\rangle(x) := \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x))$$

is defined on a set of full measure. It is an integrable function and satisfies (wherever defined)

$$\int_X \langle f\rangle(x)\,d\mu = \int_X f(x)\,d\mu\,.$$

The proof of this theorem requires a more substantial technical mastery of analysis and we will postpone it to Chapter 14.

Definition 9.11. *Let T be transformation T of a measure space (X,Σ,μ) to itself. A set $S\in\Sigma$ is called* <u>*(weakly)*</u> <u>*invariant*</u> *if $T^{-1}(S)=S$ except possibly for a set of measure zero. We will use the term* <u>*strictly*</u> <u>*invariant*</u> *if $T^{-1}(S)=S$.*

If no misunderstanding is likely, we may drop the word "weakly".

Definition 9.12. *A transformation T of a measure space X to itself is called* <u>*ergodic*</u> *(with respect to μ) if it preserves the measure μ and if every (weakly) T-invariant set has measure 0 or 1.*

This is a slight departure from most texts. Usually, ergodicity means that only *strictly* invariant sets have measure 0 or 1. It turns out that these notions are equivalent (see exercise 14.18). This slight change allows us to give some interesting examples of ergodicity in Section 9.4.

Corollary 9.13. *A measure preserving transformation $T:X\to X$ is ergodic with respect to a probability measure μ if and only if for every integrable function f*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} f(T^i(x)) = \int_X f(x)\,d\mu$$

for all x except possibly on a set of measure 0.

The proof of this corollary will also be given in Chapter 14. Somewhat confusingly, this last result is often also called the Birkhoff ergodic theorem. We will also adhere to that usage, just so that we can avoid saying "the corollary to the Birkhoff ergodic theorem" on many occasions. This corollary really says that a transformation is ergodic if and only if *time averages equal spatial averages*. This is a very important result because, as we will see, spatial averages are often much easier to compute. This has major implications in physics.

One needs to be careful, because it can happen that a transformation is ergodic with respect to two (or more) different measures.

Definition 9.14. *Two probability measures μ and ν are mutually singular if there is a measurable set S with $\mu(S) = 1$ and $\nu(S) = 0$, and vice versa.*

Corollary 9.15. *If T is ergodic with respect to two distinct probability measures μ and ν, then those measures are mutually singular.*

Proof. If μ and ν are distinct measures, we can choose f such that

$$c_1 = \int_X f\,d\mu \neq \int_X f\,d\nu = c_2\,.$$

By Corollary 9.13, the time average $\langle f\rangle(x)$ must be c_1 for μ almost every x and so the x for which the average is c_2 has μ measure 0. The reverse also holds. ∎

One can furthermore prove that the set of invariant probability measures is non empty and every invariant measure is a convex combination of ergodic measures [**48**][chapter 8]. This says that, in a sense, ergodic measures are the building blocks of chaotic dynamics. If we find ergodic behavior with respect to some measure μ, then we understand the statistical behavior for almost all points with respect to μ. There may be other complicated behavior but this is “negligible” if you measure it with μ.

9.4. Examples of Ergodic Measures

In this section, we consider the piecewise linear map T with derivative equal to 2, depicted in Figure 50. To fix our thoughts, we set $A = [0,1]$ and $B = [1,2]$. In this section, we will exhibit uncountably many invariant probability measures μ with respect to which T is ergodic. Note that any two such measures must be mutually singular (Definition 9.14). This situation is by no means exceptional.

We start with the measure δ_0 that assigns (full) measure 1 to the point 0 and measure 0 to any (measurable) set not containing 0. As we can see in Figure 50, for any set S

$$0 \in S \quad \Longleftrightarrow \quad 0 \in T^{-1}(S)\,.$$

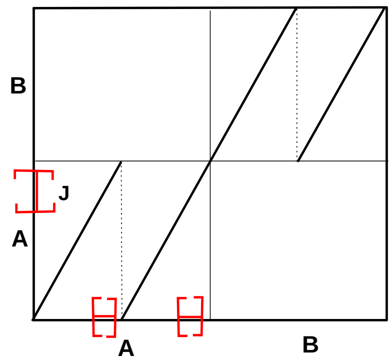


Figure 50. This map has many ergodic measures

Thus $\delta_0(S) = \delta_0\left(T^{-1}(S)\right)$, that is: δ_0 is (weakly[7]) T-invariant. Since any T-invariant set either contains the point 0 or not, such a set trivially has measure either zero or one. By Definition 9.12, T is ergodic with respect to δ_0. Let us check the conclusion of Corollary 9.13. For some very small $\varepsilon > 0$, set

$$f(x) = \begin{cases} 0 & \text{if } x \in [0,\varepsilon) \\ \alpha & \text{if } x \in [\varepsilon,2] \end{cases}$$

Take some arbitrary x. Under iteration by T, it will most likely randomly bounce around either in the interval $[0,1]$ or in the other interval, $[1,2]$. If $x_0 \in (0,1)$, then $f(T^i(x_0))$ will nearly always be α, and if $x_0 \in [1,2]$, it will always be α. Either way, the time-average is close to α. However, the integral $\int_X f(x)\,d\delta_0)$ gives $f(0) = 0$! What is going on? See this footnote[8].

Note that, in this example, the set $\{0\}$ has pre-image $\{0\} \cup \{1/2\}$. The second point (1/2) carries no measure. Thus $\{0\}$ is invariant but *not* strictly invariant! Similarly, we can put a discrete measure on any q-periodic orbit by giving each point of the orbit a measure $1/q$. With respect to that measure, the transformation will be ergodic, because the only invariant with positive measure is the entire orbit. Again, the invariant set is not strictly invariant (see exercise 14.19).

[7] But not strictly!

[8] The set $(0,2]$ has measure 0 with respect to δ_0. Corollary 9.13 tells us to neglect such sets. Thus we must take $x = 0$, and then the summation also gives 0.

The next example is the uniform measure μ_A in $A=[0,1]$. Each measurable subset of A has a measure equal to its Lebesgue measure. It is easy to see that this is a probability measure (one that integrates to 1). From Figure 50, we see that the inverse image of an interval $J \subseteq A$ equals two disjoint intervals of half its length. This shows that μ_A is invariant under T. We will show in Chapter 10 that each T-invariant set has μ_A measure either 0 or 1 (that is: μ_A is ergodic), but here is a partial result.

Proposition 9.16. *Let T be the map $x \to 2x$ mod 1 on $A=[0,1]$ and suppose μ is the Lebesgue measure. If $S \subseteq A$ is a T-invariant set with $\mu(S)>0$, then S must be dense in A.*

Proof. Note that T restricted to the interval $A=[0,1]$ is just the doubling map. By the discussion above, the Lebesgue measure μ is T-invariant.

The inverse image $T^{-1}(S)$ is:

$$S_0 \cup S_1 := \frac{S+0}{2} \cup \frac{S+1}{2},$$

According to Proposition 9.4 (ii), we can cover S with an open set O of measure less than $\mu(S)+\varepsilon$ for any $\varepsilon>0$. Using exercise 9.4, we see that we can cover S with disjoint open intervals of measure less than $\mu(S)+\varepsilon$. Thus we can cover each of S_0 and S_1 with open intervals of length no more than half that. By Definition 9.3, $\mu(S_0)$ and $\mu(S_1)$ are at most $\frac{1}{2}\mu(S)$. Since, however, S is invariant and so $\mu(S_0)+\mu(S_1)=\mu(S)$, we conclude that $\mu(S_0)=\mu(S_1)=\frac{1}{2}\mu(S)$.

Iterating this procedure, we get

$$T^{-2}(S) = S_{00} \cup S_{01} \cup S_{10} \cup S_{11} := \frac{S+0}{4} \cup \frac{S+2}{4} \cup \frac{S+1}{4} \cup \frac{S+3}{4}.$$

Each of these contains 2^{-2} of the measure of S. Similarly, the nth iterate gives a collection of 2^n regularly spaced copies of $2^{-n}S$. Clearly, the union of these over n is dense and each little copy must contain a set of positive measure belonging to S. ■

Note that the complement of an invariant set is also invariant. Thus result implies that if $S \subset [0,1]$ is an invariant set whose complement S^c has positive measure, then both are dense. This is equivalent to the following.

Corollary 9.17. *Let T be the map $x \to 2x$ mod 1 on $A=[0,1]$ and suppose μ is the invariant Lebesgue measure. If T is not ergodic, then there must be* *an invariant S such that both it and its complement have positive measure and are dense in A.*

We will see in Chapter 10 that in fact this is not possible, and so T is ergodic with respect to μ_A.

For now note that both A and B are T-invariant sets and $\mu_A(A) = 1$ while $\mu_A(B) = 0$. We check Corollary 9.13 again. Let f be

$$f(x) = \begin{cases} \alpha & \text{if } x \in [\frac{1}{2}, 1) \\ 0 & \text{else} \end{cases}$$

For arbitrary x in $[0,1]$, we expect $T^i(x)$ to hit the interval $[0, \frac{1}{2}]$ half the time on average. So the sum should give $\frac{\alpha}{2}$. Indeed, if we compute the integral $\int f\, d\mu_A$, that is what we obtain.

Now we turn to an at first sight very strange and counter-intuitive example. In the unit interval, we consider the set of x with all possible binary expansions, but now we construct a measure ν_p that assigns a measure $p \in (0,1)$ to "0", and $1-p$ to "1". In effect this amounts to assigning a measure p to the interval $[0, \frac{1}{2}]$ and $1-p$ to $[\frac{1}{2}, 1]$. The interesting case is of course when $p \neq \frac{1}{2}$. So that is what we will assume.

Continuing the construction of the measure ν_p, the set of sequences starting with 00 get assigned a measure p^2; the ones starting with 01, a measure $p(1-p)$; 10, a measure $(1-p)p$; and 11, a measure $(1-p)^2$. The sum of these is 1. We now keep going ad infinitum, always keeping the sum of the measures equal to 1, see Figure 51. So ν_p is a probability measure.

The same reasoning as in Proposition 9.16 shows that an interval I consisting of points whose binary expansion starts with $a = a_1 a_2 \cdots a_n$ has pre-image $I_0 \cup I_1$, where I_0 consists of the points whose expansion starts with $0a$ and I_1, those starting with $1a$.

$$\nu_p(I_0) + \nu_p(I_1) = p\nu_p(I) + (1-p)\nu_p(I) = \nu_p(I),$$

and so the measure ν_p is T-invariant.

This gives us an uncountable set of T-invariant measures ν_p (one for each $p \in (0,1)$). For each of these measures, we are in the same situation as Corollary 9.17: if ν_p is not ergodic, there must be very strange invariant sets. (And in fact, those measures *are* ergodic.

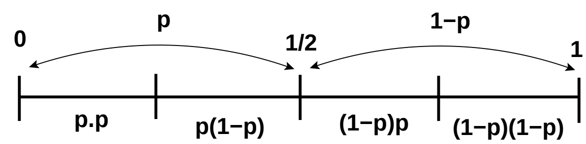


Figure 51. The first two stages of the construction of the singular measure ν_p.

9.5. The Lebesgue Decomposition

The examples of invariant measures of Section 9.4 also help to illustrate the following fact [**40**] which we mention without proof (but see [**8**]).

Theorem 9.18 (Lebesgue Decomposition). *Let μ be a given measure. An arbitrary measure ν has a unique representation as the sum*

$$\nu = \nu_{ac} + \nu_d + \nu_{sc}\,.$$

where ν_{ac} absolutely continuous with respect to the Lebesgue measure μ, ν_d is a discrete measure, and ν_{sc} is singular continuous.

We now define these notions somewhat informally. A measure ν_{ac} is absolutely continuous with respect to μ if for all measurable sets A, $\mu(A) = 0$ implies that $\nu_{ac}(A) = 0$. It is usually written as $\nu_{ac} \ll \mu$. The Radon-Nikodym theorem theorem then implies that ν_{ac} has a non-negative, integrable density with respect to μ. This means that if $\nu_{ac} \ll \mu$, we can write $d\nu_{ac} = \rho(x)d\mu$ (see [**40**]). The density ρ is also called the *Radon-Nikodym derivative* of ν_{ac} (relative to μ) and it is often written as

$$\frac{d\nu_{ac}}{d\mu} = \rho\,.$$

We can use the density to change variables under the integral. For any integrable f

$$\int f(x)\,d\nu_{ac}(x) = \int f(x)\rho(x)\,d\mu(x)\,.$$

Thus ρ is the density of ν_{ac} (with respect to μ). Often, μ is the Lebesgue measure so that $d\mu(x) = dx$. This is usually the case when we think of common probability measures in statistics, such as the Beta distribution on $[0,1]$,

$$d\nu(x) = Cx^{a-1}(1-x)^{b-1}\,dx\,.$$

This is an example of a measure that is absolutely continuous with respect to the Lebesgue measure. In this case, ρ is called the probability density, and its integral is $\nu(x)-\nu(0)$, the cumulative probability distribution. The constant C is needed to normalize the integral $\int d\nu = 1$.

The *discrete measure* ν_d is concentrated on a finite or countable set of μ-measure zero. The measure δ_0 is an example of this.

Finally, the measure ν_p for $p \neq \frac{1}{2}$ is an example of a *singular continuous measure* with respect to the Lebesgue measure μ. This is a measure that is singular with respect to μ, but, still, single element sets $\{x\}$ that satisfy $\mu(\{x\}) = 0$ also have ν_p-measure zero.

Recall that Corollary 9.15 says that if $p \neq q$ are two numbers in $[0,1]$, then the measures ν_p and ν_q are mutually singular, even though they are clearly continuous with respect to one another by the above informal definition. Since this is maybe more than a little counter-intuitive, let us verify that again.

Lemma 9.19. *Let p, q distinct numbers in* $[0,1]$*. The measures* ν_p *and* ν_q *are mutually singular.*

Proof. As we saw in Section 9.4, the angle doubling transformation given by T restricted to the interval $[0,1]$ is ergodic with respect to each of the two measures. So let $f(x) = 1$ on $[0, \frac{1}{2}]$ and 0 elsewhere. Birkhoff's theorem implies that for x in a set of full ν_p-measure, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x)) = \int_X f(x)\, d\nu_p = p\,.$$

This means that ν_p-almost all x land in $[0, \frac{1}{2}]$ a fraction p of the time on average. Thus the set of points that land in $[0, \frac{1}{2}]$ on average a fraction q of the time has ν_p measure zero. But those have full ν_q measure. ∎

Note that the binary expansion of the ν_p typical (that is: in a subset having full measure) x has on average a fraction of exactly p zeroes.

9.6. Exercises

Exercise 9.1. Let S, A_n be sets in a space X, and I any (possibly uncountable) index set. See Figure 52.
a) Show that $\left(\bigcap_{n\in I} A_n\right)^c = \bigcup_{n\in I} A_n^c$.
b) Show that $\left(\bigcup_{n\in I} A_n\right)^c = \bigcap_{n\in I} A_n^c$.
c) Show that every closed set $C \subseteq S$ is the complement of an open set $O \supseteq S^c$.
d) Show that for O and S as in (c): $S\backslash O^c = O\backslash S^c$.
(*Note: the first two statements are known as the De Morgan laws .*)

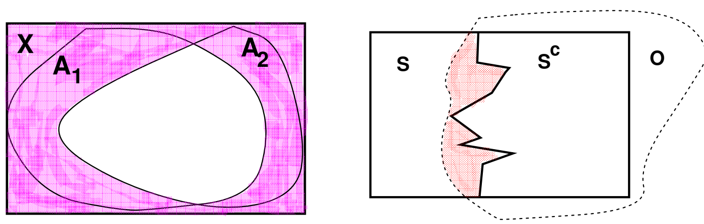


Figure 52. The left figure illustrates that $(\cap A_i)^c = \cup_i A_i^c$. The right figure illustrates that for an open set O containing S^c, $S\backslash O^c = O\backslash S^c$ (shaded in red).

Exercise 9.2. Reformulate the counter example in Section 9.1 as a counter example in $\mathbb{R}$. *(Hint: two numbers in* $[0,1]$ *are equivalent if their difference is rational. Let* $V \subseteq [0,1]$ *be a set that contains exactly one representative of each class. Let* R *be the set of rationals in* $[-1,1]$. *Then consider the union* $U := \cup_{r\in R} V + r$. *Show that* $[0,1] \subseteq U \subseteq [-1,2]$.*)*

Exercise 9.3. a) Show that the rational numbers in $[0,1]$ can be contained in an open set of arbitrarily small measure. (*Hint: for some* $\lambda > 1$, *put the number* p/q *in an open interval of length* $C\varphi(q)^{-1}\lambda^{-q}$, *where* φ *is the totient function.*)
b) Use (a) to show that the rational numbers in $\mathbb{R}$ an be contained in an open set of arbitrarily small measure. (*Hint: in each unit interval, choose an appropriate* C *as defined in (a).*) c) Show there is a closed set in $[0,1]$ of measure greater than $1-\varepsilon$ that contains *only* irrational numbers.

Exercise 9.4. Show that any open set O in $\mathbb{R}$ is a finite or countable union of disjoint open intervals. (*Hint: for every* $x \in O$ *there is an open interval* $(a,b) \subseteq O$ *that contains* x. *Now let* $\alpha = \inf\{a : (a,b) \subseteq O\ ,\ x \in (a,b)\}$ *and similar for* β. *This way we obtain a partitioning of* O *into open intervals. Each such interval must contain a rational number of which there are countably many.*)

In the next exercise, we prove the following Lemma.

Lemma 9.20. *i) Any set in a probability space X with outer measure zero is Lebesgue measurable with Lebesgue measure zero.*
ii) A countable union of measure 0 sets has measure 0.

Exercise 9.5. a) Show that the empty set has measure zero. (*Hint: see Definition 9.5.*)
b) Prove part (i) of the lemma for a non empty set. (*Hint: a non empty set contains a point which is a Borel set; now apply Definition 9.3.*)
c) Prove part (ii) of the lemma. (*Hint: use equation* (9.1).)

Exercise 9.6. Let $X = [0,1]$, E the set of irrational numbers in X, and μ the Lebesgue measure.
a) Show that $\int_E d\mu = 1$. *(Hint: approximate the Lebesgue integral as in Section 9.1.)*
b) Show that the Riemann integral $\int_E dx$ is undefined. *(Hint: look up the exact definition of Riemann integral)*

Exercise 9.7. Construct the middle third Cantor set $C \subseteq [0,1]$ in the following way (Figure 53). At stage 0, take out the open middle third interval of the unit interval. At stage 1, take out the open middle third interval of the two remaining intervals. At stage n, take out the open middle third interval of each of the 2^n remaining intervals. The set C consists of the points that are not removed. See also exercise 1.10.
a) Show that C consists of all points $x = \sum_{i=1}^{\infty} a_i 3^{-i}$ where $\{a_i\}_{i=1}^{\infty}$ are arbitrary sequences in $\{0,2\}^{\mathbb{N}}$.
b) Show that the Lebesgue measure of C is zero.
c) Show that C is uncountable. (*Hint: look at the proof of Theorem 1.24.*)

Figure 53. The first two stages of the construction of the middle third Cantor set. The shaded parts are taken out.

Exercise 9.8. Construct the set $C \subseteq [0,1]$ in the same way as in exercise 9.7, but now at stage n, take out (open intervals of) an arbitrary fraction $m_n \in (0,1)$ from the middle of each of the remaining intervals.
a) Show that C is non-empty. (*Hint: find a point that is never taken out.*)
b) Let $m_i = 1 - e^{-\alpha^i}$ for some $\alpha \in (0,1)$. Compute the Lebesgue measure of C and its complement. (*Hint: at every stage, consider the length of the set that is left over. You should get that $\mu(C) = e^{-1/(1-\alpha)}$.*)

We remark that Cantor sets with positive measure such as those in exercise 9.8 are sometimes called fat Cantor sets.

Exercise 9.9. a) Show that the Borel sets contain the closed sets. (*Hint: a closed set is the complement of an open set.*)
b) Show that the middle third Cantor set (see exercise 9.7) is a Borel set.
c) Show that the Cantor sets of exercise 9.8 are Borel sets.
d) Show the sets in (c) are Lebesgue measurable.

Exercise 9.10. Construct the Cantor function $c:[0,1]\to[0,1]$, also called Devil's staircase as follows. See also exercise 9.7. See Figure 54.
Start with stage 0: $c(0)=0$ and $c(1)=1$. At stage 1, set $c(x)=\frac{1}{2}$ if $x\in[\frac{1}{3},\frac{2}{3}]$. At stage 2, set $c(x)=\frac{1}{4}$ if $x\in[\frac{1}{9},\frac{2}{9}]$ and $c(x)=\frac{3}{4}$ if $x\in[\frac{7}{9},\frac{8}{9}]$.
Use a computer program to draw 5 or more stages. $c(x)$ is the continuous function that is the limit of this process.

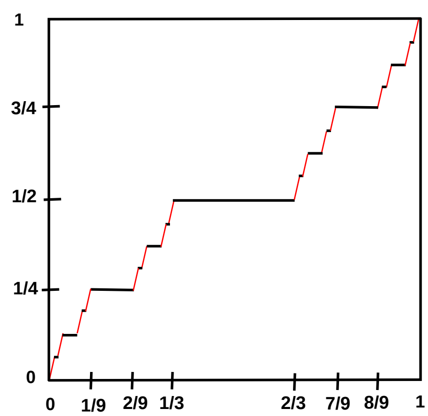


Figure 54. An impression of the Cantor function of exercises 9.10 and 9.11. The first four stages are drawn in black, the red segments are affine interpolations.

Exercise 9.11. Recall the definition of the Cantor function, $c : [0,1] \to [0,1]$ (exercise 9.10).
a) Use exercise 9.7 (a) to show that for x in the Cantor set

$$x = \sum_{i=1}^{\infty} a_i 3^{-i} \quad \Longrightarrow \quad c(x) = \sum_{i=1}^{\infty} \frac{a_i}{2} 2^{-i}.$$

b) Show that on any interval not intersecting the Cantor set c is constant.
c) Show that $c : [0,1] \to [0,1]$ is onto.
d) Show that c is non-decreasing.
e) Show that $c(x)$, is continuous. (*Hint: find a proof that a non-decreasing function from an interval onto itself is continuous.*)

Since c is increasing from 0 to 1, we can interpret it as a cumulative distribution function. The measure μ of $[a,b] \subseteq [0,1]$ equals $c(b) - c(a)$. If $[a,b]$ is inside any of the flat parts, then its measure equals zero. Thus the measure of the complement of the Cantor set is zero, and all measure is concentrated on the Cantor set.

Exercise 9.12. Find the Lebesgue decomposition (Theorem 9.18) of c in exercise 9.11 interpreted as a measure. Explain!

Exercise 9.13. a) Show that the derivative c' of the Cantor function c of exercise 9.10 equals 0 almost everywhere.
b) Show that Lebesgue integration gives $\int_0^1 c'(t)\,dt = 0$. (*Hint:* $c'(t) = 0$ *on a set of full measure. Then use the informal definition of Lebesgue integration in Section 9.2.*)
c) Conclude that in this case $c(1) - c(0) = \int_0^1 c'(t)\,dt$ is false.

The equation in item (c) of exercise 9.13 holds in the case where the function c admits a derivative *everywhere*. The interested student should recall the second fundamental theorem of calculus (e.g. [**53**] [Section 7.1]).

Exercise 9.14. Consider the map $t : [0,1] \to [0,1]$ given by $T(x) = \{10x\}$, the fractional part of $10x$.
a) Show that the Lebesgue measure dx is invariant under T.
b) Prove Corollary 9.17 for this map.
c) Show that the frequency with $T^n(x)$ visits the interval $I = [0.358, 0.359)$ equals the frequency with which 358 occurs (if that average exists).
d) Assuming ergodicity of T, show that for Lebesgue almost every x, that average equals 10^{-3}. (*Hint: use the corollary to Birkhoff's theorem with* $f(x) = 1$ *on I and 0 elsewhere.*)

Exercise 9.15. a) Show that there exist x in whose decimal expansion the word "358" occurs more often than in almost all other numbers (see exercise 9.14 (d)).
b) Show that the frequency of occurrences of "358" in the decimal expansion of a number x does not necessarily exist.
c) What is the Lebesgue measure of of set of numbers referred to in (a) and (b). (*Hint: see exercise 9.14 (d).*)

Exercise 9.16. In an interview, Yakov Sinai explained ergodicity as follows. Suppose you live in a city above a shoe store. One day you decide you want to buy a perfect pair of shoes. Two strategies occur to you. You visit the shoe store downstairs every day until you find the perfect pair. Or you can rent a car to visit every shoe store in the city and find the best pair that way. The system is ergodic if both strategies give the same result. Explain Sinai's reasoning.

Exercise 9.17. a) Fix an integer $b > 1$ and let w be any finite word in $\{0, 1, \cdots b-1\}^{\mathbb{N}}$ of length n. Show that for almost all x, the frequency with which that word occurs in the expansion in base b equals b^{-n}. (*Hint: assume ergodicity of $x \to \{bx\}$ and follow the reasoning in exercise 9.14.*)
b) The measure of the set of x for which that frequency is not b^{-n} is zero.

Definition 9.21. *Let $b \geq 2$ an integer. A real number in $[0, 1]$ is called <u>normal</u> <u>in</u> <u>base</u> <u>b</u> if its infinite expansion in the base b has the property that all words of length n occur with frequency b^{-n}. A number is called <u>absolutely</u> <u>normal</u> if the property holds for every integer $b > 2$.*

Exercise 9.18. a) Use exercise 9.17, Corollary 1.25, and Lemma 9.20 to show that the set of words not normal in base b has measure 0.
b) Show that the set of absolutely normal numbers has full measure.

Exercise 9.19. a) Show that the set of numbers that are not normal in base $b > 2$ is uncountable. (*Hint: words with a missing digit are a subset of these; see exercise 9.7.*)
b) Repeat (a), but now for base 2. (*Hint: rewrite in base 4 with digits 00, 01, 10, and 11; follow (a).*)

Exercise 9.20. a) Show that the set of absolutely normal numbers is dense. (*Hint: follows from exercise 9.18.*)
b) Show that numbers with finite expansion in base b are non-normal in base b.
c) For any $b > 1$, show that the set of non-normal numbers in base b is also dense. (*Hint: rational numbers.*)

Exercise 9.21. Show that a rational number is non-normal in any base. (*Hint: generalize proposition 5.8 to show that the expansion of a rational number in base b is eventually periodic.*)

Exercise 9.22. a) In base 2, construct a number C_2 whose expansion is the list of all finite words. Start with the string s_1 consisting of all length 1 words in ascending order, "0" and "1". So $s_1 = 01$. Then obtain s_2 consisting of all length 2 words in ascending order. So $s_2 = 00011011$. And so forth. The binary expansion of C_2 is: $0.s_1s_2s_2s_3\cdots$. (*Note: this number in base 2 and its generalizations to base b are usually called* <u>*Champernowne numbers*</u> [**17**]*[Chapter 4].*)
b) Challenge[a]: Show that the number C_b whose expansion in base b is the list of all finite words constructed following the method in (a) is normal in base b. (*Hint: pick a word w of length n. Show that w occurs in 1 out of* b^n *times in every "level"* $k \geq n$*.*)
c) Numerically compute the first 6 continued fraction convergents for $C_{10} = 0.1234567890\cdots$. (*Hint: the fourth continued fraction coefficient equals 135678.*)

[a] Though intuitively "obvious", the details of this proof are very tricky! If you know a simple proof, let me know.

Definition 9.22. *A real sequence* $\{x_i\}_{i=1}^{\infty}$ *is* <u>*equidistributed*</u> *modulo 1 (with respect to Lebesgue measure) if its fractional values* $\{a_i\}_{i=1}^{\infty}$ *are such that for each subinterval* $[a,b]$ *of* $\mathbb{R}/\mathbb{Z}$

$$\lim_{n\to\infty} \frac{|\{a_1,a_2,\cdots a_n\}\cap[a,b]|}{n} = b-a\,.$$

In other words: the frequency of hitting a set is proportional to the Lebesgue measure of that set.

Exercise 9.23. Show that x is normal in base $b > 2$ ($b \in \mathbb{N}$) if and only if the sequence $a_n = \{xb^n\}$ is equidistributed modulo 1, where $\{\cdot\}$ means fractional part. (*Hint: one direction is obvious; for the other direction, note that for all epsilon* > 0 *and any interval* $[c,d]$*, there are b-adic intervals I and J such that* $I \subseteq [c,d] \subseteq J$ *and* $|j\backslash I| < \varepsilon$*.*)

As with so many issues in number theory, for any of the numbers we care about — such as e, π, $\sqrt{2}$, et cetera — it is not known (in 2021) whether they are normal in any base.

Exercise 9.24. a) Show that a rotation on $\mathbb{R}/\mathbb{Z}$ preserves the Lebesgue measure. (*Hint: Corollary 9.6 (iii).*)
b) Show that a rotation on $\mathbb{R}/\mathbb{Z}$ by a rational number is not ergodic. (*Hint: start with the identity which is a rotation by 0.*)

Chapter 10

Three Maps and the Real Numbers

Overview. In this chapter, we consider the three maps from $[0,1)$ to itself that are most important for our understanding of the statistical properties of real numbers. They are: multiplication by an integer n modulo 1, rotation by an irrational number, and the Gauss map that we discussed in Chapter 6. In doing this, we review three standard techniques to establish ergodicity. In this chapter we restrict all measures, transformations, and so on to live in one dimension ($[0,1)$ or $\mathbb{R}/\mathbb{Z}$).

10.1. Invariant Measures

If we wish to prove that a measurable transformation $T : X \to X$ is ergodic, we first need to find an invariant measure. Recall the notions of pushforward of a measure (Definition 9.8) and invariant measure (Definition 9.9).

Lemma 10.1. *Let $T : X \to X$ a measurable transformation and μ a T-invariant measure on X. Then for every μ-integrable function f, we have*

$$\int f(x)\, d\mu(x) = \int f(T(x))\, d\mu(x)\,.$$

Proof. Setting $y = Tx$, the definition of the pushforward gives

$$\int_X f(y)\, dT_*\mu(y) = \int_X f(T(x))\, d\mu(x)\,.$$

On the other hand, since μ is invariant, we also have

$$\int_X f(y)\, dT_*\mu(y) = \int_X f(y)\, d\mu(y)\,.$$

Putting the two together gives the lemma. ■

In most cases, and certainly in this text, we are interested in invariant measures ν that are absolutely continuous with respect to the Lebesgue measure (see Section 9.3). Thus $d\nu = \rho(x)dx$. It is often easier to compute with densities than it is with measures. We formulate the pushforward for densities.

Lemma 10.2. *The pushforward $T_*\rho$ by $T : [0,1) \to [0,1)$ of a density ρ is given by*

$$T_*\rho(y) = \sum_{Tx=y} \frac{\rho(x)}{|T'(x)|}\,.$$

This is called the Perron-Frobenius operator.

Proof. The measure of the pushforward $T_*\rho$ contained in the small interval dy is $\tilde{\rho}(y)dy$. By Definition 9.8, it is equal to $\sum_{Tx=y} \rho(x)dx$ where dx is the length of the interval $T^{-1}(dy)$ (see Figure 55). Now the length of $T^{-1}(dy)$

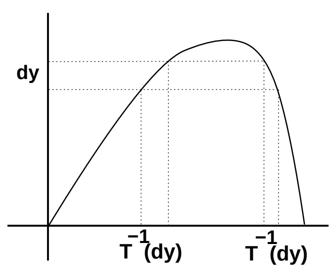


Figure 55. The inverse image of a small interval dy is $T^{-1}(dy)$

is of course equal to the length of $\left|\frac{d}{dy}T^{-1}(y)\right| dy$. Since

$$\left|\frac{d}{dy}T^{-1}(y)\right| dy = \frac{dy}{|T'(x)|}\,,$$

the result follows. ■

Thus T preserves an absolute continuous (with respect to the Lebesgue measure) measure with density ρ if and only if

$$\rho(y) = \sum_{Tx=y} \frac{\rho(x)}{|T'(x)|}\,. \tag{10.1}$$

The first, and simplest, of the three transformations are the rotations. A rotation T is invertible and $T'(x) = 1$. Therefore, if $\rho(x) = 1$, Lemma 10.2 also yields 1 for its pushforward $T_*\rho$, and thus equation (10.1) is satisfied. If instead T is defined as $x \to \tau x$ modulo 1, where τ is any integer other than ± 1 or 0, the situation is different, but still not very complicated. We will call these transformations angle multiplications for short. Now each y has $|\tau|$ inverse images $\{x_1, \cdots x_\tau\}$ and $T'(x_i) = \frac{1}{\tau}$. So if $\rho(x) = 1$, Lemma 10.2 yields $T_*\rho(x) = 1$ for the pushforward.

The situation is slightly more complicated for the Gauss map of Definition 6.1.

Proposition 10.3. *i) Rotations and angle multiplying transformations on* $\mathbb{R}/\mathbb{Z}$ *preserve the Lebesgue measure.*
ii) The Gauss map preserves the probability measure

$$d\nu = \frac{1}{\ln 2}\frac{dx}{1+x}\,.$$

Proof. We already proved item (i). For item (ii), notice that

$$\nu([0,x]) = \frac{1}{\ln 2}\int_0^x \frac{1}{1+s}\,ds = \frac{1}{\ln 2}\ln(1+x)\,,$$

so $\nu([0,1]) = 1$ and ν is as probability measure. It is easy to check that that the inverse image under T of $[0,x]$ is the union of the intervals $\left[\frac{1}{a+x}, \frac{1}{a}\right]$ (see

Figure 56), and so

$$\begin{aligned}\nu(T^{-1}([0,x])) &= \nu\left(\bigcup_{a=1}^{\infty}\left[\tfrac{1}{a+x},\tfrac{1}{a}\right]\right)\\ &= \tfrac{1}{\ln 2}\sum_{a=1}^{\infty}\left\{\ln\left(1+\tfrac{1}{a}\right)-\ln\left(1+\tfrac{1}{a+x}\right)\right\}\\ &= \tfrac{1}{\ln 2}\sum_{a=1}^{\infty}\left\{\ln\left(\tfrac{a+1}{a}\right)-\ln\left(\tfrac{a+1+x}{a+x}\right)\right\}\\ &= \tfrac{1}{\ln 2}\sum_{a=1}^{\infty}\left\{\ln\left(\tfrac{a+x}{a}\right)-\ln\left(\tfrac{a+1+x}{a+1}\right)\right\}\\ &= \tfrac{1}{\ln 2}\ln(1+x) \quad .\end{aligned}$$

The last equality follows because the sum telescopes.

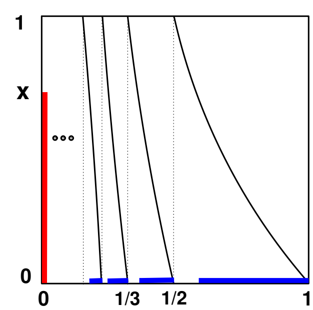


Figure 56. The interval $[0,x)$ (shaded red) and its pre-image under the Gauss map (shaded blue).

This computation shows that the measure on intervals of the form $[0,x]$ or $(0,x)$ is invariant. Taking a difference, we see that the measure an any interval (x,y) is invariant. Therefore, the same is true for any open set (see 9.4). Thus it is true any Borel set (Definition 9.1). Since Lebesgue measurable sets can be approximated by Borel sets (Proposition 9.4), the result follows. ■

At the end of this last proof, we needed to jump through some hoops to get from the invariance of the measure of simple intervals to that of all Borel sets. This can be avoided if we prove the invariance of the density directly via equation (10.1). But to do that, you first need to know a tricky sum, see exercise 10.5.

With the invariant measures in hand, we can now turn to proving the ergodicity of the three maps starring in this chapter.

10.2. The Lebesgue Density Theorem

Proposition 10.4. *Given a measurable set $E \subseteq [0,1]$ with $\mu(E) > 0$, then for all $\varepsilon > 0$ there is an interval I such that*

$$\frac{\mu(E \cap I)}{\mu(I)} > 1 - \varepsilon\,.$$

We will say that the density of E in I is greater than $1 - \varepsilon$.

Proof. By Proposition 9.4, there are open sets O_n containing E such that $\mu(O_n \backslash E) = \delta_n$, where δ_n tends to 0 as n tends to infinity. Using property (4) of Corollary 9.6, we see that

$$\mu(O_n) = \mu(O_n \backslash E) + \mu(E) = \mu(E) + \delta_n\,. \tag{10.2}$$

According to exercise 9.4, for each n, there is a collection of disjoint open intervals $\{I_{n,i}\}$ such that

$$O_n = \cup_i I_{n,i}\,.$$

Now fix an $\varepsilon > 0$ and suppose that $\mu(E \cap I) \leq (1-\varepsilon)\mu(I)$ for all intervals. In particular this holds for those intervals belonging to the collection of intervals $\{I_{n,i}\}$. So for any n, we have

$$\mu(E \cap O_n) = \mu\left(E \cap (\cup_i I_{n,i})\right) = \sum_i \mu(E \cap I_{n,i}) \leq \sum_i (1-\varepsilon)\mu(I_{n,i})\,.$$

The middle equality follows again from property (4) of Corollary 9.6. Notice that the left-hand side equals $\mu(E)$, since O_n contains E, and the right-hand side equals $(1-\varepsilon)\mu(O_n)$ by definition of the intervals $I_{n,i}$. Together with equation (10.2), this gives

$$\mu(E) \leq (1-\varepsilon)\mu(O_n) = (1-\varepsilon)(\mu(E) + \delta_n)\,.$$

If n tends to infinity, δ_n tends to 0, and thus $\mu(E)$ must be 0. ∎

This is a weak version of a much better theorem. We do not actually need the stronger version, but its statement is so much nicer, it is probably best to remember *it* and not the proposition. A proof can be found in [**62**].

Theorem 10.5 (Lebesgue Density Theorem). *If E is a measurable set in $\mathbb{R}^n$ with $\mu(E) > 0$, then for almost all $x \in E$*
$$\lim_{\varepsilon\to 0} \frac{\mu(E \cap B_\varepsilon(x))}{\mu(B_\varepsilon(x))} = 1\,,$$
where $B_\varepsilon(x) := \{y \in \mathbb{R}^n \,:\, |y-x| < \varepsilon\}$, the open ε ball centered at x. That is: this holds for all x in E, except possibly for a set of μ measure 0.

10.3. Rotations and Multiplications on $\mathbb{R}/\mathbb{Z}$

In this section, we will invoke the Lebesgue density theorem, to prove the ergodicity of multiplications by $\tau \in \{\pm 2, \pm 3, \cdots\}$ modulo 1 and translations by an irrational number ω modulo 1 on $\mathbb{R}/\mathbb{Z}$. In each case, however, Proposition 10.4 is sufficient. We denote the Lebesgue measure by μ.

Lemma 10.6. *Every orbit of an irrational rotation R_r is dense in $\mathbb{R}/\mathbb{Z}$.*

Proof. We want to show that for all x and y, the interval $[y-\delta, y+\delta]$ contains a point of the orbit starting at x. Denote by $\frac{p_n}{q_n}$ the continued fraction convergents of r (of Definition 6.4). By Lemma 6.13
$$\lim_{n\to\infty} x + q_n r - p_n = \lim_{n\to\infty} x + d_n = x\,.$$
Fix n large enough enough, so that the distance (on the circle) between x and $x + q_n r - p_n$ is less than δ. Then the points $x_i := x + i q_n r$ modulo 1 advance (or recede) by less than δ. And thus at least one must land in the stipulated interval (see Figure 57). ∎

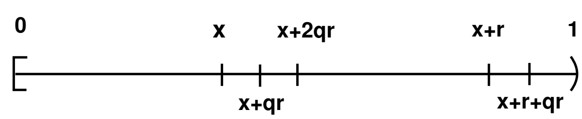


Figure 57. r is irrational and $\frac{p}{q}$ is a convergent of r. Then $x+qr$ modulo 1 is close to x. Thus adding qr modulo 1 amounts to a translation by a small distance.

Theorem 10.7. *Irrational rotations modulo 1 are ergodic with respect to the Lebesgue measure.*

Proof. By Proposition 10.3, the Lebesgue measure is invariant.

Suppose the conclusion of the theorem is false. Then there is an invariant set A such that both it and its complement A^c — which is also invariant — have strictly positive measure. By Proposition 10.4, for every ε there are intervals I and J where A, respectively A^c, have density greater than $1-\varepsilon$. Suppose that the length $\ell(I)$ of I is less than $\ell(J)$. Then there is an $n \geq 1$ so that

$$n\ell(I) \leq \ell(J) < (n+1)\ell(I).$$

By Lemma 10.6, there is i such that $R_\omega^{-i}(I)$ falls in the first $\frac{1}{n}$-fraction of J, another one in the second, and so forth (see Figure 58). In all cases, this

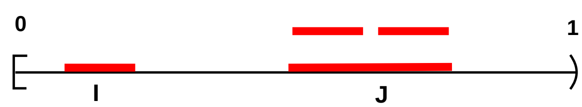


Figure 58. $\ell(I)$ is between $\frac{1}{3}$ and $\frac{1}{2}$ of $\ell(J)$. So there are two disjoint images of I under R_ω^{-1} that fall in J.

means that at least half of J is covered by images of I. By invariance, the images of I have A density greater than $1-\varepsilon$. That means that A has density at least $\frac{1}{2}(1-\varepsilon)$ in J, which is a contradiction. The case where $\ell(I) = \ell(J)$ is easy: split I into 2 equal intervals; at least one of these must have density greater than $1-\varepsilon$. ■

In the proof of the next theorem, we employ the same strategy as in the proof of Proposition 9.16 and Corollary 9.17. But this time, the Lebesgue density theorem helps us to get a much stronger result. We wish to prove that angle multiplications are ergodic. But it turns out that with the same effort we can prove the result for a larger class of maps. An example of such a map is given in Figure 59.

Definition 10.8. *Let $\{I_i\}$ be a finite or countable partition of $[0,1]$ of intervals of positive length ℓ_i so that $\sum_i \ell_i = 1$. On each interval I_i, define $f_i : I_i \to [0,1]$ to be an affine map onto $[0,1]$. $T = \cup_i f_i$ is called a complete affine interval map .*

To make the exposition a little clearer, let us first define nth level intervals. These are the domains I_0 where $T^n : I_0 \to [0,1)$ is a bijection. For

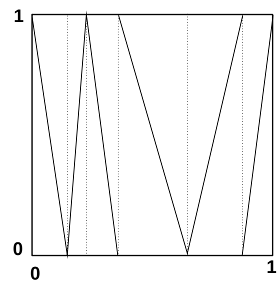


Figure 59. An example of the transformation $T : [0,1) \to [0,1)$ described in Theorem 10.9.

example, the 1st level intervals of the Gauss map are $(1/(a+1), 1/a]$ for $a \in \mathbb{N}$; the third level intervals for the angle doubling map are $[k/8, (k+1)/8)$ for k in $\{0, \cdots 7\}$.

Theorem 10.9. *Complete, affine interval maps preserve the Lebesgue measure and are ergodic with respect to that measure.*

Proof. By hypothesis we have $|f_i'| = 1/\ell_i$ and so $\sum_i |f_i'|_i^{-1} = \sum_i \ell_i = 1$ and so the Perron-Frobenius equation (10.1) immediately implies that the Lebesgue measure is preserved.

Suppose that the set A is T-invariant and has positive μ measure and let J be an arbitrary interval. We will prove that

$$\mu(A \cap J) = \mu(a) \cdot \mu(J) \qquad \text{or} \qquad \mu(A^c \cap J) = (1 - \mu(A))\mu(J)\,. \tag{10.3}$$

This implies that A^c has measure 0, otherwise a contradiction with the Lebesgue density would result. And so $\mu(A) = 1$, and thus T is ergodic.

Let $b : I_0 \to [0,1)$ be an arbitrary branch of T^n. Since b is affine $b^{-1}(A)$ occupies the same proportion in I_0 as A does in $[0,1)$. See Figure 60 All of the $|f_i'|$ are bounded away from 1, say, larger than $d > 1$. It follows that the length of any nth level interval is at most d^{-n}. Thus we can approximate J arbitrarily well with a collection of nth level intervals, provided n is large enough. Since in each nth level interval, the inverse image of A occupies a fraction of $\mu(A)$ of that interval, we have

$$\lim_{n\to\infty} \mu(T^{-n}(A) \cap J) = \mu(A) \cdot \mu(J)\,. \tag{10.4}$$

But since A is invariant, this proves (10.3), and thus the theorem. ∎

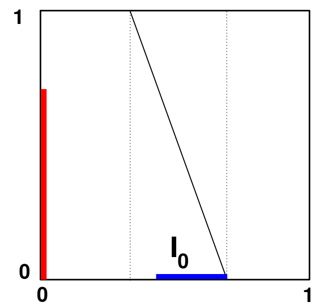


Figure 60. The pre-image in the nth level interval of A (in red) is the blue interval. This branch of T^n is affine and so there is no distortion. As a consequence, the proportions that A and its pre-image occupy are the same.

For completeness, we record the result in the important special case of angle multiplications.

Corollary 10.10. *Multiplication by $\tau \in \mathbb{Z}$ with $|\tau| > 1$ modulo 1 is ergodic.*

The results in this section have interesting consequences. One of the most important ones is the following.

Corollary 10.11. *Suppose $T : [0,1) \to [0,1)$ is ergodic with respect to Lebesgue measure, then $T^i(x)$ is equidistributed (see Definition 9.22) modulo 1 for almost every x.*

Proof. Define f as $f(x) = 1$ if x is in the interval $[a,b]$ and 0 elsewhere. By Corollary 9.13 to f, for almost all x

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x)) = \int_X f(x)\,dx = b-a\,.$$

■

In particular, this applies to both rotations by irrational angles and angle multiplications, as well as the maps defined in Theorem 10.9. In the case where T is an irrational rotation, the sequences $\{f(T^i(x))\}_{i=1}^{\infty}$ and $\{f(T^i(x'))\}_{i=1}^{\infty}$ differ only by a translation on the circle. So if one is equidistributed, then the other must be too. So in this case, the result holds for *all* x.

The principal consequence of the ergodicity of angle multiplication is the *absolute normality* (see Definition 9.21) of almost all numbers. This was discussed at length in the exercises of Chapter 9.

10.4. The Return of the Gauss Map

Our next aim is to show that the Gauss map T of Definition 6.1 is ergodic. Thanks to Proposition 10.3, we know the invariant measure. It might seem that Theorem 10.9 proves the rest. It almost does! The only problem is that the that the branches of the Gauss map are not affine. Here is what the problem with that is.

We suppose again that $A \subseteq [0,1)$ is an invariant set of positive μ measure. Just as before, we are given an arbitrary interval J and we want to prove that the density of A^c (A's complement) cannot be very close to 1. Then, by the Lebesgue density theorem, we can conclude that $\mu(A^c) = 0$. This proves that T is ergodic.

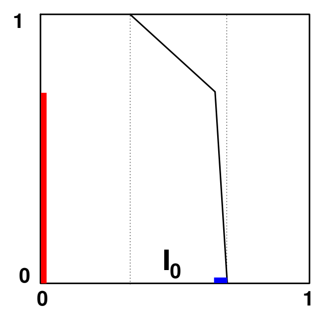


Figure 61. The pre-image in the nth level interval of A (in red) is the blue interval. This branch of T^n is far from affine and so the distortion is large. As a consequence, the proportions that A and its pre-image occupy are very different.

We need to make sure of two things. The first is that the lengths of the nth level intervals tend to zero, so that we can cover J efficiently with them. This is relatively easy. The second task is more subtle. The inverse image of A in any nth level interval have density bounded away from 0 to prevent its complement A^c from having density close to 1 in I_0. What is the problem? That branch is not affine. Especially for large n, it might have a much bigger derivative in $A \cap I_0$ than it does in $A^c \cap I$ (see Figure

61). This would distort the image under T^n in such a way that $\frac{\mu(A\cap I_0)}{\mu(I_0)}$ is much smaller than $\mu(A)$, and could even tend to zero as n grows. The solution to this dilemma lies in controlling that distortion. If we can prove that for that particular branch $\left|\frac{\partial T^n(x_0)}{\partial T^n(y_0)}\right|$ for x_0 and y_0 in I_0, is contained in, say, the interval $[1/K, K]$, $K > 0$, independent of n, then the argument of the proof of Theorem 10.9 gives that the densities of A on $[0,1)$ and of A in I_0 cannot differ too much. This ensures that $\frac{\mu(A\cap I_0)}{\mu(I_0)}$ is bounded away from 0 independently of n (the level).

The exposition in the remainder of this section and the next closely follows [**73**].

Definition 10.12. *Let I_0 be an interval. The* distortion *D of T^n on that interval is defined as*

$$D := \sup_{x_0, y_0 \in I_0} \left| \ln \left| \frac{\partial T^n(x_0)}{\partial T^n(y_0)} \right| \right| .$$

Here, ∂ stands for the derivative with respect to x.

Proposition 10.13. *Let T be the Gauss map. The distortion of T^n on any nth level interval I_0 is uniformly bounded in n.*

Proof. Denote the forward images of I_0 by I_1, I_2, et cetera. Similarly for x_0 and y_0. Set $I_n = [0,1]$. The chain rule gives

$$\partial T^n(x_0) = \partial T(x_0) \cdot \partial T(x_1) \cdots \partial T(x_{n-1}) .$$

Substitute this into the definition of the distortion to get

$$D \leq \sum_{i=0}^{n-1} \sup_{x_i, y_i \in I_i} \left| \ln |\partial T(x_i)| - \ln |\partial T(y_i)| \right| .$$

(This is an inequality due to the fact that we take the supremum over all x_i and y_i instead of just the initial points x_0 and y_0 in the definition.) By the mean value theorem, there is $z_i \in I_i$ such that the right-hand side of this expression equals

$$\sum_{i=0}^{n-1} |\partial \ln |\partial T(z_i)|| \cdot |y_i - x_i| \leq \sum_{i=0}^{n-1} \sup_{z_i \in I_i} |\partial \ln |\partial T(z_i)|| \cdot |I_i| .$$

Now we note that $\partial \ln |\partial T|$ equals $\left|\frac{\partial^2 T}{\partial T}\right|$. Furthermore, the mean value theorem (once again) gives $|I_i| = \frac{|I_{i+1}|}{|\partial T(u_i)|}$ for some $u_i \in I_i$. Substituting this into the last equation, we get

$$D \le \sum_{i=0}^{n-1} \sup_{z_i,u_i \in I_i} \left| \frac{\partial^2 T(z_i)}{\partial T(z_i)\, \partial T(u_i)} \right| \cdot |I_{i+1}|\,. \tag{10.5}$$

We need to estimate the expression in the right-hand side. Recall that we are analyzing a single branch of T^n. That implies that each interval I_i lies in one of the basic — or first level — intervals $(\frac{1}{a_i+1}, \frac{1}{a_i}]$ depicted in figure 25, where a_i is the continued fraction coefficient associated with that particular branch. Since $\partial T = -x^{-2}$ and $\partial^2 T = \frac{1}{2}x^{-3}$, for that particular branch, we have for $x \in (\frac{1}{a_i+1}, \frac{1}{a_i}]$

$$\left| \frac{\partial^2 T(z_i)}{\partial T(z_i)} \right| \le 2(a_i+1) \qquad \text{and} \qquad \left| \frac{1}{\partial T(u_i)} \right| \le \frac{1}{a_i^2}\,.$$

Next we estimate the length on nth level interval $|I|$. In figure 25, one can see that the only place where $|\partial T(x)|$ is small is when x is close to 1. These points are then mapped by T to a neighborhood of zero where they pick up a large derivative. It follows that the derivative of T^2 is positive and bounded by some $d > 1$ and thus the length of the intervals I_{n-i} decays as $Kd^{-i/2}$.

Putting this together, we see that (10.5) gives

$$D \le \sum_{i=0}^{n-1} \frac{2(a_i+1)}{a_i^2}\, Kd^{(i+1-n)/2}\,.$$

Since $a_i \in \mathbb{N}$, this tells us that the expression in (10.5) is uniformly bounded in n. ■

As explained in the introduction to this section, our main result follows immediately.

Corollary 10.14. *The Gauss map is ergodic with respect to* $d\nu = \frac{1}{\ln 2}\frac{dx}{1+x}$.

10.5. Number Theoretic Implications

Finally, it is pay-back time! We have seen some rewards for our efforts to understand ergodic theory in terms of understanding *normality* in the exercises of Chapter 9 (Definition 9.21). But the biggest pay-off is in understanding some basic properties of the continued fraction expansion of "typical" real numbers. That is what we do in this section.

In this section, T denotes the the Gauss transformation and ν its invariant measure (see Proposition 10.3) while μ will denote the Lebesgue measure. Note that a set has μ measure zero if and only if it has ν measure zero (exercise 10.1). For the continued fraction coefficients a_n and the continued fraction convergents $\frac{p_n}{q_n}$, see Definition 6.4.

We start with a remarkable result that says that the arithmetic (usual) mean of the continued fraction coefficients diverges (item (i)) for almost all numbers, but their geometric mean is almost always converges (item (ii)).

Theorem 10.15. *For almost all numbers x, the continued fraction coefficients $a_n = a_n(x)$ satisfy:*
i) $\lim_{n\to\infty}\left(\dfrac{a_1+\ldots+a_n}{n}\right)=\infty$ *and*
ii) $\lim_{n\to\infty}(a_1\cdot\ldots\cdot a_n)^{1/n}=\prod_{a=1}^{\infty}\left(1-\frac{1}{(a+1)^2}\right)^{-\log_2 a}<\infty.$
This last constant is approximately equal to $2.86542\cdots$ *is called Khinchin's constant.*

Proof. i) Define $f_k:[0,1]\to\mathbb{N}$ by

$$\text{For } a\in\{1,\cdots k\}:\quad f_k(x)=a \text{ if } x\in\left(\tfrac{1}{a+1},\tfrac{1}{a}\right]$$
$$f_k(x)=0\quad\text{elsewhere}\ .$$

Denote the pointwise limit by f_∞. We really want to use Corollary 9.13 to show that the "time average"

$$\lim_{n\to\infty}\left(\frac{a_1+\ldots+a_n}{n}\right)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}f_\infty(T^i(x))$$

is unbounded. But f_∞ is not integrable and so cannot be used. However the sum on the left hand side is bounded from below by the right-hand side if we replace f_∞ by f_k (which is integrable). Proposition 10.3 and Corollary 9.13 say that the time average of f_k equals

$$\frac{1}{\ln 2}\int_0^1\frac{f_k(x)}{1+x}dx=\frac{1}{\ln 2}\sum_{a=1}^{k}\int_{\frac{1}{a+1}}^{\frac{1}{a}}\frac{a}{1+x}dx \tag{10.6}$$

The integral of $1/(1+x)$ is of course $\ln(1+x)$ and so the above gives

$$\frac{1}{\ln 2}\sum_{a=1}^{k}a\left(\ln\left(\frac{a+1}{a}\right)-\ln\left(\frac{a+2}{a+1}\right)\right)\ . \tag{10.7}$$

This sum telescopes and the student should verify (see exercise 10.12) that this gives

$$\frac{1}{\ln 2}\left(\ln(k+1)-k\ln\left(1+\frac{1}{k+1}\right)\right), \tag{10.8}$$

which diverges as $k\to\infty$ and proves the first statement.

ii) This proof is very similar to that of (i), except that now we want to compute the "time average"

$$\lim_{n\to\infty}\left(\frac{\ln a_1+\ldots+\ln a_n}{n}\right).$$

The exponential of this will give us the result we need. So this time, we define

$$\text{For } a\in\mathbb{N}: \quad g_\infty(x)=\ln a \text{ if } x\in\left(\frac{1}{a+1},\frac{1}{a}\right]. \tag{10.9}$$

This time around, g_∞ is ν-integrable (as we will see below) and we get

$$\frac{1}{\ln 2}\int_0^1\frac{g_\infty(x)}{1+x}dx=\sum_{a=1}^{\infty}\frac{\ln a}{\ln 2}\left(\ln\left(\frac{a+1}{a}\right)-\ln\left(\frac{a+2}{a+1}\right)\right). \tag{10.10}$$

(Note that $\frac{\ln a}{\ln 2}=\log_2 a$.) Since we can write

$$\ln\left(\frac{a+1}{a}\right)-\ln\left(\frac{a+2}{a+1}\right)=-\ln\left(1-\frac{1}{(a+1)^2}\right), \tag{10.11}$$

we finally get the result (as well as the assertion that g_∞ is ν-integrable) by taking the exponential of the sum in (10.10). See exercise 10.13. ■

An example of a sequence $\{a_n\}_{n=1}^\infty$ that has a diverging running average but whose running geometric average converges, is given by $a_n=1$, except when $n=2^{2k}$ we set $a_{2^{2k}}=2^{2^k}$. For $n=2^{2k}$, we have

$$\frac{a_1+\ldots+a_n}{n}>\frac{a_n}{n}=2^{2^k-2k},$$

which clearly diverges as $k\to\infty$. Meanwhile, the geometric average at that point is (after taking the logarithm and noting that $ln1=0$):

$$\frac{\ln a_1+\ldots+\ln a_n}{n}=\frac{\sum_{j=1}^{n}\ln 2^{2^k}}{2^{2k}}=\frac{2^k\ln 2}{2^{2k}}.$$

The latter converges to 0, which makes the geometric average 1.

Theorem 10.16. *For almost all numbers x, the convergents* $p_n(x)/q_n(x)$ *satisfy*
i) $\lim_{n\to\infty} \frac{\ln q_n}{n} = \frac{\pi^2}{12\ \ln 2}$, *and*
ii) $\lim_{n\to\infty} \frac{1}{n} \ln \left| x - \frac{p_n}{q_n} \right| = -\frac{\pi^2}{6\ \ln 2}$.

Remark 10.17. The constant $\frac{\pi^2}{12\ \ln 2} \approx 1.1866\cdots$ is called Lévy's constant.

Proof. Item (ii) follows very easily from (i), see exercise 10.17. So here we will prove only (i).

To simplify notation in this proof, we will write $x_i := T^i(x_0)$ where T is the Gauss map. For the nth approximant of $x_0 \in (0,1)$, see Definition 6.4, we will write $\frac{p_n(x)}{q_n(x)}$. From that same definition, we conclude

$$\frac{p_n(x_0)}{q_n(x_0)} = \frac{1}{a_1(x_0) + p_{n-1}(x_1)/q_{n-1}(x_1)} = \frac{q_{n-1}(x_1)}{a_1(x_0)q_{n-1}(x_1) + p_{n-1}(x_1)}\,.$$

See also exercise 10.2 (a). By Corollary 6.8 (ii)), $\gcd(p_n, q_n) = 1$, and so using exercise 10.2 (b), we see that $p_n(x_0)$ equals $q_{n-1}(x_1)$. More generally, we have by the same reasoning

$$p_n(x_j) = q_{n-1}(x_{j+1})\,. \tag{10.12}$$

This implies that

$$\frac{p_n(x_0)}{q_n(x_0)} \cdot \frac{p_{n-1}(x_1)}{q_{n-1}(x_1)} \cdot \frac{p_{n-2}(x_2)}{q_{n-2}(x_2)} \cdots \frac{p_1(x_{n-1})}{q_1(x_{n-1})} = \frac{1}{q_n(x_0)}\,,$$

since $p_1 = 1$ by Theorem 6.6. Now we take the logarithm of the last equation. This yields

$$-\frac{1}{n}\ln q_n(x_0) = \frac{1}{n}\sum_{i=0}^{n-1} \ln x_i - \frac{1}{n}\sum_{i=0}^{n-1}\left(\ln x_i - \ln \frac{p_{n-i}(x_i)}{q_{n-i}(x_i)}\right)\,. \tag{10.13}$$

Two more steps are required. The first is showing that the last average on the right side of (10.13) tends to zero. This not difficult, because

$$\sum_{i=0}^{n-1} \ln \frac{q_{n-i}(x_i)x_i}{p_{n-i}(x_i)} = \sum_{i=0}^{n-1} \ln\left(1 + \frac{q_{n-i}(x_i)x_i - p_{n-i}(x_i)}{p_{n-i}(x_i)}\right)\,.$$

Corollary 6.7 or, more precisely, exercise 6.12 yields that

$$\frac{|q_{n-i}(x_i)x_i - p_{n-i}(x_i)|}{p_{n-i}(x_i)} < \frac{1}{p_{n-i}(x_i)q_{n-i+1}(x_i)} < 2^{-(n-i)}\sqrt{2}\,,$$

where the last inequality follows from Corollary 6.8 (i). The fact that for small x, $\ln(1+x) \approx x$ concludes the first step (see also exercise 10.10).

Since the second term on the right side of (10.13) tends to zero and $x_i = T^i(x_0)$, we take a limit to get

$$\lim_{n\to\infty} -\frac{1}{n}\ln q_n(x_0) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \ln T^i(x_0)\,.$$

The second step is then to compute the right-hand side of this expression. Naturally, the ergodicity of the Gauss map invites us to employ Birkhoff's theorem in the guise of Corollary 9.13 with $f(x)$ set equal to $\ln(x)$.

$$\frac{1}{n}\sum_{i=0}^{n-1} \ln T^i(x_0) = \int_0^1 \frac{\ln x}{(1+x)\ln 2}\,dx\,.$$

The integral is evaluated in exercise 10.16. ∎

10.6. Exercises

Exercise 10.1. a) Show that for a measurable set A: $\mu(A) = 0$ (Lebesgue measure) if and only if $\nu(A) = 0$ (invariant measure of the Gauss map). (*Hint: using Lebesgue integrals, write* $\nu(A) = \frac{1}{\ln 2}\int_A (1+x)^{-1}\,dx$.)
b) Let ν be absolutely continuous with respect to the Lebesgue measure μ. Show that if a set has full μ measure then it has full ν measure. (*Hint: see comments after Definition 9.14.*)

Exercise 10.2. To reacquaint ourselves with continued fractions, consider

$$x_0 = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}} \stackrel{\text{def}}{\equiv} [a_1, a_2, a_3, \cdots]\ . \tag{10.14}$$

a) Show that $\lfloor x_0^{-1} \rfloor = a_1$ and that

$$T(x_0) = x_0^{-1} - a_1 = \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4 + \cdots}}} \stackrel{\text{def}}{\equiv} [a_2, a_3, \cdots]\ .$$

b) Show that if $\gcd(p,q) = 1$, then $\gcd(p+aq, q) = 1$. (*Hint: use Lemma 2.5.*)
c) Prove equation (10.12).

Exercise 10.3. a) Show that every probability density ρ on $\mathbb{R}/\mathbb{Z}$ gives rise to an invariant measure under the identity.
b) What are the absolutely continuous measures — i.e. with a density, see Section 9.5 — that are invariant under rotation by $1/2$? (*Hint: consider densities with period 1/2.*)
c) The same for rotation by p/q for p and q in $\mathbb{N}$.
d) Show that the *uniform* density — with density $\rho(x) = 1$ — is invariant under $x \to nx$ modulo 1 (where $n \in \mathbb{N}$).

Exercise 10.4. Let R_0 be identity on $\mathbb{R}/\mathbb{Z}$.
a) Show that for any x, the delta measure δ_x is an invariant measure for R_0, and that R_0 is ergodic with respect to that measure.
b) Show that for any of the invariant measures in exercise 10.3 (a), R_0 is not ergodic.
c) Show that R_0 is not ergodic with respect to any of the measures of exercise 10.3 (c).

Exercise 10.5. a) Show that $\sum_{i=1}^{\infty} \frac{1}{(y+i)(y+i+1)} = \frac{1}{y+1}$ for all $y \in \mathbb{R}$ except the negative integers. (*Hint: use partial fractions, then note that the resulting sum telescopes.*)
b) Show that the inverse images of $y \in [0,1)$ under the Gauss map T are $\cup_{i \in \mathbb{N}} \{1/(y+i)$.
c) Let $\rho(y) = c/(1+y)$ a density and compute ρ and $|T'|$ at these inverse images.
d) Use (a), (b), and (c) to prove directly via equation (10.1) that the Gauss map preserves the measure of Proposition 10.3. Do not use the computation in the proof of that proposition.

Exercise 10.6. Show that $\rho(x) = 1$ is the only continuous invariant probability density of an irrational rotation R. (*Hint: if ρ is invariant under R, it must be invariant under R^i for all positive i. Use equation* (10.1) *and Lemma 10.6.*)

Exercise 10.7. a) Show that $\rho(x) = 1$ is the only continuous invariant density for the angle doubling map. (*Hint: as in exercise 10.6: use that inverse images of a point are dense, see Proposition 9.16.*)
b) Check that the same is true for the map $x \to \tau x$ modulo 1 where $\tau \in \mathbb{Z}$ and $|\tau| > 1$.

Exercise 10.8. The orbit of any irrational rotation is uniformly distributed. So why do we encounter specifically the golden mean in phyllotaxis — the placement of leaves? Research this and try to include illustrations. (*Hint: the wikipedia page on "golden ratio" is a good start. You can find even more information in* [**43**]. *The slightly tongue-in-cheek paper* [**33**] *shows that the golden mean does indeed show up in the most unexpected places. Words of caution on the 'cult of the golden ratio' can be found in* [**27**].)

Exercises 10.9 and 10.10 discuss some very useful properties of the logarithm for later reference. In fact, they are useful in a much wider context than discussed here. For instance, exercise 10.9 comes up in any discussion of entropy [**22**] or in deciding the stability of Lotka-Volterra dynamical systems [**67**]. Exercise 10.10 is important for deciding the convergence of products of the form $\prod(1+x_i)$.

Exercise 10.9. a) Show that

$$x > -1 \quad \Longrightarrow \quad \ln(1+x) \leq x\,.$$

with equality iff $x=0$. (*Hint: draw the graphs of* $\ln(1+x)$ *and* x.)
b) Let p_i and q_i positive and $\sum_i p_i = \sum_i q_i < \infty$. Use (a) to show that

$$-\sum_i p_i \ln p_i \leq -\sum_i p_i \ln q_i\,.$$

(*Hint:* $-\sum_i p_i(\ln p_i - \ln q_i) = \sum_i p_i \ln \frac{q_i}{p_i} \leq \sum_i (q_i - p_i)$ *by (a).*)
c) Let S_n be the open n-dimensional simplex $p_i > 0$ and $\sum_{i=1}^n p_i = 1$. Use (b) to show that

$$-\sum_i p_i \ln p_i \leq \ln n\,.$$

(*Hint: set* $q_i = 1/n$ *in (b).*)
d) Show $h : S_n \to \mathbb{R}$ given by $h(p) = -\sum_i p_i \ln p_i$ has a unique global maximum at $p_i = \frac{1}{n}$. (*Hint: The constraint is* $C(p) := \sum_i p_i = 1$. *Deduce that at the maximum, the gradients of* h *and* C *with respect to* $p = (p_1, \cdots, p_n)$ *must be parallel. If in doubt, look up Lagrange multiplier in, for example,* [**53**].)

In the next exercise, we prove this lemma.

Lemma 10.18. *Suppose that* $x_n > -1$ *and* $\lim_{n\to\infty} x_n = 0$. *Then* $\sum_n \ln(1+x_n)$ *converges absolutely if and only if* $\sum_n x_n$ *converges absolutely. Also* $\sum_n \ln(1+x_n)$ *diverges absolutely if and only if* $\sum_n x_n$ *diverges absolutely.*

Exercise 10.10. a) Show that $\lim_{x\to 0} \frac{\ln(1+x)-x}{x^2} = -\frac{1}{2}$. (*Hint: use L'Hôpital twice.*)
b) From (a), conclude that if $x_n > -1$ and $\lim_{n\to 0} x_n = 0$, then $\exists b_1 > 0$ such that for all n large enough $|\ln(1+x_n)| \leq b_1 |x_n|$. (*Hint: use the direct comparison test.*)
c) From (a), conclude that if $x_n > -1$ and $\lim_{n\to 0} x_n = 0$, then $\exists b_2 > 0$ such that for all n large enough $|x_n| \leq b_2 |\ln(1+x_n)|$.
d) Show that (b) and (c) imply Lemma 10.18.

The next four exercises provide some computational details of the proof of Theorem 10.15.

Exercise 10.11. Compute the frequency with which the coefficient $a_n(x) = a$ occurs in the continued fraction expansion of almost all x. (*Hint: set $f(x) = 1$ on $(1/(1+a), 1/a]$. Then use Birkhoff.*)

Exercise 10.12. a) Show that the right-hand side of (10.6) gives (10.7).
b) Show that (10.7) gives (10.8). (*Hint: write out the first few terms explicitly.*)
c) Use exercise 10.9 (a) to bound the second term of (10.8).
d) Conclude that (10.8) is unbounded. (*Hint: $(1+1/k)^k \to e$.*)

Exercise 10.13. a) Show the equality in (10.10) holds.
b) Show the equality in (10.11) holds.
c) Show that (10.11) implies part (ii) of Theorem 10.15.

Exercise 10.14. a) Show that instead of (10.11), we also have
$$\ln\left(\frac{a+1}{a}\right) - \ln\left(\frac{a+2}{a+1}\right) = \ln\left(1 + \frac{1}{a^2+2a}\right).$$
b) Use exercise 10.9 (a) to show that
$$\ln\left(1 + \frac{1}{a^2+2a}\right) \leq \frac{1}{a^2}.$$
c) Use (a) and (b) and equation (10.10) to show that
$$\frac{1}{\ln 2}\int_0^1 \frac{g_\infty(x)}{1+x}\,dx \leq \frac{1}{\ln 2}\sum_{a=1}^{\infty} \frac{\ln a}{a^2}.$$
(*Hint: indeed, this is equivalent to the fact that g_∞ is integrable. Can you explain that?*)
d) Show that (c) implies that Khinchin's constant is bounded. (*Hint: find the maximum of $\ln a - 2\sqrt{a}$. Then use Figure 11.*)

Exercise 10.15. Use exercise 10.14 (a) to show that Khinchin's constant equals $\prod_{a=1}^{\infty}\left(1 + \frac{1}{(a^2+2a)^2}\right)^{\log_2 a}$.

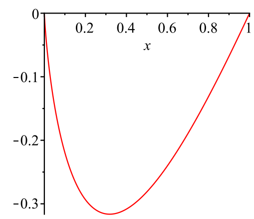


Figure 62. Plot of the function $\ln(x)\ln(1+x)$

.

Exercise 10.16. a) Show that $\lim_{x\to 0} \ln(x)\ln(1+x) = 0$ (Figure 62). (*Hint: for the limit as $x \to 0$, first use exercise 10.10 (b) and then substitute $x = e^y$.*)
b) Use (a) to show that $I := \int_0^1 \frac{\ln x}{(1+x)}\,dx = -\int_0^1 \frac{\ln(1+x)}{x}\,dx$. (*Hint: integration by parts.*)
c) Show that for $|x| < 1$, $\ln(1+x) = \sum_{i=1}^{\infty} \frac{(-1)^{n+1}x^n}{n}$.
d) Substitute (c) into I and integrate term by term to get $I = \sum_{n=1}^{\infty} (-1)^n n^{-2}$.
e) The sum in (d) equals $-\frac{\pi^2}{12}$. Show that that gives the result advertised in Theorem 10.16. (*Observation: we sure took the cowardly way out in this last step; to really work out that last sum from first principles is elementary but very laborious. The interested student should look this up on the web.*)

In exercise 10.16, note the curious fact that $\sum_{n=1}^{\infty} (-1)^n n^{-2} = -\frac{\pi^2}{12}$ while from exercise 2.26 we have that $\zeta(2) = \sum_{n=1}^{\infty} n^{-2} = \frac{\pi^2}{6}$.

Exercise 10.17. a) Show that
$$\lim_{n\to\infty} \frac{\ln q_{n+1}}{n} = \lim_{n\to\infty} \frac{\ln q_{n+1}}{n+1}\,\frac{n+1}{n} = \lim_{n\to\infty} \frac{\ln q_{n+1}}{n+1} = \lim_{n\to\infty} \frac{\ln q_n}{n}\,.$$
b) Use (a), exercise 6.12, and Theorem 10.16 (i) to show that for almost all $x \in [0,1]$
$$\lim_{n\to\infty} \frac{1}{n} \ln\left|x - \frac{p_n}{q_n}\right| = -\frac{\pi^2}{6\ln 2}\,.$$
c) What do you in (b) get if x is rational? Is that a problem?

Exercise 10.18. a) Use Corollary 6.7 to show that

$$\ln \left| x - \frac{p_n}{q_n} \right| < \ln q_n^{-2}$$

b) Use Theorem 10.16 to show that

$$\lim_{n\to\infty} \frac{1}{n} \ln \left| x - \frac{p_n}{q_n} \right| = \lim_{n\to\infty} \frac{1}{n} \ln q_n^{-2}.$$

Exercise 10.19. a) What the equivalent of Theorem 10.16 for decimal expansions of irrational numbers? (*Note: I haven't worked this out myself yet.*)
b) Use (a) to compare decimal approximation with continued fraction approximation.

Definition 10.19. *Given a one dimensional smooth map* $T : [0,1] \to [0,1]$, *the* Lyapunov exponent $\lambda(x)$ *at a point* x *is given by*

$$\lambda(x) := \lim_{n\to\infty} \frac{1}{n} \ln |\partial T^n(x)|,$$

assuming that the limit exists. Here $\partial T^n(x)$ *stands for* $\frac{d}{dx}(T^n)\,|$ *evaluated at* x.

Remark 10.20. Although we do not need it here, we remark that the generalization to higher dimension of this notion is not completely straightforward [**6**].

Exercise 10.20. Suppose that $\lambda(x) = \lambda_0$, a constant. What does Definition 10.19 tell you about how fast $|T^n y - T^n x|$ if y is very close to x?

Exercise 10.21. a) Let T be the Gauss map and μ its invariant measure. Show that the Lyapunov exponent at x satisfies

$$\lambda(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \ln \left| \partial T(T^j(x)) \right|.$$

(*Hint: think chain rule.*)
b) Show that Birkhoff's theorem (Corollary 9.13) implies that for almost all $x \in [0,1]$

$$\lambda(x) = \int_0^1 \frac{-2\ln x}{\ln 2\,(1+x)}\, dx.$$

c) Use the last part of the proof of Theorem 10.16 and exercise 10.16 to show that for almost all x, the Lyapunov exponent equals $\frac{\pi^2}{6\ln 2} \approx 2.3731$.

Exercise 10.22. a) See exercise 10.21. Let T be the Gauss map and $x = [n, n, \cdots]$. Determine the Lyapunov exponent at x. (*Hint: see also exercise 6.2.*)
b) Why does it not contradict Birkhoff's theorem that these exponents different from the one computed in exercise 10.21?

Exercise 10.23. Let T be the map given in Theorem 10.9. We will assume that T has n branches.
a) Show that for almost all points x, the Lyapunov exponent is given by $\lambda(x) = -\sum_i \ell_i \ln \ell_i$. (*Hint: see also exercise 10.21.*)
b) Show that $\lambda(x) > 0$.
c) Show that $\lambda(x) \leq \ln n$ (*Hint: exercise 10.9 (c).*)
d) Show that $\lambda(x) = \ln n$ if and only if the slopes of all branches have the same absolute value. (*Hint: exercise 10.9 (d).*)

Exercise 10.24. a) Show that if $k \in \mathbb{N}$ is such that $\log_{10} k$ is rational, then $k = 10^r$, $r \in \mathbb{N}$. (*Hint: use prime factorization on $k^q = 10^p$.*)
b) From now on, suppose that $\log_{10} k$ is irrational. Show that $T : x \to x + \log_{10} k$ modulo 1 is ergodic with respect to the Lebesgue measure.
c) Let $f(x) = 1$ when $x \in [\log_{10} 7, \log_{10} 8]$ and 0 elsewhere. Compute $\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} f(T^i(x))$.
d) Explain how often, for almost every x, 1 through 9 occur in $\{k^i x\}_{i=0}^{\infty}$ as first digits.
e) How often does any combination of any 2 successive digits, say 36, occur as first digits?

Stock prices undergo multiplicative corrections, that is: each day their price is multiplied by a factor like 0.99 or 1.01. On the basis of the previous problem, it seems reasonable that the distribution of their first digits satisfies the logarithmic distribution of exercise 10.24. In fact, a much wider range of real world data satisfies this distribution than this "multiplicative" explanation would suggest. This phenomenon is called Benford's law and appears to be only partially understood [**13**].

Definition 10.21. *A map T that preserves the measure μ is called mixing if*

$$\lim_{n\to\infty} \mu(T^{-n}(A) \cap B) = \mu(A) \cdot \mu(B)$$

for any two measurable sets A and B.

We prove the following result in exercise 10.25

Proposition 10.22. *If T is mixing with an invariant probability measure, then it is ergodic.*

Exercise 10.25. a) Prove Proposition 10.22. (*Hint: let A be the invariant set and choose $B = A$ in Definition 10.21; get a quadratic equation for* $\mu(A)$)
b) Show that ergodic does *not* imply mixing. (*Hint: irrational rotation.*)
c) Show that the transformation T in Theorem 10.9 is mixing. (*Hint: a measurable set in* $[0,1)$ *can be approximated by finitely many open intervals, then see equation* (10.4).)

Part 3

Topics in Number Theory

Chapter 11

The Cauchy Integral Formula

Overview. Again, we need to venture very far, apparently, from number theory to make progress. In the mid 19th century, the main insight in number theory came from Riemann, who realized that the distribution of the primes was intimately connected to the properties of the (analytic continuation of the) Riemann zeta function to the complex plane. In this chapter, we develop the necessary complex analysis tools — essentially the Cauchy integral formula — to study the convergence of a certain improper integral (Theorem 11.18), which is the key to the proof of the prime number theorem in the next chapter (Theorem 12.15). For more detailed introductions to complex analysis, we refer to [**3**, **26**, **45**].

11.1. Analyticity versus Isolated Singularities

Definition 11.1. *A set is open if it contains no points lying on its boundary and connected if it is not the disjoint union of two non-empty open sets. A* <u>*domain*</u> *or* <u>*region*</u> *is an open, connected set in* $\mathbb{C}$.

An excellent source for information on topological notions such as connectedness is [**50**].

Definition 11.2. *Let A be a domain. A function $f : A \to \mathbb{C}$ is analytic at z_0 if*

$$f'(z) := \lim_{\delta \to 0} \frac{f(z_0 + \delta) - f(z_0)}{\delta}$$

exists in a neighborhood of z_0. The word holomorphic is synonymous with analytic. If f is analytic on all of $\mathbb{C}$, it is also called an entire function.

We will use the fact that this says that analyticity is an *open* condition.

Corollary 11.3. *If f is analytic at z_0, then it is analytic in an open neighborhood of z_0.*

This creates, as it were, a loophole which will be crucial in the proof of Theorem 11.18. Suppose we know f is analytic some closed set S. Then in fact, f must be analytic in some open set containing S. Because if not, there must be a sequence of points z_i converging to $z \in S$ where f is not analytic. Then at z, f would not be analytic! See Figure 63.

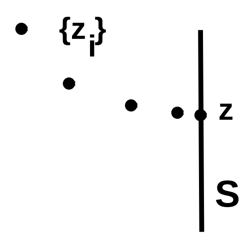


Figure 63. If f is analytic on the closed set S, then f must be analytic on some open set containing S.

Naturally, not all functions are analytic everywhere. What happens at or near a point z where f is not analytic? We say that f is singular at such a point z_0. If there is a neighborhood[1] in which it is the only singularity, we call it an isolated singularity. One can prove that every isolated singularity is one of the ones defined below.

Definition 11.4. *These are the types of isolated singularities at z_0.*
i) f may have a removable singularity. In this case, $f(z_0)$ can be given a value in such a way that f becomes analytic at z_0. An example is $\frac{\sin(z-z_0)}{z-z_0}$.

[1] An open neighborhood of z_0 minus the point z_0 itself is often called a punctured neighborhood of z_0.

ii) f may have a pole *of order $k \in \mathbb{N}$. An example of this is $a_k(z-z_0)^{-k} + a_{k-1}(z-z_0)^{-(k-1)}$ with $a_k \neq 0$. A pole of order 1 is called a* simple pole.
iii) f may have an essential singularity*. This is a pole of "infinite order". An example is $e^{1/(z-z_0)}$. (Expand as e^u and substitute $u = (z-z_0)^{-1}$.)*

One might be tempted to say that the example in item (ii) above consists of two singularities, one of order k and one of order $k-1$. However, we have

$$\frac{a_k}{(z-z_0)^k} + \frac{a_{k-1}}{(z-z_0)^{k-1}} = \frac{(a_k + a_{k-1})z - a_{k-1}z_0}{(z-z_0)^k} .$$

The numerator does not vanish at z_0, and so we have one singularity of order k. A pole of "infinite order" in item (iii) means that the expansion contains infinitely many non-zero terms $a_k(z-z_0)^{-k}$ with $k \in \mathbb{N}$.

Remark 11.5. A subtle — but sometimes important — point that is the observation that branch points like the origin for $z \to (z-z_0)^{1/2}$ or $z \to \ln(z-z_0)$ are *not* isolated singularities. The reason is that in any *punctured neighborhood* of the origin these "functions" are not one-valued. In other words, they are not functions, and therefore *a fortiori* they are not analytic functions. Even if you redefine the function in this neighborhood so that it describes a single branch of that function, then still there is a line of discontinuities (the branch cut) with the branch point as its endpoint.

For completeness, we mention the only other types of singularities: cluster points , these are limit points of other singularities; and natural boundaries , entire sets where singularities are dense. An example of the latter is the unit circle for the function $\sum_{n=1}^{\infty} z^{n!}$. Needless to say, these singularities are not isolated.

All singularities mentioned in this remark are non-isolated, and if z_0 is the locus of such a singularity, it is not possible to approximate its behavior in terms of integral powers of $(z-z_0)$.

Definition 11.6. *A function is* meromorphic *in a domain if it has only isolated poles in the domain. It is meromorphic if this holds on all of $\mathbb{C}$.*

We need a criterion for uniform convergence.

Lemma 11.7 (Weierstrass M test). *Let $A \subseteq \mathbb{C}$ and $g_n : A \to \mathbb{C}$ a sequence of functions. Suppose that $|g_n(z)| \leq m_n$ on A and that $\sum_n m_n$ converges (*uniform absolute convergence *). Then for all z in A:*

i) $\sum_{n=1}^{\infty} |g_n(z)|$ *converges (absolute convergence), and*
ii) For all $\varepsilon > 0$*, there is* n_0 *so that for all* $n > n_0$*:* $\left|\sum_{n+1}^{\infty} g_i(z)\right| < \varepsilon$ *(uniform convergence).*

Proof. Item (i) follows immediately from the hypotheses. Item (ii) follows from the fact that $\left|\sum_{n+1}^{\infty} g_i(z)\right| \leq \sum_{n+1}^{\infty} |g_i(z)| \leq \sum_{n+1}^{\infty} m_i$ and the convergence of $\sum_n m_n$ (so the partial sums of $\{m_n\}$ form a Cauchy sequence). ■

11.2. The Cauchy Integral Formula

First we set the scene with some notation. Let $[a,b]$ be an interval in $\mathbb{R}$ of positive length. A curve is a piecewise differentiable function $\gamma : [a,b] \to \mathbb{C}$. Its orientation is the direction of increasing $t \in [a,b]$. A simple, closed curve is a curve without self-intersections and whose endpoints are identical (or $\gamma(a) = \gamma(b)$. It follows that the complement of γ consists a well-defined "inside" component and a "outside" component (see Figure 64). A line integral evaluated along the curve γ is denoted by $\int_\gamma$. If the curve is simple and closed, one often writes $\oint_\gamma$ or simply $\oint$.

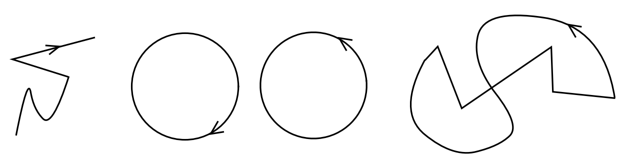


Figure 64. Left, a curve. Then two simple, closed curves with opposite orientation. The curve on the right is a union of two simple, closed curves.

Proposition 11.8 (**Cauchy's Theorem**)**.** *Let* γ *be a simple, closed curve and assume* f *is analytic on* γ *and in its interior with at most finitely removable singularities. Then we have* $\oint_\gamma f(z)\, dz = 0$

For students familiar with differential forms and Stokes' theorem, we give a very simple proof. Students unfamiliar with that material can skip the first paragraph of the proof. A full proof without assuming Stokes is more laborious and can be found in [**3**] and in [**45**]. A proof of Stokes' theorem can be found in [**58**] Chapter 5, Section 9.

Proof. By assumption γ bounds an 'inside' region D: $\gamma = \partial D$. First assume f is analytic in D (including boundary). As usual, we write $f = u + iv$ and $z = x + iy$ to relate the complex notation to calculus in $\mathbb{R}^n$.

$$\begin{aligned}\oint_{\partial D} f\,dz &= \int_{\partial D} u(dx + idy) + \int_{\partial D} iv(dx + idy) \\ & \int_{\partial D} udx - vdy + i\int_{\partial D} udy + vdx\,.\end{aligned}$$

By Stokes' theorem, for any differential form ω on a region D with a piecewise differentiable boundary as specified by the proposition, we have $\int_{\partial D} \omega = \int_D d\omega$, where d stands for the exterior derivative. Now

$$d(udx - vdy) = -(\partial_y u + \partial_x v)dxdy \quad \text{and} \quad d(udy + vdx) = (\partial_x u - \partial_y v)dxdy\,.$$

both of which are zero by the Cauchy-Riemann equations of Proposition 11.23 (exercises 11.11, 11.12, and 11.13). in the exercises. Hence,

$$\oint_{\partial D} f\,dz = \oint_D d(f\,dz) = 0$$

if f is analytic.

Since f has only finitely many singularities, they cannot accumulate. Now suppose that f has an isolated singular point z_0 at which it is, however, continuous. Let c be a circular path of small radius ε around z_0 so that the ε-disk around z_0 does not contain any other singular points or points of γ (see Figure 65). Let p be a path that connects γ to c. Now the curve Γ

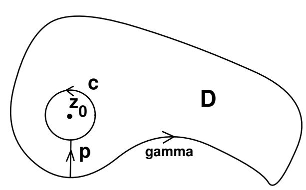


Figure 65. In the interior of the curve obtained by concatenating γ, p, c, and $-p$, f is analytic. Therefore $\oint_\gamma f\,dz - \oint_c f\,dz = 0$. If f is also bounded inside c, we also have $\oint_c f\,dz = 0$.

obtained by concatenating γ, p, c, and $-p$ is a simple closed curve and thus $\oint_\Gamma f = 0$. The integrals along p and $-p$ cancel one another. By continuity, $|f|$ is bounded by some M and so $|\oint_c f|$ is bounded by $2\pi\varepsilon M$. We can choose ε as small as we want, and so $|\oint_c f|$ must be 0. Therefore, $\oint_\gamma f = \oint_\Gamma f - \oint_c f = 0$. ■

It is also instructive to compare this with calculus on the real line. If $f : \mathbb{R} \to \mathbb{R}$ is piecewise differentiable and continuous, then from calculus, we know that

$$\int_a^b f\,dx = F(b) - F(a)\,.$$

This does not depend on the path we choose to get from a to b. Let $y_i : [0,1] \to [a,b]$ be different parametrizations of the segment $[a,b]$. Then

$$\oint_{\gamma_1-\gamma_2} f = \left(\int_{\gamma_1} - \int_{\gamma_1}\right) f = \int_0^1 f(y_1(t))y_1'(t)\,dt - \int_0^1 f(y_2(t))y_2'(t)\,dt = 0\,.$$

It is this statement that Cauchy's theorem generalizes.

Theorem 11.9 (**Cauchy's Integral Formula**). *Let γ be a simple, closed curve going around z once in counter-clockwise direction and suppose that f is analytic on and inside γ. Then*

$$f(z) = \frac{1}{2\pi i}\int_\gamma \frac{f(w)}{w-z}\,dw\,.$$

Proof. Define the function g

$$g(w) = \begin{cases} \frac{f(w)-f(z)}{w-z} & w \neq z \\ f'(z) & w = z \end{cases}$$

The function g is continuous and therefore analytic (also at z). So $\oint_\gamma g = 0$. By linearity,

$$\int_\gamma \frac{f(w)}{w-z}\,dw = \int_\gamma g(w)\,dw + f(z)\int_\gamma \frac{1}{w-z}\,dw\,.$$

The first integral in the right-hand side is zero by Cauchy's theorem (Proposition 11.8). Now let c be the curve $w = z + re^{it}$ with $t \in [0, 2\pi]$. The same construction as in the second part of the proof of Proposition 11.8 shows that $\oint_\gamma - \oint_c = 0$ (see Figure 65) and thus $\oint_\gamma = \oint_c$. So

$$\int_\gamma \frac{1}{w-z}\,dw = \int_0^{2\pi} \frac{ire^{it}}{re^{it}}\,dt = 2\pi i\,.$$

Substituting this into the earlier equation yields the statement. ■

Remark. The surprising aspect of this formula is that the value of an analytic function at z_0 is determined by the values of that function on a simple, closed curve that encircles z_0.

11.3. Corollaries of the Cauchy Integral Formula

Cauchy integral formula can be used to show the remarkable result that a function that is analytic at z_0 has derivatives of all orders at that point. These derivatives are denoted by $f^{(k)}(z_0)$. The simplest way of proving this is by actually calculating an expression for these derivatives.

Lemma 11.10. *Suppose that $w-z \neq 0$, then for $|d|$ small enough $w-z-d \neq 0$ and for some K we have*

$$\frac{1}{d}\left[\frac{1}{(w-z-d)^k}-\frac{1}{(w-z)^k}\right]=\left[\frac{k(w-z)^{k-1}+R(d)d}{(w-z-d)^k(w-z)^k}\right],$$

with $|R(d)| \leq K$.

Proof. First set

$$\frac{1}{(w-z-d)^k}-\frac{1}{(w-z)^k}=\frac{(w-z)^k-[(w-z)-d]^k}{(w-z-d)^k(w-z)^k}.$$

According to the binomial theorem (Theorem 5.30), there is a K such that

$$-[(w-z)-d]^k=-(w-z)^k+k(w-z)^{k-1}d+R(d)d^2.$$

with $|R(d)| \leq K$. Inserting this and canceling d in the left-hand side yields the lemma. ■

Theorem 11.11. *Let γ be a simple, closed curve going around z once in counter-clockwise direction and suppose that f analytic on and inside γ (see Figure 66). Then for the kth derivative of f at z_0, or $f^{(k)}(z_0)$, we have*

$$i) \qquad \frac{f^{(k)}(z)}{k!}=\frac{1}{2\pi i}\oint_\gamma \frac{f(w)}{(w-z)^{k+1}}\,dw.$$

$$ii) \qquad \frac{\left|f^{(k)}(z)\right|}{k!}\leq\frac{M\ell(\gamma)}{r^{k+1}},$$

where $M=\max_{w\in\gamma}(|f(w)|)$, $\ell(\gamma)$ is the length of γ, and r is a lower bound for the distance of z to γ.

Proof. Cauchy's integral formula establishes the result for $k = 0$. The induction step proceeds as follows. Suppose we are given

$$f^{(k-1)}(z) = \frac{(k-1)!}{2\pi i} \oint_\gamma \frac{f(w)}{(w-z)^k}\, dw\,.$$

Since z lies inside γ, so does $z+d$ if d is small enough (Figure 66). We use

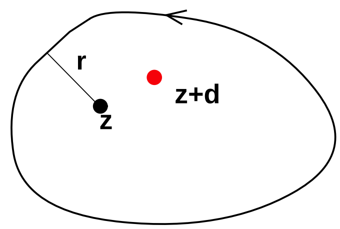


Figure 66. The curve γ goes around z exactly once in counter-clockwise direction. If d is small enough, $z+d$ also lies inside γ.

the induction hypothesis to compute the next derivative as $\lim_{d\to 0} \frac{f^{(k-1)}(z+d) - f^{(k-1)}(z)}{d}$. This equals

$$\begin{aligned}
\cdots &= \lim_{d\to 0} \frac{(k-1)!}{2\pi i d}\left[\oint_\gamma \frac{f(w)}{(w-z-d)^k}\, dw - \oint_\gamma \frac{f(w)}{(w-z)^k}\, dw\right] \\
&= \lim_{d\to 0} \frac{(k-1)!}{2\pi i d} \oint_\gamma f(w)\left[\frac{1}{(w-z-d)^k} - \frac{1}{(w-z)^k}\right] dw \\
&= \lim_{d\to 0} \frac{(k-1)!}{2\pi i} \oint_\gamma f(w)\left[\frac{k(w-z)^{k-1} + R(d)d}{(w-z-d)^k (w-z)^k}\right] dw\,.
\end{aligned}$$

The first and second equalities above follow by linearity of integration. The final equality uses Lemma 11.10. The limit can now be taken safely, because the denominator is never zero, and so everything is nice and continuous.

$$\cdots = \frac{k!}{2\pi i} \oint_\gamma \left[\frac{f(w)}{(w-z)^{k+1}}\right] dw\,.$$

This establishes (i). Item (ii) follows immediately. ■

This has the remarkable implication that an analytic function — defined as having one derivative, Definition 11.2 — has derivatives *of all orders*. In particular, we have the following result.

Corollary 11.12. *The derivative of an analytic function is again analytic.*

Proposition 11.13 (Morera's Theorem). *If f is continuous and if always $\oint f\,dz = 0$ in some region A, then f is analytic in A.*

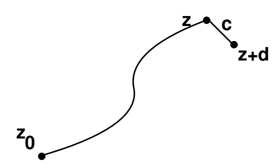


Figure 67. $F(z)$ does not depend on the path. So $F(z+d)-F(z) = \int_c f \approx f(z)d$

Proof. Pick a point z_0 and set $F(z) := \int_{z_0}^{z} f(w)\,dw$. Because $\oint f(w)\,dw = 0$, $F(z)$ does not depend on the path from z_0 to z and so is uniquely defined. Thus $F(z+d)-F(z) = \int_c f \approx f(z)d$, where c is a short, linear path from z to $z+d$ (see Figure 67). Then $F'(z) = f(z)$ and so f is the derivative of an analytic function and therefore is itself analytic. ■

Proposition 11.14. *Let $\{g_i\}$ be a sequence of functions that are analytic in a region A and suppose that $\sum_{i=1}^{\infty} g_i(z)$ converges uniformly on every closed disk contained in A. Then*
i) For any curve γ in A: $\int_\gamma \lim_n \sum_{i=1}^{n} g_i = \lim_n \int_\gamma \sum_{i=1}^{n} g_i$.
ii) $\lim_n \sum_{i=1}^{n} g_i$ is analytic in A.
iii) $\frac{d}{dz} \lim_n \sum_{i=1}^{n} g_i(z) = \lim_n \frac{d}{dz} \sum_{i=1}^{n} g_i(z)$.

Proof. Write $f_n = \sum_{i=1}^{n} g_i$ and call the limit f. Then for all $n > N$

$$\left| \int_\gamma f_n - \int_\gamma f \right| = \left| \int_\gamma f_n - f \right| \leq \int_\gamma |f_n - f| \leq \varepsilon \ell(\gamma)\,.$$

where $\ell(\gamma)$ is the length of γ (a curve whose image is a compact set). The fact that $|f_n(z) - f(z)| \leq \varepsilon$ for all $z \in \gamma$ is due to uniform convergence. This proves (i).

Next, we prove (ii). Pick $z_0 \in A$ and let $B = B_r(z_0)$ be an open disk whose closure $\bar{B}$ is contained in A. By assumption, $f_n \to f$ uniformly on $\bar{B}$ and thus f is continuous on $\bar{B}$ (see exercise 11.17). Now let γ be any simple, closed curve in $\bar{B}$. Then by Cauchy's theorem, $\oint_\gamma f_n = 0$. Item (i) implies that $\oint_\gamma f = 0$. Finally, Morera's theorem implies that f is analytic at z_0.

For part (iii), we have to show that $|f_n'(z) - f'(z)|$ tends to zero as n tends to infinity. We use Theorem 11.11 to do that. Fix some small r and so that $\gamma(t) := z_0 + re^{it}$ is contained in A. Then

$$|f_n'(z_0) - f'(z_0)| \leq \frac{1}{2\pi} \oint_\gamma \left| \frac{f_n(z) - f(z)}{(z-z_0)^2} \right| |dz| \, .$$

By uniform convergence, for large n, $|f_n(z) - f(z)|$ is less than ε on γ while $|z - z_0| = r$ and the length of γ is $2\pi r$. ■

Lemma 11.15. *If $|z - z_0| < |w - z_0|$, then*

$$\sum_{k=0}^{\infty} \frac{(z-z_0)^k}{(w-z_0)^{k+1}} = \frac{1}{w-z} \, .$$

Proof. $\sum_{k=0}^{\infty} \left[\frac{z-z_0}{w-z_0}\right]^k$ is a geometric series that can be written as $\sum_{k=0}^{\infty} x^k$, where $|x| < 1$. This equals $\frac{1}{1-x}$. Substituting this in the right-hand side of the lemma gives the result. ■

Theorem 11.16 (Taylor's Theorem). *Suppose f is analytic in a region A and let D be any open disk centered on z_0 whose closure is contained in A. Then for all $z \in D$ we have*

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!} \, (z-z_0)^n \, ,$$

which converges on D. This is called the Taylor series *of f at z_0.*

Proof. Let $\overline{D}$ be the disk bounded by the curve γ given by $w(t) = z_0 + re^{it}$. Take z inside D (see Figure 68) so that $|z - z_0| < |w - z_0|$. By Theorem 11.9 and Lemma 11.15, we have

$$f(z) = \frac{1}{2\pi i} \oint_\gamma \frac{f(w)}{(w-z)} \, dw = \frac{1}{2\pi i} \oint_\gamma \sum_{k=0}^{\infty} f(w) \, \frac{(z-z_0)^k}{(w-z_0)^{k+1}} \, dw \, .$$

Again because $|z - z_0| < |w - z_0|$, the sum converges uniformly, and so Proposition 11.14 (i) implies that the sum and integral can be interchanged. To the expression that then results, we apply Theorem 11.11 to get

$$\cdots = \frac{1}{2\pi i} \sum_{k=0}^{\infty} \oint_\gamma f(w) \, \frac{(z-z_0)^k}{(w-z_0)^{k+1}} \, dw = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!} (z-z_0)^k \, .$$

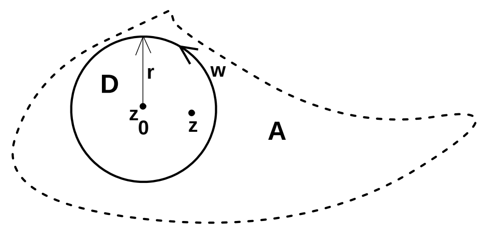


Figure 68. The curve w goes around z_0 exactly once in counter-clockwise direction. .

By Theorem 11.11 (ii), the last expression is bounded by $M\sum_{k=0}^{\infty}\frac{|z-z_0|^k}{r^k}$. Uniform convergence on compact sets contained in the open disk of radius r follows from Lemma 11.7. The series is analytic by Proposition 11.14. ■

Remark 11.17. Note that it follows that the Taylor series of an entire function (Definition 11.2) converges in all of $\mathbb{C}$.

11.4. A Tauberian Theorem

There is no formal definition of what a Tauberian theorem is, but generally it is something along the lines of the theorem below: we know that some transform $TF(z)$ of $f(t)$ converges for $\mathrm{Re}\, z > 0$, but we want to know that it converges for $z = 0$. The price we pay is some extra information on f as in the case below, where we stipulate a bound on $|f(t)|$. The reader is strongly encouraged to first have a look at the examples in exercise 11.23.

Theorem 11.18. *Let $f : [0,\infty) \to \mathbb{R}$ be integrable on compact intervals in $[0,\infty)$ and bounded by $|f| \le F$ for some $F > 0$ and define*

$$g(z) := \int_0^\infty f(t)e^{-zt}\,dt\,.$$

If $g(z)$ has an analytic continuation defined on $\mathrm{Re}\, z \ge 0$, then $\int_0^\infty f(t)\,dt$ exists and equals $g(0)$.

Remark 11.19. The function g in Theorem 11.18 is called the Laplace transform of f.

Proof. First define

$$g_T(z) := \int_0^T f(t)e^{-zt}\,dt\,.$$

Note that g'_T exists (exercise 11.24) and so g_T is entire. Pick any $\varepsilon > 0$, we will prove that for any $\varepsilon > 0$, we can choose T such that

$$\lim_{T\to\infty} |g_T(0) - g(0)| < \varepsilon\,. \tag{11.1}$$

Since $g_T(0)$ is finite, this implies that $g(0)$ also exists. So, fix $\varepsilon > 0$.

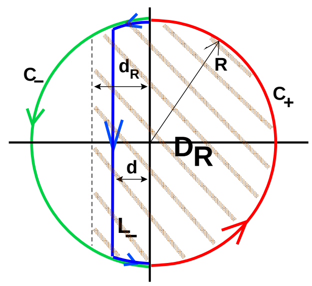


Figure 69. g is analytic in $D_R := \{\mathrm{Re}\,z \geq -d_R\} \cap \{|z| \leq R\}$ (shaded). The red curve is given by $C_+(s) = Re^{is}$ with $s \in (-\frac{\pi}{2}, \frac{\pi}{2})$. The green curve is given by $C_+(s) = Re^{is}$ with $s \in (\frac{\pi}{2}, \frac{3\pi}{2})$. The blue L_- consists of 2 small circular segments plus the segment connecting their left endpoints at a distance $0 < d < d_R$ to the left of the the imaginary axis.

For the definition of the region D_R and the curves C_+, C_-, and L_-, we refer to Figure 69. Because g is analytic on $\mathrm{Re}\,z \geq 0$, Corollary 11.3 says that for any R, there is a d_R so that g is analytic in the compact region D_R. Since g_T is analytic on all of $\mathbb{C}$, the Cauchy integral formula (Theorem 11.9) tells us that[2]

$$g(0) - g_T(0) = \frac{1}{2\pi i} \oint_{C_+ \cup L_-} \frac{(g(z) - g_T(z))\left(1 + \frac{z^2}{R^2}\right) e^{zT}}{z}\, dz\,. \tag{11.2}$$

We will show that $|g(0) - g_T(0)| < \varepsilon$ by cleverly splitting up this integral.

First compute the full integral along C_+ where $z = Re^{is} = R(\cos s + i \sin s)$. We will abbreviate $\cos s$ by c. For $\underline{c > 0}$, we first estimate the three factors in the integrand of (11.2).

$$|g(z) - g_T(z)| = \left| \int_T^\infty f(t) e^{-zt}\, dt \right| \leq F \int_T^\infty e^{-Rct}\, dt = \frac{Fe^{-RcT}}{Rc}\,. \tag{11.3}$$

[2]The factor $(1 + \frac{z^2}{R^2})$ in the integrand, introduced by Newman [**51**], may seem artificial and unnecessary at this point, but is in fact essential, see exercise 11.25.

Furthermore,

$$\left|\frac{1}{z}\left(1+\frac{z^2}{R^2}\right)\right| = \frac{1}{R}\left|1+e^{2is}\right| = \frac{1}{R}\left|e^{-is}+e^{is}\right| = \frac{2|c|}{R}\,. \tag{11.4}$$

And finally

$$\left|e^{zT}\right| = \left|e^{RTc+iRT\sin s}\right| = e^{RTc}\,. \tag{11.5}$$

Since the length of C_+ is πR, we thus obtain from (11.2) that

$$\left|\frac{1}{2\pi i}\int_{C_+}\right| \leq \frac{1}{2\pi}\cdot\frac{Fe^{-RcT}}{Rc}\cdot\frac{2c}{R}\cdot e^{RcT}\cdot\pi R = \frac{F}{R}\,. \tag{11.6}$$

For the second step, analyticity of g_T and Theorem 11.9 imply that

$$\frac{1}{2\pi i}\int_{C_-}\frac{g_T(z)\left(1+\frac{z^2}{R^2}\right)e^{zT}}{z}\,dz = \frac{1}{2\pi i}\int_{L_-}\frac{g_T(z)\left(1+\frac{z^2}{R^2}\right)e^{zT}}{z}\,dz\,,$$

allowing us to evaluate the integral along C_-. We have, now for $\underline{c<0}$,

$$|g_T(z)| = \left|\int_0^T f(t)e^{-zt}\,dt\right| \leq F\int_0^T e^{-Rct}\,dt = \frac{Fe^{-RcT}}{R|c|}\,.$$

Substituting this into the integral over C_- and using (11.4) and (11.5) gives

$$\left|\frac{1}{2\pi i}\int_{C_-}\frac{g_T(z)\left(1+\frac{z^2}{R^2}\right)e^{zT}}{z}\,dz\right| \leq \frac{1}{2\pi}\,\frac{Fe^{-RcT}}{R|c|}\,\frac{2|c|}{R}\,e^{RcT}\,\pi R = \frac{F}{R}\,. \tag{11.7}$$

The third (most painful) step is the evaluation of the remaining integral,

$$\int_{L_-} G(z)e^{zt}\,dz\,,$$

(see again Figure 69) where $G(z) := g(z)(1+\frac{z^2}{R^2})/(2\pi iz)$. On the two (compact) circular segments $z = Re^{is}$ with $\mathrm{Re}\,z \in [-d_R, 0]$, $|G|$ is maximized by the constant $M_h(R,d_R)$. The combined length of these segments is less than $4d$. Thus the integral over these pieces contributes at most $M_h(R,d_R)\,4d$. On the vertical segment, $|G|$ is bounded by another constant, $M_v(R,d)$. This may very well increase as d decreases, since, with decreasing d, the path passes very close to the origin. We have that $\left|e^{zT}\right| = e^{-dT}$ and the path length is less than $2R$. So the contribution of the vertical segment is at most $M_v(R,d)\,e^{-dT}\,2R$. Summarizing, this gives

$$\left|\int_{L_-} G(z)e^{zt}\,dz\right| \leq 4d\,M_h(R,d_R) + 2R\,M_v(R,d)\,e^{-dT}\,. \tag{11.8}$$

Now we add up the contributions of equations (11.6), (11.7), and (11.8).

$$\frac{1}{2\pi}\left|\oint_{C_+\cup L_-}\right| \leq \frac{2F}{R} + 4d\,M_h(R,d_R) + 2R\,M_v(R,d)\,e^{-dT}\,.$$

There are now three parameters, R, d, and T, whose values have not been fixed yet. We use these to "talk" the right hand side into being less than ε. Start by choosing R so that the first term is less than $\varepsilon/3$. Then choose $d \in [0,d_R]$ so that $4d\,M_h(R,d_R) < \varepsilon/3$. Finally, we choose T so that the last term is also less than $\varepsilon/3$. ∎

11.5. A Polynomial Must Have a Root

While we are on the topic of complex analysis, we take advantage of the opportunity to fill a gap in our proof of the fundamental theorem of algebra (Theorem 3.19).

Proposition 11.20. *Every polynomial of degree $d \geq 1$ has a root in $\mathbb{C}$.*

Proof. Let $p(z) = \sum_{i=0}^{d} a_i z^i$ be a non-constant polynomial (with non-zero leading coefficient a_d). The proof consists of showing that $|p(z)|$ has a minimum and that that minimum equals zero.

We write $z = re^{i\varphi}$ (polar coordinates) and immediately obtain

$$p(re^{i\varphi}) = a_d r^d e^{di\varphi}\left(1 + \frac{a_{d-1}}{a_d}\,r^{-1}e^{-i\varphi} + \cdots + \frac{a_0}{a_d}\,r^{-d}e^{-di\varphi}\right).$$

The term in parentheses can be written as $1 + r^{-1}A(r)$, where $A(r)$ can be bounded from above by a geometric series in $1/r$. Thus for r greater than some R, $A(r)$ is bounded by $A_0 \geq 0$. We then get for $r > R$

$$|p(re^{i\varphi})| = |a_d|r^d\left(1 + A(r)r^{-1}\right) \quad \text{where} \quad |A(r)| \leq A_0\,.$$

Thus for R large enough, $|p(2Re^{i\varphi})|$ is larger than $|p(Re^{i\varphi})|$. The closed disk D of radius $2R$ is compact and p is continuous, so it follows that $|p(z)|$ must have a minimum in in the interior of that disk (see Figure 70).

Let z_0 be this minimum. Take δ in the ball $|\delta| < \varepsilon$, and ε small so that the ε-disk around z_0 is in the interior of D (see Figure 70). Now expand $p(z_0+\delta) = \sum_{i=0}^{d} a_i(z_0+\delta)^i$. The expansion must contain non-trivial terms, because otherwise p would be constant. So for some $0 < k \leq d$,

$$p(z_0+\delta) = p(z_0) + b_k\delta^k + b_{k+1}\delta^{k+1} + \cdots + b_d\delta^d\,,$$

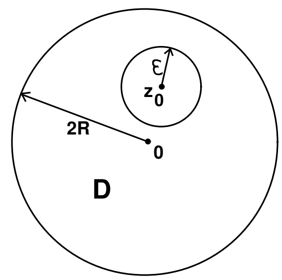


Figure 70. In the proof of Proposition 11.20, $|p(z)|$ must have a minimum z_0 in the interior of the disk $|z| < 2R$ and it cannot have a minimum unless at z_0 unless it is zero.

where $b_k \neq 0$. Thus

$$p(z_0+\delta) = p(z_0) + b_k \delta^k \left(1 + \delta B(\delta)\right) ,$$

where again for ε small enough $|B(\delta)|$ is bounded and so $p(z_0+\delta) \approx p(z_0) + b_k\delta^k$. By choosing the *phase* of δ appropriately and $|\delta|$ small enough, one make sure that if $|p(z_0)| > 0$, then $|p(z_0)+b_k\delta^k| < |p(z_0)|$. ■

Lest one might think that every complex function must have a zero, we warn the reader that e^z has no zero (see also exercise 11.16).

Together with exercise 3.24, the last result establishes the fundamental theorem of algebra (Theorem 3.19), which we repeat verbatim here.

Theorem 11.21 (**Fundamental Theorem of Algebra**). *A polynomial in* $\mathbb{C}[x]$ *(the set of polynomials with complex coefficients) of degree* $d \geq 1$ *has exactly* d *roots, counting multiplicity.*

11.6. Exercises

Exercise 11.1. Which of the following sets are regions or domains in $\mathbb{C}$?
a) $\mathbb{C}\backslash\{0\}$.
b) $\mathbb{C}\backslash\mathbb{N}$.
c) $\mathbb{C}$ minus the negative real axis.
d) $\mathbb{C}$ minus the real axis.
e) The union of the closed unit disks with centers at 1 and -1.
f) The same as (d), but minus the boundary.
g) The same as (e), but now add the imaginary axis.

In exercise 11.2 briefly discuss two "bad" (non-isolated) singularities. Around such a singularity no power series expansions can be used to approximate the functions. Pictures of the two singularities can be found in [**26**][Sections 2.4 and 3.1].

Exercise 11.2. On $D = \{z : |z| < 1\}$, define

$$f(z) = \sum_{n=1}^{\infty} z^{n!} \text{ and } g(z) = \frac{1}{sin(1/z)}\,.$$

a) Let p and q be co-prime integers and set $z = re^{2\pi ip/q}$. Show that

$$|f(z)| \geq -q + \sum_{n \geq q} r^{n!}\,,$$

and that this diverges as $r \nearrow 1$. (*Note: the unit circle is a natural boundary.*)
b) Conclude that the singularities of f are dense on the unit circle.
c) Show that g has a cluster point at the origin.

Exercise 11.3. a) Show that on $(0,1)$, $\sum_{n=1}^{\infty} x^n$ is absolutely convergent but not uniformly convergent.
b) Show that on $(0,1)$, $\sum_{n=1}^{\infty} \frac{(-1)^n x}{n}$ is uniformly convergent but not absolutely convergent. (*Hint: the sum is* $-x\ln 2$.)

Exercise 11.4. a) Let $z = x+iy$ and show that for $n \in \mathbb{N}$, $n^{-z} = n^{-x}e^{-iy\ln n}$.
b) From (a), show that $\left|n^{-z}\right| = n^{-x}$.
c) Use (b) to show that $\zeta(z) = \sum_{n=1}^{\infty} n^{-z}$ is uniformly convergent on compact disks in $\operatorname{Re} z > 1$. (*Hint: use Lemma 11.7 and exercise 2.25 (e).*)

Exercise 11.5. Let f analytic at z_0 and suppose furthermore that there is a sequence $\{z_n\}$ converging to z_0 such that $f(z_n) = 0$.
a) Show that f has all derivatives at z_0. (*Hint: Theorem 11.11.*)
b) Show that if at least one of $f^{(n)}(z_0) \neq 0$, then for z close enough to z_0, $f(z) \neq 0$. (*Hint: the first non-zero term in the Taylor expansion dominates as in Section 11.5.*)
c) Use (a) and (b) to show that for all $n \geq 0$, $f^{(n)}(z_0) = 0$.
d) Use Taylor's theorem to show that f is zero in an open disk containing z_0.

Exercise 11.6. Let A be a region (that is: an open, connected set) containing a sequence $\{z_n\}$ converging to z_0. Let f and g be analytic functions on A such that $f(z_n) = g(z_n)$ for all n.
a) Show that $h := f - g$ is analytic of A and satisfies $h(z_n) = 0$.
b) Use exercise 11.5 to show that $h = 0$ in an open disk containing z_0.
c) Write A as the disjoint union of

$$A_0 := \{z_0 \in A : h(z) = 0 \text{ on an open neighborhood of } z_0\} \quad \text{and} \quad A_1 := A \backslash A_0 .$$

Show that A_0 is open in A. (*Hint: by definition of* A_0.)
d) Show A_1 is open in A. (*Hint: consider* $z \in A_1$, *if* $h(z) \neq 0$, *use continuity of h; if* $h(z) = 0$, *use exercise 11.5 that h is not zero in a neighborhood of z.*)
e) Show that one of A_0 or A_1 must be empty. (*Hint: use Definition 11.1.*)
f) Conclude that the analytic continuations of f and g in A coincide. (*Note that this was remarked more informally in Section 2.5.*)

The last result of exercise 11.6 will be relevant when we discuss the analytic continuation of the zeta function. We isolate the result here.

Theorem 11.22 (**Uniqueness of Analytic Continuation).** *Suppose f and g are analytic in a region or domain $A \in \mathbb{C}$. Let Z be the set of points such that $f(z) - g(z) = 0$ and suppose that Z has a limit point in A. Then the analytic continuation of f and g coincide on A.*

Exercise 11.7. For $z \in \mathbb{C}$, define

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

a) Assume or prove[a] that the sum converges uniformly on every closed disk. Conclude that e^z is entire. (*Hint: Proposition 11.14 (ii).*)
b) Use exercise 11.6 to show that it is the unique analytic continuation of the real function e^x.
c) Compare the expansion of e^{iy} with those of $\cos y$ and $\sin y$ and conclude that $e^{iy} = \cos y + i \sin y$.
d) Use $e^{a+b} = e^a e^b$ to establish that

$$e^{x+iy} = e^x(\cos y + i \sin y) .$$

e) Use (a) and (d) to show that e^z is entire but never equal to 0.

[a]The factorial always wins out.

Exercise 11.8. a) Use exercise 11.7 (c) to show that (Figure 71) for $y \in \mathbb{R}$

$$\cos y = \frac{1}{2}\left(e^{iy} + e^{-iy}\right) \quad \text{and} \quad \sin y = \frac{1}{2i}\left(e^{iy} - e^{-iy}\right).$$

b) Use exercise 11.6 to show that

$$\cos z = \frac{1}{2}\left(e^{iz} + e^{-iz}\right) \quad \text{and} \quad \sin z = \frac{1}{2i}\left(e^{iz} - e^{-iz}\right)$$

are the unique extensions of the sine and cosine functions to the complex plane.
c) Find a formula with only exponentials for $\tan z$. (*Hint:* $\tan x = \frac{\sin x}{\cos x}$.)

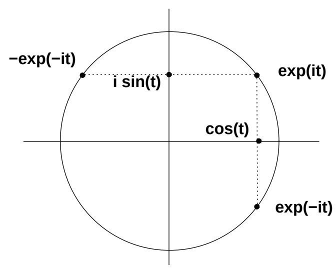


Figure 71. The complex plane with e^{it}, $-e^{-it}$ and e^{-it} on the unit circle. $\cos t$ is the average of e^{it} and e^{-it} and $i\sin t$ as the average of e^{it} and $-e^{-it}$.

Exercise 11.9. Use $e^{it}\left(e^{-it} + e^{it}\right) = 1 + e^{2it}$ to show that
a) $2\cos^2(t) = 1 + \cos(2t)$, and
b) $2\sin t \cos t = \sin 2t$.

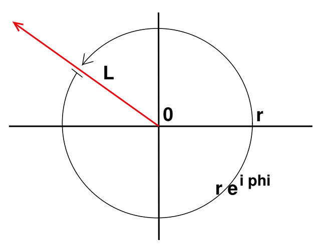


Figure 72. Moving around the origin once in the positive direction increases φ, and thus $\ln z$, by 2π. Discontinuities can be avoided if we agree never to cross the half line or branch cut L.

Exercise 11.10. The complex logarithm $\ln z$ is the (local) inverse of e^z. See Figure 72.
a) Use "polar" coordinates, i.e. $z = re^{i\varphi}$, to show that $\ln z = \ln r + i\varphi$.
b) Fix r and increase φ from 0 to 2π. Assuming that you do not encounter discontinuities, show that $\ln z$ has increased by $2\pi i$ while its real part remained constant.
c) Conclude that $\ln z$ is multivalued .
d) Let L be any half line from the origin to infinity. Show that $\ln z$ is analytic of $\mathbb{C}$ minus L. L is called a branch cut .

For any function $f : \mathbb{C} \to \mathbb{C}$, we can always write $z = x + iy$ and $f(z) = u(x+iy) + iv(x+iy)$. In the next three exercises, we prove the following result.

Proposition 11.23. *$f : \mathbb{C} \to \mathbb{C}$ is analytic (see Definition 11.2) at z_0 if and only if in a neighborhood of z_0, f is differentiable*[3] *as a function from $\mathbb{R}^2$ to itself and the Cauchy-Riemann equations hold:*

$$\partial_x u = \partial_y v \qquad \text{and} \qquad \partial_x v = -\partial_y u\,.$$

Exercise 11.11. a) Show that if f is analytic at z_0, then in a neighborhood of z_0, $f'(z) = \lim_{\delta \to 0} \frac{f(z+\delta)-f(z)}{\delta}$ does not depend on δ (as long as it tends to 0).
b) Compute the derivative in (a) for δ real and δ imaginary.
c) Use (a) to show these two are equal.
d) Use (c) to prove that analyticity implies that u and v satisfy the Cauchy-Riemann equations.

Exercise 11.12. For real a and b, let $A = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ and $z = \begin{pmatrix} x \\ y \end{pmatrix}$.
a) Show that multiplication by A of z in $\mathbb{R}^2$ acts exactly like multiplication by $a+ib$ of $x+iy$ in $\mathbb{C}$.
b) Write the matrix A as $Re^{i\theta}$. (*Hint: $R = \sqrt{a^2+b^2}$. What is θ?*).
c) Use (b) to show that a non-zero derivative at a point z_0 of an analytic function is a dilatation composed with a rotation.
d) Explain that if $f'(z_0)$ is non-zero, f "locally" preserves angles.

Definition 11.24. *A map f from a region $A \subset \mathbb{C}$ to $\mathbb{C}$ is conformal at z_0 if its derivative at z_0 exists and is non-zero.*

[3] This means that the partial derivatives exist and are continuous.

Exercise 11.13. Write $z = x+iy$ and $f(z) = u(x+iy)+iv(x+iy)$, where u and v are real functions. In a neighborhood of (x_0+iy_0), suppose that the matrix of (continuous) derivatives $Df(x,y)$ satisfies Cauchy-Riemann.
a) Use exercise 11.12 to show that this implies that $Df(x,y)$ acts like a complex number.
b) Use (a) to imply that f is analytic.

Exercise 11.14. Write $z = x+iy$ and $f(z) = u(x+iy)+iv(x+iy)$, where u and v are real functions.
a) Given that $u(x+iy) = e^{-y}\cos x$, compute v and $f(z)$. (*Hint: use the Cauchy-Riemann equations to compute $\partial_x v$ and $\partial_y v$. Integrate both to get v. Finally, express $u+iv$ as $f(z)$.*)
b) Given that $v(x+iy) = -y^3+3x^2y-y$, compute u and f.
c) Given that $f(z) = \tan z$, compute u and v. (*Hint: use exercise 11.8 (c).*)

An interesting result — though we will not prove it — is the following. A weaker version of this is called the Casorati-Weierstrass Theorem and has an easy proof [**26**][chapter 4] [**45**][chapter 3].

Theorem 11.25 (Picard Theorem). *Let f have an isolated essential singularity at z_0. Then the image of any punctured neighborhood of z_0 contains all values infinitely often with at most one exception.*

The next results are important corollaries (proof in exercise 11.15).

Corollary 11.26 (Little Picard). *Let f be entire and not constant. Then the image of f contains all values with at most one exception.*

Corollary 11.27 (Liouville's theorem). *A bounded entire function must be constant.*

Exercise 11.15. Assume f is entire and not constant.
a) Show that f has an expansion $\sum_{i=0}^{\infty} a_n x^n$ that converges in all of $\mathbb{C}$. (*Hint: see Taylor's theorem.*)
b) Show that if f is a polynomial (only finitely many non-zero a_n), then it has a pole at infinity. (*Hint: effect a coordinate change that moves ∞ to 0, i.e. set $w = 1/z$. What does f look like in terms of the new coordinate?*)
c) Show that in case (b), for all $z_0 \in \mathbb{C}$, $f(z)-z_0$ has a zero. (*Hint: the fundamental theorem of algebra (Theorems 3.19 and 11.21).*)
d) Show that if f is a not a polynomial, then it has an essential singularity at infinity.
e) Show that (c) and (d) and the Picard Theorem imply little Picard.
f) Show that Little Picard implies Liouville's theorem.

The function e^z is a good illustration of little Picard (see exercise 11.7 (e)). The next problem illustrates the Picard Theorem (Theorem 11.25).

Exercise 11.16. a) Show that if $z = x + iy$, then

$$\frac{1}{z} = \frac{x}{x^2+y^2} - i\frac{y}{x^2+y^2}\,.$$

b) Show that

$$f(z) := e^{\frac{1}{z}} = e^{\frac{x}{x^2+y^2}}\left(\cos\left(\frac{y}{x^2+y^2}\right) + i\sin\left(\frac{y}{x^2+y^2}\right)\right).$$

c) Show that if $y = 0$ and $x \searrow 0$, then $f(z)$ is real and tends to infinity.
d) Show that if $y = 0$ and $x \nearrow 0$, then $f(z)$ is real and tends to zero.
e) Show that if you approach 0 in any other direction, f has arbitrarily large oscillations. (*Hint: fix t and set $y = tx$ and let $x \searrow 0$.*)
f) Show that $f(z) \neq 0$ for all z.

Exercise 11.17. Let $\{f_n\}$ be a sequence of continuous functions on a compact set S in $\mathbb{R}^n$ or $\mathbb{C}$. Suppose $f_n \to f$ uniformly on S and let $x, y \in S$.
a) Show that there is an n such that $|f_n(x) - f(x)| < \varepsilon/3$.
b) Given n as in (a), show that there is a δ such that for all x with $|y - x| < \delta$, we have $|f_n(y) - f_n(x)| < \varepsilon/3$.
c) Show that (a) and (b) imply that $|f(y) - f(x)| < \varepsilon$. (*Hint: this is called the "$\varepsilon/3$ trick".*)
d) Show that (c) implies that f is continuous.

Exercise 11.18. We give an easy informal "proof" of Theorem 11.11 by interchanging differentiation and integration without justification.
a) Let k a non-negative integer. Suppose that

$$f^{(k)}(z_0) = \frac{k!}{2\pi i}\oint_\gamma \frac{f(z)}{(z-z_0)^{k+1}}\,dz\,.$$

Change the order of integration and differentiation to show that

$$\frac{d}{dz_0}f^{(k)}(z_0) = \frac{k!}{2\pi i}\oint_\gamma \frac{d}{dz_0}\frac{f(z)}{(z-z_0)^{k+1}}\,dz = \frac{(k+1)!}{2\pi i}\oint_\gamma \frac{f(z)}{(z-z_0)^{k+2}}\,dz\,.$$

b) Use (a) to give a proof by induction of Theorem 11.11.

Integrals and limits cannot always be exchanged, and the same holds for derivatives. The following exercise provides examples (see Figure 73). For uniformly converging analytic functions, the changes can be made (Proposition 11.14).

Exercise 11.19. On $[0,1]$, consider the functions

$$g_k(x) = k^2 x^k (1-x) \qquad \text{and} \qquad h_k(x) = \frac{\sin(k\pi x)}{k}\,.$$

a) Show that $\lim_{k\to\infty} g_k(x) = 0$.
b) Show that $\int_0^1 \lim_{k\to\infty} g_k(x)\,dx = 0$ while $\lim_{k\to\infty} \int_0^1 g_k(x)\,dx = \lim_{k\to\infty} \frac{k^2}{(k+1)(k+2)} = 1$.
c) Show that $\lim_{k\to\infty} h_k(x) = 0$.
d) Show that $\frac{d}{dx} \lim_{k\to\infty} h_k(x) = 0$ while $\lim_{k\to\infty} \frac{d}{dx} h_k(x) = 0$ does not exist at $x = 1/2$ (for example).

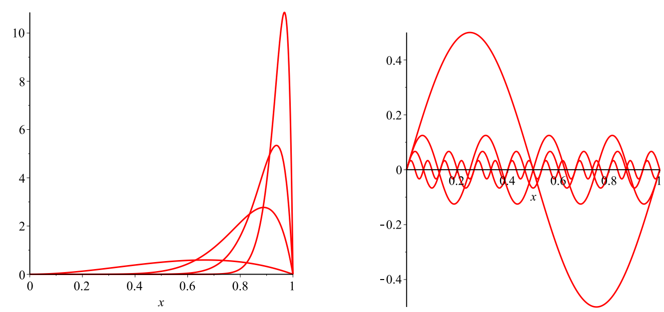


Figure 73. The functions g_k and h_k of exercise 11.19 for $k \in \{2,8,15,30\}$.

Exercise 11.20. Set $\alpha = a + ib$ where a and b real and greater than zero and let $f(z) = (z-\alpha)^{-1}$.
a) Show that f is analytic inside and on the contour C given in Figure 74.
b) Show $\oint_C f = 0$.
c) Show that $\int_{b_i} f$ tends to 0 as R tends to infinity. (*Hint:* $|f| \to 0$ *while the path length remains finite.*)
d) Show that $\int_r f$ tends to πi as R tends to infinity. (*Hint: set* $z(t) = ib + Re^{it}$ *with* $t \in [0,\pi]$.)
e) Show that $\oint_p f$ tends to $-2\pi i$ as R tends to infinity. (*Hint: set* $z(t) = \alpha + re^{-it}$ *with* $t \in [0,2\pi]$.)
f) Conclude that $\lim_{R\to\infty} \int_{-R}^{+R} f(z)\,dz = \pi i$. (*Hint: use (a).*)

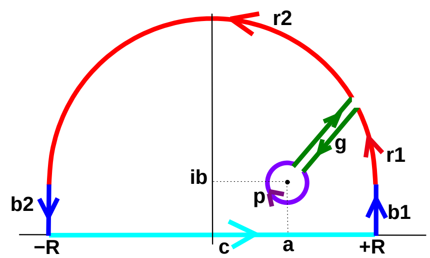


Figure 74. The contour C is the concatenation of c (celeste), b_1 (blue), r_1 (red), g (green), p (purple), $-g$, r_2, and b_2. The path r is a semi-circle of radius R. The path p is a small circle of radius r. See exercise 11.20.

Exercise 11.21. We check the outcome of exercise 11.20 by direct integration. We use the notation of that problem.

a) Show that

$$\int_{-R}^{+R} f(z)\,dz = \int_{-R}^{+R} \frac{x-a+ib}{(x-a)^2+b^2}\,dx\,.$$

b) Sustitute $s = x - a$ and show that

$$\int_{-R}^{+R} f(z)\,dz = \int_{-R-a}^{+R-a} \frac{s+ib}{s^2+b^2}\,ds\,.$$

c) Show that the real part of this integral tends to zero as $R \to \infty$. (*Hint: it is odd plus something that tends to zero.*)

d) Show that $\lim_{R\to\infty} \int_{-R-a}^{+R-a} \frac{ib}{s^2+b^2}\,ds = \pi i$. (*Hint: substitute $bt = s$ and use that the derivative of* $\arctan x$ *equals* $1/(x^2+1)$.)

Exercise 11.22. Let $f(z) = \sum_{n\geq -k} a_k (z-z_0)^k$ with $k > 0$.

a) Compute

$$\operatorname{Res}(f, z_0) := \frac{1}{2\pi i} \oint f(z)\,dz$$

along the path $\gamma(t) = z_0 + \varepsilon e^{it}$, $t \in [0, 2\pi]$ for small $\varepsilon > 0$. This is called the <u>residue</u> of f at z_0.

b) Let Γ be *any* piecewise smooth contour winding exactly once around z_0 in the anti-clockwise direction. Show that

$$\oint_\Gamma f(z)\,dz = 2\pi i \operatorname{Res}(f, z_0)\,.$$

(*Hint: consider a contour that narrowly avoids the singularity such as the contour C in Figure 74.*)

Exercise 11.23. a) Let $f(t)=1$. Show that its Laplace transform as defined in Theorem 11.18 does not have an analytic continuation to the imaginary axis.
b) In (a), show that the conclusion of Theorem 11.18 does not hold.
c) Repeat (a) and (b), but now for $f(t)=e^{i\omega t}$.

Exercise 11.24. Consider $g_T(z)$ as in the proof of Theorem 11.18.
a) Write out $H_\varepsilon := \frac{1}{\varepsilon}\left(g_T(z+\varepsilon)-g_T(z)\right)$.
b) Use linearity of integration to show that $\lim_{\varepsilon\to 0} H_\varepsilon = \int_0^T -tf(t)e^{-zt}\,dt$.
c) Show that the integral in (b) exists.
d) Conclude that g_T is entire.

Exercise 11.25. a) Explain why it is crucial in the proof of Theorem 11.18 that $g(z)$ is analytic on the imaginary axis.
b) Explain why the factor $(1+\frac{z^2}{R^2})$ is essential to the proof of Theorem 11.18.

Chapter 12

The Prime Number Theorem

Overview. In 1850, it seemed that Chebyshev was awfully close to proving the prime number theorem (Theorem 2.21). But to bridge that last brook, a whole new approach to the problem was needed. That approach was the connection with analytic functions in the complex domain pioneered by Riemann in 1859 [**60**]. A very weak version of the Riemann hypothesis (Conjecture 2.22), namely the absence of zeroes of $\zeta(z)$ in $\operatorname{Re} z \geq 1$ turns out to be an essential step. We look at this in Section 12.3 and in particular Lemma 12.12. It would take another 37 years after Riemann's monumental contribution before the result was finally proved by De La Vallée Poussin and Hadamard in 1896. The version we prove is a highly streamlined derivative of that proof, the last stage of which was achieved by Newman in 1982 [**51**]. We made heavy use of Zagier's rendition of this proof [**75**] and of [**63**].

12.1. Preliminaries

Recall that $\pi(x)$ denotes the number of primes in the interval $[2,x]$. So $\pi(2) = 1$, $\pi(3.2) = 2$, and so on. The reason that the variable x is real is that it simplifies the formulas to come. The Riemann zeta function is denoted by $\zeta(s)$, see Definition 2.19 and Proposition 2.20. In this chapter, we will frequently encounter sums of the form $\sum_p$. For example see Definition 12.1

below. Such sums will always be understood to be over all positive primes. On the other hand, $\sum_{p\le x}$ indicates a sum over all positive primes p less than or equal to x. A similar convention holds for products $\prod_p$ and $\prod_{p\le x}$. The letter z will always denote a complex variable.

We now define a couple of new functions.

Definition 12.1. *The first Chebyshev function is defined as*

$$\theta(x) := \sum_{p\le x} \ln p\,.$$

The function $\Phi : \{z\in\mathbb{C} : \operatorname{Re} z > 1\} \to \mathbb{C}$ *is defined as*

$$\Phi(z) := \sum_p \frac{\ln p}{p^z}\,.$$

It is analytic in $\operatorname{Re} z > 1$.

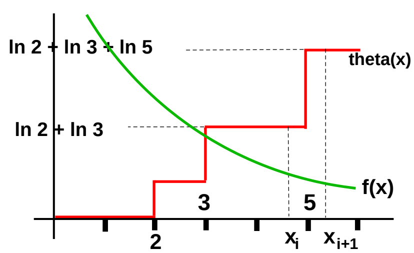


Figure 75. The Riemann-Stieltjes integral (12.1) near $x = 5$ picks up the value $f(5)(\theta(x_{i+1}) - \theta(x_i))$.

In what follows, we will need to integrate expressions like

$$I(x) := \int_1^x f(t)\,d\theta(t)\,, \tag{12.1}$$

where f is differentiable. If we partition the interval $[1,x]$ by $1 = x_0 < x_1 \cdots x_n = x$, then $I(x)$ can be approximated as

$$I(x) \approx \sum_{i=1}^{n} f(c_i)(\theta(x_{i+1}) - \theta(x_i))\,,$$

where $c_i \in (x_i, x_{i+1})$ and then the appropriate limit (assuming it exists) can be taken. This is a Riemann-Stieltjes integral. It is very similar to the Riemann integral from calculus, except that instead of the increments $x_{i+1} - x_i$, we look at increments of a function: $\theta(x_{i+1}) - \theta(x_i)$ (see [**40**]), see

Figure 75. Now, $\theta(t)$ is constant except at the values $t = p$ (a prime) where it has a jump discontinuity of size $\ln p$. Thus, in this case, $I(x)$ simplifies to

$$I(x) = \int_1^x f(t)\, d\theta(t) = \sum_{p \le x} f(p) \ln(p)\,. \tag{12.2}$$

On the other hand, we can find a different expression for $I(x)$ by *integration by parts* (sometimes called *partial integration*)

$$I(x) = \int_1^x d\,f(t)\theta(t) - \int_1^x \theta(t)\, d f(t) = f(t)\theta(t)\big|_1^x - \int_1^x f'(t)\theta(t)\, dt\,. \tag{12.3}$$

The point of this operation is usually that now we have expressed the integral in (12.2) as fixed expression plus another integral which has better convergence properties than the original integral. For instance if $f(t) = t^{-k}$, then $f'(t) \propto t^{-k-1}$ and so the integral converges faster.

Lemma 12.2. *We have for* $x \ge 2$

$$\pi(x) = \frac{\theta(x)}{\ln x} + \int_2^x \frac{\theta(t)}{t\,(\ln t)^2}\, dt\,.$$

Proof. First note that since 2 is the smallest prime, equation (12.2) gives

$$\pi(x) = \int_{2-\varepsilon}^x \frac{d\,\theta(t)}{\ln t}\,.$$

Apply integration by parts (12.3) to obtain

$$\pi(x) = \frac{\theta(x)}{\ln x} - \int_{2-\varepsilon}^x \theta(t)\, d\frac{1}{\ln t}\,.$$

Using $d\frac{1}{\ln t} = -\frac{dt}{t(\ln t)^2}$ to work out the last term yields the lemma with lower limit $2-\varepsilon$ in the integral. But since $\theta(t) = 0$ for $t < 2$, we may replace that limit by 2. ■

Lemma 12.3. *For* $\mathrm{Re}\, z > 1$, *we have*

$$\begin{aligned} \frac{\Phi(z)}{z} - \frac{1}{z-1} &= \int_1^\infty \left(\frac{\theta(x)}{x} - 1\right) x^{-z}\, dx \\ &= \int_0^\infty \left(\theta(e^t)e^{-t} - 1\right) e^{-zt+t}\, dt\,. \end{aligned}$$

Proof. Using (12.2), we can write $\Phi(z)$ as $\int_1^\infty x^{-z} d\theta(x)$. Then apply (12.3) (partial integration) to obtain

$$\Phi(z) = x^{-z}\theta(x)\Big|_1^\infty + z\int_1^\infty x^{-z-1}\theta(x)\,dx\,.$$

We will see in equation (12.6) that for $\operatorname{Re} z > 1$, the boundary term $x^{-z}\theta(x)\Big|_1^\infty$ vanishes. This gives

$$\frac{\Phi(z)}{z} = \int_1^\infty \frac{\theta(x)}{x} x^{-z}\,dx\,.$$

Noting that $1/(z-1) = \int_1^\infty x^{-z}\,dx$, the first equality follows. The second equality follows by substitution of x by t where $e^t = x$. ■

12.2. Chebyshev's Theorem

We prove Theorem 12.7, an approximate version of the prime number theorem (Theorem 2.21). Recall that $\lfloor x \rfloor$ is the greatest integer less than or equal to x (see Definition 2.1), whereas $\binom{a}{b}$ indicates the binomial factor $\frac{a!}{b!(a-b)!}$ (see Theorem 5.30).

We start with a remarkable lemma. Let a, b, and $k > 0$ be integers. We introduce the notation $a^k \parallel b$ to mean that $a^k \mid b$ but not $a^{k+1} \mid b$. In words, this is expressed by saying that a^k *divides* b *exactly*.

Lemma 12.4. *Let p prime and suppose that $p^k \parallel \binom{n}{m}$ with $n > m > 0$. Then we have $p^k \le n$.*

Proof. Let p prime and suppose that $p^k \parallel (1 \cdot 2 \cdots n)$. We want to find k. Any multiple $ap \le n$ in the product $1 \cdot 2 \cdots n$ contributes one factor p to p^k. The number of multiples ap less than or equal to n equals $\left\lfloor \frac{n}{p} \right\rfloor$. So these contribute $\left\lfloor \frac{n}{p} \right\rfloor$ to k. If ap is also a multiple of p^2 then it contributes two factors to k. Thus we need to add another factor in the form of $\left\lfloor \frac{n}{p^2} \right\rfloor$. Continuing like that, we find

$$p^k \parallel n! \quad \Longrightarrow \quad k = \sum_{j=1}^\infty \left\lfloor \frac{n}{p^j} \right\rfloor . \tag{12.4}$$

As a consequence, we obtain for the binomial factor

$$p^k \parallel \binom{n}{m} \quad \Longrightarrow \quad k = \sum_{j=1}^\infty \left(\left\lfloor \frac{n}{p^j} \right\rfloor - \left\lfloor \frac{m}{p^j} \right\rfloor - \left\lfloor \frac{n-m}{p^j} \right\rfloor \right) . \tag{12.5}$$

Consider the expression $E = \lfloor x_1 + x_2 \rfloor - \lfloor x_1 \rfloor - \lfloor x_2 \rfloor$. By substituting $x_1 = a_1 + \omega_1$ and $x_2 = a_2 + \omega_2$, where a_i are integers and $\omega_i \in (0,1)$, one sees that $E \in \{0,1\}$. Going back to the expression in equation (12.5), we see that if $p^j > n$, then the contribution is always zero. Thus if $n > m > 0$, the last positive contribution occurred for $j = k$ such that $p^k \leq n$. ■

The crux of the proof of Chebyshev's theorem is contained in two simple, yet very clever, lemmas.

Lemma 12.5. *For $n \geq 2$, we have* $\frac{2^n}{n+1} < \binom{n}{\lfloor n/2 \rfloor} \leq 2^{n-1}$.

Proof. We prove the right-hand side first. Since $\binom{2}{1} = 2^1$ and for $n > 0$

$$\binom{2n+2}{n+1} = \frac{(2n+1)(2n+2)}{(n+1)^2}\binom{2n}{n} < 4\binom{2n}{n},$$

The result follows in the even case. The odd case is similar. For the left-hand side, we note that $\binom{n}{\lfloor n/2 \rfloor}$ is the largest of the $n+1$ numbers $\binom{n}{i}$ and so

$$(n+1)\binom{n}{\lfloor n/2 \rfloor} > \sum_{i=0}^{n}\binom{n}{i} = 2^n .$$

■

Lemma 12.6. *i) For all $n \geq 2$, we have* $\binom{n}{\lfloor n/2 \rfloor} \leq n^{\pi(n)}$.
ii) For $n \geq 2$ a power of 2, we have $e^{\theta(n)-\theta(n/2)} \leq \binom{n}{n/2}$.

Proof. For the first inequality, use unique factorization (Theorem 2.11) and the definition of $\pi(n)$ to write

$$\binom{n}{\lfloor n/2 \rfloor} = \prod_{i=1}^{\pi(n)} p_i^{k_i} .$$

By Lemma 12.4, $p_i^{k_i} \leq n$. Thus $\prod_{i=1}^{\pi(n)} p_i^{k_i} \leq n^{\pi(n)}$, which yields the inequality.

For the second inequality, we start by noticing that n is even and so any prime p in the interval $\left(\frac{n}{2}, n\right]$ is a divisor of $n!$ but not of the denominator of $\binom{n}{n/2}$. Therefore any such p divides $\binom{n}{n/2}$. This implies that

$$\prod_{\frac{n}{2} < p \leq n} p \leq \binom{n}{n/2} .$$

Noting that $p = e^{\ln p}$ and inserting the definition of $\theta(x)$ (Definition 12.1) yields the last inequality. ■

Theorem 12.7 (Chebyshev's Theorem). *For all $a < \ln 2$ and $b > \ln 4$, there is a large enough K such that*

$$\forall x \geq K \ : \ \frac{\pi(x)}{x/\ln x} \in [a,b]\,.$$

Proof. Putting Lemmas 12.5 and 12.6 together gives

$$\frac{2^n}{n+1} \leq n^{\pi(n)} \qquad \text{and} \qquad e^{\theta(n)-\theta(\frac{n}{2})} \leq \frac{1}{2}2^n \text{ (if } n \text{ a power of 2)}\,.$$

Taking the logarithm of the first of these inequalities gives

$$\left(\ln 2 - \frac{\ln(n+1)}{n}\right)\frac{n}{\ln n} < \pi(n)\,,$$

which yields an estimate for a.

For n a power of 2, we get from the second inequality

$$\theta(n) - \theta\left(\frac{n}{2}\right) \leq \frac{n}{2}\ln 2 \quad \text{and} \quad \theta\left(\frac{n}{2}\right) - \theta\left(\frac{n}{4}\right) \leq \frac{n}{4}\ln 2 \quad \text{and} \quad \cdots$$

and so on. Thus $\theta(n) \leq n\ln 2$. For $x \geq 2$, there is an n' that is a power of 2 in the interval $[x, 2x)$. Thus $\theta(x) \leq \theta(n') \leq 2x\ln 2$. Therefore,

$$\theta(x) \leq x\ln 4\,. \tag{12.6}$$

Substituting this into Lemma 12.2 gives that

$$\pi(x) \leq \ln 4\,\frac{x}{\ln x} + \ln 4 \int_2^x (\ln t)^{-2}\,dt\,. \tag{12.7}$$

L'Hôpital's rule implies that

$$\lim_{x\to\infty} \frac{\int_2^x (\ln t)^{-2}\,dt}{x(\ln x)^{-2}} = 1\,. \tag{12.8}$$

Thus the integral in (12.7) can be replaced by $x(\ln x)^{-2}$. The dominant term of the right-hand side of that equation is the first one. Thus for any $b > \ln 4$, we have for x large enough that $\pi(x) < b\frac{x}{\ln x}$. ■

Remark 12.8. In fact, Chebyshev was able to prove the sharper result that if $\lim_{x\to\infty} \frac{\pi(x)}{x/\ln x}$ exists, it has to be 1.

Equations (12.6) and (12.8) will also play an important role in the proof of the (full) prime number theorem.

12.3. Zeroes and Poles of the Zeta Function

The proof of the prime number theorem relies in part on a careful study of the analytic extensions of some functions related to the zeta function. We do that in this section and the next.

Remark 12.9. From Chapter 4, equation (4.8) (see also exercise 4.27), we know that $\zeta(z)$ and $1/\zeta(z)$ both converge for $\mathrm{Re}\, z > 1$. Therefore neither has zeroes or poles in that region.

Here we prove a stronger statement.

Lemma 12.10. *For* $\mathrm{Re}\, z > 1$, *we have that*

$$\ln \zeta(z) = -\sum_p \ln\left(1 - e^{-z\ln p}\right) = \sum_p \sum_{n=1}^{\infty} \frac{e^{-zn\ln p}}{n} = \sum_p \sum_{n=1}^{\infty} \frac{1}{np^{nz}}\,.$$

is analytic.

Proof. First, set $w := p^{-z} = e^{-z\ln p}$. Using that the Taylor series at 0

$$-\ln(1-w) = \sum_{n\geq 1} \frac{w^n}{n}\,,$$

converges uniformly on $|w| < 1$ on compact subsets, we see from Proposition 11.14 (ii) that

$$-\ln\left(1-p^{-z}\right) = -\ln\left(1 - e^{-z\ln p}\right) = \sum_{n=1}^{\infty} \frac{e^{-zn\ln p}}{n} = \frac{1}{np^n} \qquad (12.9)$$

is analytic on $\mathrm{Re}\, z > 0$ (and thus also on $\mathrm{Re}\, z > 1$).

Next, from Proposition 2.20, we conclude that $\ln \zeta(z) = -\sum_p \ln\left(1-p^{-z}\right)$. By Lemma 10.18, this converges absolutely iff $\sum_p p^{-z}$ converges absolutely. But if we set $z = x + iy$, then

$$\sum_p \left|p^{-z}\right| \leq \sum_n \left|n^{-z}\right| = \sum_n \left|n^{-x}\right|\,,$$

which converges absolutely by exercise 2.25 (e), and thus uniformly on closed disks in $\mathrm{Re}\, z > 1$. Therefore, by Proposition 11.14, $-\sum_p \ln\left(1-p^{-z}\right)$ is analytic on $\mathrm{Re}\, z > 1$. ∎

We saw in see exercise 2.24 (c) that $\zeta(z)$ diverges as $z \searrow 1^+$. Here is a more precise statement. Recall that analytic continuations are well-defined (i.e. unique) in domains with only isolated singularities (see Theorem 11.22).

Proposition 12.11. *i) The functions* $(z-1)\zeta(z)$ *and* $(z-1)\zeta'(z)+z\zeta(z)$ *have well-defined analytic continuations on* $\mathrm{Re}\, z > 0$.
ii) (The analytic continuation of) $(z-1)\zeta(z)$ *evaluated at* $z=1$ *equals 1.*

Remark: The factor $(z-1)$ precisely cancels the simple pole in ζ at $z=1$.

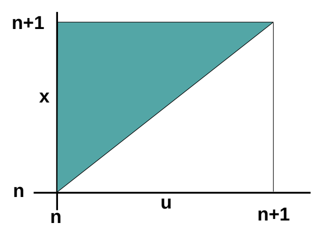


Figure 76. Integration over the shaded triangle of area $1/2$ in equation (12.11).

Proof. We have that $\int_1^\infty x^{-z}\,dx = 1/(z-1)$ and $\int_n^{n+1} n^{-z}\,dx = n^{-z}$. Using the definition of the zeta function (Definition 2.19), we define in $\mathrm{Re}\, z > 1$

$$h(z) := \zeta(z) - \frac{1}{z-1} = \sum_{n=1}^{\infty} n^{-z} - \int_1^\infty x^{-z}\,dx = \sum_{n=1}^{\infty} \int_n^{n+1} \left(n^{-z} - x^{-z}\right) dx\,. \tag{12.10}$$

Next, since $n^{-z} - x^{-z} = \int_n^x -zu^{-z-1}\,du$, we also have

$$\cdots \quad = \sum_{n=1}^{\infty} \int_n^{n+1} \int_n^x -zu^{-z-1}\,du dx\,. \tag{12.11}$$

Each term of the sum is an integral over a triangular domain of area $1/2$ (Figure 76). The maximum of the integrand is

$$\left|zn^{-z-1}\right| = \sqrt{\sigma^2+\tau^2}\;\, n^{-\sigma-1}\,,$$

where $z = \sigma + i\tau$ (with σ, τ real). So, each summand has absolute value less than half that. Thus (12.11) converges uniformly on compact disks in $\sigma > 0$ (see also exercise 11.4) and so h has an analytic continuation to $\mathrm{Re}\, z > 0$.

To prove (i), note that, by the above, $(z-1)h(z)+1=(z-1)\zeta(z)$ is analytic. Therefore so is its derivative — by Corollary 11.12. The second function of part (i) is the sum of these two. Finally, evaluating $(z-1)\zeta(z)=(z-1)h(z)+1$ at $z=1$ establishes part (ii). ■

Now follows a lemma that is brilliant and an essential step in proving the prime number theorem. It will make its appearance in Proposition 12.13.

Lemma 12.12. *$\zeta(z)$ has no zeroes on the line* $\mathrm{Re}\, z \geq 1$.

Proof. By remark 12.9, ζ has no zeroes in $\mathrm{Re}\, z > 1$. So we only need to prove that ζ has no zeroes on $\mathrm{Re}\, z = 1$.

Let $z = \sigma + i\tau$ with $\sigma > 1$ and $\tau \neq 0$ real. We start by computing the admittedly strange expression $E := \ln\big(\zeta(\sigma)^3\zeta(\sigma+i\tau)^4\zeta(\sigma+2i\tau)\big)$. By Proposition 12.11, ζ has a simple pole at 1 and no poles in $\mathrm{Re}\, z > 1$. Thus if ζ has a *zero* at $1+i\tau$, it *cannot* be compensated by a pole at $1+4i\tau$ and the pole of order 1 at $z=1$. Thus in this case, the expression e^E evaluated at $\sigma + i\tau$ where σ is *slightly* greater than 1, would yield a number that is very close to zero. We now show that this cannot happen.

Combining the fact that $\ln(ab) = \ln a + \ln b$ and Lemma 12.10, we get

$$\begin{aligned} E &= 3\ln\zeta(\sigma) + 4\ln\zeta(\sigma+i\tau) + \ln\zeta(\sigma+2i\tau) \\ &= \sum_p \sum_{n\geq 1} \frac{e^{-\sigma n \ln p}}{n} \left(3 + 4e^{-i\tau n\ln p} + e^{-2i\tau n \ln p}\right). \end{aligned}$$

Now consider the real part of this expression:

$$\mathrm{Re}\, E = \sum_p \sum_{n\geq 1} \frac{3 + 4\cos(\tau n \ln p) + \cos(2\tau n \ln p)}{n p^{n\sigma}}.$$

Noting that $1 + \cos 2x = 2\cos^2 x$ (exercise 11.9), we obtain

$$\begin{aligned} \mathrm{Re}\, E &= \sum_p \sum_{n\geq 1} \frac{2 + 4\cos(\tau n \ln p) + 2\cos^2(\tau n \ln p)}{n p^{n\sigma}} \\ &= \sum_p \sum_{n\geq 1} \frac{2\left(1 + \cos(\tau n \ln p)\right)^2}{n p^{n\sigma}} > 0. \end{aligned}$$

But $\mathrm{Re}\, E > 0$ yields $|e^E| > 1$, which implies the lemma. ■

12.4. The Function $\Phi(z)$

The proof we will give of the prime number theorem (Theorem 12.15) really consists of inserting an analyticity property of the function Φ into Theorem 11.18 to prove the convergence of an improper integral. Here is the analyticity property we need.

Proposition 12.13. $\frac{\Phi(z)}{z} - \frac{1}{z-1}$ *has an analytic continuation in the closed half plane* $\operatorname{Re} z \geq 1$.

Proof. Taking a derivative with respect to z on both sides of the first equality of Lemma 12.10, we obtain

$$\frac{-\zeta'(z)}{\zeta(z)} = \sum_p \frac{\ln p \quad e^{-z\ln p}}{1 - e^{-z\ln p}} = \sum_p \frac{\ln p}{p^z - 1}.$$

To express this in terms of the function Φ, we use $\frac{1}{x-1} = \frac{1}{x} + \frac{1}{x(x-1)}$ to get

$$\frac{-\zeta'(z)}{\zeta(z)} = \sum_p \frac{\ln p}{p^z} + \sum_p \frac{\ln p}{p^z\,(p^z-1)}.$$

The first term on the right, of course, is $\Phi(z)$ (Definition 12.1). Subtracting the second term on the right, we see that

$$\begin{aligned}\frac{\Phi(z)}{z} - \frac{1}{z-1} &= \frac{-\zeta'(z)}{z\zeta(z)} - \frac{1}{z-1} - \frac{1}{z}\sum_p \frac{\ln p}{p^z\,(p^z-1)} \\ &= -\frac{(z-1)\zeta'(z) + z\zeta(z)}{z(z-1)\zeta(z)} - \frac{1}{z}\sum_p \frac{\ln p}{p^z\,(p^z-1)}.\end{aligned}$$

We tackle the first term on the right-hand side. From Proposition 12.11 (i), we obtain that both the numerator and the denominator are analytic on $\operatorname{Re} z > 0$. We only need to make sure the denominator does not have zeros in $\operatorname{Re} z \geq 1$. By Proposition 12.11 (ii), we know that it does not have a zero at $z = 1$. Lemma 12.12 says that it has no zeroes if $\operatorname{Re} z \geq 1$.

Next we look at the second term on the right-hand side. Since $\ln p$ is smaller than any positive power of p, the last term on the right-hand side is comparable to p^{-2z}. Since $\left|p^{-2z}\right| \leq p^{-2}$ whenever $\operatorname{Re} z \geq 1$, it converges uniformly in that region and is thus analytic in the desired region (Proposition 11.14 (ii)). ■

12.5. The Prime Number Theorem

Here we first prove that the prime number theorem is equivalent to the existence of a certain improper integral. Then we use the Tauberian theorem to prove that that integral exists. That will finally establish the prime number theorem.

Lemma 12.14. *We have*

$$i) \qquad \int_1^\infty \frac{\theta(y)-y}{y^2}\,dy \quad \textit{exists} \implies \lim_{x\to\infty}\frac{\theta(x)}{x} = 1\,.$$

$$ii) \qquad \lim_{x\to\infty}\frac{\theta(x)}{x} = 1 \quad \iff \quad \lim_{x\to\infty}\frac{\pi(x)}{x/\ln x} = 1\,.$$

Proof. We first prove (i). Suppose that the conclusion of the lemma does not hold. Then for some $\varepsilon > 0$ either there is a sequence of x_i such that $\lim_{i\to\infty} x_i = \infty$ with $\theta(x_i) > (1+\varepsilon)x_i$ or the same holds with $\theta(x_i) < (1-\varepsilon)x_i$.

Let us assume the former. Since θ is monotone, we have for all i

$$\int_{x_i}^{(1+\varepsilon)x_i} \frac{\theta(y)-y}{y^2}\,dy > \int_{x_i}^{(1+\varepsilon)x_i} \frac{(1+\varepsilon)x_i - y}{y^2}\,dy = -(1+\varepsilon)x_i y^{-1} - \ln y\Big|_{x_i}^{(1+\varepsilon)x_i}\,.$$

The latter can easily be worked out and yields $\varepsilon - \ln(1+\varepsilon)$ for each i. Since this is strictly greater than 0 by exercise 10.9, $I(s) = \int_1^s \frac{\theta(y)-y}{y^2}\,dy$ cannot converge to a fixed value as s tends to infinity.

The proof of non-convergence if $\theta(x_i) < (1-\varepsilon)x_i$ is almost identical (exercise 12.17).

To prove (ii), we use Lemma 12.2 to establish that

$$\left|\pi(x) - \frac{\theta(x)}{\ln x}\right| = \int_2^x \frac{\theta(t)}{t\,(\ln t)^2}\,dt\,.$$

Next we use (12.6) to get rid of the $\theta(x)$ in the integrand, and subsequently (12.8) to estimate the remaining integral. For large x, this gives

$$\left|\pi(x) - \frac{\theta(x)}{\ln x}\right| \le \ln 4\,\frac{x}{(\ln x)^2}\,(1+\varepsilon)\,,$$

for any $\varepsilon > 0$. Now we multiply both sides by $\ln x/x$ to obtain the result. ■

So this lemma implies that to prove the prime number theorem at this point, we need to show that $\int_1^\infty \frac{\theta(x)-x}{x^2}\,dx = \int_0^\infty (\theta(e^t)e^{-t}-1)\,dt$ exists. We restate Theorem 2.21 in its full glory.

Theorem 12.15 (Prime Number Theorem). *We have*

$$\text{1)}\ \lim_{x\to\infty}\frac{\pi(x)}{(x/\ln x)} = 1 \qquad \text{and} \qquad \text{2)}\ \lim_{x\to\infty}\frac{\pi(x)}{\int_2^x \ln t\,dt} = 1\,.$$

Proof. The equivalence of parts (1) and (2) is due to the fact that L'Hopital's rule implies that $\lim_{x\to\infty}\frac{x(\ln x)^{-1}}{\int_2^x(\ln t)^{-1}\,dt} = 1$. Thus, for example,

$$\lim_{x\to\infty}\frac{\pi(x)}{\int_2^x \ln t\,dt} = \lim_{x\to\infty}\frac{\pi(x)}{x/\ln x}\ \frac{x/\ln x}{\int_2^x \ln t\,dt}\,.$$

The same reasoning works vice versa (exercise 12.10).

So we only need to prove part (1). Lemma 12.3 gives

$$\frac{\Phi(z+1)}{z+1} - \frac{1}{z} = \int_0^\infty \left(\theta(e^t)e^{-t}-1\right)e^{-zt}\,dt\,.$$

Proposition 12.13 says that the left-hand side has an analytic continuation in $\mathrm{Re}\,z \geq 0$ while equation (12.6) says that $\theta(e^t)e^{-t}-1$ is bounded. But then, by Theorem 11.18, $\int_0^\infty (\theta(e^t)e^{-t}-1)\,dt$ exists. Finally, Lemma 12.14 implies that then (1) holds. ■

12.6. Exercises

Exercise 12.1. Write out in full the computations referred to in the proofs of Lemmas 12.2 and 12.3.

Proposition 12.16 (Abel Summation). *For the sequence $\{a_n\}_{n=1}^\infty$, denote $A(x) = \sum_{n\leq x} a_n$. Then for any differentiable f, we have*

$$\sum_{n\leq x} a_n f(n) = A(x)f(x) - \int_1^x A(t)f'(t)\,dt\,.$$

Exercise 12.2. a) Show that for any small $\varepsilon > 0$,

$$\sum_{n\le x} a_n f(n) = \int_{1-\varepsilon}^{x} f(t)\, dA(t)\,.$$

b) Apply integration by parts to (a), to get

$$\sum_{n\le x} a_n f(n) = A(x)f(x) - \int_{1-\varepsilon}^{x} A(t)f'(t)\, dt\,.$$

(*Hint: you need that* $A(1-\varepsilon)=0$.)
c) Prove Proposition 12.16. (*Hint: you need that* $A(t)f'(t)$ *is finite and continuous at* $t=1$.)

Exercise 12.3. Recall the notation $\lfloor x\rfloor$ (floor) and $\{x\}$ (fractional part) from Definition 2.1.
a) Use Abel summation to show that

$$\sum_{n\le x} \frac{1}{n} = \frac{x-\{x\}}{x} + \int_1^x \frac{t-\{t\}}{t^2}\, dt\,.$$

(*Hint: set* $a_n = 1$ *and* $f(x)=\frac{1}{x}$.)
b) Use (a) to show that

$$\sum_{n\le x} \frac{1}{n} \; -1-\ln x = -\frac{\{x\}}{x} - \int_1^x \frac{\{t\}}{t^2}\, dt\,.$$

c) Use (a) to show that

$$\lim_{x\to\infty} \left| \sum_{n\le x} \frac{1}{n} \; -1-\ln x + \int_1^x \frac{\{t\}}{t^2}\, dt \right| = 0\,.$$

d) Show that the Euler-Mascheroni constant $\gamma := 1-\lim_{x\to\infty}\int_1^x \frac{\{t\}}{t^2}\, dt$ satisfies $1-\frac{\pi^2}{12} < \gamma < 1$. (*Hint: show that* $\int_n^{n+1} \frac{t-n-1/2}{t^2}\, dt$ *is negative. Then use exercise 2.24 (c) and the fact that* $\zeta(2)=\frac{\pi^2}{6}$. *Note: in fact* $\gamma \approx 0.577\cdots$. *At the time of this writing (2021), it is unknown whether* γ *is irrational.*)

Exercise 12.4. a) Follow exercise 12.3 to show that

$$\sum_{n\le x} \ln n = \lfloor x\rfloor \, \ln x - (x-1) + \int_1^x \frac{\{t\}}{t}\, dt\,.$$

b) Show that (a) implies that

$$\frac{1}{n}\,\frac{n^n}{e^n} < n! < n\,\frac{n^n}{e^n}\,.$$

(*Hint: use that the absolute value of the integral in (a) is less than* $\ln x$.)

Exercise 12.4 proves part of what is known as Stirling's formula , namely:

$$n! = \sqrt{2\pi n}\,\frac{e^n}{n^n}\left(1+\frac{1}{12n}+\cdots\right).$$

Exercise 12.5. a) Use Proposition 12.16 to show that for $\mathrm{Re}\, z > 1$

$$\zeta(z) = z \int_1^\infty \frac{\lfloor t \rfloor}{t^{z+1}}\, dt\,.$$

(*Hint: write* $a_n = 1$ *and* $f(x) = x^{-z}$.)
b) Use that $\lfloor t \rfloor = t - \{t\}$ to show that

$$\zeta(z) = \frac{z}{z-1} - z \int_1^\infty \frac{\{t\}}{t^{z+1}}\, dt = \frac{1}{z-1} + 1 - z \int_1^\infty \frac{\{t\}}{t^{z+1}}\, dt\,.$$

c) Use (b) to reprove Proposition 12.11. (*Hint: you need to prove analyticity of h in* $\mathrm{Re}\, z > 1$.)

Exercise 12.6. a) How many trailing zeros does 400! (in decimal notation) have? (*Hint: use the proof of Lemma 12.4 with* $p = 5$ *and* $p = 2$.)
b) How about $\binom{400}{200}$?

Exercise 12.7. Consider $E(x_1, x_2) := \lfloor x_1 + x_2 \rfloor - \lfloor x_1 \rfloor - \lfloor x_2 \rfloor$ as in the proof of Lemma 12.4 and show that $E \in \{0, 1\}$.

Exercise 12.8. a) In Theorem 12.7, show that we can take

$$a = \ln 2 - \frac{1}{2} \ln 3 \approx 0.14\,,$$

for all $x \geq 2$.
b) Establish numerically that

$$\frac{\ln x}{x} \int_2^x (\ln t)^{-2}\, dt < 1\,.$$

(*Note: an analytic estimate of this expression is tricky and the reward is modest. But enthusiastic students can try the following. Show that* $\int_2^x (\ln t)^{-2}\, dt - \frac{\ln x}{x}$ *has a maximum at* $x = e^2$. *Then give a rough estimate of the expression in (b) for that value of x. You will likely get a much worse estimate than 1.*)
c) Use (b) and equation 12.7 to show that $b = 5 \ln 2$ works for all $x \geq 2$.

Exercise 12.9. Suppose we had an "perfect" estimate for Lemma 12.5 of the form $\binom{n}{n/2} = c\, \frac{2^n}{\sqrt{n}}$ for some $c > 0$. Can you improve Theorem 12.7? (*Hint: no. Conclusion: we need a different method to make further progress.*)

In the next exercise, we prove the equivalence of Theorem 12.15 (a) and (b).

Exercise 12.10. a) Compute the derivative of $x(\ln x)^{-r}$ for $r > 0$.
b) Use (a) and L'Hopital to prove that for $r > 0$

$$\lim_{x\to\infty} \frac{\int_1^x (\ln t)^{-r}\, dt}{x(\ln x)^{-r}} = 1\,.$$

c) Use (b) to show that parts (a) and (b) of Theorem 12.15 are equivalent.
d) Compare (b) to (12.8).

In the next two problems we prove the following result.

Proposition 12.17. *Let p_n denote the nth prime. The prime number theorem is equivalent to*

$$\lim_{n\to\infty} \frac{p_n}{n \ln n} = 1\,.$$

Exercise 12.11. For this exercise, assume that $\lim_{x\to\infty} \frac{y}{x/\ln x} = 1$ and that $x \to \infty$ if and only if $y \to \infty$. (In fact, y stands for $\pi(x)$, and we know that $x \to \infty$ if and only if $\pi(x) \to \infty$, see Theorem 2.17.)
a) Suppose $\lim_{x\to\infty} f_i(x) = \infty$ and $\lim_{x\to\infty} \frac{f_1(x)}{f_2(x)} = 1$. Show that

$$\lim_{x\to\infty} \frac{\ln f_1(x)}{\ln f_2(x)} = 1\,.$$

(*Hint: for x large, $(1-\varepsilon) < \frac{f_1(x)}{f_2(x)} < (1+\varepsilon)$, multiply by $f_2(x)$, and take logarithms.*)
b) Show that $\lim_{x\to\infty} \frac{\ln\ln x}{\ln x} = 0$. (*Hint: substitute $x = e^{e^t}$.*)
c) Use the hypotheses and (a) to show that

$$\frac{x}{y\ln y} = \frac{x}{y\ln y}\;\frac{y}{x/\ln x}\;\frac{\ln y}{\ln x - \ln\ln x} = \frac{1}{1-\frac{\ln\ln x}{\ln x}}\,.$$

d) Use (b) to show that the limit in (c) as $x \to \infty$ tends to 1. Use the hypotheses to change to the limit as $y \to \infty$.
e) Show that (d) implies one way of Proposition 12.17.

Exercise 12.12. For this exercise, assume that $\lim_{y\to\infty} \frac{x}{y\ln y} = 1$ and that $x \to \infty$ if and only if $y \to \infty$. See exercise 12.11.
a) Follow exercise 12.11 in reverse to show that

$$\lim_{x\to\infty} \frac{y}{x/\ln x} = \lim_{x\to\infty} \frac{x}{y\ln y}\;\frac{y}{x/\ln x}\;\frac{\ln y}{\ln x - \ln\ln x} = \lim_{x\to\infty} \frac{1}{1-\frac{\ln\ln x}{\ln x}} = 1\,.$$

b) Show that (b) implies the other direction of Proposition 12.17.
c) Whereabouts is the nth prime located?

Exercise 12.13. In this exercise, we fix any $K > 1$ and $\{x_i\}_{i=1}^{\infty}$ is a sequence such that $\lim_{i\to\infty} x_i = \infty$. We also set $x' = Kx$ for notational ease.
a) Show that if $\pi(x_i') = \pi(x_i)$ and $\lim_{i\to\infty} \frac{\pi(x_i)}{x_i/\ln x_i}$ exists, then

$$\lim_{i\to\infty} \frac{\pi(x_i')}{x_i'/\ln x_i'} = \frac{1}{K} \lim_{i\to\infty} \frac{\pi(x_i)}{x_i/\ln x_i}\,.$$

b) Show that (a) and the prime number theorem imply that for large enough x, there are primes in $(x, x']$. (*Hint: if (a) holds, then there are no primes in* $[x_i, Kx_i]$.)
c) Show that in fact, the prime number theorem implies

$$\lim_{i\to\infty} \frac{\pi(x_i')}{\pi(x_i)} = K\,.$$

d) Show that (c) implies that for large enough x, there are approximately $(K-1)\pi(x)$ primes in $(x, x']$.

In fact, the following holds for all n. We omit the proof, which involves some careful computations. It can be found in [**1**].

Proposition 12.18 (Bertrand's Postulate). *For all $n \geq 2$ there is a prime in the interval $[n, 2n)$.*

The same reference [**1**] also mentions an open (in 2018) problem in this direction: *Is there always a prime between n^2 and $(n+1)^2$?*

Exercise 12.14. a) Show for every $m \in \mathbb{N}$, the set $\{m!+2, \cdots, m!+m\}$ contains no primes. (*Hint: for $2 \leq j \leq m$ we have $j \mid (m!+j)$.*)
b) Show that from Proposition 12.17, we might reasonably expect the "expected" prime gap $p_{n+1} - p_n$ to be equal to

$$G_n := (n+1)\ln(n+1) - n\ln n \approx \ln((n+1)e)\,,$$

if n large.
c) Use the prime number theorem to show that

$$G_n \approx \ln p_{n+1} - \ln\ln p_{n+1} + 1 \approx \ln p_{n+1}\,.$$

d) Assume the twin prime conjecture to show that $\frac{p_{n+1}-p_n}{\ln p_{n+1}}$ does *not* converge to a limit. See also Figure 77.
d) Use lemma 12.14 to show that the prime number theorem is equivalent to saying that the sum of the first n "expected" prime gaps equals p_{n+1}.

Exercise 12.15. a) Show for every $m \in \mathbb{N}$, the set $\{m!+2, \cdots, m!+m\}$ contains no primes. (*Hint: for $2 \leq j \leq m$ we have $j \mid (m!+j)$.*)
b) Compare that prime gap at $p_n \sim m!$ with the gap you expect from exercise 12.14. (*Hint: use exercise 12.4 (b).*) (

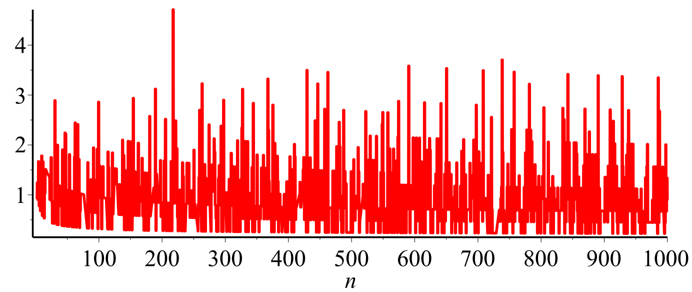


Figure 77. The prime gaps $p_{n+1} - p_n$ divided by $\ln p_{n+1}$ for n in $\{1, \cdots, 1000\}$.

Exercise 12.16. We give a different proof of Lemma 12.14 (ii) (following [**75**]).
a) Show that $\theta(x) \leq \pi(x) \ln x$.
b) Show that

$$(1-\varepsilon)\ln x \sum_{x^{1-\varepsilon} \leq p \leq x} 1 \leq \sum_{x^{1-\varepsilon} \leq p \leq x} \ln p \leq \theta(x)\,.$$

c) Show that for all $\varepsilon > 0$

$$\pi(x) - x^{1-\varepsilon} \leq \sum_{x^{1-\varepsilon} \leq p \leq x} 1\,.$$

d) Use (a), (b), and (c) to show that for all $\varepsilon > 0$

$$(1-\varepsilon)\frac{(\pi(x) - x^{1-\varepsilon})\ln x}{x} \leq \frac{\theta(x)}{x} \leq \frac{\pi(x)\ln x}{x}\,.$$

e) Use (d) to prove Lemma 12.14 (ii). (*Hint: show that* $\lim_{x\to\infty} x^{-\varepsilon} \ln x = 0$ *by substituting* $x = e^t$.)

Exercise 12.17. a) Suppose that in the proof of Lemma 12.14 there is a sequence of x_i such that $\lim_{i\to\infty} x_i = \infty$ with $\theta(x_i) < (1-\varepsilon)x_i$ for some $\varepsilon > 0$. Show that the integral in the lemma cannot converge.
b) How about if *both* occur and alternate?

We define two new functions. This definition usually accompanies Definition 12.1.

Definition 12.19. *The von Mangoldt function is given by*

$$\Lambda(n) := \begin{cases} \ln p & \text{if } n = p^k \text{ where } p \text{ is prime and } k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

The second Chebyshev function is given by

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

Just like the first Chebyshev function $\theta(x)$, the second Chebyshev function $\psi(x)$ is often used as a more tractable version of the prime counting function $\pi(x)$. In particular, in exercises 12.18 and 12.19, we will prove a lemma similar to Lemma 12.14, namely

Lemma 12.20. *We have*

$$\lim_{x\to\infty} \frac{\psi(x)}{x} = 1 \iff \lim_{x\to\infty} \frac{\pi(x)}{x/\ln x} = 1\,.$$

Exercise 12.18. a) Show that $\psi(x) = \sum_{p^k\leq x} \ln p$. (*Hint: from Definition 12.19. Note that this means that ψ counts all prime* powers *no greater than x.*)
b) Show that $\psi(x) = \sum_{p\leq x} \ln p \left\lfloor \frac{\ln x}{\ln p} \right\rfloor$. (*Hint: this expression only increases at x a power of a prime.*)
c) Show that $\psi(x) \leq \sum_{p\leq x} \ln x$. (*Hint: $\lfloor a \rfloor \leq a$.*)
d) Show that (c) implies that $\psi(x) \leq \pi(x)\ln x$.

Exercise 12.19. a) Show that Definitions 12.1 and 12.19 imply that $\theta(x) \leq \psi(x)$.
b) Use (a) and exercises 12.16 (d) and 12.18 (d) to show that

$$(1-\varepsilon)\frac{(\pi(x)-x^{1-\varepsilon})\ln x}{x} \leq \frac{\theta(x)}{x} \leq \frac{\psi(x)}{x} \leq \frac{\pi(x)\ln x}{x}\,.$$

c) Use (b) and Lemma 12.14 (ii) to prove Lemma 12.20.

Exercise 12.20. Plot $\theta(x)/x$, $\psi(x)/x$, and $\pi(x)\ln x/x$ in one figure. (See for example, Figure 78). Compare with exercise 12.18. b) Show that all three tend to 1 as x tends to infinity.

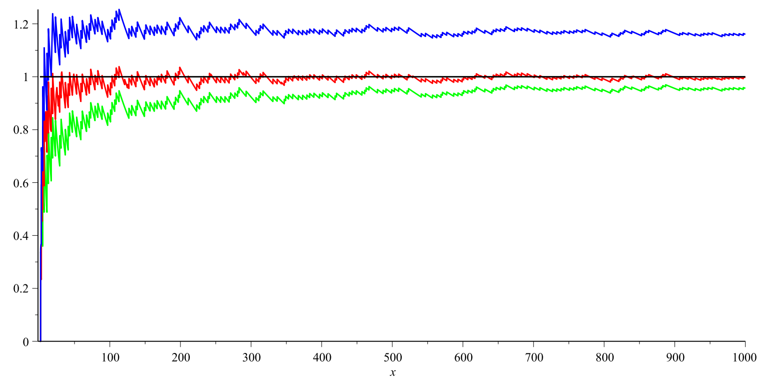


Figure 78. The functions $\theta(x)/x$ (green), $\psi(x)/x$ (red), and $\pi(x)\ln x/x$ (blue) for $x \in [1, 1000]$. All converge to 1 as x tends to infinity. The x-axis is horizontal.

Exercise 12.21. The analysis of this exercise should be compared to the proof of Proposition 12.13.
a) Use Lemma 12.10 and Corollary 11.12 to show that

$$\frac{-\zeta'(z)}{\zeta(z)} = \sum_p \sum_{n=1}^{\infty} \frac{\ln p}{p^{nz}}$$

is analytic for $\operatorname{Re} z > 1$.
b) Use Definition 12.19 to show that for $\operatorname{Re} z > 1$

$$\frac{-\zeta'(z)}{\zeta(z)} = \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^z}\,.$$

c) Use Abel summation (Proposition 12.16) and Definition 12.19 to show that for $\operatorname{Re} z > 1$

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

(*Hint: in the proposition, set $f(x) = x^{-z}$ and $A(x) = \psi(x)$. Then use that in the boundary term, $\psi(x)/x$ converges to 1.*)
d) Subtract $z/(z-1)$ from (c) and divide by z to conclude that for $\operatorname{Re} z > 1$

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

Exercise 12.22. Show that $\lim_{x\to\infty} a(x)/x = 1$ is equivalent to the following. For all $\varepsilon > 0$, we have $|a(x) - x| < \varepsilon x$ for x large enough.

Exercise 12.23. For this problem, we assume that there is a $\theta \in (1/2, 1)$ so that $|\psi(x) - x| \le Kx^{\theta}$.
a) Note (exercise 12.22) that this is stronger than $\lim_{x\to\infty} \psi(x)/x = 1$.
b) Use exercise 12.21 to show that

$$\frac{-\zeta'(z)}{z\zeta(z)} - \frac{1}{z-1} = \sum_{n\ge 1} \int_n^{n+1} \frac{\psi(x)-x}{x^{z+1}}\,dx\,.$$

c) Show that our hypothesis for this exercise implies that

$$\left|\int_n^{n+1} \frac{\psi(x)-x}{x^{z+1}}\,dx\right| \le 2Kn^{\theta-\mathrm{Re}\,z-1}\,.$$

d) Use Proposition 11.14 to show that the right hand side of (b) is analytic for $\mathrm{Re}\,z > \theta$.
e) Show that (d) implies that $\zeta(z)$ has no zeros in $\mathrm{Re}\,z > \theta$.

In the next two problems, we prove a second version of the Tauberian theorem in Chapter 11. This is essentially just a reformulation of Theorem 12.15 (2), but with $\theta(n)$ replaced by an arbitrary sequence a_n satisfying certain conditions. The proof is also essentially the same.

Theorem 12.21. *Suppose $a_n \ge 0$ so that there is a $K > 0$ with $A(x) := \sum_{n\le x} a_n \le Kx$. Define*

$$G(z) := \sum_{n=1}^{\infty} \frac{a_n}{n^z}\,.$$

G is analytic on $\mathrm{Re}\,z > 1$. Assume also that G admits an analytic continuation to $\mathrm{Re}\,z \ge 1$ except for a simple pole at 1 with residue 1. Then

$$\lim_{x\to\infty} \frac{A(x)}{x} = 1\,.$$

Exercise 12.24. a) Show that G is analytic on $\mathrm{Re}\,z > 1$. (*Hint: use the condition on $A(x)$ and Proposition 11.14 (ii).*)
b) Use (a) and Abel summation to show that

$$G(z) = z\int_1^{\infty} A(x)x^{-1-z}\,dx\,.$$

c) Show that

$$G(z) - \frac{z}{z-1} = z\int_1^{\infty} \frac{A(x)-x}{x^{1+z}}\,dx = z\int_0^{\infty} \left(A(e^t) - e^t\right) e^{-zt}\,dt\,.$$

d) In (c), set $z' + 1 = z$ and then drop the prime to show that

$$H(z) := \frac{G(1+z)}{z+1} - \frac{1}{z} = \int_0^{\infty} \frac{A((x)-x}{x^{2+z}}\,dx = \int_0^{\infty} \left(A(e^t)e^{-t} - 1\right) e^{-zt}\,dt\,.$$

Exercise 12.25. a) Show that the function $H(z)$ of exercise 12.24 (d) has an analytic continuation to $\operatorname{Re} z \geq 0$. (*Hint: the pole at $z = 0$ has been canceled by the subtraction of $1/z$.*)
b) Use Theorem 11.18 to show that $\int_0^\infty \frac{A(x)-x}{x^2}\,dx$ converges.
c) Use Lemma 12.14 (i) to show that $\lim_{x\to\infty} \frac{A(x)}{x} = 1$.

Exercise 12.26. a) Show that

$$\lim_{n\to\infty} \left(\prod_{p\leq n} p \right)^{\frac{1}{n}} = e$$

if and only the prime number theorem holds. (*Hint: see Lemma 12.14 (ii).*)
b) See Figure 79). Show that

$$\lim_{n\to\infty} \left(\operatorname{lcm}(1,2,\cdots,n)\right)^{\frac{1}{n}} = e$$

if and only the prime number theorem holds. (*Hint: see Lemma 12.20.*)

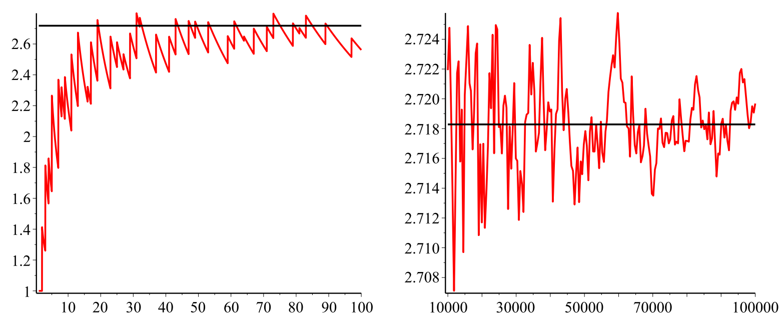


Figure 79. Plot of the function $f(n) := (\operatorname{lcm}(1,2,\cdots,n))^{\frac{1}{n}}$ for n in $\{1,\cdots,100\}$ (left) and in $\{10^4,\cdots,10^5\}$ (right). The function converges to e indicated in the plots by a line.

Chapter 13

Primes in Arithmetic Progressions

Overview. An arithmetic progression is a set S of the form $S := \{a + kq \mid k \in \mathbb{N}\}$. If $\gcd(a,q) = d > 1$, then any two distinct numbers in S are not coprime. Thus, in that case, S can contain at most one prime. We will see that asymptotically the primes are distributed equally over the remaining arithmetic progressions, namely the sets $\{a+kq \mid k \in \mathbb{N}\}$ such that $\gcd(a,q) = 1$. One of the more accessible introductions to the material in this chapter is [**4**]. For section 13.6, we used [**65**] and [**25**].

13.1. Finite Abelian Groups

Definition 13.1. *Two groups g and H are isomorphic if there exists a bijective homomorphism $f : G \to H$ (see also exercise 13.3).*

Definition 13.2. *A cyclic group is a group generated by a single element.*

The proof of the following proposition is loosely based on the analogous proof in [**30**][appendix 3C]. It uses the simple observation that every element g of a finite Abelian group generates a cyclic group. This is evident, because the sequence $\{g^i\}$ can have finitely many distinct elements, and so the smallest value of $i \geq 0$ where a repeated value occurs must be the order o of the element g.

Proposition 13.3 (Fundamental Theorem of Finite Abelian Groups). *Any finite Abelian group G of order n is isomorphic to a cartesian product of finite cyclic groups $\mathbb{Z}_{o_1}^+ \times \cdots \times \mathbb{Z}_{o_r}^+$. Furthermore, $\prod_{i=1}^r o_i = n$.*

Proof. There are finitely many ways of choosing a non-empty subset S of elements of G. Since each element has order at most n, for each of these subsets, we can find out whether it generates G. Let r be the minimal cardinality of the subsets that generate G and denote by $\mathscr{S}_r$ the (non-empty) collection of all such sets of generators of cardinality r.

Pick S in $\mathscr{S}_r$, denote its elements by g_i, and the order of $g_i \in S$ by $o_i(S)$. By construction, there is a map $\sigma(S)$ from $\prod_{i=1}^r \{0,1,\cdots,o_i(S)-1\} = \mathbb{Z}_{o_1}^+ \times \cdots \times \mathbb{Z}_{o_r}^+$ *onto* G given by

$$\sigma \,:\, (a_1,\cdots,a_r) \to \prod_{i=1}^r g_i^{a_i}\,.$$

Now let us *assume* that there is a non-empty set $\overline{\mathscr{S}}_r \subseteq \mathscr{S}_r$ so that for S in $\overline{\mathscr{S}}_r$, $\sigma(S)$ is *not* a bijection. We will show that this leads to a contradiction.

For S in $\overline{\mathscr{S}}_r$, there are i and $0 \leq a_i, a_i' < o_i(S)$ such that

$$\prod_{i=1}^r g_i^{a_i} = \prod_{i=1}^r g_i^{a_i'} \quad \Longleftrightarrow \quad \prod_{i=1}^r g_i^{(a_i-a_i') \bmod o_i(S)} = \prod_{i=1}^r g_i^{c_i} = 1\,,$$

where we have set c_i equal to the least residue of $(a_i - a_i')$ modulo $o_i(S)$. Note that in this expression at least two of the coefficients c_i are greater than 0. Now let

$$s(S) := \min_{c_i \in \{0,\cdots,o_i(S)-1\}} \left\{ \sum_{i=1}^r c_i \,:\, \prod_{i=1}^r g_i^{c_i} = 1 \right\} \geq 2\,.$$

Finally, minimize $s(S)$ over $\overline{\mathscr{S}}_r$

$$s_- := \min_{S \in \overline{\mathscr{S}}_r} s(S) \geq 2\,. \tag{13.1}$$

Let $\{g_i\}_{i=1}^r$ be the collection of generators at which this minimum is assumed. At least two of the c_i's are greater than 0, say, $c_2 \geq c_1 > 0$. Define $f_1 = g_1 g_2$ and $f_i = g_i$ for all $i > 1$. This change of variables is invertible, so $\{h_i\}_{i=1}^r$ still generate G. A simple calculation gives

$$1 = \prod_{i=1}^r g_i^{c_i} = f_1^{c_1} f_2^{c_2-c_1} f_3^{c_3} \cdots f_r^{c_r}\,.$$

Thus s_- has decreased, which contradicts (13.1). This shows that $\overline{\mathscr{S}}_r$ is empty and thus for all S in $\mathscr{S}_r$, $\sigma(S)$ is a bijection.

It is also a homomorphism, because for a and a' in $\mathbb{Z}^+_{o_1} \times \cdots \times \mathbb{Z}^+_{o_r}$

$$\sigma(a+a') = \prod_{i=1}^{r} g_i^{a_i} \cdot \prod_{i=1}^{r} g_i^{a'_i}.$$

Thus σ is an isomorphism (see exercise 13.3). Clearly, the number of elements in G equals $\prod_{i=1}^{r} o_i(S)$ which must therefore be equal to n. ■

13.2. The Hermitian Inner Product

Later on, we will briefly need to consider $V = \mathbb{C}^n$ as an inner product space. The Hermitian inner product generalizes the dot product of $\mathbb{R}^n$.

Definition 13.4. *The (standard) Hermitian inner product on $\mathbb{C}^n$ is given by*

$$(x,y) = \bar{x}_1 y_1 + \cdots + \bar{x}_n y_n\,,$$

where $\bar{y}$ indicates the complex conjugate of y.

One easily checks that this binary operation satisfies the requirements that for all x, y, and z in V and α in $\mathbb{C}$

1)	$(x,x) \geq 0$	positivity
2)	$(x,x) = 0 \Longleftrightarrow x = 0$	definiteness
3)	$(x, \alpha u + v) = \alpha(x,u) + (x,v)$	linearity
4)	$(x,y) = \overline{(y,x)}$	conjugate symmetry

More generally, any function $V \times V$ that satisfies these requirements is called an inner product, but we will not be needing that generality here.

Definition 13.5. *A set $\{e_i\}_{i=1}^n$ of vectors in V is an orthonormal basis if for all $i \neq j$ and all x in V*

1)	$(e_i, e_i) = 1$	*unit vectors*
2)	$(e_i, e_j) = 0$	*orthogonality*
3)	$\exists\, \alpha_i$ such that $x = \sum_{i=1}^n \alpha_i e_i$	*basis*

The property that is crucial for us is that the α_i in item (3) of this definition can be computed easily, namely

$$x = \sum_{i=1}^{n} (e_i, x) e_i \,. \tag{13.2}$$

For more details and a good general introduction, see [**7**][Chapter 6].

13.3. Characters of Finite Abelian Groups

Definition 13.6. *A character of a group G is a complex-valued homomorphism $f : G \to \mathbb{C}^\times$.*

For example, the identity e of a group G satisfies $e^2 = e$ and since f is a multiplicative homomorphism, we have that $f(e)^2 = f(e)$ and so $f(e) = 0$ or $f(e) = 1$. The former is excluded because 0 is not in the domain of $\mathbb{C}^\times$. Thus $f(e) = 1$ for any character. An example of a character of G is the constant function, $f(g) = 1$, also called the principal character. We indicate it by f_0.

Remark 13.7. From now on, we will denote the field of units in $\mathbb{Z}_b$ by $\mathbb{Z}_b^\times$. Sometimes we will consider it as multiplicative groups (as directly below), and other times we may consider them a field. See Section 5.4.

Before continuing, let us look at a few examples of characters, namely $G = \mathbb{Z}_5^\times$ and $G = \mathbb{Z}_8^\times$.

mod 5	f_0	f_1	f_2	f_3
1	1	1	1	1
2	1	i	-1	-i
3	1	-i	-1	i
4	1	-1	1	-1

mod 8	$f_{(0,0)}$	$f_{(0,1)}$	$f_{(1,0)}$	$f_{(1,1)}$
1	1	1	1	1
3	1	1	-1	-1
5	1	-1	1	-1
7	1	-1	-1	1

(13.3)

The table on the left lists the characters of $\mathbb{Z}_5^\times$. Each column corresponds to a different character. The table on the right lists the characters of $\mathbb{Z}_8^\times$. Note that each of these groups has four characters, but they are not the same.

How do we determine these characters? The short answer is: exploit multiplicativity. First look at $\mathbb{Z}_5^\times$. We note it is a cyclic group generated by the element 2, namely 2^k mod 5 cycles through the values 2, 4, 3, and 1 for

$k \in \{1,2,3,4\}$. Thus $f(2^4) = (f(2))^4 = 1$, and so for any character f, the value of $f(2)$ must be a 4th root of unity. So choose (as in the left table of (13.3))

$$f_m(2) = e^{2\pi i \frac{m}{4}}.$$

For any choice of m, we can obtain a multiplicative function as follows

$$e^{2\pi i(k+\ell)\frac{m}{4}} = e^{2\pi ik\frac{m}{4}} e^{2\pi i\ell\frac{m}{4}} \implies f_m(2^{k+\ell}) = f_m(2^k) f_m(2^\ell). \tag{13.4}$$

This example is wonderful, because it turned out that $\mathbb{Z}_5^\times$ is isomorphic to $\mathbb{Z}_4^+$ which simplifies things: we get something very reminiscent of a discrete Fourier transform (see Definition 13.27).

The group $\mathbb{Z}_8^\times$ also has 4 elements, namely $\{1,3,5,7\}$. But none of these elements has order 4, for $3^2 =_8 5^2 =_8 7^2 =_8 1$. Thus for any character f, each of $f(3)$, $f(5)$, and $f(7)$ must be square roots of unity. This group is therefore *not* isomorphic to $\mathbb{Z}_4^+$. However, consider

(a_1,a_2)	(0,0)	(0,1)	(1,0)	(1,1)
$3^{a_1} \cdot 5^{a_2}$ mod 8	1	5	3	7

This gives a bijection $h : \mathbb{Z}_8^\times \to \mathbb{Z}_2^+ \times \mathbb{Z}_2^+$. But in $\mathbb{Z}_2^+ \times \mathbb{Z}_2^+$,

$$\begin{aligned} h(3^{a_1}5^{a_2})h(3^{b_1}5^{b_2}) &= (a_1,a_2)(b_1,b_2) = (a_1+b_1, a_2+b_2) \\ h(3^{a_1}5^{a_2} \cdot 3^{b_1}5^{b_2}) &= (a_1+b_1, a_2+b_2). \end{aligned}$$

It also shows that h is a homomorphism, and thus $\mathbb{Z}_8^\times$ is isomorphic to $\mathbb{Z}_2^+ \times \mathbb{Z}_2^+$. So let $m = (m_1, m_2)$ where $m_i \in \{0,1\}$ and set

$$f_m(3) = e^{2\pi i \frac{m_1}{2}} \quad \text{and} \quad f_m(5) = e^{2\pi i \frac{m_2}{2}}.$$

So that (as illustrated in the right table of (13.3))

$$f_m(3^{a_1}5^{a_2}) = e^{2\pi i \frac{a_1 m_1}{2}} e^{2\pi i \frac{a_2 m_2}{2}} = e^{2\pi i\left(\frac{a_1 m_1}{2} + \frac{a_2 m_2}{2}\right)}. \tag{13.5}$$

f_m is multiplicative by the same calculation as done in (13.4), but now separated out in 'components' to prove that

$$f_m(3^{k_1+k_2}5^{\ell_1+\ell_2}) = f_m(3^{k_1}5^{\ell_1}) f_m(3^{k_2}5^{\ell_2}). \tag{13.6}$$

The student is asked to provide a few more details in exercise 13.1.

These computations tell us what is going on. We first simplify the notation, and then formulate the relevant theorem.

Definition 13.8. *For the remainder of this chapter, we abbreviate:*

$$\begin{aligned} \mathbb{Z}_o^+ &:= \mathbb{Z}_{o_1}^+ \times \cdots \times \mathbb{Z}_{o_r}^+ \,; \\ n &:= \prod_{i=1}^{r} o_i \,; \end{aligned}$$

and for a and m in $\mathbb{Z}_o^+$, we set $m/o := (\frac{m_1}{o_1}, \cdots, \frac{m_r}{o_r})$ and

$$\begin{aligned} g^a &:= \prod_{j=1}^{r} g_j^{a_j} \,; \\ a \cdot (m/o) &:= \sum_{i=1}^{r} \frac{a_j m_j \mod o_j}{o_j} \,. \end{aligned}$$

Theorem 13.9. *Let G be an n element Abelian group. With the notation of Definition 13.8, we have:*
i) The characters f_m of G are given by

$$f_m(g^a) = e^{2\pi i a \cdot (m/o)} \,.$$

ii) The characters f_m are all orthogonal to one another in the sense that:

$$\forall\, m\,,\ell\, \in \mathbb{Z}_o^+ \;:\; \sum_{a \in \mathbb{Z}_o^+} \overline{f_m(g^a)} f_\ell(g^a) = \begin{cases} n & \text{if } m = \ell \\ 0 & \text{if } m \neq \ell \end{cases}$$

(see Figure 80). Thus the n characters are all distinct.

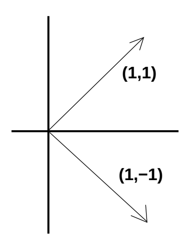


Figure 80. The two characters modulo 3 illustrate the orthogonality of the Dirichlet characters.

Proof. By complete multiplicativity (because f is a homomorphism), any character f is completely determined by its value $f(g_i)$ on the r generators of G. So

$$f\left(\prod_{j=1}^{r} g_j^{a_j}\right) = \prod_{j=1}^{r} f(g_j)^{a_j}\,.$$

$f(g_j)$ must be an o_jth root of unity, and is thus equal to $\exp(2\pi i m_j/o_j)$. The second statement follows using Definition 13.8.

$$\sum_{a\in\mathbb{Z}_o^+} \overline{f_\ell(g^a)} f_m(g^a) = \sum_{a\in\mathbb{Z}_o^+} e^{2\pi i a\cdot((m-\ell)/o)}\,. \tag{13.7}$$

If $m=\ell$, the exponent is zero and so $\sum_{a\in\mathbb{Z}_o^+} 1 = n$. The above sum is really a sum of products (Definition 13.8) which can be converted into a product of sums (exercise 4.3) of the form

$$\sum_{a_j\in\mathbb{Z}_{o_j}^+} e^{2\pi i \frac{a_j(m_j-\ell_j)k_j}{o_j}}\,.$$

So if $m\neq\ell$, then there is an j such that $m_j-\ell_j\neq 0$. The above sum then has the form

$$\sum_{a=0}^{o-1} e^{2\pi i \frac{aK}{o}} = \frac{e^{2\pi i \frac{oK}{o}}-1}{e^{2\pi i \frac{K}{o}}-1} = 0\,,$$

and so the product of the sums also reduces to zero. ■

Theorem 13.9 implies that there is an injection from $\mathbb{Z}_o^+$ to the characters given by $F: m \to f_m$. It is actually a bijection, because an injection between sets of the same size — namely $\{m\}$ and $\{f_m\}$ — must be a bijection. A slight variation on equation (13.7) allows us to go a little further, namely

$$f_{m+\ell}(g^a) = f_m(g^a) f_\ell(g^a)\,.$$

Thus the bijection becomes a group homomorphism. Using Proposition 13.3, we obtain the following corollary.

Corollary 13.10. *The characters of a finite Abelian group G together with the multiplication $(f_m f_n)(g^a) = f_m(g^a) f_n(g^a)$ form a group that is isomorphic to G which in turn is isomorphic to $\mathbb{Z}_o^+$.*

There is another interesting way to look at these characters. Order the elements of $\mathbb{Z}_o^+$ by defining some bijection, or counter, φ from $\mathbb{Z}_o^+$ to $\{1,\cdots,n\}$. We can then think of $f_m(g^a)$ as the $\varphi(a)$th component of the vector f_m in $\mathbb{C}^n$. This is what we did in the tables (13.3). Theorem 13.9 implies that the *vectors* f_m now form an orthogonal basis of $\mathbb{C}^n$ equipped with the Hermitian inner product (Definition 13.4). Reformulating the theorem gives yet another corollary. See Definition 13.27 and the exercises that follow it for more details.

Corollary 13.11. *If we define the vectors e_m as $n^{-1/2} f_m$, then the set $\{e_m\}_{m\in\mathbb{Z}_o^+}$ is an orthonormal basis (Definition 13.5) of $\mathbb{C}^n$.*

13.4. Dirichlet Characters and L-functions

The Dirichlet characters are essentially the characters of $\mathbb{Z}_q^\times$, the multiplicative group of the reduced residues of $\mathbb{Z}_q$ ($\mathbb{Z}$ modulo q) with identity element 1. Since we will use Dirichlet characters as the coefficients in Dirichlet series, we need to convert them into arithmetic functions.

Definition 13.12. *Corresponding to each character $f : \mathbb{Z}_q^\times \to \mathbb{C}^\times$, we define a q-periodic arithmetic function χ_f, the Dirichlet character modulo q, as follows:*

$$\begin{cases} \chi_f(n) = f(\mathrm{Res}_q(n)) & \text{if } \gcd(n,q) = 1 \\ \chi_f(n) = 0 & \text{if } \gcd(n,q) > 1 \end{cases}$$

By Corollary 13.10, these characters form a multiplicative subgroup of $\mathbb{C}$ that we will denote by X_q.

Recall that the *principal* Dirichlet character evaluates to 1 on numbers relatively prime to q and equals 0 elsewhere. It will be denoted by χ_{f_0} or $\underline{\chi_1}$.

More generally, it is easy to see that the Dirichlet characters are completely multiplicative (Definition 4.2) arithmetic (Definition 4.1) functions. For if $\gcd(ab,q) > 1$, then $\gcd(a,q) > 1$ or $\gcd(b,q) > 1$ (or both). And so from Definition 13.12, we see that then $\chi(ab) = \chi(a)\chi(b) = 0$. On the other hand, if both $\gcd(a,q) = 1$ and $\gcd(b,q) = 1$, then since f is a homomorphism, $\chi(ab) = \chi(a)\chi(b)$. That means that for any Dirichlet character χ, we get $\chi(1) = 1$.

Remark 13.13. Since from now on, we will only deal with Dirichlet characters modulo $q \in \mathbb{N}$, we will, in the interest of brevity, refer to these simply as characters from now on.

Definition 13.14. *The Dirichlet L-series associated to a Dirichlet character χ is defined as*

$$L(\chi,z) := \sum_{n=1}^{\infty} \frac{\chi(n)}{n^z}\,.$$

The Dirichlet L-function associated to a Dirichlet character χ is the function defined by the analytic continuation of the Dirichlet L-series.

Often these are abbreviated to L-series and L-function , though some authors reserve those names for generalizations of those notions.

These L-function have the "feel" of a zeta function as the next result indicates. We will use a complicated combination of L-functions as a "new" zeta function to prove our main theorem. In the remainder of this chapter, we abbreviate *the function f has a well-defined analytic continuation in the region S* by *f is analytic in S*.

Proposition 13.15. *If ψ is bounded and completely multiplicative, then $L(\psi,z)$ is analytic in* $\operatorname{Re} z > 1$ *and*

$$\ln\left(\sum_{n=1}^{\infty} \psi(n)\,n^{-z}\right) = \ln L(\psi,z) = -\sum_{p \text{ prime}} \ln(1-\psi(p)p^{-z}) = \sum_{p}\sum_{n=1}^{\infty} \frac{\psi(p^n)}{n p^{nz}}\,.$$

If ψ is periodic and has average zero, then $L(\psi,z)$ is analytic in $\operatorname{Re} z > 0$.

Proof. The first equality follows from the definition of L. We paraphrase the second proof of Proposition 2.20. Using the complete multiplicativity of ψ, we obtain

$$\psi(2)2^{-z}L(\psi,z) = \sum_{n=1}^{\infty} \psi(2)\psi(n)\,2^{-z}n^{-z} = \sum_{n=1}^{\infty} \psi(2n)\,(2n)^{-z}\,.$$

Thus

$$\left(1-\psi(2)2^{-z}\right)L(\psi,z) = \sum_{2\nmid n} \psi(n)n^{-z}\,.$$

Subsequently we multiply this expression by $(1-\psi(3)3^{-z})$. This has the effect of removing multiples of 3 from the remaining terms. Continuing like this, it follows that eventually[1]

$$\left(\prod_{p \text{ prime}} (1-\psi(p)p^{-z})\right)L(\psi,z) = 1\,.$$

[1]Note that we use factorization in terms of primes here

Upon taking the logarithm, we arrive at the second equality. The third one — and analyticity — follows from Lemma 12.10.

To prove the last part, we use Proposition 12.16 and compute

$$L(\psi,z) = \sum_{n\le x} \psi(n)\, n^{-z} = \Psi(x)x^{-z} + z\int_1^x \Psi(t)t^{-z-1}\,dt\,,$$

where $\Psi(x) = \sum_{n\le x}\psi(n)$. Since ψ has period, say, q with average 0, we have $\Psi(x+q) = \Psi(x)$, and so Ψ is bounded. Thus both terms in the above equation converge for $\mathrm{Re}\, z > 0$. ■

13.5. Preliminary Steps

The way we want to prove the prime number theorem for arithmetic progressions is by defining an arithmetic function $h_{q,a} : \mathbb{N} \to \mathbb{C}$ — the so-called indicator function — that equals 1 when n is equal to a modulo q and 0 elsewhere. With that function in hand, we then define $\sum_p h_{q,a}(n)\, n^{-z}$ and use the machinery in chapter 12 to compute the density of the primes in the arithmetic progression $(a, a+q, a+2q, \cdots)$. But there is a problem here. The function h is not multiplicative: $h_{q,a}(a^2)$ is not generally equal to $(h_{q,a}(a))^2$ — by way of example, $h_{3,2}(2) = 1$ while $h_{3,2}(2^2) = 0$. So we have to be more careful.

Lemma 13.16. *Let* $\gcd(a,q) = 1$. *We have*

$$\sum_{\chi\in X_q} \overline{\chi(a)}\chi(n) = \begin{cases} \varphi(q) & \text{if} \quad n =_q a \\ 0 & \text{else} \end{cases}.$$

Thus the indicator function $h_{q,a}$ *equals* $(\varphi(q))^{-1}\sum_{\chi\in X_q}\chi(a^{-1})\chi(n)$.

Proof. Since $\chi(a)$ has unit modulus, we have that $\overline{\chi(a)}\chi(a) = 1$. Because there are $\varphi(q)$ characters, the first equality follows.

The second equality is automatic if either a or n is not co-prime to q. If a and n are *distinct* co-primes, then recall that the characters form an orthogonal basis. Thus there must be another character $\chi^* \in X_q$ so that $\chi^*(a^{-1}n) \neq 1$. Since the reduced residues mod q form a field, from the above we must have that $\overline{\chi(a)} = \chi(a^{-1})$. Using multiplicativity, we obtain that $\sum_{\chi\in X_q} \overline{\chi(a)}\chi(n)$ equals

$$\sum_{\chi\in X_q} \chi(a^{-1}n) = \sum_{\chi\in X_q} (\chi^*\chi)(a^{-1}n) = \chi^*(a^{-1}n) \sum_{\chi\in X_q} \chi(a^{-1}n) = 0\,.$$

The first equality holds because $\chi^*\chi$ runs through the entire group (essentially the same argument as Lemma 5.3). The second by multiplicativity. ■

We will define quantities that allow us to mimic the proof of the prime number theorem. To facilitate this, we use uppercase letters of the corresponding notation we used earlier. So ζ becomes Z, π becomes Π, θ becomes Θ, and Φ stays the same. We will then proceed to give a proof of the prime number theorem for arithmetic progressions that follows the proof of Theorem 12.15 as closely as possible. As in Chapter 12, $\sum_p$ and $\prod_p$ mean sum or product over the (positive) primes.

The following definition should be compared with the definition of the Riemann zeta function (Definition 2.19), of the prime counting function (in Theorem 2.21), and Definition 12.1.

Definition 13.17. *We introduce a new zeta function $Z_{q,a}$, a function $\Pi_{q,a}$ that counts the primes congruent to a mod q, and two auxiliary functions.*

$$Z_{q,a}(z) := \prod_{\chi\in X_q} L(\chi,z)^{\overline{\chi(a)}} = \exp\left(\sum_{\chi\in X_q} \overline{\chi(a)}\ln(L(\chi,z))\right)\,.$$

$$\Pi_{q,a}(x) := \sum_{\substack{p\leq x\\ p=_q a}} 1\,.$$

$$\Theta_{q,a}(x) := \varphi(q)\sum_{\substack{p\leq x\\ p=_q a}} \ln p \qquad \text{and} \qquad \Phi_{q,a}(z) := \varphi(q)\sum_{p=_q a} \frac{\ln p}{p^z}\,.$$

From now on, we restrict a to $\mathbb{Z}_q^\times$, the reduced residues modulo a.

Remark 13.18. Recall that there is at most 1 prime in each congruence class that is *not* co-prime with q.

Note that $\Theta_{q,a}(x) \leq \varphi(q)\theta(x)$. Our first inequality follows from (12.6).

$$\exists C > 0 \quad \text{such that} \quad \Theta_{q,a}(x) \leq Cx\,. \tag{13.8}$$

The factor $1/\varphi(q)$ that figures so prominently in our main result, Theorem 13.26, shows up in the following lemma.

Lemma 13.19. *We have for $x \geq 2$*

$$\Pi_{q,a}(x) = \frac{\Theta_{q,a}(x)}{\varphi(q)\ln x} + \frac{1}{\varphi(q)} \int_2^x \frac{\Theta_{q,a}(t)}{t\,(\ln t)^2}\,dt\,.$$

Proof. First note that since 2 is the smallest prime, equation (12.2) gives

$$\Pi_{q,a}(x) = \frac{1}{\varphi(q)} \int_{2-\varepsilon}^x \frac{d\,\Theta_{q,a}(t)}{\ln t}\,.$$

The rest follows as in Lemma 12.2 ■

Lemma 13.20. *For* $\mathrm{Re}\,z > 1$*, we have*

$$\begin{aligned}\frac{\Phi_{q,a}(z)}{z} - \frac{1}{z-1} &= \int_1^\infty \left(\frac{\Theta_{q,a}(x)}{x} - 1\right) x^{-z}\,dx \\ &= \int_0^\infty \left(\Theta_{q,a}(e^t)e^{-t} - 1\right) e^{-zt+t}\,dt\,.\end{aligned}$$

Proof. Using (12.2), we can write $\Phi_{q,a}(z)$ as $\int_1^\infty x^{-z} d\Theta_{q,a}(x)$. Then apply (12.3) (partial integration). The proof follows that of Lemma 12.3, except that (12.6) is replaced by (13.8) ■

13.6. Primes in Arithmetic Progressions

Now we follow the reasoning of Sections 12.3 to 12.5 as closely as possible.

Lemma 13.21. *For* $\mathrm{Re}\,z > 1$*, we have that*

$$\ln Z_{q,a}(z) = -\sum_{\chi \in X_q} \sum_p \overline{\chi(a)} \ln\left(1 - \chi(p)e^{-z\ln p}\right) = \varphi(q) \sum_{p \nmid q} \sum_{\substack{n=1 \\ p^n =_q a}}^\infty \frac{1}{np^{nz}}\,.$$

and is analytic in that region.

Proof. The first equality follows from Proposition 13.15. Then we follow the reasoning of Lemma 12.10 to get

$$-\ln\left(1 - \chi(p)e^{-z\ln p}\right) = \sum_{n=1}^\infty \frac{\chi(p^n)}{np^{nz}}\,,$$

where we used complete multiplicativity of χ. Since $|\chi| = 1$, this is analytic on $\mathrm{Re}\,z > 1$. Substitute this back into the lemma. Analyticity then allows us to perform the finite sum over χ first. By Lemma 13.16, this gives a contribution $\varphi(q)$ if both $p^n =_q a$ and $\gcd(p^n, q) = 1$, and else zero. This proves the second equality of the lemma. Now the proof follows verbatim the second paragraph of the proof of Lemma 12.10. ■

Proposition 13.22. *i) The functions* $(z-1)Z_{q,a}(z)$ *and* $(z-1)Z'_{q,a}(z)+zZ_{q,a}(z)$ *have well-defined analytic continuations on* $\mathrm{Re}\, z>0$.
ii) (The analytic continuation of) $(z-1)Z_{q,a}(z)$ *evaluated at* $z=1$ *does not equal 0.*

Proof. Since $\overline{\chi_1(a)}=1$, Definition 13.17 gives

$$(z-1)Z_{q,a}(z)=(z-1)L(\chi_1,z)\cdot\exp\left(\sum_{\substack{\chi\in X_q\\ \chi\neq\chi_1}}\overline{\chi(a)}\ln(L(\chi,z))\right).$$

We need to show that $(z-1)L(\chi_1,z)$ and $L(\chi,z)$, $\chi\neq\chi_1$, are analytic in $\mathrm{Re}\, z>0$, and therefore so is $(z-1)Z_{q,a}(z)$. Adding this function to its derivative gives $(z-1)Z'_{q,a}(z)+zZ_{q,a}(z)$.

Since $\chi_1(n)$ equals 0 or 1, we can define

$$h(z):=L(\chi_1,z)-\frac{1}{z-1}.$$

The same argument presented in Proposition 12.11, shows that also here, h is analytic in $\mathrm{Re}\, z>0$. Therefore, the same holds for

$$(z-1)L(\chi_1,z)=(z-1)h(z)+1. \tag{13.9}$$

Recall that any non-principal χ is orthogonal to the principal character χ_1. Since χ_1 is always 1 (on co-primes), χ must have average zero. All characters are periodic by construction, so Proposition 13.15 implies that $\ln L(\chi,z)$ is analytic in $\mathrm{Re}\, z>0$. This proves part (i).

Part (ii) is implied by the fact that (13.9) implies that $(z-1)L(\chi_1,z)$ evaluated at $z=1$ gives 1 and that the exponential in the above expression for $(z-1)Z_{q,a}(z)$ cannot give zero. ■

Lemma 13.23. $Z_{q,a}(z)$ *has no zeroes on the line* $\mathrm{Re}\, z\geq 1$.

Proof. By Lemma 13.21, we only need to check at $z=1+i\tau$ for τ real. Define $E:=\ln(Z_{q,a}(\sigma)^3Z_{q,a}(\sigma+i\tau)^4Z_{q,a}(\sigma+2i\tau))$. By Proposition 13.22, $Z_{q,a}$ has a simple pole at 1 and no poles in $\mathrm{Re}\, z>1$. Thus if $Z_{q,a}$ has a *zero* at $1+i\tau$, then the expression e^E evaluated at $\sigma+i\tau$ where σ is *slightly* greater than 1, would yield a number that is very close to zero. The rest of the proof follows that of Lemma 12.12 verbatim. ∎

Proposition 13.24. *$\frac{\Phi_{q,a}(z)}{z} - \frac{1}{z-1}$ has an analytic continuation in the closed half plane* $\operatorname{Re} z \geq 1$.

Proof. By Lemma 13.21,

$$\frac{-Z'_{q,a}(z)}{Z_{q,a}(z)} = \sum_{\chi \in X_q} \sum_p \frac{\overline{\chi(a)}\chi(p)p^{-z}\ln p}{1-\chi(p)p^{-z}} = \sum_{\chi \in X_q} \sum_p \frac{\overline{\chi(a)}\chi(p)\ln p}{p^z - \chi(p)}\,.$$

To express this in terms of the function $\Phi_{q,a}$, we use $\frac{1}{x-k} = \frac{1}{x} + \frac{k}{x(x-k)}$ to get

$$\frac{-Z'_{q,a}(z)}{Z_{q,a}(z)} = \sum_\chi \sum_p \frac{\overline{\chi(a)}\chi(p)\ln p}{p^z} + \sum_p \frac{\overline{\chi(a)}\chi(p)^2 \ln p}{p^z\,(p^z - \chi(p))}\,.$$

Now we note that by Lemma 13.21, in the region $z > 1$, we may do the summation over χ first. We then see that by Lemma 13.16, the first term on the right hand side equals $\Phi_{q,a}(z)$. The rest of the proof follows that of Lemma 12.13 ∎

Lemma 13.25. *For all $q \geq 2$ and a such that* $\gcd(a,q) = 1$*:*

$$i) \qquad \int_1^\infty \frac{\Theta_{q,a}(y) - y}{y^2}\,dy \quad \textit{exists} \quad \Longrightarrow \lim_{x\to\infty} \frac{\Theta_{q,a}(x)}{x} = 1\,.$$

$$ii) \qquad \lim_{x\to\infty} \frac{\Theta_{q,a}(x)}{x} = 1 \iff \lim_{x\to\infty} \frac{\Pi_{q,a}(x)}{x/\ln x} = \frac{1}{\varphi(q)}\,.$$

(If $\gcd(a,q) > 1$*, the density of primes is 0.)*

Proof. The proof of (i) is entirely parallel to that of Lemma 12.14. For the proof of (ii), we use Lemma 13.19 and (13.8) instead of Lemma 12.2 and (12.6). So,

$$\left|\Pi_{q,a}(x) - \frac{\Theta_{q,a}(x)}{\varphi(q)\ln x}\right| = \frac{1}{\varphi(q)} \int_2^x \frac{\Theta_{q,a}(t)}{t\,(\ln t)^2}\,dt \leq \frac{1}{\varphi(q)}\,\frac{Cx}{(\ln x)^2}\,(1+\varepsilon)\,.$$

for any $\varepsilon > 0$. Multiply both sides by $\ln x/x$ to obtain the result. ∎

Theorem 13.26 (Prime Number Theorem for Arithmetic Progressions). *We have*

$$1)\ \lim_{x\to\infty} \frac{\Pi_{q,a}(x)}{(x/\ln x)} = \frac{1}{\varphi(q)} \qquad \text{and} \qquad 2)\ \lim_{x\to\infty} \frac{\Pi_{q,a}(x)}{\int_2^x \ln t\,dt} = \frac{1}{\varphi(q)}\,.$$

Proof. The equivalence of (1) and (2) is the same as in Theorem 12.15.

So we only need to prove part (1). Lemma 13.20 gives

$$\frac{\Phi_{q,a}(z+1)}{z+1} - \frac{1}{z} = \int_0^\infty \left(\Theta_{q,a}(e^t)e^{-t} - 1\right) e^{-zt}\, dt\,.$$

Proposition 13.24 says that the left-hand side has an analytic continuation in $\operatorname{Re} z \geq 0$ while equation (13.8) says that $\Theta_{q,a}(e^{-t})e^{-t} - 1$ is bounded. But then, by Theorem 11.18, $\int_0^\infty \left(\theta(e^t)e^{-t} - 1\right) dt$ exists. Finally, Lemma 13.25 implies that then (1) holds. ■

13.7. Exercises

Exercise 13.1. a) Finish the computation of (13.6) to show that f_m is multiplicative. (*Hint: see equation* (13.4).)
b) Check that the entries table on the right in (13.3) correspond to (13.5).

Exercise 13.2. a) Show that $\mathbb{Z}_5^\times$ as a group is isomorphic to $\mathbb{Z}_4^+$. In other words, find a bijection $f : \mathbb{Z}_5^\times \to \mathbb{Z}_4^+$ such that for all a, b in $\mathbb{Z}_5^\times$, $f(ab) = f(a) + f(b)$.
b) Show that $\mathbb{Z}_7^\times$ is isomorphic to $\mathbb{Z}_6^+$.
c) Show that $\mathbb{Z}_6^+$ is isomorphic to $\mathbb{Z}_2^+ \times \mathbb{Z}_3^+$.

Exercise 13.3. Let $f : G \to H$ a bijective homomorphism between groups. Use multiplicative notation.
a) Show that for every a and b in H, there are unique x and y in G such that

$$x = f^{-1}(a) \qquad \text{and} \qquad y = f^{-1}(b)\,,$$

where f^{-1} is the inverse of f.
b) Show that (a) implies that

$$xy = f^{-1}(a) f^{-1}(b)\,.$$

c) Show that (b) implies that $f(xy) = ab$ and thus $xy = f^{-1}(ab)$.
d) Conclude that f^{-1} is also a homomorphism.

Exercise 13.4. a) Show that $\mathbb{Z}_{16}^\times$ is isomorphic to $\mathbb{Z}_2^+ \times \mathbb{Z}_4^+$.
b) Show that $\mathbb{Z}_{16}^\times$ is *not* isomorphic to $\mathbb{Z}_8^+$. (*Hint: find the elements of order 8.*)
c) Consider the residues modulo 16 with addition *and* multiplication and verify that it is a ring.
d) Find the units (Definition 5.25) of this ring.
e) Show that the units of a (commutative) ring form a multiplicative Abelian group.

Exercise 13.5. a) Find a primitive root a modulo 26 (see Definition 5.5).
b) Find a primitive root b modulo 13.
c) Show that $\mathbb{Z}_{26}^{\times}$ is isomorphic to $\mathbb{Z}_{13}^{\times}$. (*Hint: let h map a^i to b^i and show that h is a bijective homomorphism.*)
d) Use Theorem 5.7 to prove that for odd primes, $\mathbb{Z}_{p^k}^{\times}$ is isomorphic to $\mathbb{Z}_{2p^k}^{\times}$.

Definition 13.27. *Given $x = (x_0, x_1, \cdots, x_{n-1})^T \in \mathbb{C}^n$. The discrete Fourier transform is defined as*

$$\widehat{x}_m = \sum_{k=0}^{n-1} x_k e^{-2\pi i \frac{km}{n}} ,$$

for $m \in \{0, \cdots, n-1\}$. The inverse discrete Fourier transform is given by

$$x_k = \frac{1}{n} \sum_{m=0}^{n-1} \widehat{x}_m e^{2\pi i \frac{km}{n}} .$$

Exercise 13.6. a) What are the characters of the group $\mathbb{Z}_n^+$?
b) Show that the composition of the discrete Fourier transform and the inverse discrete Fourier transform of Definition 13.27 is the identity (i.e. they are inverses of one another). (*Hint: use Theorem 13.9 and equation* (13.2).)
c) Set $\alpha := e^{2\pi i \frac{1}{n}}$. Let F be the n by n matrix whose (m,k) entry is $\alpha^{-(k-1)(m-1)}$. Show that the discrete Fourier transform is:

$$\widehat{x} = Fx .$$

d) From Definition 13.27, deduce the inverse F^{-1} of the matrix F.

Exercise 13.7. a) What are the characters of the group $\mathbb{Z}_n^+ \times \mathbb{Z}_m^+$?
b) What are the formulas in this case for the discrete Fourier transform and its inverse? (*Hint; think of this as a two-dimensional version of the Fourier transform.*)

Exercise 13.8. a) Use Theorem 13.9 and exercise 13.5 to construct the characters of $\mathbb{Z}_{13}^{\times}$ and $\mathbb{Z}_{26}^{\times}$.
b) Show that these characters basically correspond to the Fourier transform of Definition 13.27, except that the x_k are re-ordered (see also exercise 13.10).

Exercise 13.9. Proposition 2.20 is very similar to Proposition 13.15, but the former was proved in two different ways. Give the "other" proof of Proposition 13.15.
b) Is it sufficient for χ to be multiplicative (i.e. not *completely* multiplicative)?

Exercise 13.10. a) For any odd prime p denote by g its smallest primitive root. Show that there is a bijection $\operatorname{ind}_p : \mathbb{Z}_p^\times \to \mathbb{Z}_{p-1}^+$ given by

$$\operatorname{ind}_p(g^a) = a\,.$$

The value $\operatorname{ind}_p(x)$ is called the index of x relative top p. The prime root g is called the base.
b) For every odd prime less than 20, choose the smallest primitive root as base, and determine the indices of $\{1,2,\cdots,p-1\}$. *Hint: as an example, for $p=17$ with base 3, we obtain the following table*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
16	14	1	12	5	15	11	10	2	3	7	13	4	9	6	8

c) Prove that the indices behave like logarithms, that is:

$$ind_p(ab) =_{\varphi(p)} \operatorname{ind}_p(a) + \operatorname{ind}_p(b) \text{ and } \operatorname{ind}_p(a^k) =_{\varphi(p)} k \operatorname{ind}_p(a)\,.$$

Exercise 13.11. In this exercise, we use indices (exercise 13.10) to solve

$$9x^8 =_{17} 8\,.$$

a) Use exercise 13.10 (c) to show that the equation above is equivalent to

$$\operatorname{ind}_{17}(9) + 8\operatorname{ind}_{17}(x) =_{16} \operatorname{ind}_{17}(8)\,.$$

b) Use exercise 13.10 (b) to show that (a) is equivalent to

$$8\operatorname{ind}_{17}(x) =_{16} 8\,.$$

c) Use Corollary 3.8 to find the solutions to this equation. (*Hint: there are 8 solutions.*)

Exercise 13.12. Show that for any $k > 0$ there are infinitely many primes ending in k consecutive 9's.

There are useful relations between the newly minted functions in this chapter and their counterparts in Chapter 12. We prove the following lemma in exercise 13.13

Lemma 13.28. *Let $q = \prod_{i=1}^r p_i^{k_i}$. We have the following equalities:*

$$i) \qquad L(\chi_1, z) = \zeta(z) \prod_{p|q} \left(1 - p_i^{-z}\right)\,,$$

$$ii) \quad \prod_{a \in \mathbb{Z}_q^\times} Z_{q,a}(z) = \zeta(z)^{\varphi(q)} \prod_{p|q} \left(1 - p_i^{-z}\right)^{\varphi(q)}\,.$$

Exercise 13.13. a) Using the Euler product of Proposition 13.15, show that

$$\ln L(\chi_1, z) = -\sum_p \ln(1-p^{-z}) + \sum_{p|q} \ln(1-p^{-z}),$$

b) Show that (a) implies item (i) of Lemma 13.28.
c) Show that

$$\prod_a Z_{q,a}(z) = L(\chi_1, z)^{\varphi(q)} \prod_{\chi \neq \chi_1} L(\chi, z)^{\sum_a \overline{\chi(a)}}.$$

d) Show that (c) implies item (ii).

Many special cases of Dirichlet's theorem can be proved without using the machinery we have developed chapters 11 and 12 and applied in the current chapter. We discuss these cases in the next three problems.

Exercise 13.14. Define $S := \{3+4k \mid k \in \mathbb{N}\}$. *Assume* there are finitely many primes in S, namely $\{p_1, \cdots, p_k\}$ and derive a *contradiction*. Denote

$$P = 4\prod_{i=1}^{k} p_i \quad \text{and} \quad D = P - 1.$$

a) Show that D is not prime. (*Hint: $D =_4 3$.*)
b) Use (a) to show that D must have a *prime* divisor p_i in S. (*Hint: $xy =_4 3$ iff one of x or y is congruent to 3.*)
c) Use (a) and (b) to show that there is a k such that

$$D = kp_i = -1 + 4p_i \prod_{j \neq i}^{k} p_j.$$

d) Use (c) to derive that $p_i \mid 1$, a contradiction.

Exercise 13.15. Define $S := \{1+3k \mid k \in \mathbb{N}\}$. <u>Assume</u> there are finitely many primes in S, namely $\{p_1, \cdots, p_k\}$ and derive a <u>contradiction</u>. Denote

$$P = 3\prod_{i=1}^{k} p_i \quad \text{and} \quad D = P^2 + P + 1.$$

a) Show that D must have a non-trivial prime divisor $r \neq 3$ and $r \notin S$. (*Hint: $D =_3 1$ and $D =_{p_i} 1$.*)
b) Show that $P^3 =_r 1$. (*Hint: $P^3 - 1 = (P-1)D$.*)
c) Show that $\mathrm{Ord}_r^{\times}(P) = 3$. (*Hint: if $P^2 =_r 1$, then $P =_r 1$ by (b) and so $D =_r 3$; the latter is impossible, because by (a), $D =_r 0$ and $r \neq 3$.*)
d) Use (a) to show that $\gcd(P, r) = 1$ and so $P^{r-1} =_r 1$. (*Hint: Fermat's little theorem.*)
e) Use (c) and (d) to show that $3 \mid (r-1)$.
f) Point out the contradiction. (*Hint: if $r \notin S$, then $r \neq_3 1$.*)

Exercise 13.16. For any $q > 1$, define $S := \{1 + qk \mid k \in \mathbb{N}\}$. Assume there are finitely many primes in S, namely $\{p_1, \cdots, p_k\}$ and derive a contradiction. Denote

$$P = q \prod_{i=1}^{k} p_i \quad \text{and} \quad D = \sum_{i=0}^{q-1} P^i .$$

a) Show that D must have a *prime* divisor $r \nmid q$ and $r \notin S$. (*Hint: for any divisor* $e > 1$ *of* q, $D =_e 1$ *and similarly* $D =_{p_i} 1$. *Recall Bézout.*)
b) For r as in (a), show that $P^q =_r 1$. (*Hint:* $P^q - 1 = (P-1)D$ *and* $D = xr$.)
c) For r as in (a), show that $\operatorname{Ord}_r^{\times}(P) = q$. (*Hint: by (b),* $d := \operatorname{Ord}_r^{\times}(P) \mid q$; *if* $q = de$ *with* $d, e > 1$, *then* $D = \left(\sum_{i=0}^{d-1} P^i\right)\left(\sum_{i=0}^{e-1} P^{id}\right)$ *and so* $D =_e 0$; *the latter is impossible by (a).*)
d) Use (a) to show that $\gcd(P, r) = 1$ and so $P^{r-1} =_r 1$. (*Hint: unique factorization and Fermat's little theorem.*)
e) Use (c) and (d) to show that $q \mid (r-1)$.
f) Point out the contradiction.

Dirichlet proved a weaker version of Theorem 13.26 that does not use the Tauberian convergence argument of Theorem 11.18. We discuss the proof in exercise 13.17 below.

Theorem 13.29 (Dirichlet's Theorem). *Define* $S := \{n \in \mathbb{N} : n =_q a\}$. *Then*

$$\lim_{z \to 1^+} \frac{\sum_{p \in S} p^{-z}}{\sum_p p^{-z}} = \frac{1}{\varphi(q)} .$$

Exercise 13.17. a) Use Proposition 13.15 to show that for real $z \geq 1$ and $\chi \neq \chi_1$, $\sum_p \frac{\overline{\chi(a)}\chi(p)}{p^z}$ is bounded.
b) Use Lemma 13.16 to show that for $\operatorname{Re} z > 1$

$$\sum_{p =_q a} \frac{1}{p^z} = \frac{1}{\varphi(q)} \sum_{\chi \in X_q} \sum_p \frac{\overline{\chi(a)}\chi(p)}{p^z} = \frac{1}{\varphi(q)} \left(L(\chi_1, z) + \sum_{\chi \neq \chi_1} \sum_p \frac{\overline{\chi(a)}\chi(p)}{p^z} \right) .$$

c) Use Proposition 12.11 (ii) and (13.9) to show that $\lim_{z \searrow 1^+} \frac{L(\chi_1, z)}{\zeta(z)} = 1$.
d) Show that (a), (b), and (c) imply Dirichlet's theorem.

Definition 13.30. *The* natural density *of a set* $S \subseteq T$ *relative to* T *is*

$$\lim_{x \to \infty} \frac{S(x)}{T(x)} ,$$

where $S(x) = \operatorname{card}(S \cap [1,x])$ *and* $T(x) = \operatorname{card}(T \cap [1,x])$. *The Dirichlet density of a set* $S \subseteq T$ *relative to* T *is*

$$\lim_{z \searrow 1^+} \frac{\sum_{n \in S} n^{-z}}{\sum_{n \in T} n^{-z}}\,.$$

Usually, the set T *is understood to be the set of primes in* $\mathbb{N}$ *or* $\mathbb{N}$ *itself. The function* $\sum_p p^{-z}$ *is sometimes called the prime zeta function.*

Exercise 13.18. a) Show that for $n \geq 2$

$$\sum_p \frac{1}{np^{nz}} < \frac{1}{n} \int_1^\infty x^{-nz}\,dx = \frac{1}{n(nz-1)}\,.$$

b) Use (a) and Lemma 12.10 to show that as $z \searrow 1^+$

$$\ln \zeta(z) = \sum_p p^{-z} + \text{bounded}\,.$$

c) Use (12.10) to show that as $z \searrow 1^+$

$$\ln \zeta(z) = -\ln(z-1) + \text{bounded}\,.$$

(*Hint: h is analytic near* $z=1$ *and from* (12.10)*, one easily sees that it is negative for z near 1 and real.*)
d) Use (b) and (c) to show that as $z \searrow 1^+$

$$\sum_p p^{-z} = -\ln(z-1) + \text{bounded}\,.$$

e) Therefore

$$\lim_{z \searrow 1^+} \frac{\sum_p f(p)p^{-z}}{\sum_p p^{-z}} = \lim_{z \searrow 1^+} \frac{\sum_p f(p)p^{-z}}{-\ln(z-1)}\,.$$

The relation between natural density and Dirichlet density (Definition 13.30) is somewhat subtle. If the natural density exists then so does the Dirichlet density, but not vice versa. To establish the former, we prove Lemma 13.31 below in exercise 13.19. The other direction of this statement is not so easy; it is established by way of a counter-example in exercises 13.20 and 13.21.

Lemma 13.31. *Let A and B be non-empty subsets of* $\mathbb{N}$ *and* a_n *and* b_n *are their indicator functions. That is:* a_n *equals 1 if* $n \in A$ *and 0 elsewhere, and similar for* b_n*. Furthermore,* $A(x) = \sum_{n \leq x} a_n$ *and similar for* $B(x)$*. Now we have for* $\operatorname{Re} z > 1$*:*

$$\lim_{x \to \infty} \frac{A(x)}{B(x)} = \mu \quad \Longrightarrow \quad \lim_{z \searrow 1^+} \frac{\sum_{n=1}^\infty a_n n^{-z}}{\sum_{n=1}^\infty b_n n^{-z}} = \mu\,.$$

Exercise 13.19. a) Use Abel summation to show that for $\mathrm{Re}\, z > 1$

$$\sum_{n=1}^{\infty} a_n n^{-z} = z \int_1^{\infty} A(t) t^{-z-1}\, dt\,.$$

(*Hint: also use that* $A(x) \leq x$.)
b) Show that the hypothesis of Lemma 13.31 implies that for all $\varepsilon > 0$, we have $|A(x) - \mu B(x)| < \varepsilon B$.
c) Show that under the hypothesis of that lemma, we have that for all $\varepsilon > 0$,

$$\left| \frac{\int_1^{\infty} A(t) t^{-z-1}\, dt}{\int_1^{\infty} B(t) t^{-z-1}\, dt} - \mu \right| < \varepsilon\,.$$

(*Hint: write* μ *as* $\frac{\int_1^{\infty} \mu B(t) t^{-z-1}\, dt}{\int_1^{\infty} B(t) t^{-z-1}\, dt}$ *and use (b).*)

Definition 13.32. *The* <u>*logarithmic*</u> <u>*density*</u> *of a set* $S \subseteq T$ *relative to* T *is*

$$\lim_{x \to \infty} \frac{\sum_{k \in S, k \leq x} k^{-1}}{\sum_{k \in T, k \leq x} k^{-1}}\,.$$

Usually, the set T *is understood to be the set of primes in* $\mathbb{N}$ *or* $\mathbb{N}$ *itself.*

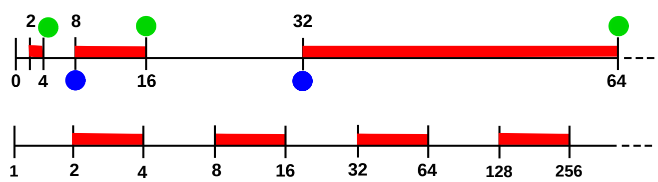


Figure 81. The set S consists of the natural numbers contained in intervals shaded in the top figure of the form $[2^{2n-1}, 2^{2n})$. The bottom picture is the same but with a logarithmic horizontal scale.

Exercise 13.20. We show that the set S depicted in top of Figure 81 does not have a natural density (relative to $\mathbb{N}$), but that it *does* have a logarithmic density.
a) Show that the limsup of the natural density is at least $5/8$ while the liminf of the density is at most $3/8$. (*Hint: first, take the average up to the green points in the figure, and then up to the blue points*)
b) Use Figure 82 to show that

$$\sum_{j=0}^{2^m-1} \frac{1}{2^m+j} = \sum_{j=0}^{2^m-1} \frac{2^{-m}}{1+j\,2^{-m}} = \int_0^1 \frac{1}{1+x}\,dx + r_m = \ln 2 + r_m\,,$$

where $r_m \in [0, 2^{-m+1}]$. (*Note: for Riemann sum, see* [**53**]*chapter 6.*)
c) Use (a) to show that

$$\sum_{k\in S, k\le n} k^{-1} = \frac{1}{2}\,\log_2 n\,\ln 2 + R(n) = \frac{1}{2}\ln n + R(n)\,,$$

where $|R(n)| \le 2(1+\ln 2)$.
d) Use (b) and exercise 12.3 (c) to show that the logarithmic density of S is 1/2. (*Note: a much simpler heuristic argument gives that according to exercise 12.3 the logarithmic density corresponds to redrawing S with the horizontal coordinate logarithmic as in the bottom picture of Figure 81, and then computing the density.*)

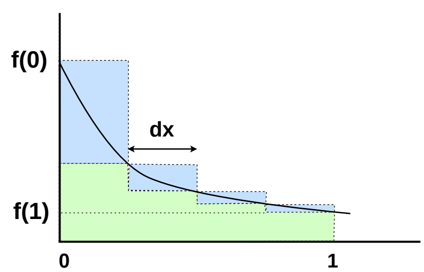


Figure 82. Proof that $\int_0^1 f(x)\,dx$ is between $\sum_{j=1}^{k} f(j\,dx)dx$ and $\sum_{j=0}^{k-1} f(j\,dx)dx$ if f is strictly decreasing.

Exercise 13.21. We show that if the logarithmic density of a set S (Definition 13.32) exists, then its Dirichlet density equals the logarithmic density.
a) Denote the elements of S by $\{n_1, n_2, \cdots\}$ and show that for $\mathrm{Re}\, z > 1$

$$\begin{aligned}
&\textstyle\sum_{n=1}^{\infty} \left(\sum_{k\in S, k\le n} k^{-1}\right) \left(n^{1-z} - (n+1)^{1-z}\right) = \\
&\qquad n_1^{-1} \left(n_1^{1-z} - (n_1+1)^{1-z} + \cdots + (n_2-1)^{1-z} - n_2^{1-z}\right) \\
&\quad + \left(n_1^{-1} + n_2^{-1}\right) \left(n_2^{1-z} - (n_2+1)^{1-z} + \cdots + (n_3-1)^{1-z} - n_3^{1-z}\right) \\
&\quad + \left(n_1^{-1} + n_2^{-1} + n_3^{-1}\right) \left(n_3^{1-z} - (n_3+1)^{1-z} + \cdots + (n_4-1)^{1-z} - n_4^{1-z}\right) \\
&\quad + \cdots \\
&\qquad = \textstyle\sum_{n\in S} n^{-z}\,.
\end{aligned}$$

(*Hint:* n_1^{-1} *gets multiplied by* $(n^{1-z} - (n+1)^{1-z})$ *for* $n \ge n_1$, n_2^{-1} *by* $(n^{1-z} - (n+1)^{1-z})$ *for* $n \ge n_2$*, and so on. The sums as given telescope to* $n_1^{-1}(n_1^{1-z} - n_2^{1-z})$, $(n_1^{-1} + n_2^{-1})(n_2^{1-z} - n_3^{1-z})$*, and so forth.*)
b) Show that if the logarithmic density of S (with respect to $\mathbb{N}$) equals μ, then, by (a), we have

$$\begin{aligned}
\sum_{n\in S} n^{-z} &= \sum_{n=1}^{\infty} \left(\sum_{k\in S, k\le n} k^{-1}\right) \left(n^{1-z} - (n+1)^{1-z}\right) = \\
&= \sum_{n=1}^{\infty} \left(\mu \sum_{k\le n} k^{-1}\right) \left(n^{1-z} - (n+1)^{1-z}\right) = \\
&= \mu \sum_{n\in\mathbb{N}} n^{-z}.
\end{aligned}$$

c) Use (b) to demonstrate the statement heading this exercise.

To emphasize once again the similarity between our generalized zeta functions and ζ of Chapter 12, we show that $Z_{q,a}$ has no zeroes in $\mathrm{Re}\, z > 1$. The proof can be copied from exercise 4.24, provided you make the requisite substitutions.

Definition 13.33. *The function* $M_{q,a} : \mathbb{N} \to \mathbb{Z}$ *is given by:*

$$M_{q,a}(n) = \begin{cases} 1 & \text{if} \quad n = 1 \\ 0 & \text{if} \quad \exists p \text{ prime} \;\text{ with }\; p \neq_q a \text{ and } p \mid n \\ 0 & \text{if} \quad \exists p \text{ prime} \;\text{ with }\; p^2 \mid n \\ (-1)^r & \text{if} \quad n = p_1 \cdots p_r \;\text{ and }\; p_i =_q a \end{cases}.$$

This the counterpart of the Möbius function of Definition 4.6.

Exercise 13.22. a) Show that $M_{q,a}$ is a multiplicative function. (*Hint: compare with the Möbius function in Chapter 4.*)
b) Use Euler's product formula and Definition 13.33 to show that in $\mathrm{Re}\, z > 1$

$$\frac{1}{Z_{q,a}(z)} = \prod_{p =_q a} \left(1 - p^{-z}\right) = \prod_{p \text{ prime}} \left(\sum_{i \geq 0} M_{q,a}(p^i) p^{-iz}\right).$$

c) Without using equation (4.7), prove that the expression in (b) equals $\sum_{n \geq 1} M_{q,a}(n) n^{-z}$. (*Hint: since $M_{q,a}$ is multiplicative, you can write a proof re-arranging terms as in the first proof of Euler's product formula.*)

Exercise 13.23. Show that for $q > 1$ in $\mathbb{N}$:

$$\lim_{n \to \infty} \left(\prod_{p \leq n,\, p =_q a} p\right)^{1/n} = e^{1/\varphi(q)}$$

if and only the prime number theorem for arithmetic progressions holds. (*Hint: see Lemma 13.25 (ii). See also exercise 12.26.*)

In exercises 13.24 and 13.25, we prove partial versions of some remarkable results knowns as Mertens' theorems. These were proved 22 years before the prime number theorem [**47**]. More details can be found in [**32**] [Section 22]. The version we give summarizes the statements given in [**46**].

Theorem 13.34 (**Mertens' Theorems).**

$$i) \quad \lim_{x \to \infty} \left(\sum_{p \leq x} \frac{\ln p}{p} - \ln x\right) = -B_3 \approx -1.3326\,.$$

$$ii) \quad \lim_{x \to \infty} \left(\sum_{p \leq x} \frac{1}{p} - \ln \ln x\right) = B_1 \approx 0.2615\,.$$

$$iii) \quad \lim_{x \to \infty} \left(\sum_{p \leq x} \ln\left(1 - p^{-1}\right) - \ln \ln x\right) = -\gamma\,.$$

B_1 and B_3 are sometimes called Mertens constants , but also go by other names. γ is the Euler-Mascheroni constant (see exercise 12.3).

Exercise 13.24. a) Deduce from (12.4) and unique factorization that

$$\frac{1}{n}\ln(n!) = \sum_{p^k \le n} \frac{1}{n}\left\lfloor \frac{n}{p^k} \right\rfloor \ln p\,.$$

(*Note: we sum over both the relevant integers k and primes p.*)
b) Show that

$$\frac{1}{n}\left\lfloor \frac{n}{p^k} \right\rfloor \le \frac{1}{p^k}\,.$$

c) Show that (a) and (b) imply that for some $K_1 > 0$

$$\left| \frac{1}{n}\ln(n!) - \sum_{p \le n} \frac{\ln p}{p} \right| < K_1\,.$$

(*Hint:* $\sum_{k \ge 2} p^{-k} = 1/(p(p-1))$.)
d) Use exercise 12.4 (a) to show that there is a K_2 so that

$$\left| \frac{1}{n}\ln(n!) - \ln n \right| < K_2\,.$$

e) Conclude that $R(x)$ is bounded where

$$R(x) := \sum_{p \le n} \frac{\ln p}{p} - \ln n\,.$$

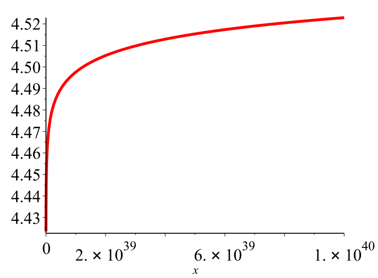


Figure 83. The function $\ln(\ln(x))$ for $x \in [1, 10^{40}]$.

Exercise 13.25. a) Let $a_n = \frac{\ln p}{p}$ is $n = p$ a prime and 0 else and set $f(t) = 1/\ln t$. Now use Abel summation (Proposition 12.16) to show that

$$\sum_{n \le x} \frac{1}{p} = \frac{1}{\ln x} \sum_{p \le x} \frac{\ln p}{p} + \int_2^x \frac{1}{t(\ln t)^2} \sum_{p \le t} \frac{\ln p}{p} \, dt\,.$$

b) Use exercise 13.24 (d) applied to the previous item to show that

$$\sum_{n \le x} \frac{1}{p} = 1 + \frac{R(x)}{\ln x} + \int_2^x \frac{1}{t \ln t} \, dt + \int_2^x \frac{R(t)}{t(\ln t)^2} \, dt$$

c) Conclude that

$$\sum_{p \le x} \frac{1}{p} = \ln \ln x + o(\ln \ln x)\,.$$

d) Compare (a) with exercise 13.18(d).
e) To appreciate *how* agonizingly slow the approach of $\ln \ln x$ to infinity is, approximate $\ln \ln 10^{10^{10}}$. (*Hint: about 25*).
f) To write that number — $10^{10^{10}}$ — in full decimal notation in a series of books, how many books would you fill? Assume that you write 2000 characters on a page and that 500 pages make one book.

Chapter 14

The Birkhoff Ergodic Theorem

Overview. To fully understand and appreciate the proof of the Birkhoff ergodic theorem, we have to dig a little deeper in analysis. We give the necessary background in this chapter and then prove the theorem. It is recommended that you carefully read Sections 9.1 and 9.2 again before starting Sections 14.1, 14.2, and 14.3 below.

14.1. Measurable Sets

We recall from Section 9.2 that if we have a space X and a collection Σ of measurable sets, then the pair (X, Σ) is called a *measurable space*.

Definition 14.1. *A function $F : X \to Y$ is called* measurable function *if the inverse image under F of any measurable set in Y is measurable in X.*

A measure μ is a non-negative function from Σ to $[0, \infty]$ that is countably additive on disjoint measurable sets (Definition 9.5). A triple (X, Σ, μ) is called a *measure space*. A *probability measure* is a measure that assigns a measure 1 to the entire space. It is time to refine our understanding of those concepts.

Definition 14.2. *A sigma algebra or σ-algebra is a collection Σ of sets with the following properties:*

$\emptyset \in \Sigma$ and Σ is closed under complementation and under countable union. In any topological space, the smallest σ-algebra that contains the open sets is called the Borel sets (Definition 9.1).

Remark 14.3. Since $(\cup_{i\in\mathbb{N}} A_i)^c = \cap_{i\in\mathbb{N}} A_i^c$ (exercise 9.1), we see that a σ-algebra is also closed under countable intersection.

We are now in a position to give a more formal definition of a measure (see Definition 9.5).

Definition 14.4. *Let (X,Σ) be a measure space. A measure is a function $\mu : \Sigma \to [0,\infty]$ such that $\mu(\emptyset) = 0$ and for every countable sequence of disjoint (measurable) sets S_i:*

$$\mu(\cup_{i=1}^{\infty} S_i) = \sum_{i=1}^{\infty} \mu(S_i)\,.$$

If Σ contains the open sets, then μ is called a Borel measure .

Definition 14.5. *Let X be a set and $\mu : \Sigma \to [0,\infty]$ a measure. A set $S \subseteq X$ is measurable or, more accurately, μ-measurable if $S \in \Sigma$. A function $f : X \to \mathbb{R}$ is measurable if for every $y \in \mathbb{R}$, $f^{-1}((y,\infty))$ is in Σ. (Here, $f^{-1}(S)$ means the inverse image of the set S)*

Why is this so complicated? Recall from Section 9.1 that if we try to assign a measure μ on the real line consistent with our intuitive notion of length, that is: intervals (a,b) must have "measure" $b-a$ and the measure is translation invariant[1], we may run into trouble. In that section we showed, using the axiom of choice, that we can construct sets (Vitali sets) that *cannot* be assigned a measure. Thus we *cannot* assume that any arbitrary set is measurable. So if we define certain abstract sets and want to talk about their measure, we have to always be very careful that we didn't leave the sigma algebra of measurable sets.

In fact, the determination that certain combinations of measurable functions are still measurable plays a significant role in the proof of Birkhoff's theorem. So let us have a closer look at this.

[1] This is the Lebesgue measure.

14.2. Measurable Functions

Not all functions are measurable. For instance, the function from $\mathbb{R}$ to itself that is 1 on the Vitali set V (see Section 9.1) and zero elsewhere, is not Lebesgue measurable. For a more interesting example, see exercise 14.3.

Suppose f is measurable. Recall that the collection of measurable sets must be closed under complementation, countable intersection, and countable union. Since $f^{-1}((-\infty,y])$ is the complement of the measurable set $f^{-1}((y,\infty))$, it is also measurable. $f^{-1}([y,\infty))$ can be written as the (countable) intersection $\cap_{n\in\mathbb{N}} f^{-1}((y-\frac{1}{n},\infty))$, it, too, is measurable. Again, by complementation, $f^{-1}((-\infty,y))$ is measurable.

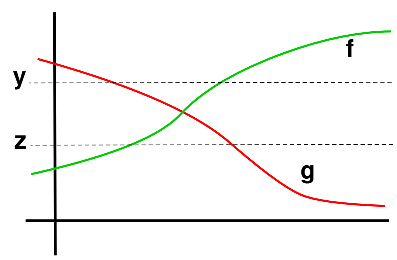


Figure 84. Illustration of the fact that if $h_+(x) = \sup\{f(x),g(x)\}$, then $h_+^{-1}((y,\infty)) = f^{-1}((y,\infty)) \cup g^{-1}((y,\infty))$. Similarly, if $h_-(x) = \inf\{f(x),g(x)\}$, then $h_-^{-1}((z,\infty)) = f^{-1}((z,\infty)) \cap g^{-1}((z,\infty))$.

It is easy to see that if f and g are measurable, then $h(x) = \sup\{f(x),g(x)\}$ is measurable because $h^{-1}((y,\infty)) = f^{-1}((y,\infty)) \cup g^{-1}((y,\infty))$ (see Figure 84). Similar for $\inf\{f(x),g(x)\}$. Almost as easy is the fact that also $f+g$ and $f\cdot g$ are measurable. For the set

$$A_{r_1,r_2} := \{x \mid f(x) > r_1\} \cap \{x \mid g(x) > r_2\}$$

is measurable for all rationals r_i, and therefore so is the (countable) union of A_{r_1,r_2} over those rationals such that $r_1+r_2 > y$ or such that $r_1 r_2 > y$.

Lemma 14.6. *Let $\{f_n\}$ be a sequence of measurable functions. Then $\sup_n f_n(x)$, $\inf_n f_n(x)$, $\limsup_n f_n(x)$, and $\liminf_n f_n(x)$ are measurable.*

Proof. Set $h_\pm$ equal to $\sup_n f_n(x)$ and $\inf_n f_n(x)$, respectively. Then

$$h_+^{-1}((y,\infty)) = \cup_{n=1}^{\infty} f_n^{-1}((y,\infty)),$$

which proves the first case. The proof for h_- is same, except that the union must be replaced by an intersection.

Set $g_\pm$ equal to $\limsup_n f_n(x)$ and $\liminf_n f_n(x)$, respectively. Since

$$g_+(x) = \lim_{n\to\infty} \sup_{i\geq n} f_i(x)$$

and $\sup_{i\geq n} f_i(x)$ is non-increasing (in n), we can replace the above limit by the infimum, and use the above results for supremum and infimum to get

$$g_+^{-1}((y,\infty)) = \cap_{n\geq 1} \cup_{i\geq n} f_i^{-1}((y,\infty)) \,.$$

And so g_+ is measurable. A similar reasoning works for g_-. ■

Remark 14.7. As a result, the pointwise limit (if it exists) of a sequence of measurable functions is also measurable.

14.3. Dominated Convergence

In this section, we prove — largely inspired by [**8**] — Lebesgue's dominated convergence theorem. This is a result of fundamental importance in its own right. It is widely used not only in analysis but also in applications of analysis to the study of partial differential equations and probability theory among others. Here, we will need it to prove the ergodic theorem.

The following theorem says that almost everywhere pointwise[2] convergence implies *nearly*[3] uniform convergence, that is: convergence is uniform, except on a set of small measure. See Figure 85.

Theorem 14.8 (**Egorov's Theorem**). *Let (X,Σ,μ) a finite measure[4] space. Suppose that $\{f_i\}$ is a sequence of measurable functions on X, so that μ almost everywhere, $f_i(x)$ converges pointwise to $f(x)$. Then for every $\varepsilon > 0$, there is a set $U \in \Sigma$ on which the convergence of $f_i \to f$ is uniform, and so that the exceptional set $\mu(X\backslash U) < \varepsilon$.*

Proof. Let

$$A_{m,n} := \left\{ x \in X \, : \, \forall i \geq m \, , \, |f_i(x) - f(x)| < \frac{1}{n} \right\} .$$

[2] Pointwise convergence means that for x fixed, $\lim_{i\to\infty} f_i(x) = f(x)$.

[3] For the usage of the word 'nearly' here, see Section 14.4.

[4] A space with $\mu(X) < \infty$

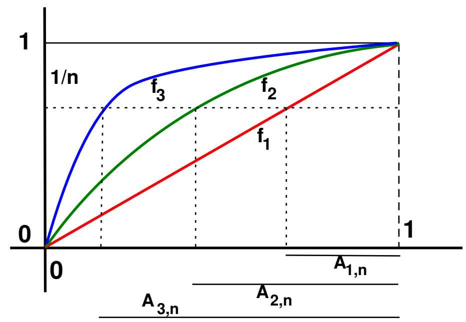


Figure 85. A sequence of functions $f_n(x) = x^{1/n}$ that converge almost everywhere pointwise to $f(x) = 1$ on $[0,1]$. The convergence is uniform on $U = [\varepsilon, 1]$ for any $\varepsilon \in (0,1)$.

We have $A_{m,n} \subseteq A_{m+1,n}$ and $\cup_m A_{m,n}$ covers all of X, except for a measure zero set Z_n (see Figure 85). Thus we can choose m_n such that

$$\mu(X \backslash A_{m_n,n}) = \mu(A^c_{m_n,n}) < \frac{\varepsilon}{2^n}, \tag{14.1}$$

where the superscript indicates the complement in X. For any x in the intersection of all $A_{m_n,n}$, we have that for $i \geq m_n$,

$$|f_i(x) - f(x)| < 1/n.$$

And thus on $U := \cap_{n \geq 1} A_{m_n,n}$, we have uniform convergence. Within X, we have

$$X \backslash U = (\cap_{n\geq 1} A_{m_n,n})^c = \cup_{n \geq 1} A^c_{m_n,n},$$

And so, by equation (14.1) and subadditivity (9.1), $\mu(X \backslash U) < \varepsilon$. ■

Next we prove first that integrable functions *nearly* live on sets of finite measure and that integrals of measurable functions over small sets are small.

Lemma 14.9. *Suppose $g : X \to [0,\infty]$ is measurable and integrable. Then:*
i) for every $\varepsilon > 0$, there is a set F of finite measure such $\int_{X \backslash F} g \, d\mu < \varepsilon$.
ii) for every $\varepsilon > 0$, there is a $\delta > 0$ such that for all small sets S with $\mu(S) < \delta$, we have $\int_S g \, d\mu < \varepsilon$.

Proof. Let $\{y_i\}$ be a countable partition of the range of an integrable function g. Denote $A_i = g^{-1}(\{y : y \geq y_{i+1}\})$ and $\Delta_i = y_{i+1} - y_i$. From the definition of the Lebesgue integral (Section 9.2 and Figure 86), we see that for every $\varepsilon > 0$ we can choose a partition so that

$$\sum_{i=1}^{\infty} \mu(A_i)\Delta_i < \int g\, d\mu < \frac{\varepsilon}{2} + \sum_{i=1}^{\infty} \mu(A_i)\Delta_i\,.$$

Since the sum must converge, we can truncate at some n to get

$$\sum_{i=1}^{n} \mu(A_i)\Delta_i < \int g\, d\mu < \varepsilon + \sum_{i=1}^{n} \mu(A_i)\Delta_i\,. \tag{14.2}$$

We may assume $\Delta_i > 0$ in this sum (the $\Delta_i = 0$ terms do not contribute). The union $F := \cup_{i=1}^{n} A_i$ must have finite measure (otherwise $\sum_{i=1}^{n} \mu(A_i)\Delta_i$ would diverge). Now we compute

$$\int_{X\setminus F} g\, d\mu = \int_X g\, d\mu - \int_F g\, d\mu < \left(\varepsilon + \sum_{i=1}^{n} \mu(A_i)\Delta_i\right) - \sum_{i=1}^{n} \mu(A_i)\Delta_i = \varepsilon\,.$$

We used *both* inequalities in (14.2) to derive the last inequality.

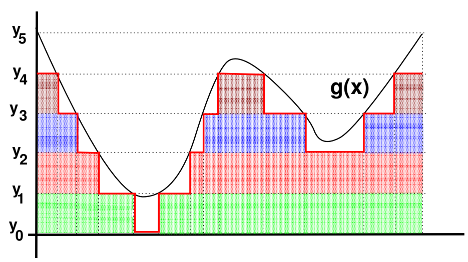


Figure 86. The definition of the Lebesgue integral. Let $\{y_i\}$ be a countable partition of the range of g. We approximate $\int g\, d\mu$ from below by $\sum_i \mu\left(g^{-1}(\{y : y \geq y_{i+1}\})\right)(y_{i+1} - y_i)$. Then g is integrable if the limit converges as the mesh of the partition goes to zero. The function y in the proof of Lemma 14.9 (ii) is indicated in red. (Here μ is the Lebesgue measure.)

To prove (ii), denote by y the function whose value equals y_i on A_i (see Figure 86). Let y_+ be the maximum of the y_i (in the definition of A_i) for which the Δ_i are positive. Then for any measurable set B with $\mu(B) < \delta$

$$\int_B g\, d\mu = \int_B (g-y)\, d\mu + \int_B y\, d\mu < \int_X (g-y)\, d\mu + \int_B y\, d\mu\,.$$

Since $\int_X y\,d\mu$ is simply $\sum_{i=1}^n \mu(A_i)\Delta_i$, the first integral is less than ε by (14.2). The same holds for the second integral if we choose δ so that $y_+\delta < \varepsilon$. ■

Theorem 14.10 (Lebesgue's Dominated Convergence Theorem). *Let $\{f_k\}$ be a sequence of real valued measurable functions on (X,Σ,μ). Suppose that the sequence converges μ almost everywhere to f and that it is dominated by an integrable function g so that for all k, $|f_k(x)| \leq g(x)$. Then*

$$\lim_{k\to\infty}\int f_k\,d\mu = \int \lim_{k\to\infty} f_k\,d\mu = \int f\,d\mu\,.$$

Proof. For any set $U \in \Sigma$, we have (using linearity of the integral)

$$\begin{aligned}\left|\int_X f_k\,d\mu - \int_X f\,d\mu\right| &= \left|\int_{X\setminus U} f_k\,d\mu - \int_{X\setminus U} f_k\,d\mu + \int_U f_k\,d\mu - \int_U f\,d\mu\right| \\ &\leq 2\left|\int_{X\setminus U} g\,d\mu\right| + \left|\int_U (f_k - f)\,d\mu\right|.\end{aligned}$$

We consider the finite measure case (where $\mu(X) < \infty$) and the infinite measure case separately.

When $\mu(X) < \infty$, we use Egorov's theorem to choose the set U on which we have uniform convergence so that $\mu(X\setminus U) < \delta$, where δ is as in Lemma 14.9 (ii). So for any $\eta > 0$, we can choose k large enough so that the above inequality becomes

$$\left|\int_X f_k\,d\mu - \int_X f\,d\mu\right| < 2\varepsilon + \eta\,\mu(U)$$

Upon choosing η small enough, the result follows because $\mu(U) < \infty$.

In the case where $\mu(U)$ is infinite, we need to do one step extra. Use Lemma 14.9 (i) to find a set $F \subseteq U$ of *finite* measure so that $\int_{U\setminus F} g\,d\mu < \varepsilon$. The first inequality is now followed by

$$\begin{aligned}\left|\int_U (f_k - f)\,d\mu\right| &= \left|\int_{U\setminus F} (f_k - f)\,d\mu + \int_F (f_k - f)\,d\mu\right| \\ &\leq 2\int_{U\setminus F} g\,d\mu + \left|\int_F (f_k - f)\,d\mu\right|.\end{aligned}$$

The second integral can now be estimated in the same way as before. ■

Remark 14.11. While we proved the theorem here for real valued functions, it also holds for complex valued functions. One simply proves the result for the real and imaginary parts separately.

14.4. Littlewood's Three Principles

The subject of real analysis, and measure theory, and Lebesgue integration in particular, overtook the older, more informal notions of length and Riemann integration in part because extremely useful theorems like the dominated convergence theorem simply do not hold in the older setting. Here is a simple example to illustrate that.

Let $f_n : [0,1] \to \mathbb{R}$ be given by $f_n(x) = 1$ if $x \in S_n$ where

$$S_n := \left\{ \frac{i}{k} : 0 < k \le n \;\text{ and }\; 0 \le i \le k \right\},$$

and $f_n(x) = 0$ elsewhere. Clearly, each f_n is Riemann integrable (having only finitely many discontinuities). Also the f_n are dominated by $g(x) = 1$. See Figure 87. However, $\lim_{n\to\infty} \int f_n\,dx = 0$ while $\int \lim_{n\to\infty} f_n\,dx$ is not

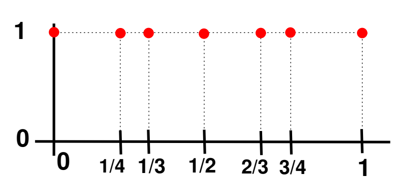


Figure 87. The non-zero values of the function f_4 in red.

defined since $\lim_{n\to\infty} f_n$ is not Riemann integrable (as it has a dense set of discontinuities). In exercises 14.5-14.8, we give other interesting "counterexamples". For now, note that if we switch to Lebesgue integration, there is no problem because then $\lim_{n\to\infty} f_n$ is integrable, and non-zero only on the rationals (measure zero) and and so its integral is zero.

Nonetheless, this more powerful mode of reasoning seems very abstract and for that reason it is difficult to develop an intuition in the subject. It is perhaps comforting to know that at least some of the masters of the subject themselves recognized this. The most famous instance of this is formed by Littlewood's three principles [**42**].

Each of these principles describes a desirable behavior that indeed holds if only one excludes sets of arbitrarily small measure. This is expressed by the word "nearly", which we encountered a few times in Section 14.3: we say that in these cases the behavior *nearly* holds.

- Every measurable set is *nearly* a finite union of disjoint open intervals.
- Every measurable function is *nearly* continuous.
- Every pointwise convergent sequence of functions is *nearly* uniformly convergent.

The first principle is in fact Proposition 9.4 (ii). The third principle is of course Egorov's theorem (Theorem 14.8). For the second principle is virtually all texts refer to Luzin's Theorem (see below). However, Littlewood himself mentions a slightly different theorem in this context, [**42**] [Section 4.1]. For completeness, we state Luzin's theorem with only a sketch of the proof.

Theorem 14.12 (Luzin's Theorem). *Let f be measurable in $(\mathbb{R},\Sigma,\mu)$ where Σ are the Borel sets and μ is the Lebesgue measure. For every $\varepsilon > 0$, there is a small open set O of (Lebesgue) measure less than ε so that f is continuous when restricted to $\mathbb{R}\backslash O$.*

Sketch of proof. We approximate f by stepfunctions $f_n$as in Figure 86, so that almost everywhere $f_n \to f$. The f_n are constant except on a exceptional set E_n of measure zero where the discontinuities are located. By Egorov, we now have that $f_n \to f$ is uniform except on a arbitrarily small set F. So we set

$$S := \mathbb{R} - F - \cup_{j\geq n} E_n \,.$$

Now on S, the continuous f_n converge uniformly to f and therefore f is continuous[5] on S. Now all we need to do, is to approximate S with a closed set C as in Proposition 9.4 (i) and let O be the complement of C. ■

One must be careful in the interpretation of this last result: it does *not* mean that the points of the $\mathbb{R}\backslash S$ are points of continuity of f. As an example, consider the function that is 1 on the rational numbers and 0 everywhere else. As a function $\mathbb{R} \to \mathbb{R}$, it is nowhere continuous, but it's restriction to

[5] This is proved in any introductory analysis course.

the irrational numbers is continuous. Luzin's theorem still goes a little further, and asserts that we can contain the rationals in an open sets of arbitrary small measure (exercise 9.3).

14.5. Weyl's Criterion

To get us in the mood for the ergodic theorem, we first look at another which is very often used in number theory. First, we need a basic result from analysis.

Theorem 14.13 (Fejér's Theorem, informal version)**.** *Let S be the unit circle parametrized by $z = e^{2\pi ix}$ in the complex plane. Let $f : S \to \mathbb{R}$ be continuous. Then for any $\varepsilon > 0$ there is $p_M(e^{2\pi ix}) := \sum_{m=-M}^{M} a_m e^{2\pi imx}$ such that $|f(z) - p_M(z)| < \varepsilon$.*

The full version of this theorem explicitly constructs the approximating polynomials, see [**41**] [Chapter 12] or [**3**] [Chapter 11]. We will not prove this result here, as that would take us too far afield.

In what follows, we let $\mathbf{1}_{[a,b]}(x)$ denote the function that is 1 if x is in $[a,b]$ and 0 elsewhere.

Theorem 14.14 (**Weyl's Criterion**)**.** *The following are equivalent.*
i) The real sequence $\{x_n\}$ is equidistributed modulo 1, i. e. for all a and b

$$\lim_{n\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{[a,b]}(x_k) = (b-a)\,.$$

ii) For every continuous function $f : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$,

$$\lim_{n\to\infty} \frac{1}{K} \sum_{k=1}^{K} f(x_k) = \int_0^1 f\,dx\,.$$

iii) For all $m \neq 0$ in $\mathbb{Z}$

$$\lim_{n\to\infty} \frac{1}{K} \sum_{k=1}^{K} e^{2\pi imx_k} = 0\,.$$

Proof. We first prove the equivalence of (i) and (ii). Since real and imaginary of f can be dealt with in the same way, it is sufficient to consider only the real case.

Since f is continuous on a compact domain, it is *uniformly* continuous (see [**53**]). So let $\{x_i\}_{i=0}^m$ is a partition of the circle $\mathbb{R}/\mathbb{Z}$ and $c_i \in [x_i, x_{i+1}]$, then for any $\varepsilon > 0$, we can choose a fine enough partition so that

$$f_m(x) = \sum_{i=0}^{m-1} f(c_i)\, \mathbf{1}_{[x_i, x_{i+1}]}(x)$$

and

$$|f(x) - f_m(x)| < \varepsilon\,. \tag{14.3}$$

Now (i) implies that

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} f(x_k) = \sum_{i=0}^{m-1} f(c_i)(x_{i+1} - x_i) = \int f_m\, dx\,.$$

With (14.3), this implies that

$$\left| \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} f_m(x_k) - \int f\, dx \right| < 2\varepsilon\,,$$

and (ii) follows. The reverse follows by approximating $\mathbf{1}_{[}a,b]$ by the continuous function g_ε as indicated Figure 88. It is easy to see that (ii) implies

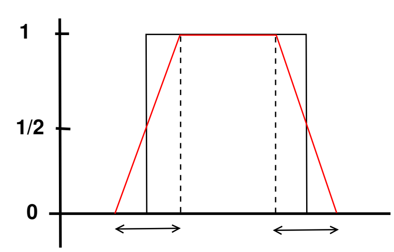


Figure 88. The functions $\mathbf{1}_{[}a,b]$ in black and g_ε in red. ε is the width (indicated by two-sided arrows) of the regions where g_ε and the step function do not agree.

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} g_\varepsilon(x_k) = \int g_\varepsilon\, dx = b - a\,.$$

Taking the limit as ε tends zero establishes (i).

It is clear that (ii) immediately implies (iii). Thus we only need to prove that (iii) implies (ii). Let $f : S \to \mathbb{R}$ be continuous. Theorem 14.13 implies that for any $\varepsilon > 0$ there is a $p_M(e^{2\pi ix}) := \sum_{m=-M}^{M} a_m e^{2\pi imx}$ so that

$$\left|\int_0^1 f(e^{2\pi ix})\,dx - \int_0^1 p_M(e^{2\pi ix})\,dx\right| = \left|\int_0^1 f(e^{2\pi ix})\,dx - a_0\right| < \varepsilon\,.$$

If we also appeal to item (iii), we see that

$$\left|\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\left[f(e^{2\pi ix_k}) - p_M(e^{2\pi ix_k})\right]\right| = \left|\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K} f(e^{2\pi ix_k}) - a_0\right| < \varepsilon\,.$$

Comparison of the last two inequalities yields the implication. ■

If T is ergodic with respect to Lebesgue, and if we set $x_i = T^i(x_0)$ and $f(x) := e^{2\pi imx}$. Now, of course, item (ii) is exactly Corollary 9.13, which says that time averages equal space averages. The standard example of this is $T(x_{k+1}) = x_k + \rho$ where ρ is irrational, as we discussed at length in Chapter 10. However, it is still amusing to give a very simple and direct proof of this, based on Weyl's criterion.

Indeed, it requires no more than than summing a geometric series to see that

$$\frac{1}{n}\sum_{k=0}^{n-1} e^{2\pi imx_k} = \frac{e^{2\pi imx_0}}{n}\sum_{k=0}^{n-1} e^{2\pi imk\rho} = \frac{e^{2\pi imx_0}}{n}\cdot\frac{e^{2\pi imn\rho}-1}{e^{2\pi im\rho}-1}\,.$$

Since ρ is irrational and $m \neq 0$, we have that $e^{2\pi im\rho} - 1 \neq 0$, and so the factor $1/n$ drives the limit to zero. (If $m = 0$ the left hand side immediately yields one).

A final remark is that in the proof of Weyl's criterion, it might seem more reasonable to prove (iii) implies (i). But there is a subtle obstruction to an easy proof. Such a proof would express $\mathbf{1}_{[a,b]}$ as a trigonometric sum that converges uniformly. However, this is not possible. It turns out that if we try that the trigonometric sum $s(x)$ approximating functions with a jump discontinuity always "overshoots" by almost 9%. This is called the Gibbs phenomenon [**3**, **26**], and also goes by the name of ringing .

14.6. Proof of Birkhoff's Ergodic Theorem

Our proof is based on [**48**] [section 9]. As before, we will denote iterates under T by subscripts.

$$T(x_0) = x_1\,, \quad T(T(x_0)) = T^2(x_0) = T(x_1) = x_2\,, \quad \cdots$$

and so on. We also define the sums

$$S_f^n(x_0) = \sum_{i=1}^{n} f(T^i(x_0))\,.$$

Remark 14.15. In this section, we work in a measure space (X,Σ,μ). We stipulate that $T : X \to X$ is a measurable transformation that preserves a probability measure μ and that $f : X \to \mathbb{R}$ (or $\mathbb{C}$) is an arbitrary μ-integrable function.

Proposition 14.16 (Maximal Ergodic Theorem). *If for μ-almost every x, there is an $n(x)$ such that $S_f^{n(x)}(x) \geq 0$ (≤ 0), then $\int f\,d\mu \geq 0$ (≤ 0).*

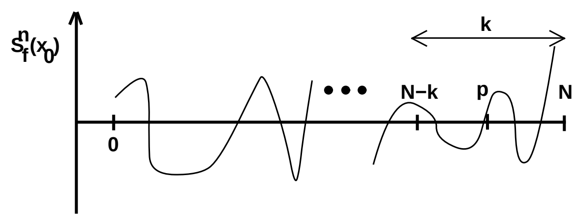


Figure 89. A plot of $S_f^n(x_0)$ for some fixed x_0 for $n \in \{0,\cdots,N\}$.

Proof. Note that this statement holds for f with "$\geq$" if and only if it holds for $g = -f$ with "$\leq$". So it is sufficient to prove only the $\geq$ version.

First assume that $n(x)$ is bounded (for almost all x) by some $k > 0$. Then no matter how large we take N, there is some $p(x)$ in $\{N-k,\cdots,N\}$ such that $S_f^{p(x)}(x) \geq 0$ (see Figure 89). We then have for μ-almost all x_0

$$S_f^N(x_0) = S_f^{p(x_0)}(x_0) + S_f^{N-p(x_0)}(x_p) \geq -S_{|f|}^{N-p(x_0)}(x_p) \geq -S_{|f|}^{k}(x_{N-k})\,.$$

Therefore, for μ-almost all x,

$$\sum_{i=1}^{N} f(T^i(x)) \geq -\sum_{i=N-k+1}^{N} \left|f(T^i(x))\right|\,.$$

Bearing in mind that μ is invariant, we integrate this inequality. So by Lemma 10.1, $\int f(T^i(x))\,d\mu = \int f(x)\,d\mu(x)$ and similarly for $|f|$. In this way we obtain, after integrating, that $N\int f\,d\mu \geq -k\int |f|\,d\mu$. But since we may take N arbitrarily large, it follows that $\int f\,d\mu \geq 0$.

Now, let k be arbitrary positive integers and define

$$f_k(x) = \begin{cases} f(x) & \text{if } n(x) \le k \\ 0 & \text{else} \end{cases}$$

We have $|f_k| \le |f|$ and so the f_k are dominated by $|f|$ and since f is μ-integrable, so are the f_k. Since the f_k converge pointwise to f, we have

$$\int f\, d\mu = \lim_{k\to\infty} \int f_k\, d\mu \ge 0\,,$$

by dominated convergence (Theorem 14.10). ■

We will need the contra-positive of this result. Here it is explicitly.

Corollary 14.17. *Suppose $\int f\, d\mu < 0$ (> 0), then there is a set P of positive μ-measure such that for all x in P, $S_f^n(x) < 0$ (> 0) for all n.*

Under the hypotheses of remark 14.15, the statement of Theorem 9.10 is as follows.

Theorem 14.18 (Birkhoff or Pointwise Ergodic Theorem). *The limit of the time average*

$$\langle f\rangle(x) := \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x))$$

is defined on a set of full measure. It is an integrable function and satisfies (wherever defined)

$$\int_X \langle f\rangle(x)\, d\mu = \int_X f(x)\, d\mu\,.$$

Proof. We want to compute the limit of the time average of f. So let

$$\langle f\rangle^+(x) = \limsup_{n\to\infty} \frac{1}{n} S_f^n(x) \qquad \text{and} \qquad \langle f\rangle^-(x) = \liminf_{n\to\infty} \frac{1}{n} S_f^n(x)\,.$$

By Lemma 14.6 and the comments immediately prior to it, $\langle f\rangle^\pm$ are measurable functions. First suppose they are bounded. Then they are also integrable, because $\mu(X) = 1$.

Suppose that the following statement is *false*:

$$\int_X \langle f\rangle^-\, d\mu \ge \int_X f\, d\mu\,.$$

Then, since $\langle f\rangle^{\pm}$ and $\mu(X)$ are bounded, there must be an $\varepsilon > 0$ so that

$$\int_X \langle f\rangle^- \, d\mu < \int_X (f-\varepsilon)\, d\mu \quad \Longrightarrow \quad \int_X (\langle f\rangle^- - f + \varepsilon)\, d\mu < 0\,.$$

By the contrapositive of the maximal ergodic theorem, this gives that there are (a positive measure of) x so that $S^n_{\langle f\rangle^- - f+\varepsilon}(x) < 0$ for all n. Now, it is easy to see that $S^n_{f+g} = S^n_f + S^n_g$ and that $\langle f\rangle^-$ is invariant along orbits. Thus for any such x, we obtain that

$$n\langle f\rangle^-(x) - S^n_f(x) + n\varepsilon < 0 \qquad \text{or} \qquad \langle f\rangle^-(x) < \frac{1}{n} S^n_f(x) - \varepsilon\,.$$

If we take the lim inf as $n \to \infty$ on both sides, we arrive at a contradiction. This establishes that

$$\int_X \langle f\rangle^- \, d\mu \geq \int_X f\, d\mu\,. \tag{14.4}$$

In a similar way (exercise 14.12), one derives that

$$\int_X f\, d\mu \geq \int_X \langle f\rangle^+ \, d\mu\,. \tag{14.5}$$

Putting (14.4) and (14.5) together shows that if $\langle f\rangle^{\pm}$ are bounded, then

$$\int_X \langle f\rangle^+ \, d\mu \leq \int_X f\, d\mu \leq \int_X \langle f\rangle^- \, d\mu\,.$$

and thus all three are equal. Since $\langle f\rangle^+(x) \geq \langle f\rangle^-(x)$, we also have that these two quantities must be equal except on a set of measure zero. This proves that for almost all x, the limit of $\frac{1}{n} S^n_f(x)$ exists.

Now we drop the assumption that $\langle f\rangle^{\pm}$ are finite. So let

$$X_n := \{x \in X \,:\, -n \leq \langle f\rangle^-(x) \leq \langle f\rangle^+(x) \leq n\}\,.$$

T maps X_n to itself and so all hypotheses hold for X_n and therefore the above conclusion holds for X_n instead of X, and thus for $X_\infty = \cup_n X_n$. We are done if $X\backslash X_\infty$ has μ-measure zero. Now, X_n is measurable because $\langle f\rangle^{\pm}$ are, and so X_∞ and its complement are also measurable. Suppose the complement has positive measure, then since f is integrable, there must be a $c > 0$ so that

$$\int_{X\backslash X_\infty} -c\, d\mu < \int_{X\backslash X_\infty} f\, d\mu < \int_{X\backslash X_\infty} c\, d\mu\,.$$

We apply again the contrapositive of the maximal ergodic theorem, to get that then there must be a (positive measure of) x in $X\backslash X_\infty$ so that for all n

$$S^n_{(f+c)}(x) > 0 \ \text{ and } \ S^n_{(f-c)}(x) < 0 \quad \Longrightarrow \quad -nc < S^n_f(x) < nc\,.$$

But this contradicts the definition of $X \backslash X_\infty$. ■

As mentioned in Chapter 9, it is frequently the following Corollary which one has in mind when referring to Birkhoff's ergodic theorem.

Corollary 14.19. *A transformation $T : X \to X$ that preserves a probability measure μ has the property that every T-invariant set has measure 0 or 1 if and only if for every integrable function f*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x)) = \int_X f(x)\, d\mu$$

for all x except possibly on a set of measure 0.

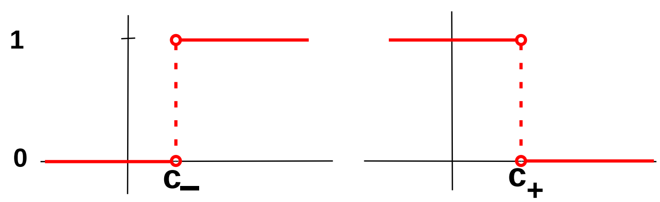


Figure 90. The functions $\mu(X_c^-)$ and $\mu(X_c^+)$.

Proof. By Theorem 14.18, $\langle f\rangle(x)$ is defined on a set of full measure. So let

$$X_c^- := \{x \in X \ : \langle f\rangle(x) < c\} \qquad \text{and} \qquad X_c^+ := \{x \in X \ : \langle f\rangle(x) > c\}\ .$$

Replacing x by an inverse image (under T) of x does not change the value of $\langle f\rangle(x)$, and so $X_c^\pm$ are invariant sets. By the ergodic assumption, $\mu(X_c^-)$ (as a function of c) must have measure 0 or 1, and is therefore an increasing step function with the step of height 1 occurring, say, at $c = c_-$. Similarly, $\mu(X_c^+)$ is a step function, with an decreasing step of height 1 occurring at $c = c_+$. See Figure 90.

If $c_- < c_+$, then for any interval $[c_1, c_2] \in (c_-, c_+)$, we obtain that $\mu(X_{c_1}^-) = \mu(X_{c_2}^+) = 1$, which is impossible, since these sets do not intersect. In the same way, if $c_+ < c_-$, then for any interval $[c_1, c_2] \in (c_+, c_-)$ $\mu(X_{c_1}^+) = \mu(X_{c_2}^-) = 0$, which contradicts the fact that the union of $X_{c_1}^+$ and $X_{c_2}^-$ is the entire space and so must have measure 1. So $c_- = c_+ = c_0$. Thus $\langle f\rangle(x) = c_0$ on a set of full measure. And therefore Theorem 14.18 implies that $\int_X f(x)\,d\mu = \int_X \langle f\rangle(x)\,d\mu = c_0$, which implies that time average equals space average.

Vice versa, if T is *not* ergodic, then there are invariant sets X_1 and its complement X_2 both of positive measure. Let $\mathbf{1}_{X_1}$ be the function that is 1 on X_1 and 0 elsewhere. The time average $\langle \mathbf{1}_{X_1}\rangle(x)$ is 1 or 0, depending on where the starting point x is. In either case, it is not equal to the spatial average $\int_X \mathbf{1}_{X_1}(x)\,d\mu \in (0,1)$. ■

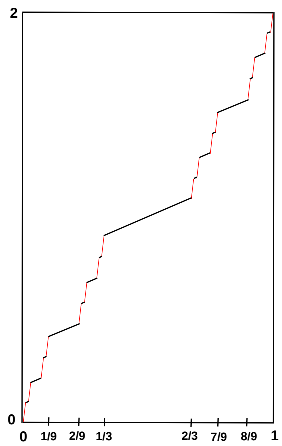
In Section 9.3, we observed that ergodic measures are the building blocks of chaotic dynamics. Thus transformations where there is a unique ergodic measure are especially interesting.

Definition 14.20. *A transformation T of a measure space is uniquely ergodic if there is a unique Borel probability measure with respect to which T is ergodic.*

14.7. Exercises

Exercise 14.1. a) Show that $g_n(X) = \sup_{i\geq n} f_i(x)$ is non-increasing (in n).
b) Let $f_n(x) = \sin nx$. Determine $\limsup_n f_n(1)$. (*Hint: use Lemma 10.6*).
c) Show that the twin prime conjecture (Conjecture 1.29) is equivalent to $\liminf_n p_{n+1} - p_n = 2$ (p_n is the nth prime).

Exercise 14.2. a) Give a definition of a measurable function f from a topological space to $\mathbb{C}$. (*Hint: split up the real and imaginary parts and use the Borel sigma algebra as measurable sets, then follow Section 14.1.*)
b) Show that if c is a constant and a real function f measurable, then cf is measurable.
c) Consider the set V in Section 9.1 and show that it is not measurable.
d) Consider the function $\mathbf{1}_V$ which is 1 on points in V and 0 elsewhere. Show that $\mathbf{1}_V$ is not measurable.



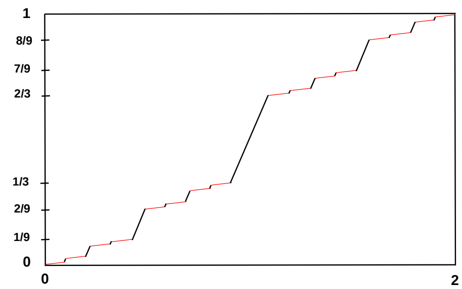


Figure 91. Left: an impression of the function $g(x) = c(x) + x$ where $c(x)$ is taken from figure 54. Left: its inverse $h := g^{-1}$ is well-defined but not Lebesgue measurable. Linear interpolated segments are colored red.

Exercise 14.3. In this exercise, we exhibit a non-trivial function g that is not Lebesgue measurable. The student should review exercises 9.8–9.13 for the properties of the Cantor function $c(x)$. Define $g(x) = c(x) + x$ so that $g : [0,1] \to [0,2]$. Its inverse is denoted by h (see Figure91). The Lebesgue measure is denoted by μ.
a) Show that g is invertible and call its inverse h. (*Hint: c is strictly increasing on the Cantor set C, so g is strictly increasing on* $[0,1]$).
b) Show that the complement O of $g(C)$ is Lebesgue measurable, and that $\mu(O) = 1$. (*Hint: O is a countable collection of open intervals whose lengths sum to 1.*)
c) Use (b) to show that $\mu(g(C)) = 1$. (*Hint: Corollary 9.6.*)
d) Show that there is a non-measurable set $W \subset g(C)$. (*Hint: recall that* $V \subseteq [0,1]$*; insert the pieces corresponding to O into* $[0,1]$. *Bits of V will be translated, but this does not affect their measure. When finished, you have a non-measurable set* $W \subset g(C) \subset [0,2]$.)
e) Show that $\mu(h(W)) = 0$, and so $h(W)$ is measurable. (*Hint:* $h(W)$ *is contained in C. Use Lemma 9.20.*)
f) Use (a) and (e) to show that there is a Lebesgue measurable set Z such that $h^{-1}(Z) = W$ and W is not measurable.

Exercise 14.4. Explain why Henri Lebesgue wrote the following about his method of integration (as cited by [**29**][page 796]):
"I have to pay a certain sum, which I have collected in my pocket. I take the bills and coins out of my pocket and give them to the creditor in the order I find them until I have reached the total sum. This is the Riemann integral. But I can proceed differently. After I have taken all the money out of my pocket I order the bills and coins according to identical values and then I pay the several heaps one after the other to the creditor. This is my integral."

Exercise 14.5. In this exercise, we show that the dominated convergence with Riemann integration cannot be saved even by restricting to continuous functions $f:[0,1]\to[0,1]$. Let f_n be given as follows, see Figure 92. For every pair (j,k) with $\gcd(j,k)=1$ and so that $j/k\in[0,1]$, define

$$h_n(j,k,x)=\max\left\{0,1-n^3\left|x-\frac{j}{k}\right|\right\} \quad\text{and}\quad f_n(x):=\sum_{\substack{j/k\in[0,1]\\ \gcd(j,k)=1\\ k\le n}} h(j,k,x)\,.$$

a) Show that f_n is continuous and dominated by $g(x)=1$. (*Hint: show that the minimal distance between the centers of any two "triangles" $h_n(j,k,x)$ defined in (a) is at least* $1/n^2$.)
b) Show that $\lim_{n\to\infty}\int f_n\,dx=0$. (*Hint: first give a rough estimate how many rationals with denominator less than* $n+1$ *there are in the unit interval.*)
c) Let $r\in[0,1]$ be an algebraic number of degree at least two. Show that Roth's theorem (Theorem 1.21) implies that for all $\varepsilon>0$

$$\exists\, c(r,\varepsilon)>0 \;\text{ such that }\; \forall\,\frac{j}{k}\in\mathbb{Q}\;:\;\left|r-\frac{j}{k}\right|>\frac{c(r,\varepsilon)k^{1-\varepsilon}}{n^3}\,.$$

d) Show that (c) implies that for every algebraic number r of degree at least two, $\lim_n f_n(r)=0$.
e) Use exercise 7.14 (e) to show that for this example dominated convergence does not hold for Riemann integration. (*Hint: show that the Riemann integral of* $\lim_n f_n$ *is not defined.*)
f) Explain that if you use Lebesgue integration, there is no problem.

Exercise 14.6. Let $f_n(x)=n$ for $x\in[0,1/n]$ and 0 elsewhere and set $g(x)=1/x$.
a) Show that g dominates the f_k. See Figure 93.
b) Show that $\lim_{k\to\infty}\int f_k\,d\mu\neq\int\lim_{k\to\infty}f_k\,d\mu$.
c) Why do (a) and (b) not contradict Theorem 14.10?

Exercises 14.7 and 14.8 provide a more interesting illustration of the dominated convergence theorem. Generalizing exercise 11.19, for fixed $r\ge$

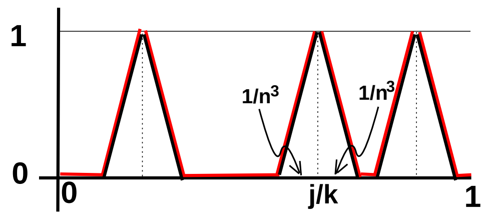


Figure 92. The function f_n (in red) in exercise 14.5 is a sum of very thin triangles with height 1. Each triangle is given by $h_n(j,k,x)$ (in black).

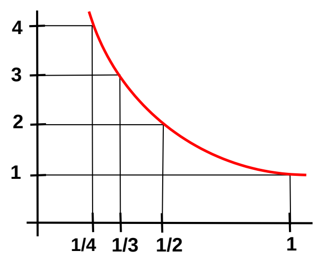


Figure 93. The functions f_n in exercise 14.6 and the function g (in red) that dominates them.

1, consider the functions $g_k(x) = k^r x^k(1-x)$ on $[0,1]$. Define $G_k(x) = \sup_{i \leq k} g_i(x)$ and $G(x) = \sup_i g_i(x)$.

Exercise 14.7. a) Show that $g_k(x)$ is increasing on $[0, \frac{k}{k+1}]$ and decreasing on $[\frac{k}{k+1}, 1]$.

b) Show that g_k has maximum $k^{r-1}\left(\frac{k}{k+1}\right)^{k+1} \approx k^{r-1}e^{-1}$. (*Hint:* $\lim_{k\to\infty}(1+a/k)^k = e^a$.)

c) Show that $g_{k-1}(x) = g_k(x)$ iff $x \in \left\{0, \left(\frac{k-1}{k}\right)^r, 1\right\}$ and that

$$g_k\left(\left(\frac{k-1}{k}\right)^r\right) = k^r\left(\frac{k-1}{k}\right)^{rk}\left(1-\left(\frac{k-1}{k}\right)^r\right) \approx k^{r-1}re^{-r}.$$

d) Show that $g_k(x) = g_{k+1}(x)$ iff $x \in \left\{0, \left(\frac{k}{k+1}\right)^r, 1\right\}$ and that

$$g_k\left(\left(\frac{k}{k+1}\right)^r\right) = (k+1)^r\left(\frac{k}{k+1}\right)^{r(k+1)}\left(1-\left(\frac{k}{k+1}\right)^r\right) \approx (k+1)^{r-1}re^{-r}.$$

e) Show that $\left(\frac{k}{k+1}\right)^r - \left(\frac{k-1}{k}\right)^r \approx rk^{-2}$. (*Hint: compute the first non-zero term in the expansion of* $(1+x)^{-r} - (1-x)^r$.)

f) Show that $\int G(x)\,dx$ is "sandwiched" between the sum S_1 of the areas of the rectangles like the one shaded red in Figure 94 and the sum S_2 of the red ones plus the green ones.

g) Use (c) through (f) to show that that there are functions c_1 and c_2 such that

$$S_1 = c_1(r)\sum k^{r-1}k^{-2} \quad \text{and} \quad S_2 = c_2(r)\sum(k+1)^{r-1}k^{-2}.$$

h) Conclude that G is integrable iff $r < 2$.

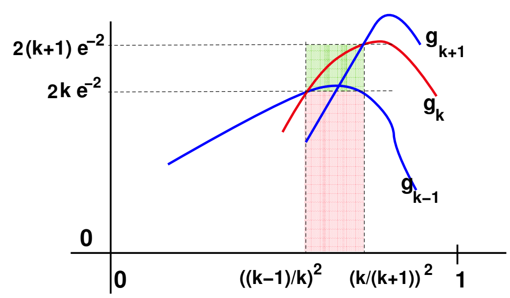


Figure 94. In this figure $r = 2$. We show the function $g_k(x)$ (red) on $[0,1]$ and its intersections. The sum of the rectangles like the one shaded in red give a lower bound for $\int G_k\,dx$ while the sum of the red and green rectangles give an upper bound.

Exercise 14.8. a) Use exercise 14.7 (f) to show that the dominated convergence theorem implies that for $r < 2$, we have

$$\int_0^1 \lim_{k\to\infty} g_k(x)\,dx = \lim_{k\to\infty}\int_0^1 g_k(x)\,dx\,.$$

b) What goes wrong for $r \geq 2$?
c) Show that

$$\int_0^1 \lim_{k\to\infty} g_k(x)\,dx = 0 \ \text{ and } \ \int_0^1 g_k(x)\,dx = \frac{k^r}{(k+1)(k+2)}\,.$$

d) Why is (c) consistent with (a) and (b)?

Exercise 14.9. Let A be a compact collection of *irrational* numbers and and $\{n_i\}$ a sequence of natural numbers whose partial sums satisfy

$$S_k = \sum_{i=1}^{k} n_i \qquad \text{where} \qquad \lim_{k\to\infty}\frac{S_k}{k} = \infty\,.$$

Create a sequence $\{x_i\}$ of real numbers as follows. Choose an x_0 and set $n_0 = 0$. For $i \in \{1,\cdots,n_1\}$, let $x_i = x_{i-1} + \alpha_1$ where $\alpha_1 \in A$; for $i \in \{S_1+1,\cdots,S_2\}$, let $x_i = x_{i-1} + \alpha_2$ where $\alpha_2 \in A$; and so on.
a) Show that for any fixed $m \neq 0$ in $\mathbb{Z}$, there is a $\varepsilon_m > 0$ so that

$$\min_{\alpha\in A}\left|e^{2\pi i m\alpha} - 1\right| > \varepsilon_m\,.$$

(*Hint: the compactness of the set of irrational numbers is crucial.*)
b) Show that

$$\frac{1}{S_k}\sum_{n=0}^{S_k-1} e^{2\pi i m x_n} = \frac{1}{S_k}\left\{e^{2\pi i m x_0}\sum_{n=0}^{n_1-1} e^{2\pi i m n\alpha_1} + \cdots + e^{2\pi i m x_{S_{k-1}}}\sum_{n=0}^{n_k-1} e^{2\pi i m n\alpha_k}\right\}.$$

c) Use the geometric series as in Section 14.5 to show that for each sum in (b), we obtain

$$\sum_{n=0}^{n_\ell-1} e^{2\pi i m n\alpha_k} = \frac{e^{2\pi i m n_\ell\alpha_k} - 1}{e^{2\pi i m\alpha_k} - 1}\,.$$

d) Use (a) and (c) to show that

$$\left|\sum_{n=0}^{n_\ell-1} e^{2\pi i m n\alpha_k}\right| < 2\varepsilon_m^{-1}\,.$$

e) Use (d) and the condition on the partial sums to show that

$$\lim_{k\to\infty}\frac{1}{S_k}\sum_{n=0}^{S_k-1} e^{2\pi i m x_n} = 0\,.$$

f) Show that (e) and Weyl's criterion imply that the sequence $\{x_i\}$ is equidistributed modulo 1.(*Hint: you need to pass from* $\lim_{k\to\infty}\frac{1}{S_k}\sum_{n=0}^{S_k-1}$ *to* $\lim_{S\to\infty}\frac{1}{S}\sum_{n=0}^{S}$*; so vary the value of the last* n_k.)

Many number theory textbooks (correctly) state that the fractional parts of $f(n) = \ln p_n$, where p_n denotes the nth prime, are not equidistributed. This is slightly misleading because an unsuspecting student could be tempted into wondering to what mysterious distribution the numbers the fractional parts of $\ln p_n$ would deign to converge to? The answer — perhaps somewhat disappointingly — is that the logarithm increases so slowly that in fact those numbers do *not* converge *at all* as we show in exercises 14.10 and 14.11. We denote the fractional part of x by $\{x\}$. For a *slowly* increasing function $f : \mathbb{N} \to \mathbb{R}$ and an interval J of the unit circle $[0,1]/\{0=1\}$, we define the "hitting frequency" as follows:

$$F_J(0,n) := \frac{\#\{\{f(i)\} \in J \quad \text{for} \quad i \in \{1,\cdots,n\}\}}{n} .$$

Note that if the fractional parts of $\{f(n)\}$ converge to any distribution whatsoever, then there is a $c \in [0,1]$ so that $\lim_{n\to\infty} F_J(0,n) = c$.

Exercise 14.10. In this exercise, we set $f(n) := \ln n$ and let $J = [\alpha, \alpha+\delta)$ be an arbitrary interval of length δ in the unit circle. For $K \in \mathbb{N}$ and n_K, choose n'_K so that

$$f(n_K) \leq K + \alpha < f(n_K + 1) \qquad \text{and} \qquad f(n'_K) \leq K + \alpha + \delta < f(n'_K + 1) .$$

a) Show that

$$\lim_{K\to\infty} \frac{n_K}{n'_K} = e^{-\delta} .$$

b) Assuming that $\lim_{n\to\infty} F_J(0,n) = c$, show that (see Figure 95)

$$n'_K F_J(0,n'_K) \approx n_K \cdot c + (n'_K - n_K) \cdot 1 .$$

c) Show that (b) implies that

$$0 = \lim_{K\to\infty} F_J(0,n'_K) - c = (1-c)(1-e^{-\delta}) .$$

d) Conclude that the fractional parts of $f(n) = \ln n$ do not converge to any distribution. (*Hint: we can only get convergence if the the hitting frequency converges to 1 for every interval J.*)

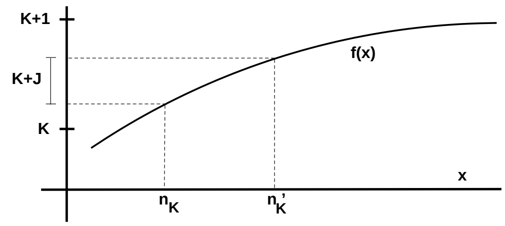


Figure 95. A schematic illustration of the quantities defined in exercises 14.10 and 14.11.

Exercise 14.11. In this exercise, we set $f(n) := \ln p_n$, where p_n are the primes. The definitions of J, n_K, and n'_K are as in exercise 14.10.
a) Recall the prime counting function (defined in Theorem 2.21) and show that

$$n_K = \pi\left(e^{K+\alpha}\right) \quad \text{and} \quad n'_K = \pi\left(e^{K+\alpha+\delta}\right).$$

b) Use Chebyshev's theorem (Theorem 12.7) to show that there are positive a and b (with $a < 1 < b$) so that for large enough K,

$$q_K := \frac{n_K}{n'_K} \in \left[\frac{a}{b}\,e^{-\delta}, \frac{b}{a}\,e^{-\delta}\right].$$

c) Assuming that $\lim_{n\to\infty} F(0,n) = c$, use the reasoning of exercise 14.10 to show that

$$0 = \lim_{K\to\infty} F_J(0,n'_K) - c = \lim_{K\to\infty} (1-c)(1-q_K).$$

d) Show that $c = 1$ leads to a contradiction.
e) Show that a and b and δ can be chosen so that $\limsup_{K\to\infty} q_K < 1$.
f) Show that for any interval with the choices as in (e), we must have $c = 1$.
g) Let $\{I_i\}$ be a finite set of intervals as in (f). Show that the hitting frequency in $J := \cap I_i$ must be be 1.
h) Show that the intervals in (f) can be chosen so that they have empty intersection.

Exercise 14.12. See the proof of Theorem 14.18.
a) Show that $S^n_{f+g} = S^n_f + S^n_g$.
b) Show that $\langle f\rangle^-$ is invariant along orbits.
c) Use (a) and (b) to show that $S^n_{\langle f\rangle^- - f+\varepsilon}(x) = n\langle f\rangle^-(x) - S^n_f(x) + n\varepsilon$.
d) Use (a), (b), and (c) to deduce a contradiction from $\varepsilon > 0$ and

$$\liminf_{n\to\infty} \langle f\rangle^-(x) < \liminf_{n\to\infty}\left(\frac{1}{n}S^n_f(x) - \varepsilon\right).$$

Exercise 14.13. See the proof of Theorem 14.18.
a) Show that T maps X_n to itself.
b) Use the results in Section 14.2 to show that $X \backslash X_\infty$ is measurable.
c) Show that under the hypotheses of the proof, there must be a $c > 0$ so that $\int_{X \backslash X_\infty} (f - c)\, d\mu > 0$.

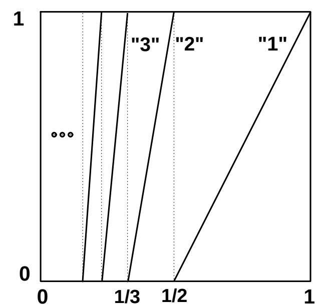


Figure 96. A few branches of the Lüroth map T of exercise 14.14. The names of the branches are as indicated in the figure. The Lüroth expansion of $x \in]0,1]$ is $[a_1, a_2, \cdots]$, where the ith digit a_i equals k if $T^{i-1}(x)$ falls in the domain of the kth branch.

Exercise 14.14. The Lüroth map $T : [0,1) \to [0,1)$ is defined by

$$T(x) = \begin{cases} n(n+1)x - n, & \text{if } x \in [\frac{1}{n+1}, \frac{1}{n}) \\ 0 & \text{if } x = 0 \end{cases}$$

where $n \geq 1$. See Figure 96.
a) Show that T preserves the Lebesgue measure and is ergodic. (*Hint: see Theorem 10.9.*)
b) Show that for almost all x, the digit k defined in Figure 96 has a frequency of $\frac{1}{k(k+1)}$ in the expansion of x for $k \geq 1$.
c) Show that almost all x have Lyapunov exponent (Definition 10.19)

$$\lambda(x) = \sum_{k=1}^{\infty} \frac{\ln k(k+1)}{k(k+1)} \approx 2.05\,.$$

(*Hint: see exercise 10.23.*)

Exercise 14.15. This exercise relies on exercise 14.14 and Section 6.6. Let $b_k(x) : I_k \to [0,1)$ be the branch of T^{k-1} such that $x \in I_k$, then the kth convergent $[a_1, \cdots, a_k]$ of x is the (unique) endpoint of I_k that maps to zero under T^k (see Proposition 6.16). The branches of T are labeled as indicated in Figure 96. For simplicity, we note (without proof) that the kth convergent is always a rational number also denoted by p_k/q_k. The Lüroth expansion of a number $x \in [0,1)$ is the list $[a_1, a_2, \cdots]$ where a_i is the label of the branch in whose domain $T^{i-1}(x)$ is located. For more details, see [**9**].

a) Show that

$$\left| x - \frac{p_k}{q_k} \right| < |I_k|,$$

where $|I_k|$ is the length of I_k.

b) Show that $T^k : I_k \to [0,1)$ is an affine bijection.

c) Show that

$$|I_{k+1}| \leq \left| x - \frac{p_k}{q_k} \right| < |I_k|,$$

(*Hint: b_k maps I_k affinely onto $[0,1)$ (see Figure 97) and so the sub-intervals of I_k have the same proportions as the sub-intervals of the unit interval in the Lüroth map of Figure 96.*)

d) Use (b) to show that

$$\ln \frac{1}{|I_k|} = \sum_{j=0}^{k-1} \ln \left| \partial T(T^j(x)) \right|.$$

(*Hint: see exercise 10.21; here we use the same notation.*)

e) Use (c) and (d) to show that

$$\lim_{k \to \infty} \frac{1}{k} \ln \left| x - \frac{p_k}{q_k} \right| = -\lambda(x),$$

where $\lambda(x)$ is the Lyapunov exponent (Definition 10.19) of T at x.

Exercise 14.16. This exercise is based on exercises 14.14 and 14.15.

a) Compare the almost everywhere convergence of the continued fraction convergents with the Luroth convergents. (*Hint: one is alternating and converges faster.*)

b) Can you venture an intuitive explanation for the faster convergence? (*Hint: look at T^2 in both cases.*)

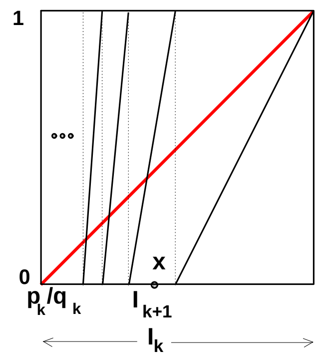


Figure 97. A few branches of the $k+1$st iterate of the Lüroth map T restricted to the interval I_k. In red a branch of T^k and in black a few branches of T^{k+1}.

Exercise 14.17. Two measures ν and μ are said to be in the same measure class if they have the same sets measure zero sets. Suppose we fix a measure class and are given that T is ergodic with respect to an (unknown) ergodic measure in this class.
a) Given a set S and its characteristic function $\{\mathbf{1}\}_\mathbf{S}$. Show that $\mu(S) = \int \{\mathbf{1}\}_\mathbf{S}\, \mathbf{d}\mu$.
b) Use (a) and Corollary 14.19 to show that

$$\mu(S) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x))\,.$$

c) Show that this determines the measure μ.
d) Show that if there was another ergodic measure ρ, then it would live entirely in the sets of μ-measure zero. (*Hint: see Corollary 9.15.*)

As noted before, in the more common definition of ergodicity (Definition 9.12) "weak invariance" is replaced by the "strict invariance". In exercise 14.18, we show that these notions are in fact equivalent. We assume that T is ergodic with respect to the (invariant) measure μ.

Exercise 14.18. a) Show that ergodicity with weak invariance implies ergodicity with strict invariance. (*Hint: this is trivial.*)
b) Now assume that T is "strict invariance" ergodic with respect to the measure μ, and let S_0 be a weakly invariant set of positive measure. Show that

$$S := \bigcap_{k=0}^{\infty} \bigcup_{i \geq k} T^{-i}(S_0)\,,$$

is strictly invariant. (*Hint: $x \in S$ if and only if $x \in T^{-k}(S_0)$ for infinitely many distinct k.)*
c) Denote the symmetric difference by $\triangle$ (see Figure 98). Show that for

$$k \geq 0 \;:\; \mu(T^{-k-1}(S_0) \triangle T^{-k}(S_0)) = 0\,.$$

(*Hint: for $k = 0$, this follows from the definition of S_0; then use the fact that T preserves μ.)*
d) Use (c) and Figure 98) to show that for

$$k \geq 0 \;:\; \mu(T^{-k}(S_0) \triangle S_0) = 0\,.$$

e) Show that

$$S \triangle S_0 \subseteq \bigcup_{k \geq 0} T^{-k}(S_0) \triangle S_0\,.$$

f) Use (c) and (d) to show that $\mu(S_0) = \mu(S)$. (*Hint: T preserves measure.*)
g) Use ergodicity to show that $\mu(S_0)$ has full measure.

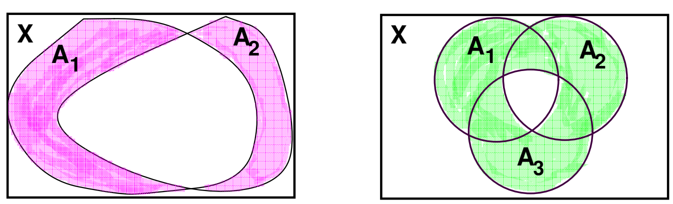


Figure 98. Left: the symmetric difference of two sets A_1 and A_2 is $A_1 \triangle A_2 := (A_1 \cup A_2) \backslash (A_1 \cap A_2)$. On the right side, we illustrate that $A_3 \triangle A_1$ is contained in $(A_3 \triangle A_2) \cup (A_2 \triangle A_1)$ (green).

Exercise 14.19. Suppose $T : X \to X$ where X is a compact, metric space. If T has a periodic orbit O, exhibit a (discrete) invariant measure μ which is ergodic. (*Hint: for any set A not intersecting O, $\mu(A) = 0$. See also Section 9.4.)*

Exercise 14.20. For this exercise, *assume* that the linear combinations of the functions $e^{2\pi imx}$ are dense in the set $L^1(\mathbb{R}/\mathbb{Z})$ of integrable function on the circle. (Let f be integrable, then for every $\varepsilon > 0$, there is a finite combination $\tilde{f}$ of multiples of $e^{2\pi imx}$ so that $\int |f - \tilde{f}| dx < \varepsilon$.)
a) Show that the Lebesgue measure is ergodic and measure preserving if and only if for all $m \neq 0$ in $\mathbb{Z}$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} e^{2\pi imT^k(x)} = 0 .$$

(*Hint: use the proof of Theorem 14.14.*)
b) Show that T in (a) is ergodic if and only if $\{T^k(x)\}$ is equidistributed.

We saw in Section 9.4 that a given transformation T may have uncountably many coexisting invariant measures. The Krylov-Bogoliubov theorem (see [**38**]) states that a continuous map T from a compact metric space to itself has an (at least one) invariant Borel probability measure. Exercise 14.21 gives a counterexample if we drop continuity.

Exercise 14.21. $T : [0,1] \to [0,1]$ is given by $T(x) = x/2$ if $x > 0$ and $T(0) = 1$. Assume that there exists an invariant probability measure μ satisfies $\mu(T^{-1}(A)) = \mu(A)$ and such that μ is defined on all open sets.
a) Show that if $\mu((1/2,1)) = p > 0$, then $\mu((0,1))$ is unbounded, a contradiction. (*Hint: use Definition 14.4.*)
b) Show that (a) implies that all measure must be concentrated on the points $\{2^{-i}\}_{i=0}^{\infty}$ and $\{0\}$.
c) Show that if any of the points in (b) carry positive measure, then we also get a contradiction. (*Hint: similar to (a).*)
d) Conclude that it is impossible to consistently assign an invariant measure to open sets.
e) Show that there *does* exist an invariant measure on the trivial sigma algebra. (*Hint: what is the smallest σ-algebra possible under Definition 14.2?*)

Exercise 14.22. $T : [0,1] \to [0,1]$ is given by $T(x) = x/2$.
a) Show that T has a unique invariant Borel probability measure, namely $\mu(\{0\}) = 1$. (*Hint: use the strategy of exercise 14.21.*)
b) Show that with respect to the measure in (a), T is ergodic.
c) Show that T is uniquely ergodic (Definition 14.20).

Proposition 14.21. *Suppose $T : X \to X$ where X is a compact, metric space. If T has a unique invariant Borel probability measure μ, then that measure is the unique ergodic measure for T.*

Exercise 14.23. We prove Proposition 14.21 as in [**38**] [Section 4.1]. For any measurable set A, define the conditional measure μ_A as

$$\mu_A(B) = \frac{\mu(A \cap B)}{\mu(A)}\,.$$

Assume that there is an invariant measurable set S with $0 < \mu(S) < 1$.
a) Show that μ_S is an invariant measure.
b) Show that $\mu_{X\setminus S}$ is an invariant measure.
c) Show that the measures in (a) and (b) are distinct. (*Hint: what is the measure of S?*)
d) Show that (c) contradicts the hypothesis of Proposition 14.21.

Chapter 15

The Unsolvability of the Quintic

Overview. In 1963, V. I. Arnold proposed a proof with a topological flavor of the insolvability by radicals of the quintic that was (supposed to be) accessible for high school students. The proof is notable in that it uses much less algebra than the traditional proof based on Galois theory. It also has the advantage that it is conceptually much clearer.

To our knowledge this proof made the (English language) press only in 2004 in [**2**]. It was the subject of a paper [**59**] in the Monthly as recently as 2022. We will give a proof loosely based on these recent publications.

15.1. Solvable Groups

We start by defining the permutation group on n symbols. Recall that a permutation

Definition 15.1. *The group of permutations of* $\{1,\cdots,n\}$ *is denoted by* S_n*. These groups are called* <u>*symmetric*</u> <u>*groups*</u>*.*

The reason for that name is as profound — in this context — as it is obvious. Given a polynomial f of degree n, it has n roots counting multiplicity. Typically these roots are distinct. They do not come in any specified order, so if we permute them, they still are the roots of that exact same polynomial f. Thus the group of permutations on n (generally) distinct roots are a groups of symmetries of the polynomial f.

The identity element $\{e\}$ of a group G is a normal subgroup. Thus if G is Abelian, then the 'quotient' $G/\{e\}$ is obviously also Abelian. In this sense, the following definition can be seen as a generalization of Abelian.

Definition 15.2. *A group G is solvable if it has subgroups G_i with*

$$\{e\} = G_m \subseteq G_{m-1} \cdots \subseteq G_0 = G\,,$$

such that G_i is normal in G_{i+1} and, furthermore, G_{i+1}/G_i is Abelian for $i \in \{0, \cdots, m-1\}$.

The denomination *solvable* is, again, no coincidence. As we shall see, the polynomials of degree 4 and less can be *solved by radicals* (defined below) precisely because the group of permutations of the 4 (or less) roots has the above property of being solvable. This is not true for the group of permutations of 5 (or more)elements, and hence, the general quintic is not solvable by such elementary means.

In an alternative definition of a solvable group, the requirement that G_{i+1}/G_i is Abelian is replaced by the requirement that G_{i+1}/G_i must be cyclic. However, by the fundamental theorem for finite Abelian groups (Proposition 13.3), this is the same thing.

We recall that a subgroup H of G gives rise to a quotient group G/H if and only the subgroup is normal (Definition 7.30). In this case, the cosets Hx partition the original group G. It follows that the order of H is a divisor of the order of G. Among other things, it also means that cosets can be multiplied and left and right cosets are the same. So for x and y in G:

$$HxHy = Hxy \qquad \text{and} \qquad Hx = xH\,. \tag{15.1}$$

Now G/H is Abelian if and only if for all x and y, $HxHy = HyHx$, or, equivalently, $HxHy(Hx)^{-1}(Hy)^{-1} = H$. This always holds if and only if elements of the form $xyx^{-1}y^{-1}$ are in the subgroup H. See exercise 15.5 and, for more details, Chapter 15 of [**54**].

Definition 15.3. *Elements of a group G of the form $aba^{-1}b^{-1}$ are called commutators of G. The subgroup generated by the commutators is called the commutator subgroup or the derived subgroup and indicated by $[G,G]$.*

Lemma 15.4. *The commutator subgroup is normal.*

Proof. We use Definition 7.30. It is easy to see that

$$xaba^{-1}b^{-1}x^{-1} = \left(xax^{-1}\right)\left(xbx^{-1}\right)\left(xa^{-1}x^{-1}\right)\left(xb^{-1}x^{-1}\right),$$

which is again a commutator and is thus in the commutator subgroup. The same works for a product of commutators (exercise 15.1). ∎

Two important remarks are in order here. The first is that the commutator subgroup contains the commutators, of course, but *may* contain other elements as well (exercises 15.2 and 15.3). The second is that group $G/[G,G]$ is, in a sense, the 'Abelian part' of G. So the smaller the derived group is, the 'more' Abelian G is.

If a finite group G "lacks" solvability, it must be because at some point, the commutator subgroup G_{i+1} of G_i stopped becoming smaller than G_i itself. In other words, the commutator subgroup of G_i is equal to itself. Thus a natural way to test if a group $G := G_0$ is solvable, is by considering the sequence of commutator subgroup $G_{i+1} := [G_i, G_i]$. If we start with a finite group, this sequence of subgroups of decreasing must eventually end with $G_{m+1} = G_m$ for some m. If G_m consists of the identity, then G is solvable. If G_m is a non-trivial group, then G is not solvable.

Definition 15.5. *The descending series of normal subgroups,*

$$G := G_0 \rhd G_1 \rhd \cdots G_m \rhd \cdots,$$

where each group is the commutator subgroup of the previous entry, is called the <u>*derived*</u> <u>*series*</u>*. Here we used the standard notation $G \rhd H$ for H a proper normal subgroup of G.*

Definition 15.6. *A* <u>*simple*</u> <u>*group*</u> *is a group with no non-trivial normal subgroup. A* <u>*perfect*</u> <u>*group*</u> *is a group that is generated by its commutators.*

Thus the commutator subgroup of a simple group must be equal to the identity or to itself. In other words: that group is perfect. The reverse, however, is false as be seen in exercise 15.4, where we show that $SL_2(\mathbb{R})$ has a non-trivial normal subgroup, even though its commutator subgroup equals the whole group.

At this point, the reader could be excused for fearing that a (finite) group G might admit several very different and non-unique descending series of proper subgroups. But it turns out that if at every stage you quotient by the maximal proper subgroup, the sequence is essentially unique. This general statement is called the Jordan-Hölder theorem , see [**10**]. Applied to cyclical groups $\mathbb{Z}_n$, this yields the unique factorization theorem (Theorem 2.11). The Jordan-Hölder theorem can, in fact, be seen as a generalization of the unique factorization theorem.

15.2. The Derived Series of S_n

Recall that a permutation on n symbols can be written as a product of a number of transpositions. The parity of the permutation indicates whether that number is even or odd and is well-defined (see exercise 15.7). In what follows, we assume the reader knows cycle notation for permutations; details can be found [**54**]. To avoid ambiguities, we only remark that $\sigma = (bdce)(abc)$ is read as follows: σ maps b first to c and then to e, so $\sigma(b) = e$, and so forth. See Figure 99.

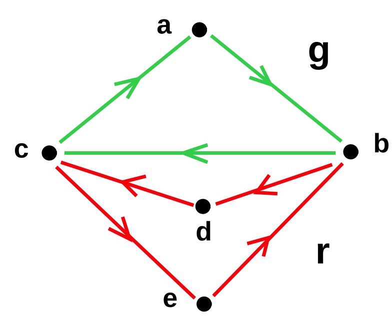


Figure 99. The permutation $\sigma = r \circ g$ consist of first applying the green cycles g and then the red cycle r.

Definition 15.7. *A alternating group of degree n is the group of even permutations on n symbols. It is denoted by A_n.*

Proposition 15.8. *A_n consists of products of 3-cycles. Furthermore, the commutator subgroup of S_n for $n > 2$ is A_n: $[S_n, S_n] = A_n$.*

Proof. Given a permutation in A_n. Since it is even, we can write it as a composition of pairs of transpositions. There are three possibilities for each pair:

$$\begin{aligned}(ab)(ab) &= I\\ (ab)(bc) &= (abc)\\ (ab)(cd) &= (abc)(bcd)\,.\end{aligned}$$

Thus every even permutation can be written as a product of 3-cycles.

On the other hand, every 3-cycle can be written as a commutator. This can be most easily seen by realizing that a transposition is its own inverse, and so

$$(ab)(ac) = (acb) \quad \Longrightarrow \quad (ab)(ac)(ab)(ac) = (abc)\,. \tag{15.2}$$

Thus every element of A_n can be written as a product of 3-cycles and therefore of commutators. Furthermore, the number of transpositions in a commutator must be a multiple of four and so their product must be in A_n. ■

Proposition 15.9. *For $n \geq 5$, A_n is perfect: $[A_n, A_n] = A_n$.*

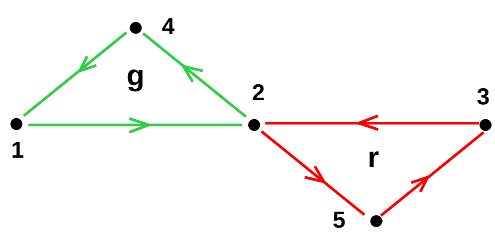


Figure 100. Denote the red 3-cycle by r and the green by r. The commutator $r^{-1}g^{-1}rg$ equals the 3-cycle (123).

Proof. The crux here is that if we have at least five symbols at our disposal, we can write every 3-cycle as a commutator of *even* permutations. From (15.2), we see that every 3-cycle is a commutator. So it is sufficient to write an arbitrary 3-cycle as a commutator of 3-cycles. From Figure 100 we conclude that the cycle $(1,2,3)$ can be written as

$$(235)(421)(532)(124) = (532)^{-1}(124)^{-1}(532)(124) = (123)\,.$$

We need to use two 'extra' symbols (4 and 5) to do this. ■

This gives us the derived series for S_n with $n \geq 5$, that leaves the cases $n \in \{2,3,4\}$ open. Now S_2 has only two permutations, namely the identity and (12). Its derived series is $S_2 \rhd \{e\}$. S_3 has $3! = 6$ permutations. By Proposition 15.8, A_3 is generated by the 3-cycles and so it must consist of (123) plus its iterates. This group has three elements. The order of any normal subgroup must be a divisor of 3, and therefore can only be A_3 or the identity $I = \{e\}$. A trivial check shows that the latter is the case. That leaves only S_4 to be analyzed.

Lemma 15.10. *Let* $V = \{I,(12)(34),(13)(24),(14)(23)\}$. *We have* $[A_4,A_4] = V$ *and* $[V,V] = \{e\}$.

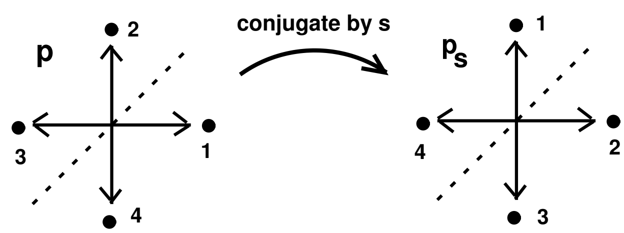


Figure 101. The effect of conjugating a permutation p by the reflection s in the dotted line.

Proof. We start by noting that the commutator subgroup V is invariant under conjugation by Lemma 15.4. Thus if $p \in V$ then for any $s \in A_4$, we have that $p_s := sps^{-1}$ is in V. It follows that (see Figure 101)

$$\text{if } \ p : i \to j \ \text{ then } \ p_s : s(i) \to s(j)\,.$$

Next, we use that A_4 is generated by 3-cycles to compute one element of $[A_4,A_4]$.

$$(123)(234)(321)(432) = (14)(23)\,.$$

But by the previous paragraph, relabeling will give another element of $[A_4,A_4]$. All relabelings plus the identity give the elements in the statement of the proposition. Since we the order of $[A_4,A_4]$ must be a non-trivial divisor of 12 (the order of A_4), we can have at most two more elements in $[A_4,A_4]$.

If $[A_4,A_4]$ had any other element, it could not be a single 2-cycle or a 4-cycle (they have the wrong parity). So it would have to be a 3-cycle.

In view of the first paragraph, this gives more than two elements (see also exercise 15.9), which completes the proof. ■

The main result of this section now follows.

Theorem 15.11. *We have the following derived series:*

$$S_2 \rhd \{e\}$$

$$S_3 \rhd A_3 \rhd \{e\}$$

$$S_4 \rhd A_4 \rhd V \rhd \{e\}$$

$$\forall n > 4 \;:\; S_n \rhd A_n\,.$$

Proof. This follows immediately from putting all results in this section together. ■

Remark 15.12. We can naturally associate a sequence of Abelian groups to each derived series, namely

$$S_2/\{e\} \cong \mathbb{Z}_2$$

$$S_3/A_3 \cong \mathbb{Z}_2\,,\; A_3/\{e\} \cong \mathbb{Z}_3$$

$$S_4/A_4 \cong \mathbb{Z}_2\,,\; A_4/V \cong \mathbb{Z}_3\,,\; V/\{e\} \cong \mathbb{Z}_2 \times \mathbb{Z}_2$$

$$\forall n > 4 \;:\; S_n/A_n \cong \mathbb{Z}_2\,.$$

15.3. Solving Cubic and Quartic Equations

We all know how to solve for the roots of linear and quadratic polynomial equations in $\mathbb{C}$ and are probably aware of the fact that there are general formulas for the roots of cubic and quartic polynomials. However, these general formulas are sufficiently complicated (see [**55**, **56**]) that they have little or no practical use. What is interesting, though, is that at $n = 4$ it stops: there exist no general formulas for $n \geq 5$ that beside continuous functions involve finitely many radicals!

To understand this better, we first look at the formulas for the roots of cubic and quartic polynomials. We consider polynomials with coefficients in a field. In fact, we will assume the field to be $\mathbb{C}$. The field operations are, of course, addition, subtraction, multiplication, and division. To these operations we add taking an nth root (specifying which one).

Definition 15.13. *An equation is solvable via radicals*[1] *if its roots are functions of the coefficients that use only the field operations and (specified) nth roots.*

We now consider a general degree n polynomial $q(z) := \sum_{i=0}^{n} c_i z^i$. We start by simplifying the problem.

Lemma 15.14. *If we can find the roots of the general depressed polynomial $p(z) := z^n + \sum_{i=0}^{n-2} a_i z^i$, then we can find the roots of the general polynomial $q(z) := \sum_{i=0}^{n} c_i z^i$.*

Proof. Since the coefficients are in $\mathbb{C}$, solving $q(z) = 0$ is equivalent to solving $z^n + \sum_{i=0}^{n-1} \frac{c_i}{c_n} z^i = 0$. Next, eliminate z in this polynomial in favor of x, where[2]

$$z = x - \frac{c_{n-1}}{nc_n}\,. \tag{15.3}$$

It is a straightforward computation to check that the term of order $n-1$ cancels and the polynomial now becomes

$$\left(x - \tfrac{c_{n-1}}{nc_n}\right)^n - \tfrac{c_{n-1}}{c_n}\left(x - \tfrac{c_{n-1}}{nc_n}\right)^{n-1} + \cdots =$$

$$\left(x^n - \tfrac{c_{n-1}}{c_n} x^{n-1} + \tbinom{n}{2}\tfrac{c_{n-1}^2}{n^2 c_n^2} x^{n-2} + \cdots\right) + \tfrac{c_{n-1}}{c_n}\left(x^{n-1} - \tfrac{(n-1)c_{n-1}}{nc_n} x^{n-2} + \cdots\right) + \cdots =$$

$$x^n + \textstyle\sum_{i=0}^{n-2} a_i x^i\,,$$

upon suitably defining the a_i in terms of the c_i. So a root $x^* = F(a_1, \cdots, a_{n-2})$ of $p(x)$ corresponds to a solution

$$z^* = x^* - \frac{c_{n-1}}{nc_n} = F(a_1, \cdots, a_{n-2}) - \frac{c_{n-1}}{nc_n}\,,$$

where the a_i are defined in terms of the c_i. ■

Lemma 15.15 (Viète's Formula). *Two numbers $x_\pm$ in $\mathbb{C}$ satisfy*

$$x_- + x_+ = -b \quad \text{and} \quad x_- x_+ = -c\,,$$

[1] A radical is an nth root; *radix* being the Latin word for *root*.

[2] It is important to note that this might not work if the coefficient are in some other field, such as as a degree p polynomial with coefficients in $\mathbb{F}_p$ (see Chapter 5). The number p satisfies $p \cdot 1 = 0$ in this field. Such a number the characteristic of the field, or ring. It is zero if there is no natural number p satisfying that equation, as in the case of $\mathbb{Z}$ or $\mathbb{C}$.

if and only if

$$x_{\pm} = -\frac{b}{2} \pm \sqrt{\frac{b^2}{4} + c}\,.$$

Proof. Both x_- and x_+ can be seen to satisfy $x(-b-x) = -c$. This has precisely the two indicated solutions. Vice versa, it is easy that the two relations hold. ∎

Now we turn to the cubic equation and set $p_3(x) = x^3 + a_1 x + a_0$. First, substitute $x = x_- + x_+$. After a slight rearrangement of the terms, we get the following equation for the roots

$$p_3(x_- + x_+) = x_-^3 + x_+^3 + (a_1 + 3x_- x_+)(x_- + x_+) + a_0 = 0\,.$$

We can solve this if

$$x_-^3 + x_+^3 = -a_0 \qquad \text{and} \qquad x_- x_+ = -\frac{a_1}{3} \quad \text{and so}$$

$$x_-^3 + x_+^3 = -a_0 \qquad \text{and} \qquad x_-^3 x_+^3 = -\frac{a_1^3}{3^3}\,.$$

We apply Viète's forula (Lemma 15.15) to this last pair of equations.

$$x_{\pm}^3 = -\frac{a_0}{2} \pm \sqrt{\frac{a_0^2}{4} + \frac{a_1^3}{3^3}}\,. \tag{15.4}$$

So $x_{\pm}$ are third roots of this expression, and the roots x^* of p_3 are

$$x^* = x_- \omega_- + x_+ \omega_+\,, \tag{15.5}$$

where $\omega_{\pm}$ are third roots of unity and must chosen so that also $x_- x_+ = -\frac{a_1}{3}$ is satisfied. There are exactly three such third roots, so if we choose ω_-, then ω_+ is uniquely determined. Thus this gives three solutions of the cubic equation $p_3(x) = 0$.

Moving along to the quartic equation, after reducing the equation as before, we set $p_4(x) = x^4 + a_2 x^2 + a_1 x + a_0$. If $a_1 = 0$, then we apply the usual quadratic formula to get two solutions for x^2 and hence four solutions for x. So assume that $a_1 \neq 0$. To solve for the roots, we try to equate $p_4(x)$ to a product of quadratic equations, as follows.

$$p_4(x) = (x^2 - yx + b)(x^2 + yx + c) = x^4 + (-y^2 + b + c)x^2 + (b - c)yx + bc\,. \tag{15.6}$$

For this to hold, we need that

$$b+c=y^2+a_2 \quad \text{and} \quad b-c=\frac{a_1}{y} \quad \text{and} \quad bc=a_0\,.$$

Note that we may assume that $y \neq 0$ (because $a_1 \neq 0$). The first two of these three equations can be used to deduce that

$$b=\frac{1}{2}\left(y^2+a_2+\frac{a_1}{y}\right) \quad \text{and} \quad c=\frac{1}{2}\left(y^2+a_2-\frac{a_1}{y}\right) \tag{15.7}$$

Substitute these into $bc=a_0$ to get

$$\left(y^2+a_2+\frac{a_1}{y}\right)\left(y^2+a_2-\frac{a_1}{y}\right)-4a_0=y^4+2a_2y^2+(a_2^2-4a_0)-\frac{a_1^2}{y^2}=0\,. \tag{15.8}$$

This is a cubic equation in y^2. We can use the cubic formula to solve and get 6 possible values for y. Pick any non-zero value for y and compute b and c from (15.7) (and check that $bc=a_0$). Finally, substitute these values into (15.6) and solve the quadratic equations. Notice the somewhat mysterious fact that *any* non-zero value for y must give the same four solutions, as there are only four solutions by the fundamental theorem of algebra (Theorem 11.21).

We will see in exercises 15.10 and 15.11 that the practical implementation of these schemes is far from trivial. In practice one either 'guesses' one or more roots of such equations or uses a numerical scheme to approximate solutions to arbitrary precision, such as Newton's method.

15.4. Monodromy

One more piece of information is needed to finalize the proof of the non-solvability of the quintic. This goes by the name of monodromy. This is a rather technical and very important concept. Luckily we only need the basic material as it relates to taking nth roots in the complex plane.

For concreteness, we let $p: z \to z^n$ for n some integer greater than 1. A loop in $\mathbb{C}$ is a continuous path $\alpha:[0,1]\to\mathbb{C}$ whose endpoints are the same (denoted as the base-point). Choose a (non-zero) base-point in $\mathbb{C}$, say 1. Consider loops with 1 as its endpoints. Furthermore since under p, every point has exactly n inverses, except 0, we want our loops to avoid 0. So we consider loops in $\mathbb{C}\backslash\{0\}$ based at, say, 1. To every loop α avoiding zero, there is a well-defined integer k we can assign to it that counts the number of

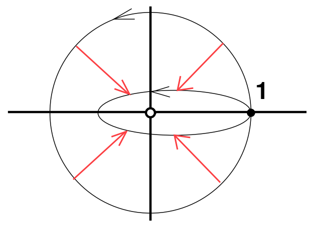


Figure 102. Two loops in $\mathbb{C}\backslash\{0\}$ based at 1 that wind around 0 exactly once. The red arrows indicate how the outer curve can be continuously deformed in $\mathbb{C}\backslash\{0\}$ to the inner curve.

times it circles around the origin in the counter-clockwise direction. This is called the winding number of the loop. It turns out that any two curves with the same k can be continuously deformed into one another. We illustrate this in Figure 102; details can be found in [**50**], Chapters 9 and 10. This continuous deformation defines an equivalence relation on loops in $\mathbb{C}\backslash\{0\}$ based at 1. This is called homotopy equivalence .

Definition 15.16. *The set of loops in* $\mathbb{C}\backslash\{0\}$ *based at 1 and up to homotopy equivalence is called the fundamental group of* $\mathbb{C}\backslash\{0\}$ *relative to the point 1. It is denoted by* $\pi_1(\mathbb{C}\backslash\{0\},1)$.

Since in this example all loops with the same winding number are (homotopy) equivalent, the fundamental group $\pi_1(\mathbb{C}\backslash\{0\},1)$ is just $\mathbb{Z}$.

It is clear how loops map under p, as p is a well-defined continuous function. The problem is when we try to define the image of a loop α around 0 under its multi-valued 'inverse', $\sqrt[n]{z}$. Informally, here is how we do it. We refer to Figure 103. The complete inverse image of the base-point under p, or $p^{-1}(1)$, is called the fiber of 1 under p. In this case, the fiber of 1 consists of the n points $a_\ell := e^{2\pi i\ell/n}$, $\ell \in \{0,\cdots,n-1\}$. Suppose the loop α has base-point 1 and runs around zero exactly once. Choose one of the points in the fiber of 1, say a_ℓ. Take some small neighborhood of 1 so that its inverse images under p around the a_i are disjoint. Then *locally* in each of these inverse images, the map is invertible. In particular, we can draw the inverse image $\tilde{\alpha}$ for some small $t \in [0,\varepsilon]$ such that $\tilde{\alpha}(0) = a_\ell$. By chopping up all of α into small pieces, this process can be continued until we have an inverse image $\tilde{\alpha}$ of α with $\tilde{\alpha}(0) = a_\ell$. This is called a lift of α and has the

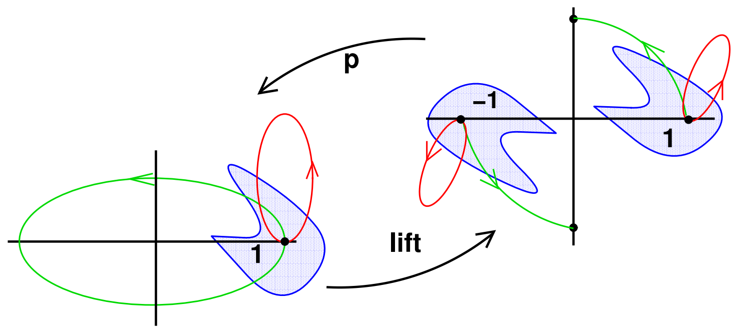


Figure 103. The curious behavior of lifts loops based at 1 through a 4th root. The fiber of 1 is $\{1, i, -1, -i\}$. The green loop on the left has winding number 1 and lifts to curves whose endpoints are not the same. We exhibit one lift based at 1 and one based at -1. The red loop on the left has winding number 0 and lifts to loops: we exhibit the ones based at 1 and -1. The blue neighborhood on the left lifts to disjoint blue neighborhoods on the right. Restricted to these local neighborhoods, p is invertible.

property that $p(\tilde{\alpha}) = \alpha$. What is important — and should be clear — is that in a lifted path $\tilde{\alpha}$, the endpoints do not necessarily coincide anymore.

What happened here is that the loop α induces a permutation on the fiber of its base-point. In the case of nth roots, it is easy to see[3] that if α has winding number k, then its lift $\tilde{\alpha}$ based at a_i is a curve from a_i to $a_{(i+k) \mod n}$. We can think of this as the action of the winding number on the points of the fiber $p^{-1}(1)$. In this case, we see that this action is a cyclic permutation of those points and they form a group. Thus the lifts of a curve give rise to a cyclic permutation of the points of the fiber.

Definition 15.17. *Let $p(z) = z^n$. The action of the winding number on the roots gives rise to a homomorphism $\pi_1(\mathbb{C}\backslash\{0\}, 1) \to S_n$. The image of that homomorphism in this case is isomorphic to $\mathbb{Z}_n$ and is called the monodromy group*[4].

Before we state the central result of this section, we need one more definition. The commutator of two loops α and β with the same base-point

[3] More general formal proofs tend to get very technical.

[4] We note here that there is no immediate connection with the similarly named 'monodromy operator' for differential equations with periodic coefficients, see [5], Section 26

is given by concatenating first $\alpha(t)$, then $\beta(t)$, then $\alpha(1-t)$, and finally $\beta(1-t)$ and finally rescaling the parameter t so that its domain is $[0,1]$. The commutator is denoted by $\underline{[\alpha,\beta]}$.

Lemma 15.18. *Let $\alpha(t)$ and $\beta(t)$ be continuous loops in $\mathbb{C}$ with the same base-point. Then $\gamma := \sqrt[n]{[\alpha,\beta]}$ is a loop. (In other words: their endpoints coincide.)*

Proof. Suppose $\alpha(t)$ and $\beta(t)$ loop around the origin a and b times respectively, the their commutator loops around $a+b-a-b=0$ times, and so its lift is a loop. The underlying reason, of course, is that $\mathbb{Z}_n$ is Abelian. ∎

The astounding simplicity of this lemma notwithstanding, it represents the essence of the proof of the non-solvability of the quintic by radicals.

15.5. The General Quintic is Not Soluble via Radicals

We now address the central question of this chapter: when can we solve for the roots of polynomials by radicals? Here is the definition.

Definition 15.19. *We say that a polynomial $p(x)$ over a field F admits <u>solution</u> <u>by</u> <u>radicals</u> if its roots can be obtained from the the coefficients of the polynomial p using the usual field operations (addition, subtraction, multiplication, and division) and application of <u>radicals</u> (nth roots for $n = 2,3,4,\cdots$).*

Consider the reverse situation where we know the roots s_i and want to compute the coefficients of the above polynomial. This is easy, because

$$p(z) = (z-s_1)\cdots(z-s_n) = z^n - (s_1+\cdots+s_n)z^{n-1} + \cdots + (-1)^n s_1 s_2 \cdots s_n\,.$$

From this, the coefficients follow as continuous functions of the roots. From now on, we will denote this function by $F(s_1,\cdots,s_n)$. The ith coefficient c_i is given by the ith component of $F:\mathbb{C}^n \to \mathbb{C}^n$.

There is one overwhelmingly important observation that will dominate this conversation. It is this: *The polynomial p is unchanged by permutations of its roots!* Note that this is very much *not* true for its coefficients. The way we are going to take advantage of this is as follows. We are going to look at continuous paths for the roots of polynomials with the property that at $t=0$ we start at roots $(s_1,s_2,\cdots,s_n)$ and at $t=1$, we end at

$(\pi(s_1), \pi(s_2), \cdots, \pi(s_n))$, where π is some permutation of the n roots. The point of this is that it gives us a *continuous* way to effect a permutation. This enables us to use topological arguments to analyze what expressions can be used for the zeros of polynomials.

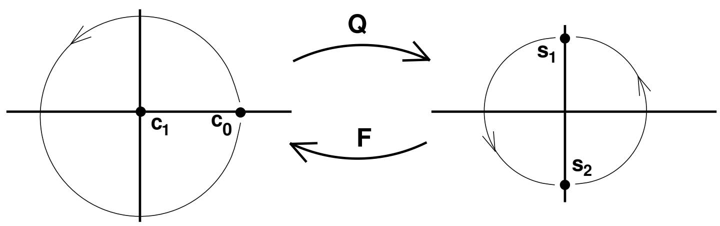


Figure 104. Left, the coefficients $c_1(t) = 0$ and $c_0(t) = e^{2\pi i t}$ of a quadratic polynomial. Right, its solutions $s_1(t)$ and $s_2(t)$. Between $t = 0$ and $t = 1$, the solutions are permuted.

To make this more concrete, we look at a quadratic case. Consider the polynomial $z^2 + 2e^{2\pi i t}$ where $t \in [0,1]$, see Figure 104. On the left, we have the coefficient space. The coefficients describe a loop in $\mathbb{C}^2$.

$$c(t) = (c_0(t), c_1(t)) = (2e^{2\pi i t}, 0)\,.$$

The corresponding lifts $\gamma(t)$ in the solution space on the right give

$$\gamma(t) = (s_1(t), s_2(t)) = (i\sqrt{2}\,e^{\pi i t}, -i\sqrt{2}\,e^{\pi i t})\,.$$

Note that this is *not* a loop, for $\gamma(0) = (i\sqrt{2}, -i\sqrt{2})$ and $\gamma(1) = (-i\sqrt{2}, i\sqrt{2})$. This gives us one way of verifying that Q must be multi-valued, because under F both these points map to $(0,2)$. See also exercise 15.13. Below we will give a different proof.

The seemingly rather trivial observation that the roots are exchangeable immediately leads to a remarkable result.

Proposition 15.20. *For $n \geq 2$, let $p(z) = z^n + \sum_{i=0}^{n-1} c_i z^i$ have roots $\{s_i\}$. There is no general solution in the form of a continuous function. That is: the solution must be in the form of an n-valued function $q : \mathbb{C}^n \to \mathbb{C}$.*

Proof. Suppose there exists a continuous function $Q : \mathbb{C}^n \to \mathbb{C}^n$ with components q_i that takes the coefficients $\{c_i\}$ to the roots $\{s_i\}$.

$$s_i = q_i(c_1, c_2, \cdots, c_n)\,.$$

This solution is general and therefore must remain the same if we arbitrarily permute the s_i. It follows that all q_i are the same: $q_i(c_1,\cdots,c_n) = q(c_1,\cdots,c_n)$. But we still have n distinct solutions, so q must be n-valued and therefore is not a function. ■

Remark 15.21. This result says that there is no *general* single-valued formula that gives the roots. However, in special cases, there may very well be such a formula. A case in point is $p(z) = (z-1)^5$. The roots are $x = 1$ (with multiplicity 5).

We will from now on concentrate on describing q. It is a multi-valued 'function' $\mathbb{C}^n \to \mathbb{C}$ and so, according to Definition 15.19, it must contain at least one radical. But roots act on complex numbers, not vectors. So at some point in the evaluation of q (see Figure 105), we must take a multi-valued nth root $\sqrt{..} : \mathbb{C} \to \mathbb{C}$. Then further continuous, single-valued operations may follow to eventually yield the zeros of p. The final form of q may be a sum, product, et cetera of such expressions. However, the interesting 'action' is where we take the nth roots, everything consists of continuous and single-valued functions. Accordingly, we will visualize Q only in terms of the nth root operation.

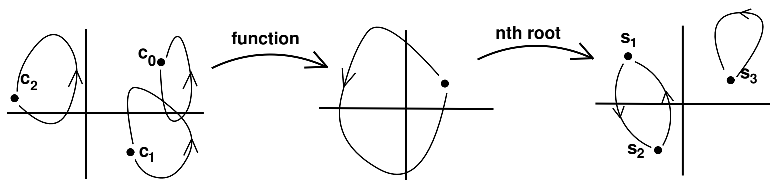


Figure 105. At some point in the evaluation of q, we must take a multi-valued nth radical $\sqrt[n]{..} : \mathbb{C} \to \mathbb{C}$.

Before stating our main results, we recall from the proof of Lemma 15.14 that it is sufficient to study the zeros of monic polynomials $p(z) = z^n + \sum_{i=0}^{n-1} c_i z^i$. The reader should compare these results with Theorem 15.11 and the remark following it.

Theorem 15.22. *Let $p(z) = z^3 + \sum_{i=0}^{2} c_i z^i$ have roots $\{s_i\}_{i=1}^3$. The general solution by radicals Q must at least contain nested radicals (i.e. $\sqrt{..\sqrt{..}}$).*

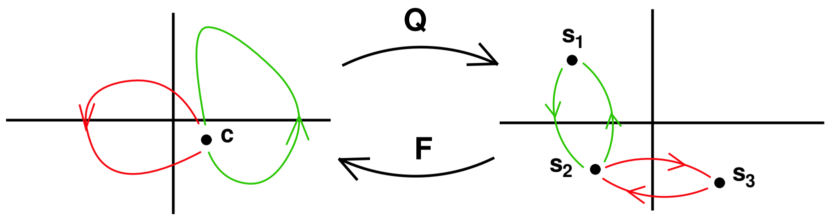


Figure 106. Right, paths $s_1(t)$ and $s_2(t)$ that swap the first two roots in green. This results in a green loop $g(t)$ on the left, upon taking the 3rd power; Similarly, in red, paths that swap the second and third loop and their 3rd power.

Proof. Suppose that the general solution is an expression that contains radicals but does not contain *nested* radicals. The radical in the description of Q is depicted in Figure 106. If we swap two of the three roots as depicted in green, then the image under $z \to z^3$ is a green loop g. Take a different pair and swap those in red. This will give a red loop r.

Now consider the commutator $r^{-1}g^{-1}rg$. If we just compute that commutator on the right directly, we obtain

$$(23)(12)(23)(12) = (123).$$

However, according to Lemma 15.18, the action of a commutator on the permutation is trivial, a direct contradiction. ∎

Theorem 15.23. *Let $p(z) = z^4 + \sum_{i=0}^{3} c_i z^i$ have roots $\{s_i\}_{i=1}^4$. The general solution by radicals Q must at least contain triply nested radicals (i.e. $\sqrt{..\sqrt{..\sqrt{..}}}$).*

Proof. Suppose that the general solution is an expression that contains nested radicals but nothing 'worse' than that. Consider 4 loops numbered from 1 to 4, as in Figure 107. Apply Lemma 15.18. Under the first root, the commutator $[1,2]$ goes to a loop, say α. Similarly the $[3,4]$ goes to β. The commutator of these commutators, $[[1,2],[3,4]]$ is mapped under the root to a commutator. Upon taking the *second* root, the same lemma ensures that the final image is a loop again. This is in direct contradiction with Lemma 15.10 which says that the commutator subgroup of the commutator subgroup is, in fact, a non-trivial group. ∎

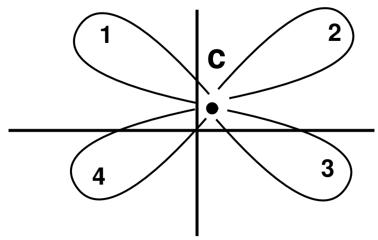


Figure 107. Four loops numbered from 1 to 4 make a commutator of commutators, namely $[[1,2],[3,4]]$.

Theorem 15.24. *Let $n \geq 5$ and $p(z) = z^n + \sum_{i=0}^{n-1} c_i z^i$ have roots $\{s_i\}_{i=1}^{n}$. There is no general solution by containing only finite nests of radicals.*

Proof. Now we consider paths that are commutators of commutators of $\cdots$ of commutators, as long as the number is finite, say M. These are constructed in a similar way as illustrated in Figure 107, but with $2M$ loops instead of just 2. After taking the M nested roots, though, this commutator must be trivial by Lemma 15.18 (again). However, this contradicts Proposition 15.9. ■

Remark 15.25. This theorem says that there is no *general* single-valued formula that gives the roots. However, in special cases, there may very well be such a formula. A case in point is $p(z) = (z-1)^5$. The roots are $x = 1$ (with multiplicity 5).

Remark 15.26. There are important differences with the more traditional treatment of this topic, which uses Galois theory to address the question of solvability by radicals [**54**, **68**]. In the Galois theory approach, solubility of individual polynomials is considered, not just 'general' solubility. In that sense, the Galois theory approach is stronger than the theory described here. However, in Galois theory, we are only allowed to apply the field operations to the coefficients— addition and multiplication and their inverses — whereas here, any continuous function — such $\ln c_i$ or $|c_i|$ — is fair game. So in that sense the theory described here is stronger.

15.6. Exercises

Exercise 15.1. Complete the proof of Lemma 15.4, that is:
a) Check that $(xax^{-1})^{-1} = xa^{-1}x^{-1}$, and
b) Check the formula in the proof for a product of commutators.

The commutator subgroup $[G,G]$ of a group G is generated by and therefore contains, of course, its commutators. What is less clear is that it may contain more. Examples of this are surprisingly hard to exhibit. In the next few exercises, we show that the commutator subgroup of $SL(2,\mathbb{R})$, the 2×2 real matrices with determinant 1 with matrix multiplication as its operation, has this property. To do so, we will show that $-I$ is not a commutator itself, but *is* a product of commutators.

Exercise 15.2. Consider the group $G := SL(2,\mathbb{R})$ and assume that $-I$ is a commutator.
a) Let $A_i \in G$ and show that $A_1A_2A_1^{-1}A_2^{-1} = -I$ implies that

$$A_1A_2A_1^{-1} = -A_2 \quad \text{and} \quad A_2A_1A_2^{-1} = -A_1\,.$$

b) Conclude that $\operatorname{Tr}A_1 = \operatorname{Tr}A_2 = 0$. (*Hint: the trace is invariant under conjugation.*)
c) Show that

$$A_i = \begin{pmatrix} a_i & b_i \\ -\frac{1+a_i^2}{b_i} & -a_i \end{pmatrix}.$$

d) Compute that

$$A_1A_2 = \begin{pmatrix} -a_1a_2 - \frac{b_1}{b_2}(1+a_2^2) & a_1b_2 - a_2b_1 \\ -\frac{a_2}{b_1}(1+a_1^2) - \frac{a_1}{b_2}(1+a_1^2) & -a_1a_2 - \frac{b_2}{b_1}(1+a_1^2) \end{pmatrix}.$$

e) Now set $A_1A_2 = -A_2A_1$ and conclude from equating the 12 and 22 entries that

$$\begin{aligned} 2a_1a_2 &= \frac{b_1}{b_2}(1+a_2^2) + \frac{b_2}{b_1}(1+a_1^2) \\ 0 &= \frac{a_2}{b_1}(1+a_1^2) + \frac{a_1}{b_2}(1+a_1^2)\,. \end{aligned}$$

f) Multiply the first equation by b_1b_2 to show that $a_1a_2b_1b_2 > 0$ and rework the second equation to get that $\frac{a_1b_1}{a_2b_2} < 0$.
g) Show that the assumption that $-I$ is a commutator, is false.

Exercise 15.3. a) For $\mu^2 \neq 1$, compute

$$\begin{pmatrix} \mu & 0 \\ 0 & \mu^{-1} \end{pmatrix} \begin{pmatrix} 1 & \frac{\lambda}{\mu^2-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu^{-1} & 0 \\ 0 & \mu \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{\lambda}{\mu^2-1}0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix},$$

and use that to show that $\begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$ is a commutator.

b) Similarly, show that $\begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix}$ is a commutator.

c) Compute that

$$\begin{pmatrix} 1 & a^{-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -a & 1 \end{pmatrix} \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & a^{-1} \\ -a & 1-ab \end{pmatrix}.$$

d) Choose a and b so that you can conclude that $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ is product of commutators.

e) Conclude that $-I$ is a product of commutators. (*Hint: the matrix in (d) represents a rotation by* $\pi/2$*. Apply that rotation twice.*)

Remark: In fact, with a little more patience, one can show that the matrices $\begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ \kappa & 1 \end{pmatrix}$ with λ and κ in $\mathbb{R}$, generate the entire group $SL(2,\mathbb{R})$. Thus this group is a perfect group (Definition 15.6).

Remark: For finite groups, it is known that the smallest group whose commutator subgroup contains more elements than just its commutators has 96 elements.

Exercise 15.4. a) Show that the commutator subgroup of a finite group G is not necessarily the largest normal subgroup. (*Hint: for primes p and q, consider the cyclic group of order pq.*)

b) Use the above remarks to show $SL_2(\mathbb{R})$ is perfect but not simple. (*Hint: consider the set* $\{I,-I\}$.)

Exercise 15.5. Let G be a group and H a normal subgroup.

a) Use (15.1) to show that $(Hx)^{-1} = Hx^{-1}$.

b) Use (a) and (15.1) to show that $HxHy(Hx)^{-1}(Hy)^{-1} = H$ if and only if $xyx^{-1}y^{-1} \in H$.

c) Conclude that G/H is Abelian iff H contains all commutators.

Exercise 15.6. In this exercise, we prove that if you write the identity of the permutation group S_n in terms of transpositions, the number of transpositions is always even. To do so, suppose that $I = \tau_1 \cdots \tau_m$ where the τ_i are transpositions. Fix any $x \in \{1, \cdots, n\}$, and let τ_{k+1} be the rightmost transposition that contains the symbol x.
a) Suppose τ_k does not contain the same two symbols as τ_{k+1}. Show that I can be written as a product of $m-2$ transpositions.
b) Suppose τ_k contains at most one symbol that also occurs in τ_{k+1}. Show that I can be written as the following product of transpositions

$$I = \tau_1 \cdots \tau_{k-1} \sigma_k \sigma_{k+1} \tau_{k+2} \cdots \tau_m ,$$

where now σ_k is the rightmost transposition containing x. (*Hint: use that* $(xa)(xb) = (xa)(ab)$, *et cetera.*)
c) Show that if the symbol x never gets cancelled as in (a), then I ends up being written as a product of transposition where only the leftmost transposition contains x.
d) Show that (c) is impossible.
e) Conclude that every x must get cancelled, and thus I must have had an even number of transpositions.

Exercise 15.7. a) Use exercise 15.6 to show that there is a well-defined surjective homomorphism $P : S_n \to \{\pm 1\}$. The image under P of a permutation σ is called the parity of σ.
b) Use Corollary 7.31 to show that A_n is a normal subgroup of S_n.

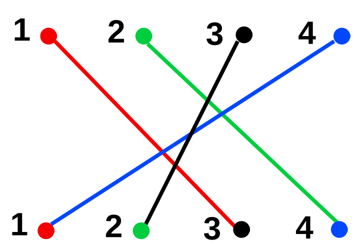


Figure 108. A permutation on n symbols can be decomposed into transpositions graphically, and thus its parity can be determined.

Exercise 15.8. A convenient way to decompose a permutation into transpositions, and indeed to find the parity defined in exercise 15.7, is to count the number of crossings of the diagram of the permutation σ as depicted in Figure 108. In this drawing only two lines can cross at the same point. Such a figure in also called a *matching diagram* .

a) Use Figure 108 to show that $\sigma(1) = 4$, $\sigma(2) = 3$, and so on.

b) Show that $\sigma = \begin{pmatrix} 1234 \\ 4312 \end{pmatrix}$.

c) Use Figure 108 to show that σ can be decomposed into transpositions as follows

$$\sigma = (13)(14)(34)(24)(23)\,.$$

d) Verify in this example that deforming the lines may lead to a different decomposition into transpositions, but does not affect the parity.

Exercise 15.9. a) Show that the group V in Lemma 15.10 is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$.

b) Show that the parity of a single 2-cycle or a 4-cycle is odd (and so they are not in A_n).

c) Show that there are four distinct 3-cycles in A_4, each of which has two distinct non-identity iterates.

The group V in exercise 15.9 and Lemma 15.10 is called the Klein four-group or simply the Klein group . It is usually denoted by V or K_4. It is the smallest non-cyclic group. It is *not* isomorphic to $\mathbb{Z}_4$ (which is cyclic).

In the following two exercises, we follow the method of Section 15.3 to solve for the roots of of a cubic and a quartic polynomial. To get a feel for the considerable subtleties in this process, the reader is encouraged to do all the steps by hand, except where otherwise indicated. Allow ample time for these exercises!

Exercise 15.10. We solve for the roots of $p_3(z) = z^3 - 34z^2 + 313z - 400$ (see Figure 109).
a) Show that it is sufficient to solve $q(x) = x^3 - \frac{217}{3}x + \frac{6370}{3^3} = 0$. (*Hint: set* $z = x + \frac{34}{3}$.)
b) Use (15.4) and $3185^2 - 217^3 = -74088$, to show that

$$x_\pm^3 = 3^{-3}\left(-3185 \pm \sqrt{3185^2 - 217^3}\right) = 3^{-3}\left(-3185 \pm \sqrt{-74088}\right).$$

c) Take x_- and x_+ to be (standard) third roots in the right half plane, and show that

$$x_- x_+ = 3^{-2}\left(3185^2 + 74088\right)^{1/3} = \frac{217}{9} = -\frac{a_1}{3}.$$

(*Hint:* $3185^2 + 74088 = 217^3$.)
d) So let $\omega := e^{2\pi i/3}$ and use (15.5) to show that the roots of q are given by

$$x_- + x_+\,, \quad x_-\omega + x_+\omega^{-1}\,, \quad x_-\omega^{-1} + x_+\omega\,.$$

e) Deduce that the roots of q are equal to $\{\frac{14}{3}, -\frac{7}{3} \pm 2\sqrt{14}\}$. (*Hint: approximate numerically or divide q by* $(x - 14/3)$.) f) Conclude that the roots of p_3 are equal to

$$\{16, 9 \pm 2\sqrt{14}\} \approx \{1.52, 16, 16.48\}.$$

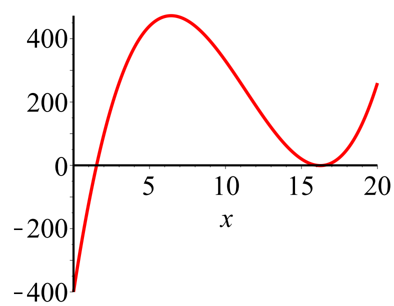


Figure 109. The graph of $p_3(z)$ of exercise 15.10 showing its three roots at approximately $\{1.52, 16, 16.48\}$.

Exercise 15.11. We solve for the roots of $p_4(z) = z^4 - 17z^2 + 20z - 6$ (see Figure 110).
a) Show that the associated 'cubic' polynomial (15.8) is

$$q(y) = y^6 - 34y^2 + 313x - 400\,.$$

b) Conclude from exercise 15.10 that this polynomial has 4 as a root.
c) Use (15.6) and (15.7) to show that p_4 satisfies

$$p_4(x) = \left(x^2 - 2x - \frac{3}{2}\right)\left(x^2 - 2x - \frac{3}{2}\right)$$

d) Conclude that the roots of p_4 are equal to

$$\{-2 \pm \sqrt{2}, 2 \pm \sqrt{7}\} \approx \{-3.41, -0.59, -0.65, 4.65\}\,.$$

e) Take another root of the 'cubic' and check that you get the same roots for p_4.

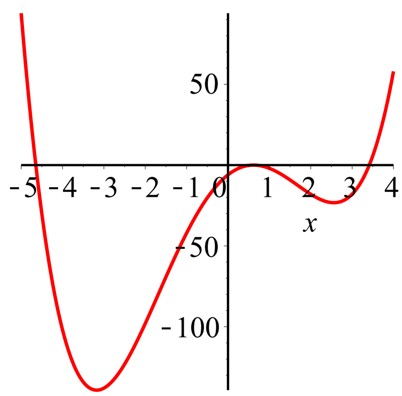


Figure 110. The graph of $p_4(z)$ of exercise 15.11 showing its roots are approximately $\{-3.41, -0.59, -0.65, 4.65\}$.

Exercise 15.12. Let $p(z) = z^4$.
a) What is the fiber of 16 under p?
b) Let $\alpha(t) = 16e^{2\pi i t}$. Define 4 different lifts of α.
c) For each of the four lifts $\tilde{\alpha}$ in (b), give the start and end point of $\tilde{\alpha}$.

Exercise 15.13. Let $p(z) = (z-s_1)(z-s_2)(z-s_3)$.
a) Show that F in Section 15.5 is given by

$$F(s_1,s_2,s_3) = (-s_1s_2s_3, s_1s_2+s_1s_3+s_2s_3, s_1+s_2+s_3)\,.$$

b) Compute $p(z)$ if the solutions s_i lie on a curve γ in $\mathbb{C}^n$ given by

$$\gamma(t) = (\sqrt[3]{2}e^{\frac{1}{3}\pi i(2t+1)}, \sqrt[3]{2}e^{\frac{1}{3}\pi i(2t-1)}, \sqrt[3]{2}e^{\frac{1}{3}\pi i(2t+3)})\,.$$

c) Draw a picture of the coefficient space and the solution space similar to Figure 104.
d) Show that $\gamma(0) \neq \gamma(1)$, but $F\circ\gamma(0) = F\circ\gamma(1)$.
e) Conclude from (d) that Q, the 'reverse' of F, is not single-valued.

Exercise 15.14. Newton's method if another problem is needed.

There are various techniques that address roots of equations that are insoluble via radicals. In exercise 15.15, we briefly discuss an example of one of these.

Definition 15.27. *The Bring radical of a real number a is the unique real root of $f(x) := x^5 + x - a$. It is denoted by $Br(a)$. See Figure 111.*

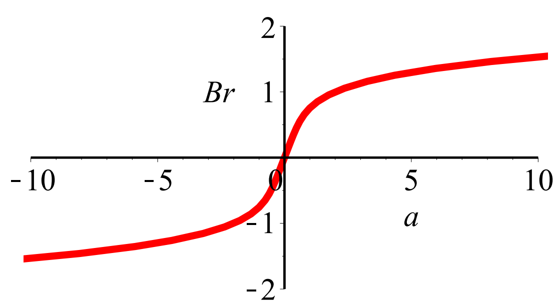


Figure 111. The graph of the Bring radical as function of (real) a.

Exercise 15.15. a) Show that $\mathrm{Br}(a)$ is a uniquely determined real if a is real.
b) Show that $\partial_a \mathrm{Br}(a) > 0$.
c) Show that $\mathrm{Br} : \mathbb{R} \to \mathbb{R}$ has a well-defined, differentiable inverse.
d) Show that a root of $x^5 + px + q$ is given by $p^{1/4}\mathrm{Br}\left(p^{-5/4}q\right)$. (*Note: once we have one root, formulas for the other roots can be derived via the standard formulas for a quartic polynomial.*)

We remark here that to actually bring an arbitrary quintic into the form $x^5 + px + q$ of exercise 15.15 so that it can be solved using the Bring radical is very complicated and requires a lot of computation (see [**23**]).

Bibliography

[1] M. Aigner and G. M. Ziegler, *Proofs from the book, 6th edition*, Springer Verlag, Providence, RI, 2018.

[2] V. B. Alekseev, *Introduction to analytic number theory*, Kluwer Academic Publishing, New York, 2004.

[3] T. M. Apostol, *Mathematical analysis, 2nd edn*, Addison-Wesley, Philippines, 1974.

[4] ______, *Introduction to analytic number theory*, Springer Verlag, New York, 1989.

[5] V. I. Arnold, *Geometrical methods in the theory of ordinary differential equations*, Springer Verlag, New York, 1983.

[6] V. I. Arnold and A. Avez, *Ergodic problems in classical mechanics*, Benjamin, New York, 1968.

[7] S. Axler, *Linear algebra done right, 3rd edition*, Springer International Publishing, Cham, Switzerland, 2015.

[8] ______, *Measure, integration & real analysis*, Springer Nature Switzerland, Cham, Switzerland, 2020.

[9] L. Barreira and G. Iommi, *Frequency of digits in the lüroth expansion*, Journal of Number Theory (2009), 1479–1490.

[10] B. Baumslag, *A simple way of proving the jordan-holder-schreier theorem*, The American Mathematical Monthly **113** (2006), 933–935.

[11] A. F. Beardon, *Continued fractions, möbius transformations and clifford algebras*, Bull. London Math. Soc. (2003), 302–308.

[12] C. M. Bender, D. C. Brody, and M. P. Müller, *Hamiltonian for the zeros of the riemann zeta function*, Phys. Rev. Lett. **118** (2017Mar), 130–201.

[13] A. Berger and T. P. Hill, *Benford's law strikes back: No simple explanation in sight for mathematical gem*, Mathematical Intelligencer (2011), 85–91.

[14] P. Berrizbeitia and B. Iskra, *Gaussian mersenne and eisenstein mersenne primes*, Mathematics of Computation **79** (2010), 1779–1791.

[15] M. V. Berry and J. P. Keating, *The riemann zeros and eigenvalue asymptotics*, SIAM Review **41** (1999), 236–266.

[16] I. Boreico, *My favorite problem: Linear independence of radicals*, The Harvard College Mathematics Review (2007), 83–87.

[17] Y. Bugeaud, *Distribution modulo one and diophantine approximation*, Cambridge University Press, Cambridge, UK, 2012.

[18] D. M. Burton, *Elementary number theory, 7th edition*, McGraw-Hill, New York, NY, 2011.

[19] G. Cantor, *Über eine elementare frage der mannigfaltigkeitslehere* **1** (1890), 75–78.

[20] J. S. Caughman, 2018. Personal communication.

[21] D. A. Clark, *A quadratic field which is euclidean but not norm-euclidean*, Manuscripta Mathematica **83** (1994), 327–330.

[22] I. P. Cornfeld, S. V. Fomin, and Ya. S. Sinai, *Ergodic theory*, Springer-Verlag, New York, NY, 1982.

[23] R. J. Drociuk, *On the complete solution to the most general fifth degree polynomial* (2000).

[24] J. C. Lagarias (editor), *The ultimate challenge: The 3x+1 problem*, AMS, Providence, RI, 2010.

[25] J. Esmonde and M. Ram Murty, *Problems in algebraic number theory*, Springer, New York, NY, 1999.

[26] B. Fornberg and C. Piret, *Complex variables and analytic functions, an illustrated introduction*, SIAM, Philadelphia, 2020.

[27] M. Gardner, *The cult of the golden ratio*, Skeptical Inquirer **18** (1994), 243–247.

[28] W. J. Gilbert, *Modern algebra with applications*, John Wiley & Sons, New York, 1976.

[29] T. Gowers, *The princeton companion to mathematics*, Princeton University Press, Princeton, NJ, 2008.

[30] A. Granville, *Number theory revealed: a masterclass*, AMS, Providence, RI, 2019.

[31] Dumas H. S., *The kam story: A friendly introduction to the content, history, and significance of classical kolmogorov-arnold-moser theory*, World Scientific, Singapore, 2014.

[32] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers, sixth edition*, Oxford University Press, London, UK, 2008.

[33] D. Harel, R. Unger, and J. L. Sussman, *Beauty is in the genes of the beholder*, Trends in Biochemical Sciences **11** (1986), 155–156.

[34] Th. W. Hungerford, *Algebra*, Springer Verlag, New York, 1974.

[35] H. S. Zuckerman I. Niven and H. L. Montgomery, *An introduction to the theory of numbers*, Wiley & Sons, New York, NY, 1991.

[36] I. Kaplansky, *Set theory and metric spaces, 2nd edition*, AMS Chelsea Publishing, Providence, RI, 2001.

[37] D. Katahdin, 2018. Personal communication.

[38] A. Katok and B. Hasselblatt, Cambridge, UK.

[39] A. Ya. Khintchine, *Continued fractions*, P. Noordhoff, Ltd, Groningen, 1963.

[40] A. N. Kolmogorov and S. V. Fomin, *Introductory real analysis*, Dover, New York, NY, 1970.

[41] S. Lang, *Undergraduate analysis, 2nd edition*, Springer Verlag, New York, NY, 1997.

[42] J. E. Littlewood, *Lectures on the theory of functions*, Oxford University Press, Oxford, UK, 1944.

[43] M. Livio, *The golden ratio: The story of phi, the world's most astonishing number*, Broadway Books, New york, NY, 2010.

[44] A. Lopez-Ortiz, *Frequently asked questions in mathematics*. Available online at: `https://cs.uwaterloo.ca/~alopez-o/math-faq/mathtext/math-faq.html`.

[45] J. E. Marsden and M. J. Hoffman, *Basic complex analysis, 3rd edn*, W. H. Freeman, New York, NY, 1999.

[46] Wolfram MathWorld, *Mertens constant*. Available online at: `https://mathworld.wolfram.com/MertensConstant.html`.

[47] F. Mertens, *Ein beitrag zur analytischen zahlentheorie*, J. reine angew. Math. **78** (1874), 46–62.

[48] J. W. Milnor, *Dynamics: Introductory lectures*, University of Stony Brook, 2001.

[49] C. M. Moore, *Ergodic theorem, ergodic theory, and statistical mechanics*, PNAS **112** (2015), 1907–1911.

[50] J. R. Munkres, *Topology, 2nd edition*, Prentice-Hall, Hoboken, NJ, 2000.

[51] D. J. Newman, *Simple analytic proof of the prime number theorem*, The American Mathematical Monthly **87** (1980), 693–696.

[52] A. M. Odlyzko, *On the distribution of spacings between zeros of the zeta function*, Mathematics of Computation **48** (1987), 1003–1026.

[53] Fitzpatrick P. M., *Advanced calculus. a course in mathematical analysis*, PWS Publishing Company, Boston, 1996.

[54] C. C. Pinter, *A book of abstract algebra, 2nd edition*, Dover, New York, 1990.

[55] PlanetMath, *Cubicformula*. Available online at: `http://planetmath.org/CubicFormula`, February 2023.

[56] ______, *Quarticformula*. Available online at: `http://planetmath.org/QuarticFormula`, February 2023.

[57] C. Pomerance, J. L. Selfridge, and S. S. Wagstaff, *The pseudoprimes to* $25 \cdot 10^9$, Mathematics of Computation **35** (1980), 1003–1026.

[58] C. C. Pugh, *Real mathematical analysis, 2nd edn*, Springer, Cham, Switzerland, 2015.

[59] P. Ramond, *The abel-ruffini theorem: Complex but not complicated*, The American Mathematical Monthly **129** (2022), 231–245.

[60] B. Riemann, *Ueber die anzahl der primzahlen unter einer gegebenen grösse*, Monatsberichte der Berliner Akademie (1859).

[61] S. Roman, *An introduction to discrete mathematics*, Harcourt Brace Jovanovich, Orlando, FL, 1989.

[62] W. Rudin, *Real and complex analysis, 3rd edn*, McGraw-Hill International, New York, NY, 1987.

[63] J. H. Shapiro, 2020. Informal Lecture Notes.

[64] C. L. Siegel, *Algebraische abhängigkeit von wurzeln. (german)*, Acta Arith. **21** (1972), 59–64.

[65] I. Soprounov, *A short proof of the prime number theorem for arithmetic progressions* (2010). Available online at: `https://academic.csuohio.edu/soprunov_i/pdf/primes.pdf`.

[66] H. M. Stark, *On the gap in the theorem of heegner*, Journal of Number Theory **1 (1)** (1969), 16–27.

[67] S. Sternberg, *Dynamical systems, revised in 2013*, Dover Publications, United States, 2013.

[68] I. Stewart and D. Tall, *Algebraic number theory and fermat's last theorem, third edition*, A. K. Peters, Natick, MA, 2002.

[69] J. Stillwell, *Mathematics and its history, third edition*, Springer, New York, NY, 2010.

[70] S. Sutherland, *V'ir Tbg n Frperg*. Available online at: `https://www.math.sunsysb.edu/~scott/papers/MSTP/crypto.pdf`.

[71] J. J. P. Veerman, *Symbolic dynamics of order-preserving sets*, Physica D **29** (1986), 191–201.

[72] ______, *Symbolic dynamics and rotation numbers*, Physica A **134** (1987), 543–576.

[73] ______, *The dynamics of well-ordered orbits*, Autonomous University of Barcelona, Barcelona, Spain, 1995.

[74] R. A. Wilson, *An example of a pid which is not a euclidean domain*. Robert A. Wilson's website, accessed in December 2021.

[75] D. Zagier, *Newman's short proof of the prime number theorem*, The American Mathematical Monthly **104** (1997), 705–708.

Index