An Introduction to Number Theory

J. J. P. Veerman
# List of Figures

1. Eratosthenes’ sieve up to \( n = 30 \). All multiples of \( a \) less than \( \sqrt{31} \) are cancelled. The remainder are the primes less than \( n = 31 \). 4
2. A directed path \( \gamma \) passing through all points of \( \mathbb{Z}^2 \). 12
3. 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. 29
4. Proof that \( \sum_{n=1}^\infty f(n) \) is greater than \( \int_1^\infty x^{-s} \, dx \) if \( f \) is positive and decreasing. 34
5. Proof that \( \sum_{n=1}^\infty f(n) \) (shaded in blue and green) minus \( f(1) \) (shaded in blue) is less than \( \int_1^\infty x^{-s} \, dx \) if \( f \) is positive and decreasing to 0. 37
6. The origin is marked by “×”. The red dots are visible from ×; between any blue dot and × there is a red dot. The picture shows exactly one quarter of \( \{-4, \ldots, 4\}^2 \setminus (0,0) \subset \mathbb{Z}^2 \). 38
7. The general solution of the inhomogeneous equation \( (\vec{r}, \vec{x}) = c \) in \( \mathbb{R}^2 \). 45
8. 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. 65
List of Figures

9 Three branches of the Gauss map. 90
10 The line $y = \omega x$ and (in red) successive iterates of the rotation $R_\omega$. Closest returns in this figure are $q$ in $\{2, 3, 5, 8\}$. 99
11 The geometry of successive closest returns. 99
12 Drawing $y = \omega_1 x$ and successive approximations $(a_{n+1}$ is taken to be 3). The green arrows correspond to $e_{n-1}, e_n,$ and $e_{n+1}$. 100
13 Black: thread from origin with golden mean slope; red: pulling the thread down from the origin; green: pulling the thread up from the origin. 103
14 Plots of the points $(n, n)$ in polar coordinates, for $n$ ranging from 1 to 50, 180, 330, and 3000, respectively. 106
15 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. 107
16 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 $\text{Re}(z)$ and $k_2$ the closest integer to $\text{Im}(z)$. In this case that is $2 + 3i$. 120
17 A depiction of $\mathbb{Z}[\sqrt{-6}]$ in the complex plane; real parts are integers and imaginary parts are multiples of $\sqrt{6}$. 121
18 The part to the right of the intersection with $\ell : y = x + 1$ (dashed) of bad path (in red) is reflected. The reflected part in indicated in green. The path becomes a monotone path from $(0, 0)$ to $(n - 1, n + 1)$. 123
19 Possible values of $\rho \gamma^{-1}$ in the proof of Proposition 7.25. 127
20 A comparison between approximating the Lebesgue integral (left) and the Riemann integral (right). 134
21 The pushforward of a measure $\nu$. 135
22 The functions $\mu(X^-)$ and $\mu(X^+)$. 136
23 This map is not uniquely ergodic 138
List of Figures

24 The first two stages of the construction of the singular measure $\nu_p$. The shaded parts are taken out. 141
25 The first two stages of the construction of the middle third Cantor set. 144
26 The inverse image of a small interval $dy$ is $T^{-1}(dy)$ 150
27 $r$ 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. 153
28 $\ell(I)$ is between $\frac{1}{3}$ and $\frac{1}{2}$ of $\ell(J)$. So there are two disjoint images of $I$ under $R_{\omega}^{-1}$ that fall in $J$. 154
29 An example of the system described in Corollary 9.9. 156
30 Illustration of the fact that for a concave function $f$, we have $f(wx + (1 - w)y) \geq wf(x) + (1 - w)f(y)$ (Jensen’s inequality). 164
31 Plot of the function $\ln(x)\ln(1 + x)$ 165
Contents

List of Figures 3

Part 1. Introduction to Number Theory 4

Chapter 1. A Quick Tour of Number Theory 4
§1.1. Divisors and Congruences 5
§1.2. Rational and Irrational Numbers 5
§1.3. Algebraic and Transcendental Numbers 6
§1.4. Countable and Uncountable Sets 7
§1.5. Exercises 8

Chapter 2. The Fundamental Theorem of Arithmetic 9
§2.1. Bézout’s Lemma 9
§2.2. Corollaries of Bézout’s Lemma 10
§2.3. The Fundamental Theorem of Arithmetic 11
§2.4. Corollaries of the Fundamental Theorem of Arithmetic 12
§2.5. The Riemann Hypothesis 13
§2.6. Exercises 16

Chapter 3. Linear Diophantine Equations 17
§3.1. The Euclidean Algorithm 17
§3.2. A Particular Solution of $ax + by = c$ 41
§3.3. Solution of the Homogeneous equation $ax + by = 0$ 43
§3.4. The General Solution of $ax + by = c$ 44
§3.5. Recursive Solution of $x$ and $y$ in the Diophantine Equation 46
§3.6. Exercises 47

Chapter 4. Number Theoretic Functions 55
§4.1. Multiplicative Functions 55
§4.2. Additive Functions 58
§4.3. Möbius inversion 58
§4.4. Euler’s Phi or Totient Function 60
§4.5. Dirichlet and Lambert Series 62
§4.6. Exercises 65

Chapter 5. Modular Arithmetic and Primes 73
§5.1. Modular Arithmetic 73
§5.2. Euler’s Theorem 74
§5.3. Fermat’s Little Theorem and Primality Testing 76
§5.4. Fermat and Mersenne Primes 79
§5.5. Division in $\mathbb{Z}_b$ 81
§5.6. Exercises 83

Chapter 6. Continued Fractions 89
§6.1. The Gauss Map 89
§6.2. Continued Fractions 90
§6.3. Computing with Continued Fractions 95
§6.4. The Geometric Theory of Continued Fractions 97
§6.5. Closest Returns 98
§6.6. Exercises 101

Part 2. Topics in Number Theory

Chapter 7. Algebraic Integers 111
§7.1. Rings and Fields 111
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2</td>
<td>Primes and Integral Domains</td>
<td>113</td>
</tr>
<tr>
<td>7.3</td>
<td>Norms</td>
<td>115</td>
</tr>
<tr>
<td>7.4</td>
<td>Euclidean Domains</td>
<td>118</td>
</tr>
<tr>
<td>7.5</td>
<td>Example and Counter-Example</td>
<td>119</td>
</tr>
<tr>
<td>7.6</td>
<td>Exercises</td>
<td>123</td>
</tr>
<tr>
<td>8</td>
<td>Ergodic Theory</td>
<td>129</td>
</tr>
<tr>
<td>8.1</td>
<td>The Trouble with Measure Theory</td>
<td>129</td>
</tr>
<tr>
<td>8.2</td>
<td>Measure and Integration</td>
<td>131</td>
</tr>
<tr>
<td>8.3</td>
<td>The Birkhoff Ergodic Theorem</td>
<td>135</td>
</tr>
<tr>
<td>8.4</td>
<td>Examples of Ergodic Measures</td>
<td>138</td>
</tr>
<tr>
<td>8.5</td>
<td>The Lebesgue Decomposition</td>
<td>140</td>
</tr>
<tr>
<td>8.6</td>
<td>Exercises</td>
<td>142</td>
</tr>
<tr>
<td>9</td>
<td>Three Maps and the Real Numbers</td>
<td>149</td>
</tr>
<tr>
<td>9.1</td>
<td>Invariant Measures</td>
<td>149</td>
</tr>
<tr>
<td>9.2</td>
<td>The Lebesgue Density Theorem</td>
<td>152</td>
</tr>
<tr>
<td>9.3</td>
<td>Rotations and Multiplications on ( \mathbb{R}/\mathbb{Z} )</td>
<td>153</td>
</tr>
<tr>
<td>9.4</td>
<td>The Return of the Gauss Map</td>
<td>156</td>
</tr>
<tr>
<td>9.5</td>
<td>Number Theoretic Implications</td>
<td>158</td>
</tr>
<tr>
<td>9.6</td>
<td>Exercises</td>
<td>161</td>
</tr>
<tr>
<td>10</td>
<td>The Cauchy Integral Formula</td>
<td>169</td>
</tr>
<tr>
<td>10.1</td>
<td></td>
<td>169</td>
</tr>
<tr>
<td>10.2</td>
<td>Exercises</td>
<td>169</td>
</tr>
<tr>
<td>11</td>
<td>The Prime Number Theorem</td>
<td>171</td>
</tr>
<tr>
<td>11.1</td>
<td>Introduction</td>
<td>171</td>
</tr>
<tr>
<td>11.2</td>
<td>Chebyshev’s Theorem</td>
<td>173</td>
</tr>
<tr>
<td>11.3</td>
<td>Exercises</td>
<td>175</td>
</tr>
</tbody>
</table>

Bibliography | 179   |

Index | 181   |
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 \( \text{lcm}(a, b) \).

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

**Definition 1.3.** A number \( a > 1 \) 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 if it has
more than 2 divisors. (The number 1 is neither.)

\[ \begin{array}{cccccc}
2 & 3 & 5 & 7 & 11 & 13 \\
17 & 19 & 23 & 29 & 31
\end{array} \]

**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 7 — these are called irreducible numbers, while the term
prime is reserved for numbers satisfying Corollary 2.12.
1.2. Rational and Irrational Numbers

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 2.4). This method is illustrated in Figure 1. When done, the primes are those numbers that are circled or unmarked in the list.

**Definition 1.4.** Let $a$ and $b$ in $\mathbb{Z}$. Then $a$ and $b$ are relatively prime if $\gcd(a, b) = 1$.

**Definition 1.5.** 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 $a \equiv_{m} b$ or $a = b \mod m$ or $a \in b + m\mathbb{Z}$.

**Definition 1.6.** The residue of $a$ modulo $m$ is the (unique) integer $r$ in $\{0, \cdots, m - 1\}$ such that $a \equiv_{m} r$. It is denoted by $\text{Res}_{m}(a)$.

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 1.7 (well-ordering principle).** Any non-empty set $S$ in $\mathbb{N} \cup \{0\}$ 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. ■
Note that any non-empty set $S$ of integers with a lower bound can be transformed by addition of an integer $b \in \mathbb{N}_0$ into a non-empty $S + b$ in $\mathbb{N}_0$. Then $S + b$ has a lower bound, 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 1.8.** Let be a non-empty set $S$ in $\mathbb{Z}$ with a lower (upper) bound. Then $S$ has a smallest (largest) element.

**Definition 1.9.** An element $x \in \mathbb{R}$ is called rational if it satisfies $qx - p = 0$ where $p$ and $q \neq 0$ are integers. Otherwise it is called an irrational number. The set of rational numbers is denoted by $\mathbb{Q}$.

The usual way of expressing this, is that a rational number 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 leads to Definition 1.12.

**Theorem 1.10.** 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 1.8 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 1.11.** $\sqrt{2}$ 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

$$2s^2 = r^2.$$
1.3. Algebraic and Transcendental Numbers

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, it is not hard to see that the irrational numbers are also dense (exercise 1.8). In exercise 1.15, we prove that the number \( e \) is irrational. The proof that \( \pi \) is irrational is a little harder and can be found in [15][section 11.17]. In Chapter 2, we will use the fundamental theorem of arithmetic, Theorem 2.15, 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.

1.3. Algebraic and Transcendental Numbers

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

**Definition 1.12.** An element \( x \in \mathbb{R} \) 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 \( \pi \) are transcendental, but the proofs are considerably more difficult (see [16]). 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. 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 1.13.** We say that \( f(x) = \sum_{i=0}^{d} a_i x^i \in \mathbb{Z}[x] \) is a **minimal polynomial** for \( \rho \) if \( f \) is a non-zero polynomial of minimal degree such that \( f(\rho) = 0 \).

**Theorem 1.14. Liouville’s Theorem** Let \( f \) be a minimal polynomial of degree \( d \geq 2 \) for \( \rho \in \mathbb{R} \). Then

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

**Proof.** Clearly, if \( \left| \rho - \frac{p}{q} \right| \geq 1 \), the inequality is satisfied. So assume that \( \left| \rho - \frac{p}{q} \right| < 1 \).

Now let \( f \) be a minimal polynomial for \( \rho \), and set

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

We know that \( f\left(\frac{p}{q}\right) \) is not zero, because otherwise \( f \) would have a factor \( x - \frac{p}{q} \). In that case, the quotient \( g \) of \( f \) and \( x - \frac{p}{q} \) 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

\[
\left| q^d f\left(\frac{p}{q}\right) \right| = \sum_{i=0}^{d} a_i p^i q^{d-i} \geq 1 \quad \Rightarrow \quad \left| f\left(\frac{p}{q}\right) \right| \geq q^{-d}.
\]

because it is a non-zero integer. Finally, we use the mean value theorem which tells us that there is a \( t \) between \( \rho \) and \( \frac{p}{q} \) such that

\[
K \geq |f'(t)| = \left| \frac{f\left(\frac{p}{q}\right) - f(\rho)}{\frac{p}{q} - \rho} \right| \geq \frac{q^{-d}}{\left| \frac{p}{q} - \rho \right|}.
\]

since \( f(\rho) = 0 \). For \( K \) as above, this gives us the desired inequality. \( \blacksquare \)

**Definition 1.15.** A real number \( \rho \) is called a **Liouville number** if for all \( n \in \mathbb{N} \), there is a rational number \( \frac{p}{q} \) such that

\[
\left| \rho - \frac{p}{q} \right| < \frac{1}{q^n}.
\]
1.4. Countable and Uncountable Sets

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

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

(1.1)

It is easy to show that this is less than \(q^{-n}\) (exercise 1.17).

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

**Theorem 1.16. Roth’s Theorem** Let \(\rho \in \mathbb{R}\) be algebraic. Then for all \(\varepsilon > 0\)

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

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

This result is all the more remarkable if we consider it in the context of the following more general result (which we will have occasion to prove in Chapter 6).

**Theorem 1.17.** Let \(\rho \in \mathbb{R}\) be irrational. Then there are infinitely many \(\frac{p}{q} \in \mathbb{Q}\) such that

$$\left| \rho - \frac{p}{q} \right| < \frac{1}{q^2}.$$

1.4. Countable and Uncountable Sets

**Definition 1.18.** A set \(S\) is countable if there is a bijection \(f : \mathbb{N} \to S\). An infinite set for which there is no such bijection is called uncountable.

**Proposition 1.19.** 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 1.20.** The set $\mathbb{R}$ is uncountable.

**Proof.** The proof is one of mathematics’ most famous arguments: Cantor’s diagonal argument \[9\]. 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,

- $t(1) = 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, \ldots$
- $t(2) = 2, 0, 2, 0, 2, 0, 2, 2, 2, 2, \ldots$
- $t(3) = 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, \ldots$
- $t(4) = 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, \ldots$
- $t(5) = 0, 0, 0, 2, 0, 0, 2, 0, 2, 0, \ldots$
- $t(6) = 2, 0, 0, 0, 0, 2, 0, 0, 2, 2, \ldots$

... ... ...

Construct $t^*$ as follows: for every $n$, its $n$th digit differs from the $n$th digit of $t(n)$. In the above example, $t^* = 2, 2, 2, 0, 2, 0, \cdots$. 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? Then you would get a direct proof that $[0, 1]$ is uncountable. That can indeed be done. But there is a problem that has to be dealt with here. 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 $x = 1.0\ldots$ which might be in the list. To circumvent that problem leads to slightly more complicated proofs (see exercise 1.10).

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 $\{1, 2, \cdots, b - 1\}^\mathbb{N}$ the set of sequences $a_1a_2a_3\cdots$ where each $a_i$ is in $\{1, 2, \cdots, b - 1\}$. Such sequences are often called words.

**Corollary 1.21.** (i) The set of infinite sequences in $\{1, 2, \cdots, b - 1\}^\mathbb{N}$ is uncountable. (ii) The set of finite sequences (but without bound) in $\{1, 2, \cdots, b - 1\}^\mathbb{N}$ is countable.

**Proof.** The proof of (i) is the same as the proof that $T$ is uncountable in the proof of Theorem 1.20. 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 1.22.** (i) The set $\mathbb{Z}^2$ is countable. (ii) $\mathbb{Q}$ is countable.

**Proof.** (i) The proof relies on Figure 2. In it, a directed path $\gamma$ is traced out that passes through all points of $\mathbb{Z}^2$. Imagine that you start at $(0, 0)$ and travel along $\gamma$ 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 2. A directed path $\gamma$ passing through all points of $\mathbb{Z}^2$.

(ii) Again travel along $\gamma$ 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 of a countable set 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 1.9 and 1.25). 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”.
1.5. Exercises

**Definition 1.23.** Two sets $A$ and $B$ are said to have the same cardinality if there is a bijection $f : A \rightarrow B$. It is written as $|A| = |B|$. If there is an injection $f : A \rightarrow B$, then $|A| \leq |B|$.

**Definition 1.24.** An equivalence relation 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 $(a, c) \in R$ (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 1.23). 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 [18].

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. Are there infinitely many such pairs? (Hint: This is an open problem; the affirmative answer is called the twin prime conjecture.)

**Exercise 1.4.** Show that small enough even integers can be written as the sum of two primes. Is this always true? (Hint: This is an open problem; the affirmative answer is called the Goldbach conjecture.)

**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 iff $a \neq 0$.

b) Show that $b + x$ is irrational.

c) Show that $ax + b$ is irrational iff $a \neq 0$.

d) Conclude that $a\sqrt{2} + b$ is irrational iff $a \neq 0$.

Exercise 1.7. Show that $\sqrt{3}$, $\sqrt{5}$, etcetera (square roots of primes) are irrational.

Exercise 1.8. Show that numbers of the form that $a\sqrt{2} + b$ are irrational and dense in the reals ($a$ and $b$ are rational).

Lemma 1.25. The countable union of countably infinite sets is countably infinite.

Exercise 1.9. a) Use an pictorial argument similar to that of Figure 2 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 countably infinite sets where $i \in I$ and $I$ countable. Show that there is a bijection $f_i : \mathbb{N} \rightarrow A_i$ for each $i$.

c) Show there is a bijection $F : \mathbb{N} \times \mathbb{N} \rightarrow \bigcup_{i \in I} A_i$ given by $F(n, m) = f_n(m)$.

d) Conclude that Lemma 1.25 holds.

Exercise 1.10. 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 $(r(1), r(2), \ldots)$ forms an exhaustive list of the numbers in $[0, 1]$, as follows:

\[
\begin{align*}
    r(1) &= 0.0000000000\ldots \\
    r(2) &= 0.10101010111\ldots \\
    r(3) &= 0.0001111111\ldots \\
    r(4) &= 0.1111100000\ldots \\
    r(5) &= 0.0001010010\ldots \\
    r(6) &= 0.10000100011\ldots \\
    \vdots & \quad \vdots \\
\end{align*}
\]

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

\[
r^* = 0.111010\ldots,
\]

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 1.11. The set \( f(T) \) in the proof of Theorem 1.20 is called the middle third Cantor set. 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 1.12. 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 a small positive integer and apply \( f \) repeatedly, eventually you fall on the orbit in (a).

c) Show that this is true for all positive integers.

(Hint: This is an open problem; the affirmative answer is called the Collatz conjecture.)

Exercise 1.13. It is known that \( 2^{11213} - 1 \) is prime. How many decimal digits does this number have? (Hint: \( \log_{10} 2 \approx 0.301029996 \).)

Exercise 1.14. This exercise prepares for Mersenne and Fermat primes, see Definition 5.11.

a) Use \( \sum_{i=0}^{p-1} 2^i = \frac{2^p-1}{2-1} \) to show that if \( 2^p - 1 \) is prime, then \( p \) must be prime.

b) Use \( \sum_{i=0}^{p} (-2)^i = \frac{(-2)^{p+1}-1}{(-2)^1-1} \) to show that if \( 2^p + 1 \) is prime, then \( p \) has no odd factor.

Exercise 1.15. In the following, we assume that \( e - 1 = \sum_{i=1}^{\infty} \frac{1}{i!} = \frac{e}{q} \) is rational, and show that this leads to a contradiction.

a) Show that the above assumption implies that

\[
\sum_{i=1}^{q} \frac{q!}{i!} + \sum_{i=1}^{m} \frac{q!}{q+i!} = p(q-1)! .
\]

(Hint: multiply both sides of by \( q! \).)

b) Show that \( \sum_{i=1}^{m} \frac{q!}{q+i!} < \sum_{i=1}^{m} \frac{1}{i} \). (Hint: write out a few terms of the sum on the left.)

c) Show that the sums 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 1.16. Show that Liouville’s theorem (Theorem 1.14) also holds for rational for rational numbers \( \rho = \frac{p}{q} \) as long as \( \frac{p}{q} \neq \frac{\xi}{\zeta} \).
Exercise 1.17. 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 equation (1.1).

Exercise 1.18. Show that the inequality of Roth’s theorem does not hold for all numbers. (Hint: Let $p$ be a Liouville number.)

Definition 1.26. 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 1.27. The cardinality of a power set is always (strictly) greater than that of the set itself.

Exercise 1.19. 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 1.23.)

Exercise 1.20. Let $A$ be an arbitrary set. Assume that there is a surjection $S : A \to P(A)$ and define
$$
R = \{a \in A \mid a \notin S(a)\}.
$$
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 1.19, to establish that $|A| < |P(A)|$. (Hint: see Definition 1.23.)

In the next two exercises we show that the cardinality of $\mathbb{R}$ equals that of $P(\mathbb{N})$. This implies that $|\mathbb{R}| > |\mathbb{N}|$, which also follows from Theorem 1.20.
1.5. Exercises

Exercise 1.21. Let $T$ be the set of sequences defined in the proof of Theorem 1.20. 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 $S : T \rightarrow P(\mathbb{N})$.
b) Use the bijection $f$ in the proof of Theorem 1.20 to show there is a bijection $K \rightarrow P(\mathbb{N})$.
c) Show that (a) and (b) imply that $|P(\mathbb{N})| = |K| = |T|$. (Hint: see Definition 1.23.)
d) Find an injection $K \rightarrow \mathbb{R}$ and conclude that $|P(\mathbb{N})| \leq |\mathbb{R}|$.

Exercise 1.22.  

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

Exercise 1.23.  

a) Show that having the same cardinality (see Definition 1.23) is an equivalence relation on sets.
b) Conclude that “cardinality” is an “equivalence class of sets”.

Exercise 1.24.  

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 the remainder 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 1.25.  

a) Show that the set of algebraic numbers is countable. (Hint: use Lemma 1.25.)
b) Conclude that the transcendental numbers form an uncountable set.
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. 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 2.8) 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.

2.1. Bézout’s Lemma

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

By the well-ordering principle, Corollary 1.8, the number \(|\theta|\) and \(|\theta|\) exist for any \(\theta \in \mathbb{R}\).

**Definition 2.2.** Given a number \(\xi \in \mathbb{R}\), we denote its norm (or absolute value) \(|\xi|\) by \(N(\xi)\).

This last definition seems clumsy and unnecessary in the present context. But it will save us much trouble later on (see Section 7.4).

**Lemma 2.3.** Given \(r_1\) and \(r_2\) with \(r_2 > 0\), then there are unique \(q_2\) and \(r_3 \geq 0\) with \(N(r_3) < N(r_2)\) such that \(r_1 = r_2q_2 + r_3\).

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

\[
\frac{r_1}{r_2} = q_2 + \varepsilon ,
\]

where \(\varepsilon \in [0, 1)\). The integer \(q_2\) is called the quotient. Multiplying by \(r_2\) gives the result. \(\blacksquare\)

Note that \(r_3 \in \{0, \cdots, r_2 - 1\}\). Thus among other things, this lemma implies that every integer has a unique residue (see Definition 1.6). If \(N(r_1) < N(r_2)\), then \(q_2 = 0\). In this case, \(\varepsilon = \frac{r_1}{r_2}\) and we learn nothing new. But if \(N(r_1) > N(r_2)\), then \(q_2 \neq 0\) and we have written \(r_1\) as a multiple of \(r_2\) plus a remainder \(r_3\).

**Definition 2.4.** Given \(r_1\) and \(r_2\) with \(r_2 > 0\), the computation of \(q_2\) and \(r_3\) satisfying Lemma 2.3 is called the division algorithm. Note that \(r_3 = \text{Res}_{r_2}(r_1)\).

**Remark 2.5.** Lemma 2.3 is also called Euclid’s division lemma. This is not to be confused with the Euclidean algorithm of Definition 3.3.

**Lemma 2.6. (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:

\[
S = \{ax + by \mid x, y \in \mathbb{Z}, ax + by \neq 0\}
\]

\[
\nu(S) = \{N(s) \mid s \in S\} \subseteq \mathbb{N}
\]
2.2. Corollaries of Bézout’s Lemma

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

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

\[
(2.1) \quad a = qd + r \quad \text{and} \quad N(r) < N(d).
\]

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

\[
r = a(1 - qx) + b(-qy).
\]

This shows that \( r \in S \). But we also know from (2.1) that \( N(r) \) is smaller than \( N(d) \). Unless \( r = 0 \), 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 \( e | (ax + by) \), and so \( e | d \). But if \( e | d \), then \( N(e) \) must be smaller than or equal to \( N(d) \). 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 \( d | c \), and thus \( c \) must be a multiple of \( d \). \( \blacksquare \)

2.2. Corollaries of Bézout’s Lemma

**Lemma 2.7. (Euclid’s Lemma)** Let \( a \) and \( b \) be such that \( \gcd(a, b) = 1 \) and \( a | bc \). Then \( a | c \).

**Proof.** By Bézout, there are \( x \) and \( y \) such that \( ax + by = 1 \). Multiply by \( c \) to get:

\[
acx + bcy = c.
\]

Since \( a | bc \), the left hand side is divisible by \( a \), and so is the right hand side. \( \blacksquare \)
Euclid’s lemma is so often used, that it will pay off to have a few of the standard consequences for future reference.

**Theorem 2.8 (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 \( b | a(x - y) \).
The conclusion follows from Euclid’s Theorem. Vice versa, if \( x =_b y \), then \( (x - y) \) is a multiple of \( b \) and \( a(x - y) \) 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 \( 21x =_3 21y \). It is tempting to say that this implies that \( x =_5 y \). But in fact, \( \gcd(21, 35) = 7 \) and the solution set is \( x =_7 y \), as is easily checked. This example is in fact a special of the following corollary.

**Corollary 2.9.** Let \( \gcd(a, b) = d \) and \( b \) positive. Then \( ax =_b ay \) if and only if \( x =_b dy \).

**Proof.** Again the statement is equivalent to \( b | a(x - y) \). But now we can divide by \( d \) to get \( \frac{b}{d} | \frac{a}{d}(x - y) \). Now we can apply the cancellation theorem. ■

**Corollary 2.10.** Let \( \gcd(a, b) = 1 \) and \( b \) positive. If \( a | c \) and \( b | c \) then \( ab | c \).

**Proof.** We have \( c = ax = by \) and thus \( ax =_b 0 \). The cancellation theorem implies that \( x =_b 0 \). ■

**Proposition 2.11.** \( ax =_m c \) has a solution if and only \( \gcd(a, m) | c \).

**Proof.** By Definition 1.5, \( ax =_m c \) means \( ax + my = c \) for some \( y \), and the result follows from Bézout. ■

**Corollary 2.12.** For any \( n \geq 1 \), if \( p \) is prime and \( p \nmid \prod_{i=1}^{n} a_i \), then there is \( j \leq n \) such that \( p | a_j \).

**Proof.** We prove this by induction on \( n \), the number of terms in the product. Let \( S(n) \) be the statement of the Corollary.
2.3. The Fundamental Theorem of Arithmetic

The statement $S(1)$ is: If $p$ is prime and $p|a_1$, then $p|a_1$, which is trivially true.

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

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

Applying Euclid’s Lemma, it follows that

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

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

Corollary 2.13. If $p$ and $q_i$ are prime and $p|\prod_{i=1}^{n} q_i$, then there is $j \leq n$ such that $p = q_j$.

Proof. Corollary 2.12 says that if $p$ and all $q_i$ are primes, then there is $j \leq n$ such that $p|q_j$. 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. But first, we introduce units and extend the definition of primes to $\mathbb{Z}$.

Definition 2.14. 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 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.

Theorem 2.15 (The Fundamental Theorem of Arithmetic). Every non-zero integer $n \in \mathbb{Z}$

(1) is a product of powers of primes (up to a unit) and

(2) that product is unique (up to the order of multiplication and up to multiplication by the units $\pm 1$).
2. The Fundamental Theorem of Arithmetic

**Remark:** The theorem is also called the *unique factorization theorem*. Its statement means that up to re-ordering of the $p_i$, 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. Statement (1):** Define $S$ to be the set of integers $n$ that are not products of primes times a unit, and the set $\nu(S)$ their norms. If the set $S$ is non-empty, then by the well-ordering principle (Theorem 1.7), $\nu(S)$ has a smallest element. Let $a$ be one of the elements in $S$ that minimize $\nu(S)$. By possibly multiplying by -1, we may assume that $a$ is positive.

If $a$ is prime, then it can be factored into primes, namely $a = a$, and the theorem would hold in this case. Thus $a$ is a composite number, $a = bc$ and both $b$ and $c$ are non-units. Thus $N(b)$ and $N(c)$ are strictly smaller than $N(a)$. If $b$ and $c$ are products of primes, then, of course, so is $a = bc$. Therefore at least one of $a$ or $b$ is not. But this contradicts the assumptions on $a$.

**Proof of (2):** Let $S$ be the set of integers that have more than one factorization and $\nu(S)$ the set of their norms. If the set $S$ is non-empty, then by the well-ordering principle (Theorem 1.7), $\nu(S)$ 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 2.13, $p_1$ equals one of the $p'_i$, say, $p'_1$. Since primes are not units, $N\left(\frac{a}{p_1}\right)$ is strictly less than $N(a)$. Therefore, by hypothesis, $\frac{a}{p_1}$ is uniquely factorizable. But then 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).
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 2.5 for an example).

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

**Proof.** If $c$ is a divisor or multiple of $a$, then so is $-c$. So $a$ and $-a$ have the same divisors and multiples. Therefore without loss of generality we may assume that $a$ and $b$ are positive.

Given two positive 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:

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

where $k_i$ and $\ell_i$ 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 $\lcm(a, b)$. 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 $\lcm(a, b)$ is similar. 

A final 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 (Infinity of Primes).** There are infinitely many primes.
The Fundamental Theorem of Arithmetic

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

\[
p_i | d \quad \text{and} \quad p_i | d + 1.
\]

But since \((d + 1)(1) + d(-1) = 1\), Bézout’s lemma implies that \( \gcd(d, d + 1) = 1 \), which contradicts our earlier conclusion that \( p_i | d \).

One of 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 1.11), 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

\[
n^{\frac{1}{k}} = \frac{a}{b} \implies n \cdot b^k = a^k.
\]

Divide out any common divisors of \( a \) and \( b \), so that \( \gcd(a, b) = 1 \). Then by the fundamental theorem of arithmetic:

\[
n \prod_{i=1}^{k} p_i^{km_i} = \prod_{i=s+1}^{r} p_i^{kJ_i}.
\]

Therefore, in the factorization of \( n \), each prime \( p_i \) must occur with a power that is a non-negative multiple of \( k \). Because \( \gcd(a, b) = 1 \), the primes \( p_i \) on the left and right side are distinct. This is only possible if \( \prod_{i=1}^{r} p_i^{km_i} \) equals 1. But then \( n \) is the \( k \)-th power of an integer.

**Lemma 2.19.** We have

\[
\forall i \in \{1, \cdots n\} : \gcd(a_i, b) = 1 \iff \gcd(\prod_{i=1}^{n} a_i, b) = 1.
\]
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 2.12) 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 \( d \).

\[ \square \]

2.5. The Riemann Hypothesis

Definition 2.20. The Riemann zeta function \( \zeta(z) \) is a complex function defined as follows on \( \{ z \in \mathbb{C} \mid \text{Re} z > 1 \} \)

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

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).

Analytic continuation is akin to replacing \( e^x \) where \( x \) is real by \( e^z \) where \( z \) is complex. Another 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.

Recall that an analytic function is a function that is differentiable. Equivalently, it is a function that is 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 some \( U \) and therefore all its power expansions are zero and so it must be zero on the the entire region. Hence, analytic conjugations are unique. That is the reason they are meaningful. For more details, see for example [12, 22].

It is customary to denote the argument of the zeta function by \( s \). We will do so from here on out. Note that \( |n^{-s}| = n^{-\text{Re} s} \), and so for \( \text{Re} s > 1 \) the series is absolutely convergent. 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 2.21 (Euler’s Product Formula). For $\Re s > 1$ we have

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

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

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

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

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

$$
\prod_{p \text{ prime}} (1 - p^{-s})^{-1} = \left( \sum_{k_1=0}^{\infty} p_1^{-k_1s} \right) \left( \sum_{k_2=0}^{\infty} p_2^{-k_2s} \right) \left( \sum_{k_3=0}^{\infty} p_3^{-k_3s} \right) \cdots
$$

Re-arranging terms yields

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

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 the left hand of Euler’s formula.

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

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

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

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

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 $\ell$ tends to infinity. Therefore, the left hand tends to 1, which implies the proposition.
2.5. The Riemann Hypothesis

The most important theorem concerning primes is probably the following. We will give a proof in Chapter 11.

Theorem 2.22 (Prime Number Theorem). Let \( \pi(x) \) denote the prime counting function, that is: the number of primes less than or equal to \( x > 2 \). Then

\[
\begin{align*}
\lim_{x \to \infty} \frac{\pi(x)}{x/\ln x} &= 1 \\
\lim_{x \to \infty} \frac{\pi(x)}{\int_2^x \ln t \, dt} &= 1,
\end{align*}
\]

where \( \ln \) is the natural logarithm.

The first estimate is the one we will prove directly in Chapter 11. It turns out the second is equivalent to it (exercise 11.4). However, it is this one that is more accurate. In Figure 3 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 green. On the right, we have \( \int_2^x \ln t \, dt - x/\ln x \) in blue, \( \pi(x) - x/\ln x \) in red.

Figure 3. 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.

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 \( \int_2^n \ln t \, dt - \pi(n) < 0 \). The first such \( n \) is called the Skewes
2. The Fundamental Theorem of Arithmetic

The Fundamental Theorem of Arithmetic

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 \(\zeta(s)\) given above.

**Conjecture 2.23 (Riemann Hypothesis).** All non-real zeros of \(\zeta(s)\) lie on the line \(\text{Re } s = \frac{1}{2}\).

In his only paper on number theory \([28]\), Riemann realized that the hypothesis enabled him to describe detailed properties of the distribution of primes in terms of the location of the non-real zero of \(\zeta(s)\). 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 \([1]\), and even physics \([6, 8]\).

### 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.

---

\(^{1}\)In 2020.

\(^{2}\)This area of research, complex analysis methods to investigate properties of primes, is now called **analytic number theory**. We take this up in Chapters 10 and 11.
2.6. Exercises

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

(Hint: let $x = 16$.)

Exercise 2.4. 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 2.5. 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 $\lcm(a, b)$.

d) Show that $a / \left(\frac{ab}{\lcm(a, b)}\right) = \frac{\lcm(a, b)}{b}$ is an integer. Thus $\frac{ab}{\lcm(a, b)}$ is a divisor of $a$.

e) Similarly, show that $\frac{ab}{\lcm(a, b)}$ is a divisor of $b$.

f) Conclude that $\frac{ab}{\lcm(a, b)} \leq \gcd(a, b)$.

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 2.6. It is possible to extend the definition of \( \gcd \) and \( \text{lcm} \) 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 \( \text{lcm}(6, 10, 15) \).

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 \( \text{lcm} \) equals \( abc \), but every pair of which has \( \text{lcm} \) less than the product of that pair. (Hint: consider \( \text{lcm}(a, b) \cdot c \).)

Exercise 2.7. 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 2.8. Use the Fundamental Theorem of Arithmetic to prove:

a) Bézout’s Lemma.

b) Euclid’s Lemma.

Exercise 2.9. For positive integers \( m \) and \( n \), suppose that \( m^\alpha = n \). Show that \( \alpha = \frac{x}{2} \) with \( \gcd(a, b) = 1 \) if and only if

\[
m = \prod_{i=1}^{x} p_i^{k_i} \quad \text{and} \quad n = \prod_{i=1}^{x} p_i^{\ell_i} \quad \text{with} \quad \forall i : \; a k_i = b \ell_i .
\]
2.6. Exercises

Exercise 2.10. 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 $L = \{p_1, p_2, \ldots, p_k\}$ where $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, and so forth, up to the last prime $p_k$.

a) Show that $\prod_{i=1}^{k} \frac{1}{p_i - 1} = M$, where $M$ is finite.

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

c) Use the fundamental theorem of arithmetic to show that there is an $\alpha(r) > 0$ such that $\prod_{i=1}^{k} \left( \sum_{j=0}^{r} \frac{1}{p_i^j} \right) = \alpha(r) \frac{1}{\ell} + R$, 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 $\sum_{\ell=1}^{\infty} \frac{1}{\ell}$ diverges or see exercise 2.11 c.)

Exercise 2.11. In this exercise we consider the Riemann zeta function for real values of $s$ greater than 1.

a) Show that for all $x > -1$, we have $\ln(1 + x) \leq x$.

b) Use Proposition 2.21 and a) to show that $\ln \zeta(s) = \sum_{p \text{ prime}} \ln \left(1 + \frac{p^{-s}}{1 - p^{-s}}\right) \leq \sum_{p \text{ prime}} \frac{p^{-s}}{1 - p^{-s}} \leq \sum_{p \text{ prime}} \frac{p^{-s}}{1 - 2^{-s}}$.

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

$$\sum_{n=1}^{\infty} n^{-1} \geq \sum_{n=1}^{\infty} n^{-s} \geq \int_{1}^{\infty} x^{-s} dx.$$ (For the last inequality, see Figure 4.)

d) Show that b) and c) imply that $\sum_{p \text{ prime}} p^{-s}$ diverges as $s \to 1$.

e) Show that — in some sense — primes are more frequent than squares in the natural numbers. (Hint: $\sum_{n=1}^{\infty} n^{-2}$ converges.)
2. The Fundamental Theorem of Arithmetic

Exercise 2.12. 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 2.13. See exercise 2.12. Show that any number in $E$ is a product of primes in $E$. (Hint: follow the proof of Theorem 2.15, part (1).)

Exercise 2.14. See exercise 2.12 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 2.15. Let $L = \{p_1, p_2, \cdots \}$ be the list of all primes, ordered according ascending magnitude. Show that $p_{n+1} \leq \prod_{i=1}^{n} p_i$. (Hint: consider $d$ as in the proof of Theorem 2.17, let $p_{n+1}$ be the smallest prime divisor of $d - 1$.)

A much stronger version of exercise 2.15 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 Erdős [2].

Exercise 2.16. Let $p$ and $q$ primes greater than or equal to 5.

a) Show that $p = 12q + r$ where $r \in \{1, 5, 7, 11\}$. (The same holds for $q$.)
b) Show that $24|p^2 - q^2$. (Hint: use (a) to reduce this to $\frac{1}{2} = 6$ cases.)
2.6. Exercises

Exercise 2.17. A square full number is a number \( n > 1 \) such that each prime factor occurs with a power at least 2. A square free number is a number \( n > 1 \) 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 number.

Exercise 2.18. 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 \( E_n = 4 \cdot 3 \). (Hint: \( E_n - 1 \) is twice an odd number.)

c) Show that for \( n \geq 3 \) the decimal representation of \( E_n \) ends in 1. (Hint: look at the factors of \( E_n \).)

Exercise 2.19. 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 2.16)

Exercise 2.20. 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.

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 2.21. 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.
2. The Fundamental Theorem of Arithmetic

**Exercise 2.22.** The set-up of the proof is the same as in exercise 2.21. 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 \( k > 2 \), \( \{2, \ldots, k\} \) can be uniquely factored. Then there are primes \( p_i \) and \( q_i \), not necessarily distinct, such that

\[
k + 1 = \prod_{i=1}^{\ell} p_i = \prod_{i=1}^{\ell} q_i .
\]

c) Show that then \( p_1 \) divides \( \prod_{i=1}^{r} q_i \) and so, Corollary 2.13 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

\[
k + 1 = \frac{k + 1}{q_1} = \prod_{i=2}^{\ell} p_i = \prod_{i=2}^{\ell} q_i .
\]

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 2.24.** (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 \( \text{lcm}(a, b) \) if and only if every common multiple of \( a \) and \( b \) is a multiple of \( d \).

**Exercise 2.23.** Use the characterization of \( \gcd(a, b) \) and \( \text{lcm}(a, b) \) given in the proof of Corollary 2.16 to prove Corollary 2.24.
2.6. Exercises

Exercise 2.24. 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 $\prod_{p \text{ prime}} (1 - p^{-2})$. (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 $d > 1$:
$$1 < \zeta(d) < 1 + \int_1^\infty x^{-d} \, dx < \infty.$$ For the middle inequality, see Figure 5.

f) Show that for large $d$, the probability that $\gcd(a_1, a_2, \cdots, a_d) = 1$ tends to 1.

![Figure 5](image_url)

Figure 5. Proof that $\sum_{n=1}^\infty f(n)$ (shaded in blue and green) minus $f(1)$ (shaded in blue) is less than $\int_1^\infty x^{-d} \, dx$ if $f$ is positive and decreasing to 0.

Exercise 2.25. This exercise is based on exercise 2.24.

a) In the $\{-4, \cdots, 4\}^2\setminus(0,0)$ grid in $\mathbb{Z}^2$, find out which proportion of the lattice points is visible from the origin, see Figure 6.

b) Show that in a large grid, this proportion tends to $1/\zeta(2)$. (We note here that $\zeta(2) = \pi^2/6$.)

c) 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 6. The origin is marked by “×”. The red dots are visible from ×; between any blue dot and × there is a red dot. The picture shows exactly one quarter of \((-4, \cdots, 4)^2 \setminus (0, 0) \subset \mathbb{Z}^2\).
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 $ax + by = c$.

3.1. The Euclidean Algorithm

Lemma 3.1. In the division algorithm of Definition 2.4, 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 $\gcd(r_1, r_2) = \gcd(r_n, 0) = r_n$.  

39
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 Euclidean algorithm.

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:

\[
\begin{align*}
188 &= 158 \cdot 1 + 30 \\
158 &= 30 \cdot 5 + 8 \\
30 &= 8 \cdot 3 + 6 \\
8 &= 6 \cdot 1 + 2 \\
6 &= 2 \cdot 3 + 0
\end{align*}
\]

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 2.3). If we call them \( q_{i-1} \), the computation looks as follows:

\[
\begin{align*}
r_1 &= r_2 q_2 + r_3 \\
r_2 &= r_3 q_3 + r_4 \\
&\quad \vdots \\
r_{n-3} &= r_{n-2} q_{n-2} + r_{n-1} \\
r_{n-2} &= r_{n-1} q_{n-1} + r_n \\
r_{n-1} &= r_n q_n + 0
\end{align*}
\]

(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 [19]) to render this computation is as follows.

\[
\begin{array}{c|c|c|c|c|c|c}
& q_n & q_{n-1} & \cdots & q_3 & q_2 \\
0 & r_n & r_{n-1} & \cdots & r_3 & r_2 & r_1
\end{array}
\]

(3.2)

Thus, each step \( r_{i+1} | r_i \) 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 \( r_{i+1} \). The example we worked out
before, now looks like this:

\[
\begin{array}{cccccc}
3 & 1 & 3 & 5 & 1 & 1 \\
0 & 2 & 6 & 8 & 30 & 158 \\
20 & 6 & 8 & 30 & 158 & 188 \\
\end{array}
\]

There is a beautiful visualization of this process outlined in exercise 3.4.

### 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.

\[
\begin{pmatrix}
q_i & 1 \\
1 & 0
\end{pmatrix}
\begin{pmatrix}
r_{i-1} \\
r_i
\end{pmatrix}
= 
\begin{pmatrix}
r_i & 1 \\
1 & 0
\end{pmatrix}
\begin{pmatrix}
r_i \\
r_{i+1}
\end{pmatrix}.
\]

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 \( r_1 x + r_2 y = r \gcd(r_1, r_2) \). Note that from Bézout’s lemma (Lemma 2.6), 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 \( r_1 x + r_2 y = r \gcd(r_1, r_2) \) is given by

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

**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

\[
\begin{pmatrix}
r_i \\
r_{i+1}
\end{pmatrix} = Q_i^{-1}
\begin{pmatrix}
r_{i-1} \\
r_i
\end{pmatrix} \quad \implies \quad \begin{pmatrix}
r_{n-1} \\
r_n
\end{pmatrix} = rQ_{n-1}^{-1} \cdots Q_2^{-1}
\begin{pmatrix}
r_1 \\
r_2
\end{pmatrix}.
\]
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 equalities follow immediately.

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,
\[
\gcd(r_1, r_2) = r_n = r_{n-2} - (r_{n-3} - r_{n-2} q_{n-2}) q_{n-1} = r_{n-2} (1 + q_{n-1} q_{n-2}) + r_{n-3} (-q_{n-1}).
\]

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

\[
\begin{array}{cccc|ccc}
| & + & | & - & | & + & | & - & | & + & | & + \\
\hline
| & q_n & q_{n-1} & q_{n-2} & q_{n-3} & q_{n-4} & | \\
0 & r_n & r_{n-1} & r_{n-2} & r_{n-3} & r_{n-4} & \cdots \\
1 & | & -q_{n-1} & | & 1 & | & -q_{n-1} & | & -q_{n-3} (1 + q_{n-1} q_{n-2}) & | & 1 + q_{n-1} q_{n-2} \\
0 & | & q_{n-1} q_{n-2} & | & -q_{n-1} & | & -q_{n-3} (1 + q_{n-1} q_{n-2}) & | & 1 + q_{n-1} q_{n-2} & | & \cdots
\end{array}
\]

(The signs added in the first line in this scheme serve only to keep track of the signs of the coefficients in lines three and below.) Applying this to the
3.3. Solution of the Homogeneous equation \( ax + by = 0 \)

example gives

\[
\begin{array}{cccccc}
| & + & - & + & - & - \\
3 & 1 & 3 & 5 & 1 & 1 \\
0 & 2 & 6 & 8 & 30 & 158 & 188 \\
1 & -1 & 1 & -1 & -20 & 4 & -21 \\
\end{array}
\]

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 \( r_1x + r_2y = 0 \) is given by

\[
x = k \frac{r_2}{\gcd(r_1, r_2)} \quad \text{and} \quad y = -k \frac{r_1}{\gcd(r_1, r_2)},
\]

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 \( \frac{r_i}{\gcd(r_1, r_2)} \) 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 \( \ell(\xi) = \begin{pmatrix} r_2 \\ -r_1 \end{pmatrix} \xi \). To obtain all its lattice points (i.e., points that are also in \( \mathbb{Z}^2 \)), both \( r_2 \xi \) and \( -r_1 \xi \) must be integers. The smallest positive number \( \xi \) for which this is possible, is \( \xi = \frac{1}{\gcd(r_1, r_2)} \).

Here is another homogeneous problem that we will run into. First we need a small update of Definition 1.2.
Definition 3.6. Let \( \{ b_i \}_{i=1}^n \) be non-zero integers. Their greatest common divisor, \( \text{lcm} (b_1, \cdots, b_n) \), is the maximum of the numbers that are divisors of every \( b_i \); their least common multiple, \( \text{gcd}(b_1, \cdots, b_n) \), is the least of the positive numbers that are multiples of every \( b_i \).

Surprisingly, for this more general definition, the generalization of Corollary 2.16 is false. For an example, see exercise 2.6.

Corollary 3.7. Let \( \{ b_i \}_{i=1}^n \) be non-zero integers and denote \( B = \text{lcm} (b_1, \cdots, b_n) \).

The general solution of the homogeneous system of equations \( x = b_i 0 \) is given by

\[
\begin{align*}
x &= B 0.
\end{align*}
\]

Proof. From the definition of \( \text{lcm} (b_1, \cdots, b_n) \), every such \( x \) is a solution. On the other hand, if \( x \neq B 0 \), then there is an \( i \) such that \( x \) is not a multiple of \( b_i \), and therefore such an \( x \) is not a solution. \( \blacksquare \)

3.4. The General Solution of \( ax + by = c \)

Definition 3.8. Let \( r_1 \) and \( r_2 \) be given. The equation \( r_1 x + r_2 y = 0 \) is called homogeneous\(^1\). The equation \( r_1 x + r_2 y = c \) 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 (inhomogeneous) equation.

It is useful to have some geometric intuition relevant to the equation \( r_1 x + r_2 y = c \). In \( \mathbb{R}^2 \), we set \( \vec{r} = (r_1, r_2) \), \( \vec{x} = (x, y) \), etcetera. The standard inner product is written as \( \langle \cdot, \cdot \rangle \). 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 \( \langle \vec{r}, \vec{x} \rangle = c \). This line is orthogonal to the vector \( \vec{r} \) and its distance to the origin (measured along the vector \( \vec{r} \)) equals \( \frac{|c|}{\sqrt{\langle \vec{r}, \vec{r} \rangle}} \). The situation is illustrated in Figure 7.

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

---

\(^1\)The word “homogeneous” in daily usage receives the emphasis often on its second syllable (“ho-MODGE-uhns”). 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 3.9.** Let \((x^{(0)}, y^{(0)})\) be a particular solution of \(r_1 x + r_2 y = c\). 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 \(r_1 x + r_2 y = 0\).

![Figure 7. The general solution of the inhomogeneous equation \((\vec{r}, \vec{x}) = c\) in \(\mathbb{R}^2\).](image)

**Proof.** Let \((x^{(0)}, y^{(0)})\) be that particular solution. Let \(m\) be the line given by \((\vec{r}, \vec{x}) = c\). Translate \(m\) over the vector \((-x^{(0)}, -y^{(0)})\) to get the line \(m'\). Then an integer point on the line \(m'\) is a solution \((z_1, z_2)\) of the homogeneous equation if and only if \((x^{(0)} + z_1, y^{(0)} + z_2)\) on \(m\) is also an integer point (see Figure 7).

Bézout’s Lemma says that \(r_1 x + r_2 y = c\) 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.
3. Linear Diophantine Equations

Corollary 3.10. 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 the particular solution of Theorem 3.4 and the general 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 3.11. Given $r_1$, $r_2$, and their successive remainders $r_3$, $\ldots$, $r_n \neq 0$, and $r_{n+1} = 0$ in the Euclidean algorithm. Then for $i \in \{3, \ldots, n\}$, 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 3.12. Given $r_1$, $r_2$, and their successive quotients $q_2$ through $q_n$ as in equation 3.1, then $x_i$ and $y_i$ of Corollary 3.11 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 follows, because

$$
\begin{align*}
    r_1 &= r_1 \cdot 1 + r_2 \cdot 0, \\
    r_2 &= r_1 \cdot 0 + r_2 \cdot 1.
\end{align*}
$$

Notice that, by definition,

$$
\begin{pmatrix}
    r_i \\
    r_{i+1}
\end{pmatrix} =
\begin{pmatrix}
    r_1x_i + r_2y_i \\
    r_1x_{i+1} + r_2y_{i+1}
\end{pmatrix} =
\begin{pmatrix}
    x_i & y_i \\
    x_{i+1} & y_{i+1}
\end{pmatrix}
\begin{pmatrix}
    r_1 \\
    r_2
\end{pmatrix}.
$$

From Corollary 3.11, we now have that

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

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

We remark that the recursion in Corollary 3.12 can also be expressed as

\[
\begin{align*}
x_{i+1} &= -q_ix_i + x_{i-1} \\
y_{i+1} &= -q iy_i + y_{i-1}.
\end{align*}
\]

3.6. Exercises

Exercise 3.1. Which of the following equations have integer solutions?

a) $1137 = 69x + 39y$.
b) $1138 = 69x + 39y$.
c) $-64 = 147x + 84y$.
d) $-63 = 147x + 84y$.

Exercise 3.2. Let $\ell$ be the line in $\mathbb{R}^2$ given by $y = \rho x$, where $\rho \in \mathbb{R}$.

a) Show that $\ell$ intersects $\mathbb{Z}^2$ if and only if $\rho$ is rational.
b) Given a rational $\rho > 0$, find the intersection of $\ell$ with $\mathbb{Z}^2$. (Hint: set $\rho = \frac{t_1}{t_2}$ and use Proposition 3.5.)

Exercise 3.3. 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, see exercise 2.1)

a) $110$, $7$.
b) $51$, $-30$.
c) $-138$, $24$.
d) $272$, $119$.
e) $2378$, $1769$.
f) $270$, $175$, $560$. 
Exercise 3.4. This problem was taken (and reformulated) from [15].
a) Tile a 188 by 158 rectangle by squares using what is called a
greedy algorithm\footnote{By “greedy” we mean that at every step, you choose the biggest square possible. In general a greedy algorithm always makes a locally optimal choice.}. The first square is 158 by 158. The remaining rec-
tangle 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).
b) Consider equation (3.1) or (3.2) and use a) to show that
\[ r_1r_2 = \sum_{i=2}^{n} q_i r_i^2. \]
(Hint: assume that \( r_1 > r_2 > 0, r_n \neq 0, \) and \( r_{n+1} = 0 \).)

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) \( 270x + 175,560y = 170. \)
f2) \( 270x + 175,560y = 150. \)

Exercise 3.6. Find the solutions for \( x \) and \( y \) of the following (homoge-
nous) Diophantine equations. If yes, find a (particular) solution. (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) \( 270x + 175,560y = 0. \)

Exercise 3.7. Find all solutions for \( x \) and \( y \) in all problems of exercise 3.5
that admit a solution.

Exercise 3.8. Use Corollary 3.11 to express the successive remainders \( r_i \)
in each of the items in exercise 3.3 as \( r_1x_i + r_2y_i \).
3.6. Exercises

Exercise 3.9. Use Corollary 3.12 to express $x_i$ and $y_i$ in the successive remainders $r_i$ in each of the items in exercise 3.3.

Exercise 3.10. Consider the line $\ell$ 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 $r_i$ are integers.

a) Show that $\ell(\xi) \in \mathbb{Z}^3 \setminus \{\vec{0}\}$ 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 $\ell$ has no non-zero points in common with $\mathbb{Z}^3$. 
3. Linear Diophantine Equations

Exercise 3.11. For this exercise, see exercises 2.2 and 2.3. Given three polynomials \( p_1 \) and \( p_2 \) with coefficients in \( \mathbb{Q}, \mathbb{R} \), or \( \mathbb{C} \). Define \( \text{gcd}(p_1, p_2) \) as the highest degree polynomial that is a divisor of both \( p_1 \) and \( p_2 \). (Remark: if you do this very carefully, you will realize that in this problem you really need to consider equivalence classes of polynomials, in the sense that two polynomials \( p \) and \( q \) are equivalent if there is a non-zero constant \( c \in \mathbb{R} \) such that \( p = cq \). While we are not pursuing the details here, it is good to keep this in the back of your mind.)

a) Describe an algorithm based on the division algorithm for polynomials to determine the greatest common divisor \( d(x) \) of \( p_1(x) \) and \( p_2(x) \). (Hint: see the Euclidean algorithm of Section 3.1. To see how to apply the division algorithm to polynomials, see exercise 2.2.)

b) Give a necessary and sufficient criterion for the existence of two polynomials \( g(x) \) and \( h(x) \) satisfying

\[
 p_1(x) g(x) + p_2(x) h(x) = d(x) .
\]

(Hint: the same as Bézout’s lemma.)

c) Apply the trick of “backward” solving in Section 3.2 to determine an algorithm find polynomials \( g(x) \) and \( h(x) \) satisfying

\[
 p_1(x) g(x) + p_2(x) h(x) = d(x) ,
\]

where now \( d = \text{gcd}(p_1, p_2) \). This is the particular solution.

d) Find the general solution of

\[
 p_1(x) g(x) + p_2(x) h(x) = 0 .
\]

(Hint: see Proposition 3.5.)

e) Find the general solution of

\[
 p_1(x) g(x) + p_2(x) h(x) = e(x) ,
\]

where now \( e \) is a polynomial times \( \text{gcd}(p_1, p_2) \). (Hint: see Corollary 3.10.)

c) Apply the above to \( p_1(x) = x^7 - x^2 + 1, p_2(x) = x^3 + x^2, \) and \( e(x) = 2 - x \).

(Hint: We list the steps of the Euclidean algorithm:

\[
 \begin{align*}
 (x^7 - x^2 + 1) &= (x^3 + x^2) (x^4 - x^3 + x^2 - x + 1) + (-2x^2 + 1) \\
 (x^3 + x^2) &= (-2x^2 + 1) (-\frac{1}{2}x - \frac{1}{2}) + (\frac{1}{2}x + \frac{1}{2}) \\
 (-2x^2 + 1) &= (\frac{1}{2}x + \frac{1}{2}) (-4x + 4) + (-1) \\
 (\frac{1}{2}x + \frac{1}{2}) &= (-1) (-\frac{1}{2}x - \frac{1}{2}) + (0)
\end{align*}
\]

Thus \( \text{gcd}(p_1, p_2) = 1, \) or in other words, \( p_1(x) \) and \( p_2(x) \) are relatively prime.)

Exercise 3.12. Let \( p(x) \) be a polynomial and \( p'(x) \) its derivative.

a) Show that if \( p(x) \) has a multiple root \( \lambda \) 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 3.11, 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 \( \text{gcd}(p, p') \).)
Theorem 3.13 (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 3.13. Assume that every polynomial $f$ of degree $d \geq 1$ has at least 1 root, prove the fundamental theorem of algebra. (Hint: let $\rho$ be a root and use the division algorithm to write $f(x) = (x - \rho)q(x) + r$ where $r$ has degree 0.)

Exercise 3.14. Let $f$ and $g$ be polynomials in $\mathbb{Q}[x]$ with root $\rho$ and suppose that $g$ is minimal (Definition 1.13). Show that $g \mid f$. (Hint: use the division algorithm of exercises 2.2 and 2.3 to write $f(x) = g(x)q(x) + r(x)$ where $r$ has degree less than $g$.)

Definition 3.14. 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 3.15. Denote the golden mean, or $\frac{1 + \sqrt{5}}{2} \approx 1.618$, by $g$.

a) Show that $g^2 = g + 1$ and thus $g^{n+1} = g^n + g^{n-1}$.

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 3.16. 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 3.17. Use exercises 3.15 and 3.16 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 3.18. 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:
\[
\begin{align*}
287 &= 28 \cdot 10 + 7 \\
28 &= 2 \cdot 10 + 8 \\
2 &= 0 \cdot 10 + 2 \\
\end{align*}
\]
Hence the number in base 10 is \(2 \cdot 10^2 + 8 \cdot 10^1 + 7 \cdot 10^0\).
b) Show that to write \(n\) in base \(b\) takes about \(\log_b n\) divisions.

Exercise 3.19. Suppose \(\{b_i\}_{i=1}^n\) are positive integers such that \(\gcd(b_j,b_i) = 1\) for \(i \neq j\). We want to know all \(z\) that satisfy
\[
z = b_i c_i \quad \text{for} \quad i \in \{1, \ldots, n\}.
\]
a) Set \(B = \prod_{i=1}^n b_i\) and show that the homogeneous problem is solved by
\[
z = B \cdot 0.
\]
(Hint: use Corollary 3.10).
c) Check that
\[
z = \sum_{i=1}^n \frac{B}{b_i} x_i c_i
\]
is a particular solution.
c) Find the general solution:
\[
z = B \sum_{i=1}^n \frac{B}{b_i} x_i c_i.
\]
(This is known as the Chinese remainder theorem.)

Exercise 3.20. Use exercise 3.19 to solve:
\[
z = 2 \quad 1 \\
z = 3 \quad 2 \\
z = 5 \quad 3 \\
z = 7 \quad 5.
\]

Exercise 3.21. Assume \(\gcd(F_n,F_{n+1}) = 1\). Use exercise 3.19 to solve:
\[
z = F_n \quad F_{n-1} \\
z = F_{n+1} \quad F_n.
\]
where \(F_n\) are the Fibonacci numbers of Definition 3.14.
3.6. Exercises

Exercise 3.22. (The Chinese remainder theorem generalized.) Suppose \( \{b_i\}_{i=1}^n \) are positive integers. We want to know all \( z \) that satisfy

\[ z = b_i \cdot c_i \quad \text{for} \quad i \in \{1, \ldots, n\}. \]

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

\[ z = B \cdot 0. \]

b) Show that if there is a particular solution then

\[ \forall i \neq j : c_i = \gcd(b_i, b_j) \cdot c_j. \]

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

Exercise 3.23. Assume \( \gcd(F_n, F_{n+1}) = 1 \). Use exercise 3.21 to solve:

\[
\begin{align*}
z & =_6 15 \\
z & =_{10} 6 \\
z & =_{15} 10.
\end{align*}
\]

See also exercise 2.6.
Chapter 4

Number Theoretic Functions

Overview. We study number theoretic functions. These are functions defined on the positive integers with values in \( \mathbb{C} \). Generally, in the context of number theory, the value 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 or arithmetic functions or sequences are functions defined on the positive integers (i.e. \( \mathbb{N} \)) with values in \( \mathbb{C} \).

Note that in other areas of mathematics, the word sequence is the only term 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,

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

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

$$F(n) = \sum_{a_1}^{\ell_1} \cdots \sum_{a_r}^{\ell_r} f(p_1^{a_1}) \cdots f(p_r^{a_r})$$

$$= \prod_{i=1}^{r} \left( \sum_{a_i=0}^{\ell_i} f(p_i^{a_i}) \right).$$

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:

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

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.

**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 $\sigma_k$ 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 $\tau$, 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 $p$:

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

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:

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

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

**Lemma 4.8.** The Möbius function $\mu$ is multiplicative.

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

$$
n = ab \quad \text{where} \quad a = \prod_{i=1}^{r} p_i^{\ell_i} \quad \text{and} \quad b = \prod_{i=r+1}^{s} p_i^{\ell_i}.
$$
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 $\mu$) is strictly additive and so $(-1)^r$ is strictly multiplicative, hence multiplicative.

**Definition 4.9.** Euler's phi function, also called Euler's totient function is defined as follows: $\varphi(n)$ equals the number of integers in $\{1, \cdots, n\}$ that are relative prime to $n$.

### 4.2. Additive Functions

Also important are the additive functions to which we will return in Chapter 11.

**Definition 4.10.** An additive function is a sequence such that $\gcd(a, b) = 1$ implies $f(ab) = f(a) + f(b)$. A completely additive function is one where the condition that $\gcd(a, b) = 1$ is not needed.

Here are some examples.

**Definition 4.11.** Let $\omega(n)$ denote the number of distinct prime divisors of $n$ and let $\Omega(n)$ denote the number of prime powers that are divisors of $n$. These functions are called the prime omega functions.

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

$$\omega(n) = s \quad \text{and} \quad \Omega(n) = \sum_{i=1}^{2} \ell_i.$$

The additivity of $\omega$ and the complete additivity of $\Omega$ should be clear.

### 4.3. Möbius Inversion

**Lemma 4.12.** 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 $\mu$ is multiplicative. Therefore, by Proposition 4.3, $\varepsilon$ is also multiplicative. It follows that $\varepsilon(\prod_{i=1}^{s} p_i^{\ell_i})$ can be calculated.
4.3. Möbius inversion

by evaluating a product of terms like \( \varepsilon(p^\ell) \) where \( p \) is prime. For example, when \( p \) is prime, we have

\[
\varepsilon(p) = \mu(1) + \mu(p) = 1 + (-1) = 0 \quad \text{and} \\
\varepsilon(p^2) = \mu(1) + \mu(p) + \mu(p^2) = 1 - 1 + 0 = 0.
\]

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

Lemma 4.13. For \( n \in \mathbb{N} \), define

\[
S_n = \left\{ (a, b) \in \mathbb{N}^2 \middle| \exists d > 0 \text{ such that } d \mid n \text{ and } ab = d \right\} \quad \text{and} \\
T_n = \left\{ (a, b) \in \mathbb{N}^2 \middle| b \mid n \text{ and } a \mid \frac{n}{b} \right\}.
\]

Then \( S_n = T_n \).

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

\[
\frac{ab = d}{d \mid n} \implies b \mid n \text{ and } a \mid \frac{n}{b}.
\]

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

\[
\frac{b \mid n}{a \mid \frac{n}{b}} \implies d \mid n \text{ and } ab = d.
\]

And so \((a, b)\) is in \( S_n \). ■

Theorem 4.14. (Möbius inversion) Let \( F : \mathbb{N} \to \mathbb{C} \) be any number theoretic function. Then the equation

\[
F(n) = \sum_{d \mid n} f(d)
\]

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

\[
f(d) = \sum_{a \mid d} \mu(a) F \left( \frac{d}{a} \right) = \sum_{\{a, b\mid ab = d\}} \mu(a) F(b).
\]

Proof. \( \iff \): We show that substituting \( f \) gives \( F \). Define \( H \) as

\[
H(n) \equiv \sum_{d \mid n} f(d) = \sum_{d \mid n} \sum_{a \mid d} \mu(a) F \left( \frac{d}{a} \right).
\]
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
\[
H(n) = \sum_{d|n} f(d) = \sum_{d|n} \sum_{ab=d} \mu(a) F(b)
\]
(4.1)
For the second step we apply Lemma 4.13 to the set over which the summation takes place. This gives:
\[
H(n) = \sum_{b|n} \left( \sum_{a|\left(\frac{n}{b}\right)} \mu(a) \right) F(b)
\]
(4.2)
Finally, Lemma 4.12 implies that the term in parentheses equals $G\left(\frac{n}{b}\right)$. It equals 0, except when $b = n$ when it equals 1. The result follows.

**Uniqueness:** 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 $i \in \{1, \ldots, k\}$, we have $f(i) = g(i)$. Then
\[
F(k+1) = \left( \sum_{d|\left(k+1\right), d\leq k} f(d) \right) + f(k+1) = \left( \sum_{d|\left(k+1\right), d\leq k} g(d) \right) + g(k+1).
\]
So that the desired equality for $k+1$ follows from the induction hypothesis.

\[
\blacksquare\]

### 4.4. Euler’s Phi or Totient Function

Recall the phi function from Definition 4.9.

**Lemma 4.15. (Gauss)** For $n \in \mathbb{N}$: $n = \sum_{d|n} \phi(d)$.

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

\[
S(d, n) = \{ m \in \mathbb{N} \mid m \leq n \text{ and } \gcd(m, n) = d \}
\]
This is equivalent to
\[
S(d, n) = \{ m \in \mathbb{N} \mid m \leq n \text{ and } \gcd\left(\frac{m}{d}, \frac{n}{d}\right) = 1 \}
\]
4.4. Euler’s Phi or Totient Function

From the definition of Euler’s phi function, we see that the cardinality $|S(d,n)|$ of $S(d,n)$ is given by $\phi \left( \frac{n}{d} \right)$. Thus we obtain:

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

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

**Theorem 4.16.** Let $\prod_{i=1}^{r} p_i^{\ell_i}$ be the prime power factorization of $n$. Then $\phi(n) = n \prod_{i=1}^{r} \left( 1 - \frac{1}{p_i} \right)$.

**Proof.** Apply Möbius inversion to Lemma 4.15:

(4.3)

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

The functions $\mu$ and $d \to \frac{1}{d}$ are multiplicative. It is easy to see that the product of two multiplicative functions is also multiplicative. Therefore $\phi$ is also multiplicative (Proposition 4.3). Thus

(4.4)

$$ \phi \left( \prod_{i=1}^{r} p_i^{\ell_i} \right) = \prod_{i=1}^{r} \phi \left( p_i^{\ell_i} \right). $$

So it is sufficient to evaluate the function $\phi$ on prime powers. Noting that the divisors of the prime power $p^\ell$ are $\{1, p, \ldots, p^\ell\}$, we get from Equation 4.3

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

Substituting this into Equation 4.4 completes the proof. ■

From this proof we obtain the following corollary.

**Corollary 4.17.** Euler’s phi function is multiplicative.

---

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 4.11.
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 4.18.** Let $f$, $g$, and $F$ be arithmetic functions (see Definition 4.1). Define the Dirichlet convolution of $f$ and $g$, denoted by $f \ast g$, as

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

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

The function $\varepsilon$ acts as the identity of the convolution. Indeed,

$$((\varepsilon \ast f))(n) = \sum_{ab=n} \varepsilon(a)f(b) = f(n).$$

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

We can now do some very cool things of which we can unfortunately give but a few examples. As a first example, we reformulate the Möbius inversion of Theorem 4.14 as follows:

$$(4.5) \quad F = 1 \ast f \quad \iff \quad f = \mu \ast F.$$  

This leads to the next example. The first of the following equalities holds by Lemma 4.15, the second follows from Möbius inversion.

$$(4.6) \quad I = 1 \ast \varphi \quad \iff \quad \varphi = \mu \ast I.$$  

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

$$(4.7) \quad \varepsilon = 1 \ast f \quad \iff \quad f = \mu \ast \varepsilon = \mu.$$  

---

2A very unusual word in mathematics textbooks.
4.5. Dirichlet and Lambert Series

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

**Definition 4.20.** Let $f(n)$ be an arithmetic function (or sequence). A **Dirichlet series** is a series of the form $F(s) = \sum_{n=1}^{\infty} f(n)n^{-s}$. Similarly, a **Lambert series** is a series of the form $F(x) = \sum_{n=1}^{\infty} f(n)x^n 1 - x^n$.

The prime example of a Dirichlet series is — of course — the Riemann zeta function of Definition 2.20, $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$.

**Lemma 4.21.** For the product of two Dirichlet series we have

\[
\left( \sum_{n=1}^{\infty} f(n)n^{-s} \right) \left( \sum_{n=1}^{\infty} g(n)n^{-s} \right) = \sum_{n=1}^{\infty} (f * g)(n)n^{-s}.
\]

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

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

We collected the terms with $ab = n$.

Since $\frac{\zeta(s)}{\zeta(s)} = 1$, equation (4.7) and the lemma immediately imply that

\[
(4.8) \quad \frac{1}{\zeta(s)} = \sum_{n \geq 1}^{\infty} \frac{\mu(n)}{n^s}.
\]

Recall from Chapter 2 that one of the chief concerns of number theory is the location of the non-real zeros of $\zeta$. At stake is Conjecture 2.23 which states that all its non-real zeros are on the line $\text{Re } s = 1/2$. The original definition of the zeta function is as a series that is absolutely convergent for $\text{Re } s > 1$ only. Thus it is important to establish that the analytic continuation of $\zeta$ is valid for $\text{Re } s \leq 1$. The next result serves as a first indication that $\zeta(s)$ can indeed be continued for values $\text{Re } s \leq 1$.

**Corollary 4.22.** Let $\zeta$ be the Riemann zeta function and $\sigma_k$ as in Definition 4.4, then

\[
\zeta(s - k)\zeta(s) = \sum_{n=1}^{\infty} \frac{\sigma_k(n)}{n^s}.
\]

---

\(^3\)The fact that this follows so easily justifies the use of the word referred to in the previous footnote.
Proof. 
\[ \zeta(s-k)\zeta(s) = \sum_{a \geq 1} a^{-s} \sum_{b \geq 1} b^k b^{-s} = \sum_{n \geq 1} \sum_{b|n} b^k. \]

Lemma 4.23. A Lambert series can re-summed as follows:
\[ \sum_{n=1}^{\infty} f(n) \frac{x^n}{1-x^n} = \sum_{n=1}^{\infty} (1*f)(n)x^n. \]

Proof. First use that 
\[ \frac{x^b}{1-x^b} = \sum_{a=1}^{\infty} x^{ab}. \]
This gives that 
\[ \sum_{b=1}^{\infty} f(b) \frac{x^b}{1-x^b} = \sum_{b=1}^{\infty} \sum_{a=1}^{\infty} f(b)x^{ab}. \]
Now set \( n = ab \) and collect terms. Noting that \( (1*f)(b) = \sum_{b|n} f(b) \) yields the result.

Corollary 4.24. The following equality holds
\[ \sum_{n \geq 1} \varphi(n) \frac{x^n}{1-x^n} = \frac{x}{(1-x)^2}. \]

Proof. We have
\[ \sum_{n \geq 1} \varphi(n) \frac{x^n}{1-x^n} = \sum_{n \geq 1} (1*\varphi)(n)x^n = \sum_{n \geq 1} I(n)x^n. \]
The first equality follows from Lemma 4.23 and the second from Lemma 4.15. The last sum can be computed as \( x \frac{d}{dx} (1-x)^{-1} \) which gives the desired expression.

The last result is of importance in the study of dynamical systems. In figure 8, the map \( f_t \) is constructed by truncating the map \( x \to 2x \mod 1 \) for \( t \in [0,1] \). Corollary 4.24 can be used to show that the set of \( t \) for which \( f_t \) does not have a periodic orbit has measure (“length”) zero [34, 35], 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 \( a = 2^k \prod_{i=1}^{m} p_i^{e_i} \) by unique factorization, where the \( p_i \) are odd primes. Similarly for \( b \).)
c) What does Proposition 4.3 say?

Exercise 4.3. a) Compute the numbers \( \sigma_1(n) = \sigma(n) \) of Definition 4.4 for \( n \in \{1, \ldots, 30\} \) 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 \( \sigma \) to check the list obtained in (a).
e) For what values of \( n \) in the list of (a) is \( n | \sigma(n) \)? (Hint: 6 and 28.)
Exercise 4.4. a) Compute the numbers $\sigma_0(n) = \tau(n)$ of Definition 4.4 for $n \in \{1, \ldots, 30\}$ 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 $\tau$ to check the list obtained in (a).

Exercise 4.5. a) Compute the numbers $\varphi(n)$ of Definition 4.9 for $n \in \{1, \ldots, 30\}$ without using Theorem 4.16.
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 $\varphi$ to check the list obtained in (a).

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

Exercise 4.7. Let $\tau(n)$ be the number of distinct positive divisors of $n$. Answer the following question without using Theorem 4.5.
a) Show that $\tau$ 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 $n \in \mathbb{N}$.

Exercise 4.8. Two positive integers $a$ and $b$ are called amicable if $\sigma(a) = \sigma(b) = a + b$. The smallest pair of amicable numbers 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 4.9. 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 4.3(c).)
c) Use exercise 1.14 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 4.3.
Exercise 4.10. 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, $\sigma(b) - b = a$ and $\sigma(a) - a = b$.\(^a\)

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.
4. Number Theoretic Functions

Exercise 4.11. In this exercise, we give a different proof of Theorem 4.16. It uses the principle of inclusion-exclusion [29]. 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 $|A|$, we have

\begin{equation}
|S - \bigcup_{i=1}^r A_i| = |S| - |S_1| + |S_2| - \cdots + (-1)^r |S_r|,
\end{equation}

where $|S_i|$ is the sum of the sizes of all intersections of $\ell$ members of \{ $A_1, \cdots, A_r$ \}.

Now, in the following we keep to the following conventions. Using prime factorization, write

\[ n = \prod_{i=1}^r p_i^{k_i}, \]

\[ A_i = \{ z \in S \mid p_i \text{ divides } z \}. \]

\[ S = \{ 1, 2, \cdots, n \} \quad \text{and} \quad R = \{ 1, 2, \cdots, r \}, \]

\[ \ell \subseteq R \quad \text{such that} \quad |\ell| = \ell. \]

a) Show that $\varphi(\ell) = |S - \bigcup_{i=1}^r A_i|$. (Hint: any number that is not co-prime with $n$ is a multiple of at least one of the $p_i$.)

b) Show that $|A_i| = \frac{n}{p_i}$.

c) Show that $|\bigcap_{i \in \ell} A_i| = n \prod_{i \in \ell} \frac{1}{p_i}$. (Hint: use Corollary 3.7.)

d) Show that $|S_\ell| = n \sum_{\ell \subseteq R} \prod_{i \in \ell} \frac{1}{p_i}$.

e) Show that the principle of inclusion-exclusion implies that $|S - \bigcup_{i=1}^r A_i| = n + n \sum_{\ell \subseteq R} (-1)^\ell \prod_{i \in \ell} \frac{1}{p_i}$.

f) Show that $n + n \sum_{\ell = 1}^r (-1)^\ell \prod_{i \in \ell} \frac{1}{p_i} = n \prod_{i=1}^r (1 - \frac{1}{p_i})$. Notice that this implies Theorem 4.16. (Hint: write out the product $\prod_{i=1}^r (1 - \frac{1}{p_i})$.)

Exercise 4.12. Let $F(n) = n = \sum_{d|n} f(n)$. Use the Möbius inversion formula (or $f(n) = \sum_{d|n} \mu(d) \varphi(\frac{n}{d})$) to find $f(n)$. (Hint: substitute the Möbius function of Definition 4.6 and use multiplicativity where needed.)

Exercise 4.13. a) Compute the sets $S_n$ and $T_n$ of Lemma 4.13 explicitly for $n = 4$ and $n = 12$.

b) Perform the resummation done in equations 4.1 and 4.2 explicitly for $n = 4$ and $n = 12$. 

4.6. Exercises

Exercise 4.14. Recall the definition of Dirichlet convolution \( f \ast g \) of the arithmetic functions \( f \) and \( g \) (Definition 4.18).

a) Show that Dirichlet convolution is commutative, that is:
\[
 f \ast g = g \ast f .
\]

b) Show that Dirichlet convolution is associative, that is:
\[
 (f \ast g) \ast h = f \ast (g \ast h) .
\]

c) Show that Dirichlet convolution is distributive, that is:
\[
 f \ast (g + h) = f \ast g + f \ast h .
\]

d) The binary operation Dirichlet convolution has an identity \( \varepsilon \), defined by
\[
 f \ast \varepsilon = \varepsilon \ast f = f .
\]
Show that the function \( \varepsilon \) of Lemma 4.12 is the identity of the convolution.

Exercise 4.15. Use exercise 4.14 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 4.16. See Definition 4.11. 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 \( 10^6 \) 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 \( \tau \) is multiplicative.)

d) Use (d) to show that \( g \) is multiplicative.

e) Show that
\[
 \tau(n^2) = \sum_{d|n} 2^{\omega(d)} .
\]

Exercise 4.17. 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 4.11.

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 4.18. Define the Liouville $\lambda$-function by $\lambda(1) = 1$ and $\lambda(n) = (-1)^{\Omega(n)}$.

a) Compute $\lambda(10^n)$ and $\lambda(6!)$.  
b) Show that $\lambda$ 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 4.19. Let $f$ be a multiplicative function.

Define $q(n) = \sum_{d|n} \mu(d)f(d)$, where $\mu$ 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 4.20. Use exercise 4.19 (e) and the definition of $\omega$ in exercise 4.16 and $\lambda$ in exercise 4.18 to show that

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

Exercise 4.21. 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 4.22. a) Use Euler’s product formula and the sequence $\mu$ of Definition 4.6 to show that

$$\frac{1}{\zeta(s)} = \prod_{p \text{ prime}} \left(1 - p^{-s}\right) = \prod_{p \text{ prime}} \left(\sum_{i=0}^{\infty} \mu(p^i)p^{-is}\right) .$$

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

Exercise 4.23. a) Use equation (4.8) to show that
\[ \frac{\zeta(s-1)}{\zeta(s)} = \sum_{a \geq 1} a \sum_{b \geq 1} \frac{\mu(b)}{b^s}. \]
b) Use Lemma 4.23 and the first equality of equation (4.3) to show that
\[ \frac{\zeta(s-1)}{\zeta(s)} = \sum_{n \geq 1} \frac{\varphi(n)}{n^s}. \]

Exercise 4.24. a) Use Corollary 4.22 to show that
\[ \zeta(s-k) = \sum_{a \geq 1} \frac{\sigma_k(a)}{a^s} \sum_{b \geq 1} \frac{\mu(b)}{b^s}. \]
b) Show that
\[ \zeta(s-k) = \sum_{n \geq 1} (\sigma_k * \mu)(n) n^{-s}, \]
where * means the Dirichlet convolution (Definition 4.18).

Exercise 4.25. Show that \( \zeta(s) \) has no zeroes and no poles in the region \( \Re(s) > 1 \). (Hint: use that \( \zeta(s) \) converges for \( \Re(s) > 1 \) 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 is no \( 0 < b < a \) so that \( a \equiv b \) 0.

5.1. Modular Arithmetic

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 1.6). 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.

Suppose we are given an operation \( * \) on a set \( G \) and that \( (G, *) \) satisfies certain requirements — namely that it forms a group — that include having an identity element \( e \). The order of an element \( g \) is the smallest positive integer \( k \) such that \( g \ast g \cdots g \), repeated \( k \) times and usually written as \( g^k \), equals \( e \). One can show that the elements \( \{e, g, g^2, g^{k-1}\} \) also form a group. To follow this up would take us too far afield. A few details are given in Chapter 7. The full story can be found in [13], [14], or [17].
In the case at hand, \( \mathbb{Z}_b \), where 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 \( \mathbb{Z}_7 \):

\[
3 + 3 + 3 + 3 + 3 + 3 + 3 = 7 \quad \text{and} \quad 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 = 71.
\]

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

**Definition 5.1.** The order of a modulo \( b \), written as \( \text{Ord}^\times_b(a) \), is the smallest positive number \( k \) such that \( a^k \equiv 1 \pmod{b} \). (If there is no such \( k \), the order is \( \infty \).)

**5.2. Euler’s Theorem**

**Definition 5.2.**
1. A complete set of residues \( C \) modulo \( b \) is a set \( b \) integers in \( \mathbb{Z} \), such \( C \) has exactly one integer in each congruence class (modulo \( b \)).
2. A reduced set of residues \( R \) modulo \( b \) is a set \( \varphi(b) \) integers in \( \mathbb{Z} \), such \( R \) has exactly one integer in each class congruent to \( p \in \{1, \cdots, b-1\} \) (modulo \( b \)) such that \( p \) is relatively prime to \( b \).

**Lemma 5.3.** Suppose \( \gcd(a, b) = 1 \). If the numbers \( \{x_i\} \) form a complete set of residues modulo \( b \) (a reduced set of residues modulo \( b \)), then \( \{ax_i\} \) is a complete set of residues modulo \( b \) (a 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 2.8.

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 2.19 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 \( a \) and \( b \) greater than 0 and \( \gcd(a, b) = 1 \). Then \( a^{\varphi(b)} \equiv 1 \pmod{b} \), where \( \varphi \) denotes Euler’s phi or totient function (Definition 4.9).
Section 5.2. Euler’s Theorem

**Proof.** Let \( \{x_i\}_{i=1}^{\phi(b)} \) be a reduced set of residues modulo \( b \). Then by Lemma 5.3, \( \{ax_i\}_{i=1}^{\phi(b)} \) be a reduced set of residues modulo \( b \). Because multiplication is commutative, we get
\[
\prod_{i=1}^{\phi(b)} x_i = b \prod_{i=1}^{\phi(b)} ax_i = b^{\phi(b)} \prod_{i=1}^{\phi(b)} x_i
\]
Since \( \gcd(x_i, a) = 1 \), Lemma 2.19 implies that \( \gcd(\prod_{i=1}^{\phi(b)} x_i, a) = 1 \). The cancelation theorem applied to the equality between the first and third terms proves the result. 

Euler’s theorem says that \( \phi(b) \) is a multiple of \( \text{Ord}_b^\times(a) \). 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 \( \text{Ord}_b^\times(a) = \phi(b) \), then \( a \) is called a primitive root of \( b \) (or a primitive root modulo \( b \)).

For example, the smallest integer \( k \) for which \( 3^k \equiv 1 \pmod{7} \) is 6. Since \( \phi(7) = 6 \), we see that 3 is a primitive root of 7. Since multiplication is well-defined in \( \mathbb{Z}_7 \), it follows that \( (3 + 7k)^6 \equiv 1 \pmod{7} \). 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.

This has interesting connections with day-to-day arithmetic, namely the expression of rational numbers in base 10.

**Proposition 5.6.** Let \( a \) and \( n \) greater than \( 0 \) and \( \gcd(a, n) = \gcd(10, n) = 1 \). The decimal expansion of \( \frac{a}{n} \) is non-terminating and eventually periodic with a period \( p \), where (1) \( p = \text{Ord}_n^\times(10) \) and (2) \( p | \phi(n) \).

**Proof.** The proof proceeds by executing a long division, each step 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 \( \frac{a}{n} \) is \( q_0 \). The next step of the long division is:
\[
10r_0 = nq_1 + r_1,
\]
where again we choose \( r_1 \in \{0, \cdots n-1\} \).
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, we have $\gcd(10r_0, n) = \gcd(r_1, n)$. In turn, this implies via Lemma 2.19 that $\gcd(r_0, n) = \gcd(r_1, n)$. And again, we see that $r_1 \neq 0$.

The process now repeats itself.

$$10 \left(10r_0 - nq_1\right) = nq_2 + r_2,$$

and we record the second digit after the decimal point, $q_2 \in \{0, \cdots 9\}$. 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

$$10^{p+r_0} = n 10^{r_0}.$$

By Theorem 2.8, 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 part (1). Part (2) follows directly from Euler’s Theorem.

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

The next result is an immediate consequence of Euler’s theorem. It follows by setting $y = x + k\varphi(b)$. It has important applications in cryptography.

**Corollary 5.7.** Let $a$ and $b$ be coprime with $b \geq 0$. If $x = \varphi(b) y$, then $a^x =_b a^y$.

### 5.3. 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 “going on too long”. Not an isolated case, it would appear!

**Corollary 5.8 (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 $a^p \equiv a \pmod{p}$. 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 \( \sqrt{n} \). By the prime number theorem (Theorem 2.22), this will give on the order of $\frac{2\sqrt{n}}{\ln n}$ primes. Then divide $n$ by all these numbers. This will take on the order of $\frac{2\sqrt{n}}{\ln n}$ divisions.

Another possibility is to use the converse of Fermat’s little theorem (Corollary 5.8). If $n$ and $p$ are distinct primes, we know that $p^{n-1} \equiv 1 \pmod{n}$. The Fermat primality test for $n$ consists of testing for example whether $2^{n-1} \equiv 1 \pmod{n}$. However, the converse of Fermat’s little theorem is not true! So even if $2^{n-1} \equiv 1 \pmod{n}$, 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 3.18, where it was shown that this takes on the order of $\log_2 340$ (long) divisions.

\[
\begin{align*}
340 & = 170 \cdot 2 + 0 \\
170 & = 85 \cdot 2 + 0 \\
85 & = 42 \cdot 2 + 1 \\
42 & = 21 \cdot 2 + 0 \\
21 & = 10 \cdot 2 + 1 \\
10 & = 5 \cdot 2 + 0 \\
5 & = 2 \cdot 2 + 1 \\
2 & = 1 \cdot 2 + 0 \\
1 & = 0 \cdot 2 + 1
\end{align*}
\]

And so

\[340 = 101010100 \pmod{2} \]

Next, compute a table of powers $11^2$ modulo 341, as done below. By doing this from the bottom up, this can be done using very few computations. For
instance, once $11^{18}_{=341}$ 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.

\[
\begin{array}{c|c|c}
1 & 11^{266}_{=341} = 340 & 330 \\
0 & 11^{128}_{=341} = 143 & 143 \\
1 & 11^{64}_{=341} = 319 & 121 \\
0 & 11^{32}_{=341} = 121 & 121 \\
1 & 11^{16}_{=341} = 330 & 330 \\
0 & 11^8_{=341} = 143 & 143 \\
1 & 11^4_{=341} = 319 & 319 \\
0 & 11^2_{=341} = 121 & 121 \\
0 & 11^1_{=341} = 11 & 11 \\
\end{array}
\]

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

$$11^{340}_{=341} = 330 \cdot 319 \cdot 330 \cdot 319 = 341 \cdot 132.$$ Again, this takes no more than $\log_2 340$ multiplications. Thus altogether, for a number $x$ and a computation in base $b$, this takes on the order of $2\log_b x$ multiplications plus $\log_b x$ divisions$^1$. For large numbers, this is much more efficient than the $\sqrt{x}$ of the naive method.

The drawback is that we can get false positives.

**Definition 5.9.** The number $n \in \mathbb{N}$ is called a pseudoprime to the base $a$ if $\gcd(a, n) = 1$ and $a^{n-1} = n$ 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 [27] that there are infinitely many of them.

**Definition 5.10.** The number $n \in \mathbb{N}$ is called a Carmichael number if it is composite and it is a pseudoprime to every base.

---

$^1$Divisions take more computations than multiplications. We do not pursue this here.
5.4. Fermat and Mersenne Primes

The smallest pseudoprime is 341, because \(2^{340} = 341 \cdot 1\) while 341 = 11 \cdot 31. In this case, one can still show that 341 is not a prime by using a different base: \(3^{340} = 341 \cdot 56\). 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.

5.4. 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 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 5.11.** (1) The Mersenne numbers are \(M_k = 2^k - 1\). A Mersenne prime is a Mersenne number that is also prime.

(2) The Fermat numbers are \(F_k = 2^{2^k} + 1\). A Fermat prime is a Fermat number that is also prime.

(3) The number \(n \in \mathbb{N}\) is called a perfect, if \(\sigma(n) = 2n\).

**Lemma 5.12.** (1) If \(a \cdot d = k\), then \((2^d - 1) | (2^k - 1)\).

(2) If \(a \cdot d = k\) and \(a\) is odd, then \((2^d + 1) | (2^k + 1)\).

**Proof.** We prove only (2). The other statement can be proved similarly. So suppose that \(a\) is odd, then

\[2^d = 2^e + 1 \rightarrow 2^{ad} = 2^e + 1 \rightarrow 2^{ad} \cdot (-1)^a = 2^e + 1 \rightarrow 2^{ad} + 1 = 2^e + 1 \cdot 0\]

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

A proof using the geometric series can be found in exercise 1.14. This lemma immediately implies the following.

**Corollary 5.13.** a) If \(2^k - 1\) is prime, then \(k\) is prime.

b) If \(2^k + 1\) is prime, then \(k = 2^e\).
So candidates for Mersenne primes are the numbers $2^p - 1$ where $p$ is prime. This works for $p \in \{0, 1, \cdots, 10\}$, but $2^{11} - 1$ is the monkey-wrench. It is equal to $23 \cdot 89$ and thus is composite. After that, the Mersenne primes become increasingly sparse. Among the first approximately 2.3 million primes, only 45 give Mersenne primes. As of this writing (2017), it is not known whether there are infinitely many Mersenne primes. In 2016, a very large Mersenne prime was discovered: $2^{74,207,281} - 1$. 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 5.14.** 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 $\sigma$ 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 5.15 (Euler’s Theorem).** Suppose $n$ is even. Then $n$ is perfect if and only if $n$ is of the form $2^{k-1}(2^k - 1)$ where $2^k - 1$ is prime.

**Proof.** One direction follows from the previous lemma. We only need to prove that if an even number $n = q2^{k-1}$ where $k \geq 2$, is perfect, then it is of the form stipulated.

Since $n$ is even, we may assume $n = q2^{k-1}$ where $k \geq 2$ and $2 \nmid q$. Using multiplicativity of $\sigma$ and the fact that $\sigma(n) = 2n$:

$$
\sigma(n) = \sigma(q)(2^k - 1) = q2^k.
$$

Thus

$$
(2^k - 1)\sigma(q) - 2^k q = 0.
$$
5.5. Division in $\mathbb{Z}_b$

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:

\[(5.2) \quad q = (2^k - 1)t \quad \text{and} \quad \sigma(q) = 2^k t,\]

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 $n = q2^{k-1} = (2^k - 1)2^{k-1}$ 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. $\blacksquare$

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

5.5. Division in $\mathbb{Z}_b$

The next result is a game changer! It tells us that there is a unique element $a^{-1}$ such that $aa^{-1} = b$ if and only if $a$ is in the reduced set of residues (modulo $b$). Thus division is well-defined in the reduced set of residues modulo $b$. This serves as a clear signal to emphasize two distinct algebraic structures. A ring is a structure with addition and its inverse subtraction plus multiplication, but where multiplication may not have an inverse. A field is almost the same, but now multiplication always has an inverse: division. A more detailed description of these algebraic construction is given in Section 7.1. The numbers 1 and $-1$ are always in the reduced set of residues modulo $b$. This set is sometimes called the set of units (see Definition 7.6) of $\mathbb{Z}_b$.

**Proposition 5.16.** Let $R$ be a reduced set of residues modulo $b$. Then

1. for every $a \in R$, there is a unique $a'$ in $R$ such that $a' = b$, $a'a' = b$ 1,
2. for every $a \not\in R$, there exists no $x \in \mathbb{Z}_b$ such that $ax = b$ 1,
3. let $R = \{x_i\}_{i=1}^{\varphi(b)}$, then also $R = \{x_i^{-1}\}_{i=1}^{\varphi(b)}$.

**Proof.** **Statement 1:** The existence of a solution follows immediately from Bézout’s Lemma, namely $a' = b x$ 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 2.8 to the difference of these equations.

**Statement 2:** By hypothesis, gcd\( (a,b) > 1 \). We have that \( ax =_b 1 \) is equivalent to \( ax + by = 1 \), which contradicts Bézout’s lemma.

**Statement 3:** 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 cancelation. ■

**Lemma 5.17.** Let \( p \) be prime. Then \( a^2 =_p 1 \) 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 2.13 says that 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 5.18 (Wilson’s theorem).** If \( p \) prime in \( \mathbb{Z} \), then \( (p - 1)! =_p -1 \).

If \( b \) is composite, then \( (b - 1)! \neq_b -1 \).

**Proof.** This is true for \( p = 2 \). If \( p > 2 \), then Proposition 5.16(3) and Lemma 5.17 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 \( \pm 1 \) with its inverse. 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 \( b = a^2 \). But then still gcd\( ((b - 1)!), b \geq a \).

By Bézout, we must have \((b - 1)! \mod b \) must be a multiple of \( a \). ■

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 \((k - 1)! =_k 0\) to determine whether \( k \) is prime, it takes only 1 multiplication and 1 division to determine whether \( k + 1 \) is prime.

Here is the take-away that will be important for Chapter 7. More in particular, we have the following result.

**Corollary 5.19.** Let \( p \) be prime.

1. For every \( a \in \mathbb{Z}_p \) there is a unique \( a' =_p -a \) such that \( a + a' =_p 0 \).
2. For every \( a \in \mathbb{Z}_p \) and \( a \neq 0 \), there is a unique \( a' = a^{-1} \) so that \( aa' =_p 1 \).

Addition and multiplication are well-defined in \( \mathbb{Z}_b \) (see exercises 5.1 and 5.2). Thus when \( p \) is prime, we can add, multiply, subtract, and divide in \( \mathbb{Z}_p \). In the words of Chapter 7, when \( p \) is a prime, then \( \mathbb{Z}_p \) is a field.

It is an interesting fact that the same is not true for a composite number \( b \). According to Proposition 5.16, we need the reduced set of residues for the multiplication to be invertible. At the same time, for the set to be closed under multiplication, we need all of \( \mathbb{Z}_b \) (think of \( 1 + 1 + \cdots \)). Thus the operations addition and multiplication in \( \mathbb{Z}_b \) collaborate with each other only if \( b \) is a 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 1.5 and 1.24.)

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 \( \text{Res}_m(a) = \text{Res}_m(b) \).

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. 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 \( Z \) modulo \( m \) by \( Z_m \) (see Definition 1.5). Prove that addition and multiplication are well-defined in \( Z_m \), using the following steps.

a) If \( a =_m a' \) and \( b =_m b' \), then \( \text{Res}_m(a) + \text{Res}_m(b) =_m \text{Res}_m(a') + \text{Res}_m(b') \). (Hint: show that \( a + b = c \) if and only if \( a + b =_m c \). In other words: the sum modulo \( m \) only depend on \( \text{Res}_m(a) \) and \( \text{Res}_m(b) \) 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 \( n = \sum_{i=1}^{k} a_i10^i \) where \( a_i \in \{0, 1, 2, \ldots, 9\} \).

a) Show that \( 10^k = 3 \) 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 \( n = \sum_{i=1}^{k} a_i10^i \) where \( a_i \in \{0, 1, 2, \ldots, 9\} \). 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 \( n = 5a_0 \).)

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) List \((n-1)! \mod n \) for \( n \in \{2, \ldots, 16\} \).

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!

Exercise 5.6. Denote by \( \{x_i\} \) the reduced set of residues modulo 16. If the reasoning similar to that of Wilson’s theorem holds, you expect that \( \prod x_i =_{16} 1 \).

a) Compute \( \prod x_i \).

b) How does the reasoning in Wilson’s theorem fail? (Hint: see lemma 5.17.)

c) List the product of the reduced set of residues modulo \( n \) for \( n \in \{2, \ldots, 16\} \).

Exercise 5.7. a) For \( i \) in \{1, 2, \ldots, 11\} and \( j \) in \{2, 3, \ldots, 11\}, make a table of \( \text{Ord}_j^i(i) \), \( i \) varying horizontally. After the \( j \)th 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, 0, 2, 3, 2\}.)
5.6. Exercises

Exercise 5.8. 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 5.6 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 5.9. a) Compute $2^n - 1 \mod n$ for $n$ odd in \{3, ..., 40\}.
b) Are there any pseudo-primes in the list?

Exercise 5.10. Assume that $n$ is a pseudoprime to the base 2.
a) Show that $2^n - 2 = n$. 0.
b) Show from (a) that $n | M_n - 1$. (See Definition 5.11.)
c) Use Lemma 5.12 to show that (b) implies that $M_n | 2^{M_n} - 1$.
d) Conclude from (c) that if $n$ is a pseudoprime in base 2, so is $M_n$.

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 $2^7 = 641 - 1$.
c) Show that $2^{32} + 1 = (2^7)^4 + 1 = 641$ 0.
d) Conclude that $F_5$ is divisible by 641.

Exercise 5.12. a) Compute $\varphi(100)$. (Hint: use Theorem 4.16.)
b) Show that $179^{121} = 100 \cdot 79^{121}$.
c) Show that $79^{121} = 100 \cdot 79^1$. (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 [33].

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 $1.14.1 \rightarrow 3.42.3 \rightarrow 3.16.3$. 
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 \( rp = q \). 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.

Exercise 5.18. a) Compute \( 7^2 \mod 13 \), using modular exponentiation.
b) Similarly for \( 484^{187} \mod 1189 \).
c) Find \( 100! + 102! \mod 101 \). (Hint: Fermat.)
d) Show that \( 1381! \equiv 1382 \mod 1189 \). (Hint: Wilson.)

Exercise 5.19. Characterize the set of \( n \geq 2 \) for which \( (n-1)! \mod n \) is
not in \( \{0, -1\} \). (Hint: Wilson.)

Theorem 5.20 (Binomial Theorem). If \( n \) is a positive integer, then
\[
(a + b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i} \text{ where } \binom{n}{i} = \frac{n!}{i! (n-i)!}.
\]
5.6. Exercises

Exercise 5.20. a) If \( p \) is prime, show that \( \binom{p}{i} \) mod \( p \) equals 0 if \( 1 \leq i \leq p - 1 \), and equals 1 if \( i = 0 \) or \( i = p \).

b) Evaluate \( \binom{4}{2} \) mod 4 and \( \binom{6}{4} \) mod 6. So where in (a) did you use the fact that \( p \) is prime?

c) Use the binomial theorem to show that if \( p \) is prime \( (a + b)^p \equiv a^p + b^p \) mod \( p \).

Exercise 5.21. Let \( p \) be prime.

a) Show that \( 0^p \equiv 0 \). 

b) Show that for \( k > 0 \), if \( k^p \equiv k \), then \( (k + 1)^p \equiv k + 1 \). (Hint: use exercise 5.20.)

(c) Conclude that for for all \( n \in \{0, 1, 2, \cdots \} \), \( n^p \equiv n \). (Hint: use induction.)

(d) Prove that for all \( n \in \mathbb{Z} \), \( n^p \equiv n \). (Hint: \( (-n)^p \equiv (-1)^p n^p \).)

(e) Use (d) to prove Fermat’s Little Theorem. (Hint: use cancellation.)

Exercises 3.19 and 3.22 show how to solve linear congruences generally. Quadratic congruences are much more complicated. As an example, we look at the equation \( x^2 \equiv p - 1 \) in the following exercise.

Exercise 5.22. a) Show that Fermat’s little theorem gives a solution of \( x^2 - 1 \equiv 0 \) whenever \( p \) is an odd prime. (Hint: consider \( a^{p-1} \).)

b) Show that Wilson’s theorem implies that for odd primes \( p \)

\[
(-1)^\frac{p-1}{2}\left[\binom{p-1}{2}\right]^2 \equiv -1. 
\]

(Hint: the right hand gives all reduced residues modulo \( p \).)

c) Show that if also \( p \equiv 1 \pmod{4} \) (examples are 13, 17, 29, etc), then

\[
\left[\binom{p-1}{2}\right]^2 \equiv -1. 
\]

satisfies the quadratic congruence \( x^2 + 1 \equiv 0 \).

d) Suppose \( x^2 + 1 \equiv 0 \) and \( p \) is an odd prime. Show that Fermat’s little theorem implies that

\[
(-1)^\frac{p-1}{2} = 1. 
\]

e) Show that \( p \) in (d) cannot satisfy \( p \equiv 3 \).

Exercise 5.23. Let \( R = \{1, 3, 5, 7, 9, 11, 13, 15\} \), the reduced set of residues modulo 16. In \( \mathbb{Z}_{16} \) and for each \( r \in R \), define multiplication by \( r^{-1} \) (or division by \( r \)).

Exercise 5.24. a) Draw up the addition and multiplication tables for \( \mathbb{Z}_2 \).

b) Do the same for \( \mathbb{Z}_8 \). Is division well-defined?

c) The same for the reduced set of residues plus 0 in \( \mathbb{Z}_8 \). What goes wrong here?
Exercise 5.25. Given \( n > 2 \), let \( R \subseteq \mathbb{Z}_n \) be the reduced set of residues and let \( S \subseteq \mathbb{Z}_n \) be the set of solutions in \( \mathbb{Z}_n \) of \( x^2 = n^1 \) (or self inverses).

a) Show that \( S \subseteq R \). (Hint: Bézout.)

b) Show that
\[
\prod_{x \in R} x = n \prod_{x \in S} x \quad (= n^1 \text{ if } S \text{ is empty}).
\]

c) Show that if \( S \) contains \( a \), then it contains \(-a\).

d) Show that if \( a = n - a \), then \( a \) and \(-a\) are not in \( S \).

e) Show that
\[
\prod_{x \in R} x = n (-1)^m \text{ some } m.
\]

f) Show that \( m = |S|/2 \).
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 [20].

6.1. The Gauss Map

Definition 6.1. The Gauss map is the transformation $T : [0, 1] \to [0, 1)$ defined by

$$T(\xi) = \frac{1}{\xi} - \left\lfloor \frac{1}{\xi} \right\rfloor = \left\{ \frac{1}{\xi} \right\}$$

and $T(0) = 0$, where we have used the notation of Definition 2.1.

Lemma 6.2. Set $q_i = \left\lfloor \frac{r_{i+1}}{r_i} \right\rfloor$ as in equation (3.1). Then the sequence $\{r_i\}$ defined by the Euclidean algorithm of Definition 3.3 satisfies:

$$\begin{align*}
\frac{r_{i+1}}{r_i} &= \frac{1}{r_{i-1}/r_i} - q_i = T \left( \frac{r_i}{r_{i-1}} \right) \quad \text{and} \\
\frac{r_i}{r_{i-1}} &= \frac{1}{q_i + \frac{r_{i+1}}{r_i}}
\end{align*}$$
6. Continued Fractions

Proof. From equation (3.1) or (3.4), we recall that \( r_{i+1} = r_i q_i + r_{i-1} \), or
\[
   r_{i+1} = r_{i-1} - q_i r_i
\]
where \( q_i = \left\lfloor \frac{r_{i-1}}{r_i} \right\rfloor \), and that \( \{r_i\} \) is a decreasing sequence. The first equation is obtained by dividing both sides by \( r_i \) and replacing \( \frac{r_i}{r_{i-1}} \) by the reciprocal of \( \frac{r_{i-1}}{r_i} \). The second equation of the lemma is obtained by inverting the first.

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. In the second equation of Lemma 6.2, write \( \omega_i = \frac{r_i}{r_{i+1}} \) 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 \( \omega \) 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
6.2. Continued Fractions

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:

\[
\begin{cases}
\omega_i = \frac{1}{\omega_{i-1} - a_{i-1}} = T(\omega_{i-1}) \quad \text{and} \\
\omega_{i-1} = \frac{1}{a_{i-1} + \omega_i}
\end{cases}
\] (6.1)

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 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 \( \omega_1 \) is rational. The \( \ell \)th branch of \( T \), depicted in Figure 9, has \( I_\ell = \left( \frac{1}{\ell+1}, \frac{1}{\ell} \right) \) 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 characterizes \( \omega_1 \).

For, in that case, we have

\[
\omega_1 = \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}}.
\] (6.2)

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  

\[
\omega_1 = \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}} \overset{\text{def}}{=} [a_1, a_2, a_3, \cdots].
\]

is called the continued fraction expansion of \( \omega_1 \). The finite truncations

\[
\frac{p_n}{q_n} = \frac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_n}}} \overset{\text{def}}{=} [a_1, a_2, \cdots, a_n].
\]

---

\(^1\)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 \( \omega_1 \), 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” know that it converges, or we would not bother with it.
are called the continued fraction approximants (or continued fraction convergents) of \(\omega\). The coefficients \(a_i\) are called the continued fraction coefficients.

Let us illustrate this definition with a few examples of continued fraction expansions:

- \(\pi - 3 = [7, 15, 1, 292, 1, 1, 2, \cdots]\),
- \(e - 2 = [1, 2, 1, 4, 1, 1, 6, 1, 1, 8, \cdots]\),
- \(\theta \equiv \sqrt{5} - 1 = [2, 2, 2, 2, \cdots]\),
- \(g \equiv \sqrt{5} - 1 = [1, 1, 1, 1, \cdots]\).

For example, \(\pi - 3\) the sequence of continued fraction approximants starts out as:

\[
\begin{align*}
1 &\quad 7 &\quad 15 &\quad 106 &\quad 16 &\quad 113 &\quad 4687 &\quad 33102 &\quad 4703 &\quad 33215 &\quad \cdots
\end{align*}
\]

The number \(g\) is also well-known. It is usually called the golden mean. Its continued fraction approximants are formed by the Fibonacci numbers defined in Definition 3.14 and given by \(\{1, 1, 2, 3, 5, 8, 13, 21, \cdots\}\). Namely, the approximants are \(1, 1, 2, 3, 5, 8, \cdots\).

We have defined continued fraction expansion only for numbers in \(\omega \in (0, 1)\). This can be easily be remedied by adding a “zeroth” digit \(a_0\) — signifying the floor of \(\omega\) — to it. Thus the expansion of \(\pi\) 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 \(\omega\) is rational.

**Proof.** If \(\omega\) 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 6.2, we see that \(\omega\) is rational. \(\blacksquare\)

**Theorem 6.6.** For the continued fraction approximants, we have

\[
\begin{align*}
p_n &= a_n p_{n-1} + p_{n-2} & q_0 &= 1, & p_0 &= 0, \\
q_n &= a_n q_{n-1} + q_{n-2} & q_{-1} &= 0, & p_{-1} &= 1,
\end{align*}
\]

or, in matrix notation,

\[
\begin{bmatrix} q_0 & p_0 \\ q_{-1} & p_{-1} \end{bmatrix} = A_n \begin{bmatrix} q_{n-1} & p_{n-1} \\ q_{n-2} & p_{n-2} \end{bmatrix} = A_n \cdots A_2 A_1,
\]
where

\[ A_i = \begin{pmatrix} a_i & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad A_i^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -a_i \end{pmatrix}. \]

**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 

\[ q_1 = a_1 \quad \text{and} \quad p_1 = 1 \]

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

\[ p_k = a_k p_{k-1} + p_{k-2}, \]

\[ q_k = a_k q_{k-1} + q_{k-2}. \]

The definition of the approximants gives:

\[ \frac{p_k}{q_k} = \frac{1}{a_1 + \cdots + \frac{1}{a_k}} \quad \text{and} \quad \frac{p_{k+1}}{q_{k+1}} = \frac{1}{a_1 + \cdots + \frac{1}{a_k + \frac{1}{a_{k+1}}}}. \]

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

\[ p_{k+1} = \left( a_k + \frac{1}{a_{k+1}} \right) p_{k-1} + p_{k-2}, \]

\[ q_{k+1} = \left( a_k + \frac{1}{a_{k+1}} \right) q_{k-1} + q_{k-2}. \]

Using equation (6.3) gives

\[ p_{k+1} = p_k + \frac{1}{a_{k+1}} p_{k-1}, \]

\[ q_{k+1} = q_k + \frac{1}{a_{k+1}} q_{k-1}. \]

The quotient \( \frac{p_{k+1}}{q_{k+1}} \) does not change if we multiply only the right hand side of these equations by \( a_{k+1} \) to insure that both \( p_{k+1} \) and \( p_{k+1} \) are integers. This gives the result. ■
Corollary 6.7. We have

\[(i)\] \[ q_n p_{n-1} - q_{n-1} p_n = (-1)^n \]

\[(ii)\] \[ \frac{P_{n-1}}{q_{n-1}} - \frac{P_n}{q_n} = \frac{(-1)^n}{q_{n-1} q_n} \]

Proof. The left hand side of the expression in (i) equals the determinant of

\[
\begin{pmatrix}
q_n & p_n \\
q_{n-1} & p_{n-1}
\end{pmatrix}
\]

which, by Theorem 6.6, must equal the determinant of

\[ A_n \cdots A_2 A_1. \]

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

\[(i)\] \[ p_n \geq 2 \frac{n-1}{2} \quad \text{and} \quad q_n \geq 2 \frac{n-1}{2} \]

\[(ii)\] \[ \gcd(p_n, q_n) = 1 \]

Proof. i) Iterating the recursion in Theorem 6.6 twice, we conclude that

\[ p_{n+1} = (a_n a_{n-1} + 1) p_{n-2} + a_n p_{n-3} \geq 2p_{n-2} + p_{n-3}, \]

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 \( \omega \), the limit \( \lim_{n \to \infty} \frac{p_n}{q_n} \) exists and equals \( \omega \).

Proof. If we replace \( n \) by \( n - 1 \) in the equality of Corollary 6.7(ii), we get another equality. Adding those two equalities gives:

\[ \frac{p_{n-2}}{q_{n-2}} - \frac{p_n}{q_n} = \frac{(-1)^n}{q_{n-1} q_n} + \frac{(-1)^{n-1}}{q_{n-1} q_{n-2}} \quad \text{or} \quad \frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{(-1)^n}{q_n} \left( \frac{1}{q_{n-2}} - \frac{1}{q_n} \right), \]

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 \( 2n \) for \( n \) in Corollary 6.7(ii), we see that the decreasing sequence \( (n \text{ odd}) \) 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 \(\omega_-\). Similarly, the increasing sequence must have a limit \(\omega_+\). 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 \(\omega_+ = \omega_- = \omega\).

\[\boxed{\text{Corollary 6.10. Suppose } \omega \text{ is irrational. For every } n > 0, \text{ we have } \frac{p_{2n-1}}{q_{2n-1}} < \omega < \frac{p_{2n}}{q_{2n}}. \text{ If } \omega \text{ is rational, the same happens, until we obtain equality of } \omega \text{ and the last approximant.}}\]

6.3. Computing with Continued Fractions

Suppose we have a positive real \(\omega_0\) and want to know its continued fraction coefficients \(a_i\). By the remark just before Theorem 6.5, we start by setting

\[a_0 = \lfloor \omega_0 \rfloor \quad \text{and} \quad \omega_1 = \omega_0 - a_0.\]

After that, we use Lemma 6.2, and get

\[a_i = \left\lfloor \frac{1}{\omega_i} \right\rfloor \quad \text{and} \quad \omega_{i+1} = \frac{1}{\omega_i} - a_i.\]

For example, we want to compute the \(a_i\) for

\[\omega_1 = \frac{1 + \sqrt{6}}{5} \approx 0.6898979 \ldots.\]

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 \(\omega_1\): 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 \(\omega_1\). 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

\[\omega_2 = \frac{4 - \sqrt{6}}{1 + \sqrt{6}} \cdot \frac{-1 + \sqrt{6}}{-1 + \sqrt{6}} = -2 + \sqrt{6} \approx 0.45 \in \left(\frac{1}{5}, \frac{1}{4}\right) \implies a_2 = 2.\]

\(^3\)This is the monotone convergence theorem, see for example [26]
Subsequently, we repeat the same steps to get
\[ \omega_3 = \frac{1}{-2 + \sqrt{6}} - 2 = \cdots = -2 + \frac{\sqrt{6}}{2} \approx 0.225 \in \left( \frac{1}{3}, \frac{1}{2} \right) \implies a_3 = 4. \]

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 \( \omega_1 \) in terms of its continued fraction coefficients:
\[ \omega_1 = \frac{1 + \sqrt{6}}{5} = [1, 2, 4, 2, 2, 4, \cdots] = [1, \overline{2, 4}]. \]

The reverse problem is also interesting. Suppose we just know the continued fraction coefficients \( \{a_i\}_{i=1}^m \) of \( \omega_1 \). We can compute the continued fraction convergents by using Theorem 6.6
\[
\begin{pmatrix} q_n \\ q_{n-1} \\ p_n \\ p_{n-1} \end{pmatrix} = A_n \cdots A_2 A_1 \quad \text{where} \quad A_i = \begin{pmatrix} a_i & 1 \\ 1 & 0 \end{pmatrix}.
\]

Theorem 6.9 assures us that the limit of the convergents \( \{p_n/q_n\}_{i=1}^\infty \) indeed equals \( \omega_1 = [a_1, a_2, \cdots] \). If also \( a_0 > 0 \), add \( a_0 \) to \( \omega_1 \) in order to obtain \( \omega_0 \).

So in our example \( \omega_1 = [1, \overline{2, 4}] \), this is easy enough to do:

<table>
<thead>
<tr>
<th>( i )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>\cdots</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_i )</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>\cdots</td>
</tr>
<tr>
<td>( p_i )</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>20</td>
<td>89</td>
<td>\cdots</td>
</tr>
<tr>
<td>( q_i )</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>13</td>
<td>29</td>
<td>129</td>
<td>\cdots</td>
</tr>
</tbody>
</table>

But, because the \( a_i \) are eventually periodic, we can also opt for a more explicit representation of \( \omega_1 \). The periodic tail can be easily analyzed. Indeed, let
\[ x = \frac{1}{2 + \frac{1}{4 + \frac{1}{x}}}, \quad \implies \quad x = \frac{1}{2 + \frac{1}{x}}. \]

After some manipulation, this simplifies to a quadratic equation for \( x \) with one root in \( [0, 1) \).
\[ x^2 + 4x - 2 = 0 \implies x = -2 \pm \sqrt{6}. \]
Select the root in \([0,1)\) as answer. Now we compute \(\omega_1\) as follows.

\[
\omega_1 = \frac{1}{1 + \frac{1}{2 + \sqrt{6}}} = \frac{1}{1 + x} = \frac{1}{-1 + \sqrt{6}} = \frac{1 + \sqrt{6}}{5},
\]

which agrees with our earlier choice of \(\omega_1\) in equation (6.4).

### 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 approximants (Theorem 6.13). Another geometric description can be found in exercise 6.15.

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

\[
e_{-1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad e_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.
\]

Note that these are row vectors. 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

\[
e_1 = a_1 e_0 + e_{-1} = \begin{pmatrix} a_1 \\ 1 \end{pmatrix},
\]

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 12). With this definition it is easy to see that in particular \(e_1 = \begin{pmatrix} a_1 \\ 1 \end{pmatrix}\) and

\[
a_1 = \left\lfloor \frac{1}{\omega_1} \right\rfloor,
\]

the same as in Definition 6.3. Note that \(\omega_1\) 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 (6.5) and (6.6), one concludes that the matrix $A_1$ satisfies

$$A_1 \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} = x_1 e_0 + x_2 e_1 \quad \text{and} \quad A_1^{-1} \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} = A_1^{-1}(x_1 e_0 + x_2 e_1).$$

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, $\omega_2$, is the one given by equation 6.1. Since $a_1$ was chosen the greatest integer so that the new slope is non-negative, we obtain that $\omega_1$ 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, $\omega_1$ always lies between $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$.

Consider the parallelogram $p(e_n, e_{n-1})$ spanned by $e_n$ and $e_{n-1}$. Define $e_n = (q_n, p_n)$. Thus, 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 $\omega x$ in the plane gives us another insight, see Figure 10. The successive intersections with the vertical unit edges are in fact the

\(^4\)Geometrically, the proof of that corollary is most easily expressed in the language of exterior or wedge products. The relevant induction step is the following computation.

$$e_n \wedge e_{n-1} = (a_n e_{n-1} + e_{n-2}) \wedge e_{n-1} = -e_{n-1} \wedge e_{n-2}.$$
6.5. Closest Returns

iterates of the rotation $R_\omega : x \rightarrow x + \omega \mod 1$ on the circle starting with initial condition 0. A natural question that arises is: when do these iterates return close to their starting point?

**Definition 6.11 (Closest Returns).** $R_\omega^q$ is a closest return if $R_\omega^q(0)$ is closer to 0 (on the circle) than $R_\omega^n(0)$ for any $0 < n < q$.

The surprise is that the continued fraction convergents correspond exactly to the closest returns (Theorem 6.13).

\[ \omega_1 q_n - p_n \]

![Figure 10](image.png)

**Figure 10.** The line $y = \omega x$ and (in red) successive iterates of the rotation $R_\omega$. Closest returns in this figure are $q$ in \{2, 3, 5, 8\}.

**Lemma 6.12.** Define $d_n \equiv \omega_1 q_n - p_n$. Then the sequence $\{d_n\}$ is alternating and its absolute value decreases monotonically. In fact, $|d_{n+1}| < \frac{1}{1 + a_{n+1}} |d_{n-1}|$.

\[ d_n, d_{n+1}, d_{n-1} \]

![Figure 11](image.png)

**Figure 11.** The geometry of successive closest returns.

**Proof.** The sequence $\{ \omega_1 - \frac{p_n}{q_n} \}$ alternates in sign by construction. Therefore, so does $\{d_n\}$. Recall that $a_{n+1}$ is the largest integer such that

\[ \omega_1 q_{n+1} - p_{n+1} = \omega_1 (a_{n+1} q_n + q_{n-1}) - (a_{n+1} p_n + p_{n-1}) = (\omega_1 q_{n-1} - p_{n-1}) + a_{n+1}(\omega_1 q_n - p_n), \]
has the same sign as \( \omega_1 q_{n-1} - p_{n-1} \). 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
\[
(1 + a_{n+1})|d_{n+1}| < |d_{n+1}| + a_{n+1} |d_n| = |d_{n-1}|.
\]
This implies the lemma.

\[\text{Figure 12. Drawing } y = \omega_1 x \text{ and successive approximations (} a_{n+1} \text{ is taken to be 3). The green arrows correspond to } e_{n-1}, e_n, \text{ and } e_{n+1}.\]

**Theorem 6.13 (The closest return property).** \( \frac{p'}{q'} \) is a continued fraction convergent if and only if
\[
|\omega_1 q' - p'| < |\omega_1 q - p| \text{ for all } 0 < q < q'.
\]
6.6. Exercises

Proof. 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 the 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 one can check that that it inherits this property (Figure 12).

Next we show, again by induction, that the \( R_q \) are closest returns, and that there are no others. It is trivial that \( R_1 \) is the only closest return for \( q = 1 \). It is easy to see that \( R_{a_1} \) is the only closest return for \( 0 < q \leq a_1 \). 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 6.12, \( d_{n+1} < d_n \). 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 12. With the exception of the origin and the endpoints of \( e_n \) and \( e_{n+1} \), the shaded regions in the figure are contained in the interior of translates of the parallelogram \( p(e_{n-1}, e_n) \), and therefore contain no lattice points. Since the vector \( c \) is parallel to and larger than \( e_n \), we have that \( b > a \). Thus there is a band of width \( d_n \) around \( y = \omega_1 x \) that contain no points in \( \mathbb{Z}^2 \) except the origin, \( e_n \), and \( e_{n+1} \).

6.6. Exercises

Exercise 6.1. Give the continued fraction expansion of \( \frac{13}{31}, \frac{21}{34}, \frac{34}{21}, \frac{a-1}{n} \) for \( n > 1, \frac{a-1}{n^2} \) for \( n > 1 \) by following the steps in Section 6.3.

Exercise 6.2. Verify the continued fraction expansion of \( \sqrt{2} \approx 1.4 \) given in the text by following the steps in Section 6.3.

Exercise 6.3. 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 6.4. Compute the continued fraction expansion for \( \sqrt{n} \) for \( n \) between 1 and 15.

---

5 By definition of \( a_1 \), the first time \( q \omega_1 \) is within \( \omega_1 \) of a natural number is when \( q = a_1 \).
Exercise 6.5. Given the following continued fraction expansions, deduce a quadratic equation for \(x\). (Hint: see Section 6.3.)

a) \(x = \left[ 8 \right] = [8, 8, 8, 8, \ldots] \).

b) \(x = \left[ 3, 6 \right] = [3, 6, 6, 6, \ldots] \).

c) \(x = [1, 2, 3] = [1, 2, 3, 1, 2, 3, \ldots] \).

d) \(x = [4, 5, 1, 2, 3] = [4, 5, 1, 2, 3, 1, 2, 3, \ldots] \).

Exercise 6.6. In exercise 6.5:

a) solve the quadratic equations (leaving roots intact).

b) give approximate decimal expressions for \(x\).

c) give the first 4 continued fraction approximants.

Exercise 6.7. Derive a quadratic equation for the number with continued fraction expansion: \([\overline{n, m, n}], [\overline{n, m}], [\overline{n, m}]\).

Exercise 6.8. From the expressions given in Section 6.2, compute the first 6 approximants of \(\pi - 3\), \(e - 2\), \(\theta\), and \(g\).

Exercise 6.9. In exercise 6.8, numerically check how close the \(n\)th approximant of \(\omega\) is to the actual value of \(\omega\).

b) Compare your answer to (a) with the decimal expansion approximation using \(i\) digits.

Exercise 6.10. In exercise 6.8, check that the increasing/decreasing patterns of the approximants satisfies the one described in the proof of Theorem 6.9.

Exercise 6.11. a) Characterize when the decimal expansion of a real number is finite. (Hint: see exercises 5.3, and 5.4.)

b) Compare (a) with Theorem 6.5.

Exercise 6.12. What does the matrix in Theorem 6.6 correspond to in terms of the Euclidean algorithm of Chapter 3?

Exercise 6.13. Use Lemma 6.12 to show that

\[
\left| \omega - \frac{P_{2n+1}}{Q_{2n+1}} \right| < \frac{1}{q_{2n+1}} \prod_{i=1}^{n} \frac{1}{1 + \alpha_{2i+1}}.
\]

Exercise 6.15. (Adapted from [4]) Consider the line \( \ell \) given by \( y = \omega x \) with \( \omega \in (0, 1) \) an irrational number. Visualize a thread lying on the line \( \ell \) 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 \( \omega \). Mark the \( n \)th of those pins as \( v_{2n} \) for \( n \in \mathbb{N} \). We will denote the points of the positive quadrant lying on or below the thread by \( A \). Repeat the same pulling 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 lying or above the thread by \( B \). See Figure 13.

a) Show that \( A \) and \( B \) are convex sets.
b) Show that \( A \) and \( B \) contain all the lattice points of the positive quadrant.
c) 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.
d) Compute the slopes of the upper boundary of the region \( A \). The same for the lower boundary of the region \( B \).
Exercise 6.16. Assume $\omega$ is irrational.

a) Use Corollary 6.7(ii) and Corollary 6.10 to show that

$$
\left| \omega - \frac{p_n}{q_n} \right| < \frac{1}{q_nq_{n+1}}.
$$

b) Use Lemma 6.12 to show that

$$
\left| \omega - \frac{p_{n+1}}{q_{n+1}} \right| < \left| \omega - \frac{p_n}{q_n} \right|.
$$

c) Use (a) and (b) to show that

$$
\frac{1}{2q_nq_{n+1}} < \left| \omega - \frac{p_n}{q_n} \right| < \frac{1}{q_nq_{n+1}}.
$$


Exercise 6.18. Use exercise 6.16 (a) to prove Theorem 1.17.

Exercise 6.19. 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 1.12).

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 6.14. 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 whether the continued fraction coefficients of algebraic numbers of degree 3 exhibit a recognizable pattern.
Exercise 6.20. A natural question that arises is whether you can formulate continued fraction for polynomials [10]. We try this for the rational function \( f(x) = \frac{x^3 + x^2}{x^2 + 1}. \) Referring to exercise 3.11 and the definition of \( a_i \) in the remark after Definition 6.3, we see that

\[
\begin{align*}
a_1 &= (x^4 - x^3 + x^2 - x + 1) \\
a_2 &= (-\frac{x}{2} - \frac{1}{2}) \\
a_3 &= (-4x + 4) \\
\text{and} \quad a_4 &= (-\frac{x}{2} - \frac{1}{2})
\end{align*}
\]

a) Compute the continued fraction approximants \( \frac{p_n}{q_n} \) for \( n \in \{1, \ldots, 4\} \) of \( f(x) \). \((\text{Hint: perform the computations as given in Theorem 6.6.})\)

b) In (a), you obtained the polynomials of exercise 3.11 up to a factor -1. Why? \((\text{Hint: The gcd we computed in that exercise is actually -1. As stated in that exercise, we neglect constants when using the algorithm for polynomials. At any rate, in the quotient, the constant cancels.})\)

c) Is there a theorem like the one in exercise 6.18? \((\text{Hint: Yes, follow the hint in that exercise.})\)

d) Solve for \( y \): \( y = [x] \). \((\text{Hint: check exercise 6.7})\)

e) Any ideas for other non-rational functions? \((\text{Hint: check the web for Padé approximants.})\)

Exercise 6.21. 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.

\[ n = \prod_{i=1}^{\infty} p_i^\ell_i \] and there are infinitely many primes. Thus we can encode the natural numbers as an infinite sequence \((\ell_1, \ell_2, \ell_3, \ldots)\) 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 \([q_1, q_2, \ldots]\) 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 6.22. Consider Figure 14. 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 \( 2\pi \).

b) Use a) to explain why we appear to see 6, 19, 25, and 44 spiral arms.

c) Why does the curvature of the spiral arms appear to (a) alternate and (b) decrease?
Exercise 6.23. The exercise depends on exercise 6.22. 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 14.

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 15.

c) Use Theorem 6.15 to show that the primes tend (as \( \max p \to \infty \)) to be equally distributed over the co-prime arms.

d) Use Theorem 4.16 to determine the number of co-prime arms. Confirm this in the left plot of Figure 15.

e) Explain the new phenomenon occurring in the right plot of Figure 15.
6.6. Exercises

Figure 15. 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.

**Theorem 6.15 (Dirichlet’s Theorem).** 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 \(\text{Res}_n(p) = r\). Then

\[
\lim_{x \to \infty} \frac{\pi(x; n, r)}{\pi(x)} = \frac{1}{\varphi(n)}.
\]

**Exercise 6.24.**

a) Visualize the continued fraction expansion of another irrational number \(\rho \in (0, 1)\) by plotting a polar plot of the numbers \((n, 2\rho \pi n)\) for various ranges of \(n\) as in exercise 6.22.

b) Check Dirichlet’s theorem as in exercise 6.23

**Exercise 6.25.** Set \(\omega_1 = e - 2 \approx 0.71828\).

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 in the following sense.

\[
\forall q \leq 2 : \left| \omega_1 - \frac{1}{2} \right| \leq \left| \omega_1 - \frac{p}{q} \right|.
\]

d) Show that \(1/2\) is not a closest return in the sense of Theorem 6.13.
Part 2

Topics in Number Theory
Algebraic Integers

Overview. We present some basic notions of algebraic number theory in a simplified way. For those of us that are not well-trained in algebra, this discipline of mathematics seems to start with a daunting barrage of definitions. Here we try to give the minimum amount of definitions, while still conveying some of the algebraic subtleties of the subject.

7.1. Rings and Fields

It will be helpful to place our previous consideration of primes in \( \mathbb{N} \) in a more general context. Here are the relevant definitions (rings and fields).

Definition 7.1. A ring is defined as a set \( R \) which is closed under two operations, usually called addition and multiplication, and has the following properties:

1) \( R \) with addition is an abelian group (that is: \( a + b = b + a \), addition is associative or \( (a+b)+c=a+(b+c) \), \( R \) has an identity element, and addition is invertible).
2) Multiplication is \( R \) is associative.
3) Multiplication is distributive over addition (that is: \( ab+bc = a(b+c) \)).

A commutative ring is a ring in which multiplication is commutative.

Remark 7.2. 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 7.3. We will assume rings to be commutative, unless otherwise mentioned.

In the light of this definition, it is useful to reflect a moment on how the two operations addition (+) and multiplication (·) mesh in rings like \(\mathbb{Z}\). Rings are closed under both operations: given any two integers \(a\) and \(b\), we can form \(a + b\) and \(ab = a \cdot b\), to get another element in \(\mathbb{Z}\). Both operations have an identity, namely 0 for addition, and 1 for multiplication. Thus, for any \(a\), we have that \(0 + a = a + 0 = a\) and \(1a = a1 = a\). The one profound difference that causes all our trouble is this: addition has an inverse, namely subtraction, but multiplication does not. For every \(a\) and \(b\), we can form \(c = a - b\), and adding \(c + b\) we get \(a\) back. But multiplication only occasionally has an inverse: we can divide 4 by 2 to get 2, and then multiply again by 2 to get 4 back. But we cannot do that with 4 and 3. It is this curious property that brings us to the study of primes, integers that have no non-trivial divisors at all. Note that the situation in \(\mathbb{Z}_p\) for prime \(p\) is very different! Here multiplication does have an inverse (see Corollary 5.19), and thus given \(a\) and \(b\) not equal to 0, we can always write

\[
a = p(ab)b^{-1}.
\]

Thus in the language of Definition 2.14, every number is a unit and there are no primes in \(\mathbb{Z}_p\), let alone unique factorization. A ring like \(\mathbb{Z}_p\) where all elements have multiplicative inverses is called a field.

In the remainder of this chapter, we will not burden the student with proofs that sets like \(\mathbb{Q}\) or \(\mathbb{Z}[\sqrt{-5}]\) are rings. This is adequately covered in algebra courses. Instead, we concentrate on and study the different types of rings.

Definition 7.4. A field is a commutative ring for which multiplication has an inverse.

Here are some examples of rings that are not fields. Numbers of the form \(a + b\sqrt{3}\) where \(a\) and \(b\) in \(\mathbb{Z}\), numbers of the form \(a + ib\sqrt{6}\) where \(a\) and \(b\) in \(\mathbb{Z}\). Other examples are \(\mathbb{Z}_6\) and the \(n\) by \(n\) matrices with coefficients in \(\mathbb{Z}\) \((n \geq 2)\). We have already seen the polynomials with rational coefficients exercises 3.11 and 6.20. They also form a ring. All of these rings have different properties. For instance, the ring of matrices is not commutative.
Here is another interesting observation. If we extend the integers to the rationals $\mathbb{Q}$, we obtain a field. Thus the problem 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.

7.2. Primes and Integral Domains

We start by defining division a little more carefully than in Definition 1.1.

**Definition 7.5.** Let $a$, $b$, and $x$ in a ring. We say that $b$ is a divisor of $a$ if there is a solution $x$ of $bx = a$. We write $b | a$.

We need to be even more careful in defining primes. We originally defined primes in the positive integers only (Definition 1.3). Since we want to deal with rings, we need to widen our scope and include the negative integers. But then we have $7 = 1 \cdot 7 = -1 \cdot -7 = -7 \cdot 1 = -1$, et cetera. Does unique factorization hold? We say yes, because we discount multiplications by unimportant factors 1 or $-1$. Such factors are called units or invertible.

**Definition 7.6.** Given a ring $R$. A unit of the ring is an element that has a multiplicative inverse in the ring. This is also called an invertible element.

**Definition 7.7.** Given a ring $R$ and a non-zero element $r$. Then $r$ is reducible if it is a product of two non-units (or non-invertible elements). If it not equal to a product of two non-invertible elements it is called irreducible.

**Definition 7.8.** Given a ring $R$. An element $p$ of the ring is a prime if it is not zero or a unit, and if whenever $p | ab$, then $p | a$ or $p | b$ (or both).
Somewhat confusingly, our earlier, intuitive notion of a prime (Definitions 1.3 and 2.14, a number with no non-trivial divisors) actually corresponds to the notion of irreducible and not prime. It turns out that in \( \mathbb{Z} \) these notions are the same. But, surprisingly, in general they are not!

**Lemma 7.9.** (i) Consider the ring \( R = \{ a + ib\sqrt{5} \mid a, b \in \mathbb{Z} \} \). Its element 3 is irreducible, but not prime.
(ii) In the ring \( \mathbb{Z}_6 \), the element 2 is prime, but not irreducible.

**Proof.** (i) Suppose the number 3 equals the product \( xy \), where \( x \) and \( y \) in \( R \). Clearly, \( x \) and \( y \) cannot both be real, because 3 is prime in \( \mathbb{Z} \). If both are non-real, then they have absolute value at least \( \sqrt{5} \), and this would lead to \( 3 \geq 5 \), 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 \( R \). But on the other hand, 

\[
(2 + i\sqrt{5})(2 - i\sqrt{5}) = 9 \quad \Rightarrow \quad 3|(2 + i\sqrt{5})(2 - i\sqrt{5}) .
\]

But 3 does not divide either of these factors. Therefore 3 is not prime.

(ii) Suppose \( 2 | ab \) in \( \mathbb{Z}_6 \). Then in \( \mathbb{Z} \), 2 divides \( ab + 6m \) for some \( m \). But that means that \( ab \) is even and thus \( a \) (or \( b \)) has a factor 2 (see Theorem 2.15). But then in \( \mathbb{Z}_6 \), 2 also 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.

**Definition 7.10.** An integral domain is a commutative ring \( R \) with no zero divisors (i.e. there are no non-zero elements \( a \) and \( b \) such that \( ab = 0 \)).

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 \( x = y \). The following is an immediate corollary.

**Corollary 7.11 (Cancellation Theorem).** In an integral domain, if \( a \neq 0 \), then \( ax = ay \) if and only if \( x = y \). (See also Theorem 2.8.)

For example in \( \mathbb{Z}_{11} \), we can do the following.

\[
x^2 + 5x + 6 =_{11} 0 \quad \Rightarrow \quad (x + 2)(x + 3) =_{11} 0 .
\]
And this implies that \( x =_{11} -2 \) or \( x =_{11} -3 \). If we work modulo 12, this does not work, because \( x \in \{1, 6\} \) are also solutions, and the problem of
7.3. Norms

\( ab = 0 \) does not simplify. The 2 by 2 matrices with coefficients in \( \mathbb{Z} \) form another ring that is not an integral domain. In fact, if \( N = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \), then \( N^2 = 0 \).

**Proposition 7.12.** Let \( R \) be an integral domain. If \( p \in R \) is prime, then \( p \) 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 7.8, \( p \mid a \) or \( p \mid b \). Assume the former. So there is a \( c \) such that \( a = pc \). We then get
\[
p = ab = pcb \quad \Rightarrow \quad p(1 - cb) = 0.
\]
This implies, of course, that \( bc = 1 \), and so \( b \) has an inverse and therefore is a unit. Similar if we assume \( p \mid b \). ■

**Remark 7.13.** The converse is false. The ring \( \mathbb{Z}[\sqrt{-5}] \) in Lemma 7.9 (i) is easily seen to be an integral domain. Yet 3 irreducible but not prime.

Here is an interesting lemma that implies (once again) that \( \mathbb{Z}_p \), \( p \) prime, is a field (see Proposition 5.16). The reason is essentially that of Lemma 5.3.

**Lemma 7.14.** 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 \( ax = by \), the cancellation theorem gives a contradiction. If they are all distinct, then, by the pigeon hole principle, there is an \( x \) such that \( ax = 1 \). Thus \( a \) has a multiplicative inverse. ■

7.3. Norms

In \( \mathbb{Z}_p \), \( p \) prime, we have that \( ab = 0 \) implies that \( p \mid ab \), which implies that \( a \) or \( b \) is zero modulo \( p \). We then see from Lemma 7.14 that there are no primes in \( \mathbb{Z}_p \). For \( b \) non-prime, \( \mathbb{Z}_b \) is not an integral domain, which complicates the study of division (and so of primes). As far as integral domains are concerned, that seems to leave only \( \mathbb{Z} \). However, we will now
look further afield to find other interesting rings. One common construction is very similar to vector spaces.

As an example, let \( j \) be a square free integer \( j \in \mathbb{Z} \) (i.e., an integer that is not divisible by the square of any prime, see exercise 2.17) and not equal to 1 or -1. It turns out that in this case, \((\sqrt{j})^2 - j = 0\) and that relation cannot be factored over the integers. As a result, we call \( \sqrt{j} \) a *algebraic integer* of degree 2.

**Definition 7.15.** An algebraic integer of degree \( d \) is a real number \( \alpha \) that is a root of a minimal polynomial in \( \mathbb{Z}[x] \) that is monic.\(^1\)

Note that the difference between this definition and that of algebraic numbers (Definition 1.12) is the use of the word *monic*. The solutions of \( x^2 - 2 = 0 \) are quadratic integers, but the solutions of \( 3x^2 - 2 = 0 \) are quadratic numbers.

Now we denote \( \mathbb{Z}[\sqrt{j}] = \{a + b\sqrt{j} \mid a, b \in \mathbb{Z}\} \).

If \( j \) is negative, we can think of \( \mathbb{Z}[\sqrt{j}] \) as a subset of the complex plane. If \( j \) is positive, then it is a subset of the real line. In both cases \( \mathbb{Z}[\sqrt{j}] \) is countable (see Theorem 1.22). Clearly, \( \mathbb{Z}[\sqrt{j}] \) is closed under addition. It is also closed under multiplication:

\[
(a + b\sqrt{j})(c + d\sqrt{j}) = ac + bd j + (ad + bc) \sqrt{j},
\]

which is again an element of \( \mathbb{Z}[\sqrt{j}] \). In fact, \( \mathbb{Z}[\sqrt{j}] \) is a ring. But we can do better! We can look at \( \mathbb{Z}[\sqrt{j}] \) as having two basis vectors

\[
\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 \quad \text{and} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \sqrt{d}.
\]

The elements of \( \mathbb{Z}[\sqrt{j}] \) are precisely the linear combinations \( a \cdot 1 + b \cdot \sqrt{d} \). Just like a two-dimensional vector space! The only difference is that the “scalars” now belong to a ring and not a field. The resulting construction is called a 2-dimensional module.

\(^1\)Monic means that the leading term has coefficient 1
7.3. Norms

**Definition 7.16.** An *module* (or left module) is a set with the same structure as a finite-dimensional vector space, except that its scalars form a commutative ring (and not a field as in a vector space). Scalars multiply the basis vectors from the left. (If scalars multiply from the right, the result is called a right module.)

This immediately gives us an alternative characterization of algebraic integers.

**Corollary 7.17.** A real number \( \alpha \) is an algebraic integer of degree \( d \) if and only if \( \mathbb{Z}[\alpha] \) is a \( d \)-dimensional module.

Next, we interpret multiplication by \( \alpha = a + b\sqrt{J} \) in \( \mathbb{Z}[^J] \) when \( \sqrt{J} \) is an algebraic integer of degree 2. Clearly, it is linear, because

\[
\alpha(c + d\sqrt{J}) = c\alpha1 + d\alpha\sqrt{J}.
\]

Therefore, \( \alpha \) 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

\[
\alpha \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} \quad \text{and} \quad \alpha \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} bj \\ a \end{pmatrix}.
\]

Thus we can use elementary linear algebra to see that

\[
(7.1) \quad \alpha = \begin{pmatrix} a & bj \\ b & a \end{pmatrix}.
\]

What is interesting here, is that the determinant of \( \alpha \)

\[
(7.2) \quad \det \alpha = a^2 - b^2j
\]

is clearly an integer and cannot be zero, because if \( j b^2 - a^2 = 0 \), then \((b\sqrt{J} - a)(b\sqrt{J} + a) = 0\). Since both factors are complex numbers (and \( \mathbb{C} \) has no zero divisors), one of the two factors equals zero. But that contradicts the fact that \( \sqrt{J} \) is an algebraic integer of degree 2.
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 phrase factorization in \( \mathbb{Z}[\sqrt{j}] \) in terms of the determinant of \( \alpha \). In number theory, this quantity is known as the norm of \( \alpha \).

We can of course consider \( \mathbb{Z}[\beta] \) where \( \beta \) is some algebraic number (Definition 1.12) of degree \( d > 2 \). We then get a \( d \)-dimensional module. But for now, we will stick to 2-dimensional modules.

One more remark gets us to the definition we need. We can extend \( \mathbb{Z}[\sqrt{j}] \) to \( \mathbb{Q}[\sqrt{j}] \) by replacing the integer coefficients (or scalars) by rational numbers. We obtain a field, but the notion of norm remains the same.

**Definition 7.18.** The field norm, or simply norm, of an element \( \alpha \) of a field \( F \) is the absolute value of the determinant of the multiplication by \( \alpha \). It will be denoted by \( N(\alpha) \).

A well-known result about determinants from linear algebra (\( \det AB = \det A \det B \)) gives a handy rule.

**Corollary 7.19.** \( N(\alpha) N(\beta) = N(\alpha \beta) \).

**Remark 7.20.** Suppose \( \alpha = a + b\sqrt{j} \) in \( \mathbb{Z}[\sqrt{j}] \). If \( j \) is negative then from equation 7.2, we also get \( N(\alpha) = \alpha \overline{\alpha} \), where \( \overline{\alpha} \) means the complex conjugate of \( \alpha \).

### 7.4. Euclidean Domains

**Definition 7.21.** A *Euclidean function* on a ring \( R \) is a function \( N : R \to \mathbb{N} \) that satisfies:

(i) For all \( \rho_1 \) and \( \rho_2 \) in \( R \setminus \{0\} \), there are \( \kappa \) and \( \rho_3 \) such that \( \rho_1 = \kappa \rho_2 + \rho_3 \) and \( N(\rho_3) < N(\rho_2) \) and

(ii) For all \( \alpha \) and \( \gamma \) in \( R \setminus \{0\} \), we have \( N(\alpha \gamma) \geq N(\alpha) \).

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 2.3 — minus the uniqueness. That lemma is also called Euclid’s division lemma\(^2\) (see also remark 2.5). We used that lemma to prove Bézout and the Euclidean algorithm of Chapter 3, which in turn led us to continued

\(^2\)Hence the name “Euclidean domain”.
fractions among other things. So the consequences of having a Euclidean function are indeed staggering!

We first need a slight re-interpretation of greatest common divisor and least common multiple of Definition 1.2.

**Definition 7.22.** Let $R$ be a Euclidean domain and $\alpha$ and $\beta$ non-zero elements. A greatest common divisor $\gcd(\alpha, \beta)$ is an element that maximizes the norm $N$ over all divisors of both $\alpha$ and $\beta$. A least common multiple $\text{lcm}(\alpha, \beta)$ minimizes the norm over all multiples of both $\alpha$ and $\beta$.

But now all statements and proofs in Chapter 2 from Bézout’s Lemma (Lemma 2.6) on, hold for Euclidean domains up to and including the infinity of primes (Theorem 2.17). The only reason to introduce the — at first sight rather clumsy — norm in that chapter (Definition 2.2) was so these proofs apply almost verbatim to Euclidean domains. We leave this as an (enormous) exercise (exercise 7.22.)

Here is an example. In a Euclidean ring $R$, we can employ the Euclidean algorithm to show that all integers can be factored into irreducible numbers in exactly the same way as in Chapter 2. Thus, in integral domains primes are irreducible by Proposition 7.12, and in Euclidean domains irreducible numbers are primes by Corollary 2.12. The following are but a few consequences taken from Chapter 2.

**Corollary 7.23.** Let $R$ be a Euclidean domain. Then $r \in R$ is prime if and only if $r$ is irreducible.

**Corollary 7.24.** A Euclidean domain has the unique factorization property.

### 7.5. 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 16). From equations 7.1

---

3 There is but one cavea in all this. As noted before, the notion of *prime* in $\mathbb{Z}$ (Definitions 1.3 and 2.14) actually corresponds to the notion of *irreducible* (Definition 7.7) in a ring, and not to the notion prime in a ring (Definition 7.8). This is done to confuse non-algebraists, so they stay away. But we’re not falling for it!
and 7.2, we can infer that \( \alpha = a + bi \) can be represented in matrix form as:

\[
\alpha = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}
\]

with \( N(\alpha) = a^2 + b^2 \).

It is easy to check that multiplication of these matrices is commutative — after all, multiplication of the underlying complex numbers is commutative.

**Proposition 7.25.** The Gaussian integers form a Euclidean domain.

**Proof.** Let \( \alpha = a + bi \) and \( \gamma = c + di \). We wish to prove that the norm \( N(\alpha) = a^2 + b^2 \) of Definition 7.18 is a Euclidean function. Since \( N \) is a positive integer, the second requirement follows immediately from Corollary 7.19.

![Figure 16. 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 \( \text{Re}(z) \) and \( k_2 \) the closest integer to \( \text{Im}(z) \). In this case that is \( 2 + 3i \).](image)

It remains to prove the first requirement is satisfied. Consider \( \alpha \) and \( \gamma \) as in the previous paragraph. We can certainly choose \( \kappa \) and \( \rho \) so that

\[
\alpha = \kappa \gamma + \rho.
\]

(For example, \( \kappa = 0 \) and \( \rho = \alpha \).) Dividing (as complex numbers) by \( \gamma \) gives

\[
\alpha \gamma^{-1} = \kappa + \rho \gamma^{-1}.
\]

It is an easy computation to see that

\[
\alpha \gamma^{-1} = \frac{ac + bd}{c^2 + d^2} + i \frac{-ad + bc}{c^2 + d^2}.
\]
7.5. Example and Counter-Example

We want to express this as a Gaussian integer \( \kappa = k_1 + ik_2 \) plus a remainder \( \varepsilon \equiv \rho \gamma^{-1} = \varepsilon_1 + i\varepsilon_2 \) whose norm is less than 1:

\[
\alpha \gamma^{-1} - \kappa = \varepsilon.
\]

It is important to realize that there might be more than one way to do that (see Figure 16). A choice that always works is the following: choose \( k_1 \) to be the integer closest to \( \frac{ac + bd}{c^2 + d^2} \), and \( k_2 \), the integer closest to \( \frac{−ad + bc}{c^2 + d^2} \). (If one of these values equals \( n + \frac{1}{2} \), then pick one integer.) With those choices, the remainders

\[
\varepsilon_1 = \frac{ac + bd}{c^2 + d^2} - k_1 \quad \text{and} \quad \varepsilon_2 = \frac{−ad + bc}{c^2 + d^2} - k_2
\]

are each not greater than \( \frac{1}{2} \) in absolute value. Thus

\[
\rho = (\varepsilon_1 + i\varepsilon_2)(c + id),
\]

with norm \((\varepsilon_1^2 + \varepsilon_2^2)(c^2 + d^2)\) by Corollary 7.19. Since the \( \varepsilon_i \) are no greater than \( \frac{1}{2} \), this is less than \( N(\gamma) = c^2 + d^2 \).

Note that in \( \mathbb{Z} \), to get a small remainder we simply choose the floor of \( \alpha \gamma^{-1} \) for the equivalent of \( \kappa \) (see the proof of Lemma 2.3). 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 7.16, 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 give an example of this in exercise 7.17.

**Proposition 7.26.** The ring \( \mathbb{Z}[\sqrt{-6}] \) does not have the unique factorization property.

![Figure 17](image-url).

Figure 17. A depiction of \( \mathbb{Z}[\sqrt{-6}] \) in the complex plane; real parts are integers and imaginary parts are multiples of \( \sqrt{6} \).
Proof. \(^4\) Let \( \alpha = a + ib\sqrt{6} \) and \( \gamma = c + id\sqrt{6} \) in the ring \( \mathbb{Z}[\sqrt{-6}] \). Suppose \( \alpha\gamma = 0 \). Recall that \( N(\alpha\gamma) = (a^2 + 6b^2)(c^2 + 6d^2) \). This is non-zero unless both \( \alpha \) and \( \gamma \) are zero. Thus there are no non-zero divisors (\( \mathbb{Z}[\sqrt{-6}] \) is an integral domain).

We show that \( \mathbb{Z}[\sqrt{-6}] \) does not have unique factorization, and therefore by Corollary 7.24 it cannot be a Euclidean domain. We do this 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 is reducible. Thus \( 2 = \alpha\gamma \), both non-units. Taking the norm (always using Corollary 7.19), we get

\[
4 = N(\alpha)N(\gamma).
\]

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 \( \alpha \) must have norm 2 and \( \beta \) must have norm 5, or vice versa. But the previous arguments show that both are impossible. \( \blacksquare \)

The list of integers \( d \) for which \( \mathbb{Z}[\sqrt{d}] \) is a Euclidean domain with the norm as Euclidean function is:

\[
d \in \{-11, -7, -3, -2, 1, 2, 3, 5, 6, 7, 11, 13, 17, 21, 29, 33, 37, 41, 57, 73\}.
\]

There are 4 other cases known where \( \mathbb{Z}[\sqrt{d}] \) admits unique factorization, namely

\[
d \in \{-163, -67, -43, -19\}.
\]

The complete list of positive \( d \) for which the quadratic integers admit unique factorization is unknown at the date of this writing\(^5\).[37]

\(^4\)The second 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 \( \mathbb{Z} \).

\(^5\)In 2018.
7.6. Exercises

Definition 7.27. The nth Catalan number $C_n$ equals $\frac{1}{n+1} \binom{2n}{n} = \binom{2n}{n} - \binom{2n}{n+1}$.

Exercise 7.1. Associativity is a property whose importance is hush-hushed a bit. As an example, the operation of exponentiation in $\mathbb{R}$ is not associative.

a) Compute $2^{3^4}$. (Hint: depending on how you place the parentheses, you get $2^{12}$ or $2^{81}$.)

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.)

c) For notational ease, indicate the non-associative operation by $\ast$. Show that the number of ways $\prod_{i=1}^n a_i$ 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: each parenthesis corresponds to a “up” move or a “right” 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 $(n-1,n+1)$ 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 $\ell$ as indicated in Figure 18 and show this is invertible.)

e) Use (b) and (c) 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 (d) to show that the number of interpretations (b) equals $C_n$ of Definition 7.27.

Figure 18. The part to the right of the intersection with $\ell : y = x + 1$ (dashed) of 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 7.2. 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 7.3. a) For $b$ non-prime in $\{4, \cdots, 20\}$, find the set of units in $\mathbb{Z}_b$.
   b) What are the primes of these rings?
   c) What is the cardinality of the set of units of $\mathbb{Z}_b$? (Hint: check Definition 4.9.)

Exercise 7.4. a) Solve $x^2 - 4 = 0$ for $b \in \{4, \cdots, 10\}$.
   b) For which $b$ can you use Definition 7.10 to set the factors to 0 in order to solve?
   c) Can you solve any others by setting the factors equal to 0?

Exercise 7.5. a) Solve $3x = 6x$ where $b$ is 11, 12, 13, 14, 15.
   b) If $b$ is such that $\mathbb{Z}_b$ is an integral domain, solve by factoring.

Exercise 7.6. 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.

Exercise 7.7. a) Which ones of the sets in exercise 7.2 are integral domains?
   b) Euclidean domains?

Exercise 7.8. Give more counter-examples as in Lemma 7.9 (i) and (ii).

Exercise 7.9. a) Find the multiplicative inverses of all elements in $\mathbb{Z}_7$.
   b) What is wrong in the following reasoning?

Suppose $2|ab$ in $\mathbb{Z}_7$ and 2 has an inverse in $\mathbb{Z}_7$ (see Theorem 2.15). But then in $\mathbb{Z}_7$, 2 also divides $a$ (and $b$). Therefore 2 is prime in $\mathbb{Z}_7$. (Hint: carefully study Definition 2.14.)
7.6. Exercises

Exercise 7.10. Consider \( \mathbb{Z}[\sqrt{-6}] \) and define \( a_{\pm} = 2 \pm \sqrt{-6} \), \( b = 2 \), and \( c = 5 \).

a) Show that \( a_{-} a_{+} = 10 \).

b) Show that \( a_{+} \mid 10 \) in \( \mathbb{Z}[\sqrt{-6}] \).

c) Show that \( a_{+} \mid 2 \cdot 5 \) in \( \mathbb{Z}[\sqrt{-6}] \).

d) Show that 2 is prime in \( \mathbb{Z}[\sqrt{-6}] \) and 5 is not.

e) Show that \( a_{+} \) does not divide 5, and so Euclid’s lemma 2.7 does not hold here.

Exercise 7.11. a) Show that irreducible numbers and primes coincide in \( \mathbb{Z} \).

b) Show that \( \mathbb{Z} \) with \( N(x) = |x| \) is a Euclidean domain.

Exercise 7.12. Let \( d \in \mathbb{Z} \) be square free. Show that \( \alpha \in \mathbb{Z}[\sqrt{d}] \) is a unit if and only if \( N(\alpha) = 1 \). (Hint: a matrix with determinant \( \pm 1 \) is invertible. Show that the inverse matrix corresponds to an element of \( \mathbb{Z}[\sqrt{d}] \).

Exercise 7.13. Let \( d \in \mathbb{Z} \) be square free. Suppose further that \( \alpha \in \mathbb{Z}[\sqrt{d}] \) has norm \( p \) where \( p \) is a prime in \( \mathbb{N} \).

a) Show that \( \alpha \) is irreducible in \( \mathbb{Z}[\sqrt{d}] \). (Hint: suppose \( \alpha = \beta \gamma \) and use Corollary 7.19 to see that \( \alpha \) or \( \beta \) must be a unit.)

b) Show that (a) implies the infinitude of irreducibles in \( \mathbb{Z}[\sqrt{d}] \).

Exercise 7.14. Given the ring \( R = \mathbb{Z}[\sqrt{-5}] \).

a) Show that 2 is irreducible. (Hint: suppose \( 2 = \beta \gamma \), where \( \beta \) and \( \gamma \) are non-units. Use Corollary 7.19 to see that \( N(\beta) = N(\gamma) = 2 \). Solve for the coefficients of \( \beta \) and \( \gamma \).

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 7.15. Given the ring $R = \mathbb{Z}[\sqrt{2}]$.

a) Show that $R$ has non zero divisors. (Hint: If $\alpha \beta = 0$, then one of the norms must be zero by Corollary 7.19. Solve for the coefficients.)

b) Suppose

$$\alpha = \kappa \gamma + \rho,$$

where $\alpha = a + b\sqrt{2}$, $\gamma = c + d\sqrt{2}$, $\kappa = \kappa_1 + \kappa_2\sqrt{2}$, and $\rho = \rho_1 + \rho_2\sqrt{2}$.

Show that

$$\alpha \gamma^{-1} = \frac{ac - 2bd}{c^2 - 2d^2} + \frac{-ad - bc}{c^2 - 2d^2}.$$

c) Choose $\kappa_1$ to be the integer closest to $\frac{ac - 2bd}{c^2 - 2d^2}$ and $\kappa_2$ the one closest to $\frac{-ad - bc}{c^2 - 2d^2}$. Show that the remainder has norm less than 1.

d) Show that the ring $\mathbb{Z}[\sqrt{2}]$ is a Euclidean domain (Hint: use Corollary 7.19.)

Exercise 7.16. We apply the Euclidean algorithm in $\mathbb{Z}[\sqrt{-1}]$ as in Section 7.5. For the notation, see the proof of Proposition 7.25. Suppose $\rho \gamma^{-1}$ falls in the unit square depicted in Figure 19. 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 $\kappa_1$ is the floor of the real part of $\kappa + \rho \gamma^{-1}$ and $\kappa_2$ 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 $\kappa$ 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 $\kappa$ so that $N(\rho) < N(\gamma)$?
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.

a) Check all computations in the following diagram.

<table>
<thead>
<tr>
<th></th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1 + i</td>
<td>-4</td>
<td>7 + 5i</td>
<td>17 + 15i</td>
</tr>
</tbody>
</table>

b) Check all computations in the following diagram.

<table>
<thead>
<tr>
<th></th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 + i</td>
<td>-1 + 3i</td>
<td>3 + 5i</td>
<td>7 + 5i</td>
</tr>
</tbody>
</table>

For 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

\[1 + i = (7 + 5i)x + (17 + 15i)y .\]

c) Compute gcd(\( 17 + 15i, 7 + 5i \)) (up to units).

e) Compute lcm(\( 17 + 15i, 7 + 5i \)) (up to units). (Hint: see 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. Show that \( \mathbb{R} = \mathbb{Z}[\sqrt{2}] \) has an infinite set of units. (Hint: Find a unit \( \alpha \) whose absolute value (as a real number) is greater than 1. Show that for every \( n \in \mathbb{Z} \), \( \alpha^n \) is a unit.)

Exercise 7.20. a) Show that the arithmetic functions (Definition 4.1) with the operations addition and Dirichlet convolution (Definition 4.18) form a commutative ring. (Hint: see exercise 4.14).

b) Show that the same does not hold for the multiplicative (Definition 4.1) arithmetic functions. (Hint: see exercise 4.15).

Exercise 7.21. Show that the functions \( f : \mathbb{R} \to \mathbb{R} \) together with the operations addition and (normal) convolution commutative ring. (Hint: follow the steps in 4.14).

Exercise 7.22. This is a HUGE problem in the sense of it being a lot of work. However, the student will benefit greatly from writing out all statements and proofs in Chapter 2 starting from Bézout’s Lemma (Lemma 2.6) up to and including the infinity of primes (Theorem 2.17) for Euclidean domains.
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 would need 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) without proof. In Chapter 9 we will then move to the implications for number theory. We remark that ergodic theory was to a large extent inspired by a problem that arose in 19th century physics [24], namely how to describe statistical behavior of a deterministic dynamical system.

8.1. 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 1.23). 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 1.20). The collection of subsets of
the line is in fact the same as the power set (Definition 1.26) \( P(\mathbb{R}) \) of the real line. And thus the cardinality of the collection of subsets is strictly larger than that of the real numbers (Theorem 1.27), 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 \( \mu \) 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) \( \mu \) is translation invariant.
4) For a countable collection of disjoint sets \( A_i \): \( \mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i) \).

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 1.24) 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 1.9) 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 \( V \). 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. Let \( r : \mathbb{N} \to \mathbb{Q} \) be a bijection between \( \mathbb{N} \) and the rationals in \( S \). Consider the union of the translates

\[ \bigcup_{i=1}^{\infty} (V + r_i) \, . \]

By definition of \( V \), this union covers the entire circle. So by requirement (2) above, its measure is 1. By requirement (3), each of the translates of \( V \) must have the same measure, \( \varepsilon \). Since by the previous paragraph, the translates
of $V$ are disjoint, requirement (4) implies that

$$1 = \sum_{i=1}^{\infty} \epsilon,$$

which is clearly impossible!

The construction of the set $V$ just outlined is 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. Does this construction $V$ really exist as an honest set? It turns out that one needs to invoke the axiom of choice\(^1\) to make sure that $V$ exists.

The consensus in current mathematics (2020) is to accept the axiom of choice. 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 measurable sets.

### 8.2. 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 [5].

Recall that a set $O \subseteq \mathbb{R}$ is an open set usually\(^2\) means that for all $x \in O$ there is an interval $(x - \epsilon, x + \epsilon)$ 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. An open set in $\mathbb{R}$ can be written as a disjoint union of open intervals (see exercise 8.4).

---

\(^1\)The **axiom of choice** states that for any set $A$, there exists a function $f : \mathcal{P}(A) \to A$ that assigns to each non-empty subset of $A$ an element of that subset. For more details, see [18].

\(^2\)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.
The outer measure\(^3\) of a set \(S\) is
\[
\mu_{\text{out}}(S) = \inf \sum_k l(I_k).
\]
where the infimum is over the countable covers of \(S\) by open intervals \(I_k\).

**Definition 8.1.** 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.

**Definition 8.2.** A set \(S\) is called Lebesgue measurable if it contains a Borel set \(B\) whose outer measure equals \(\mu_{\text{out}}(S)\).

One can work out [5] that the collection of Lebesgue measurable sets is also closed under complementation, countable intersection, and countable union. Furthermore, any open set in \(\mathbb{R}\) is a countable union of disjoint open intervals [21] (see also exercise 8.4). As a consequence of these facts, we have the following result.

**Proposition 8.3.** (i) A set \(S \subset \mathbb{R}\) is Lebesgue measurable if and only if there exist closed sets \(C_i \subseteq S\) such that
\[
\mu_{\text{out}}(S \setminus \bigcup_{i=1}^\infty C_i) = 0.
\]
(ii) A set \(S \subset \mathbb{R}\) is Lebesgue measurable if and only if there exist open sets \(O_i \supseteq S\) such that
\[
\mu_{\text{out}}(\bigcap_{i=1}^\infty O_i \setminus S) = 0.
\]

**Proof.** First observe that every closed set is the complement of an open set and vice versa. Since complementation preserves the Lebesgue measurable sets (by definition 8.2), (i) and (ii) are equivalent.

Definition 8.2 implies that for a measurable set \(S\) the following holds. For all \(\varepsilon > 0\), there are countably many disjoint open intervals \(I_i\) such that
\[
\mu_{\text{out}}(S) \leq \sum_i l(I_i) < \mu_{\text{out}}(S) + \varepsilon.
\]
Since we can make \(\varepsilon\) smaller and smaller, (ii) follows.

Vice versa, the countable union of closed sets is Borel, and thus (i) implies Definition 8.2. \(\blacksquare\)

---

\(^3\)the outer measure is not an actual measure.
Finally, we can define a general measure as follows and show that it satisfies the above characteristics, if one limits the definition to measurable sets.

**Definition 8.4.** A measure $\mu$ is a non-negative function from the measurable sets to $[0, \infty]$ such that for very countable disjoint sequence of (measurable) sets $S_i$:

$$\mu\left(\bigcup_{i=1}^{\infty} S_i\right) = \sum_{i=1}^{\infty} \mu(S_i).$$

The Lebesgue measure assigns to each Lebesgue measurable set its outer measure. (Thus the measure of an interval equals its length.)

Thus $\mu$ is a function from the measurable sets to the positive reals and the measurable sets are constructed so that properties (2), (3), and (4) in Section 8.1 hold. We summarize this as follows.

**Corollary 8.5.** The Lebesgue measure $\mu$ on $\mathbb{R}$ or $\mathbb{R}/\mathbb{Z}$ satisfies the following properties

1) $\mu :$ measurable sets $\rightarrow [0, \infty].$

2) For any interval $I$: $\mu(I)$ equals the length of $I.$

3) $\mu$ is translation invariant.

4) For a countable collection of disjoint sets $A_i$: $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$

We remark that part (4) of this result implies that in general sub-additivity holds:

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i).$$

The reason is that (4) says 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 $A_i \subseteq A_i$, we have $\mu(A_i) \leq \mu(A_i)$. Hence the sub-additivity.

We need a some more technical terms. If we have a space $X$ and a collection $\Sigma$ of measurable sets, then the pair $(X, \Sigma)$ is called a measurable space. A function $f : X \rightarrow X$ is called measurable if the inverse image under $f$ of any measurable set is measurable. A triple $(X, \Sigma, \mu)$ 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 $f$ with respect to the
Lebesgue measure $\mu$ is written as

$$I = \int f \, d\mu.$$ 

Assume $f(x)$ is non-negative. To approximate the Lebesgue integral $I$, one 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 Figure 20). The Lebesgue integral itself is defined as limit (if it exists) of these. A (not necessarily non-negative) function $f$ is called integrable if $\int |f| \, d\mu$ exists and is finite. It turns out that the Lebesgue integral generalizes the Riemann integral we know from calculus (see exercise 8.6).

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 will suppress those details here. This means we will skip some proofs in the following sections. Luckily, most of these details matter very little for the purposes of this text. However, the student does well in remembering that care needs to be taken so that manipulations employed in proofs do not cause one to leave the category of measurable sets.

---

4 Recall that the Riemann integral is approximated by partitioning the domain of $f$, see Figure 20.
8.3. The Birkhoff Ergodic Theorem

The context here is that we have a measurable transformation $T$ from a measure space $(X, \Sigma, \mu)$ to itself. The situation is quite general. The measure $\mu$ is not necessarily the Lebesgue measure, but we will assume that it is a probability measure, that is: $\int_X d\mu = \mu(X) = 1$.

![Diagram](image)

Figure 21. The pushforward of a measure $\nu$.

**Definition 8.6.** (i) Let $F : X \to Y$ be a measurable transformation and $\nu$ a measure on $X$. The pushforward $F_*\mu$ of the measure $\nu$ is a measure on $Y$ defined as
$$ (F_*\mu)(B) := \mu(F^{-1}(B)), $$
for every measurable set $B$ in $Y$ (see Figure 21).

(ii) Let $T : X \to X$ be measurable. We say that $T$ preserves the (probability) measure $\nu$ if $T_*\nu = \nu$. That is to say, if for every measurable set $B$, $\mu(T^{-1}(B)) = \mu(B)$.

**Theorem 8.7 (Birkhoff or Pointwise Ergodic Theorem).** Let $T : X \to X$ be a transformation that preserves the probability measure $\mu$. 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)
$$ \langle f \rangle(Tx) = \langle f \rangle(x) \quad \text{and} \quad \int_X \langle f \rangle(x) d\mu = \int_X f(x) d\mu. $$

The proof of this theorem requires a substantial technical mastery of measure theory and we will omit it. There are many sources for ergodic theory where details can be found. One of the more accessible ones is [23].
Definition 8.8. A transformation $T$ of a measure space $X$ to itself is called ergodic (with respect to $\mu$) if it preserves the measure $\mu$ and if every $T$ invariant set has measure 0 or 1. (A set $S \subseteq X$ is called invariant if $T^{-1}(S) = S$.)

Corollary 8.9. A measure preserving transformation $T : X \to X$ is ergodic with respect to a probability measure $\mu$ 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.

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 time averages equal spatial averages. This is a very important result because, as we will see, spatial averages are often much easier to compute.

![Figure 22. The functions $\mu(X_-^c)$ and $\mu(X_+^c)$.](image)

Proof. By Theorem 8.7, $\langle f \rangle(x)$ is defined on a set of full measure. So let

$$X_-^- := \{ x \in X : \langle f \rangle(x) < c \} \quad \text{and} \quad X_+^+ := \{ x \in X : \langle f \rangle(x) > c \}.$$  

Replacing $x$ by an inverse image of $x$ does not change the value of $\langle f \rangle(x)$, and so $X_{\pm}^\pm$ are invariant sets. By the ergodic hypothesis, $\mu(X_-^-)$ must therefore be a step function, with an increasing step of height 1 occurring at some value $c_-$. Similarly, $\mu(X_+^+)$ a step function, with an decreasing step of height 1 occurring at some value $c_+$. See Figure 22.
8.3. The Birkhoff Ergodic Theorem

If \( c_- < c_+ \), then for any interval \([c_1, c_2] \in (c_-, c_+)\), then we obtain that for any interval \([c_1, c_2] \in (c_+, c_-)\), \( \mu(X_{c_1}^+) = \mu(X_{c_2}^-) = 1 \), 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, \( \int_X f(x) \, d\mu = \langle f \rangle(x) \) 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 \( 1_{X_1} \) be the function that is 1 on \( X_1 \) and 0 elsewhere. The time average \( \langle 1_{X_1} \rangle(x) \) is 1 or 0, depending on where the starting point \( x \) is. In either case, it is not equal to the spatial average \( \int_X 1_{X_1}(x) \, d\mu \in (0, 1) \).

One needs to be careful, because it can happen that a transformation is ergodic with respect to two different measures.

**Definition 8.10.** Two measures \( \mu \) and \( \nu \) are mutually singular if there is a measurable set \( S \) with \( \mu(S) = 1 \) and \( \nu(S) = 0 \), and vice versa.

**Corollary 8.11.** If \( T \) is ergodic with respect to two distinct probability measures \( \mu \) and \( \nu \), then those measures are mutually singular.

**Proof.** Choose \( f \) such that
\[
c_1 = \int_X f \, d\mu \neq \int_X f \, d\nu = c_2.
\]
By Corollary 8.9, the time average \( \langle f \rangle(x) \) must be \( c_1 \) for \( \mu \) almost every \( x \) and so the \( x \) for which the average is \( c_2 \) has \( \mu \) measure 0. The reverse also holds.

**Definition 8.12.** A transformation \( T \) of a measure space is uniquely ergodic if there is a unique measure with respect to which \( T \) is ergodic.

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 \( \mu \), then we understand the statistical behavior for almost all points with respect to \( \mu \). There may be other complicated behavior but this is “negligible” if you measure it with \( \mu \).
8.4. Examples of Ergodic Measures

In this section, we consider the piecewise linear map $T$ with derivative equal to 2, depicted in Figure 23. To fix our thoughts, we set $A = [0, 1]$ and $B = [1, 2]$. In this section, we will exhibit uncountably many invariant probability measures $\mu$ with respect to which $T$ is ergodic. Note that any two such measures must be mutually singular (Definition 8.10). This situation is by no means exceptional.

We start with the measure $\delta_0$ 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 23, for any set $S$

$$0 \in S \iff 0 \in T^{-1}(S).$$

Thus $\delta_0(S) = \delta_0(T^{-1}(S))$, that is: $\delta_0$ is $T$-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 8.8, $T$ is ergodic with respect to $\delta_0$. Let us check the conclusion of Corollary 8.9. 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 bounce around in the interval $[0, 1]$ or in the other interval, $[1, 2]$. Thus the sum of
the corollary will give something close to $\alpha$. But the integral $\int_X f(x) \, d\delta_0$ gives 0. What is going on? See this footnote $^5$.

The next example is the uniform measure $\mu_A$ 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 23, we see that the inverse image of an interval $J \subseteq A$ equals two intervals of half its length. This shows that $\mu_A$ is invariant under $T$. We will show in Chapter 9 that each $T$ invariant set has $\mu_A$ measure either 0 or 1, but here is a partial result.

**Proposition 8.13.** If $S \subset A$ is a $T$ invariant set such that $S^c = A \setminus S$ is not empty, then both $S$ and its complement $S^c$ must be dense in $A$.

**Proof.** Note that $T$ restricted to the interval $A = [0, 1]$ is just the doubling map. Observe also that the complement $S^c$ must also be $T$ invariant.

Suppose that $S^c$, contains an interval $J$ of positive length and choose an interval $I$ so that $I \cap S$ is not empty. Since $S$ is invariant, we have that for all $n > 0$, $T^{-n}(S)$ is contained in $S$. If we can show that these pre-images are dense in $A$, then they must intersect the interval $J$ and we have a contradiction.

The inverse image $T^{-1}(I)$ is:

$$T^{-1}(I) = \left( \{0.0\} \cup \{0.1\} \right) + 2^{-1}I,$$

where the expressions 0.0 and 0.1 are binary (base 2), so that 0.1 = $\frac{1}{2}$. Iterating this procedure, we get

$$T^{-2}(I) = \left( \{0.00\} \cup \{0.01\} \cup \{0.10\} \cup \{0.11\} \right) + 2^{-2}I,$$

Similarly, the $n$th iterate gives all the expressions in base of length $n$. This is a collection of $2^n$ regularly spaced copies of $2^{-n}I$. Clearly, the union of these over $n$ is dense and so must intersect $J$. $\blacksquare$

This result implies that if $S \subset [0, 1]$ is an invariant set and its complement in $[0, 1]$, $S^c$, is not empty, then neither can contain an interval. This is equivalent to the following.

$^5$The set $\{0, 1\}$ has measure 0 with respect to $\delta_0$. Corollary 8.9 tells us to neglect such sets. Thus we must take $x = 0$, and then the summation also gives 0.
Corollary 8.14. If $S \subset A$ is a $T$ invariant set containing an interval, then $S = A$.

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 8.9 again. Let $f$ be

$$f(x) = \begin{cases} 
0 & \text{if } x \in [0, \frac{1}{2}) \\
\alpha & \text{if } x \in [\frac{1}{2}, 1]
\end{cases}$$

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 $\nu_p$ 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 $\nu_p$, 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 24. So $\nu_p$ is a probability measure.

The same reasoning as in Proposition 8.13 shows that an interval $I$ consisting of points whose binary expansion starts with $a = a_1a_2\cdots a_\alpha$ has as pre-image the interval $I_0$ consisting of points whose expansion starts with $0a$ and $I_1$ where the expansion starts with $1a$.

$$\nu_p(A_0) + \nu_p(A_1) = pv_p(A) + (1 - p)v_p(A) = v_p(A),$$

and the measure $\nu_p$ is $T$ invariant.

This gives us an uncountable set of measure (namely one for each $p \in (0, 1)$) with respect to which $T$ is ergodic.

8.5. The Lebesgue Decomposition

The examples of invariant measures of Section 8.4 also help to illustrate the following fact [21] which we mention without proof (but see [5]).
Theorem 8.15 (Lebesgue Decomposition). Let $\mu$ be a given measure. An arbitrary measure $\nu$ has a unique representation as the sum

$$\nu = \nu_{ac} + \nu_d + \nu_{sc}.$$ 

where $\nu_{ac}$ is absolutely continuous with respect to the Lebesgue measure $\mu$, $\nu_d$ is a discrete measure, and $\nu_{sc}$ is singular continuous.

We now define these notions somewhat informally. A measure $\nu_{ac}$ is absolutely continuous with respect to $\mu$ 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 then implies that $\nu_{ac}$ has a non-negative, integrable density with respect to $\mu$. This means that if $\nu_{ac} \ll \mu$, we can write $d\nu_{ac} = \rho(x)d\mu$ (see [21]). The density $\rho$ is also called the Radon-Nikodym derivative of $\nu_{ac}$ (relative to $\mu$) 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 $\rho$ is the density of $\nu_{ac}$ (with respect to $\mu$). Often, $\mu$ 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, $\rho$ 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 $\int d\nu = 1$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure24.png}
\caption{The first two stages of the construction of the singular measure $\nu_p$. The shaded parts are taken out.}
\end{figure}
The discrete measure $\nu_d$ is concentrated on a finite or countable set of $\mu$-measure zero. The measure $\delta_0$ is an example of this.

Finally, the measure $\nu_p$ for $p \neq \frac{1}{2}$ is an example of a singular continuous measure with respect to the Lebesgue measure $\mu$. This is a measure that is singular with respect to $\mu$, but, still, single element sets $\{x\}$ that satisfy $\mu(\{x\}) = 0$ also have $\nu_p$-measure zero.

Recall that Corollary 8.11 says that if $p \neq q$ are two numbers in $[0, 1]$, then the measures $\nu_p$ and $\nu_q$ 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 8.16.** Let $p, q$ distinct numbers in $[0, 1]$. The measures $\nu_p$ and $\nu_q$ are mutually singular.

**Proof.** As we saw in Section 8.4, 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 $\nu_p$-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 $\nu_p$-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 $\nu_p$ measure zero. But those have full $\nu_q$ measure. ■

Note that the binary expansion of the $\nu_p$ typical (that is: in a subset having full measure) $x$ has on average a fraction of exactly $p$ ones.

### 8.6. Exercises

**Exercise 8.1.** Reformulate the counter example in Section 8.1 as a counter example in $\mathbb{R}$. (Hint: two numbers in $[0, 1]$ are equivalent if their difference is rational. Let $V$ a set the contains exactly one representative of each class. Let $R$ be the set of rationals in $[-1, 2]$. Then consider the union $\bigcup_{r \in R} V + r$. Show that it should have measure between 1 and 3.)
8.6. Exercises

Exercise 8.2. a) Show there is an open set in [0, 1] of arbitrarily small outer measure that contains all the rationals in [0, 1].
   b) Show there is a closed set in [0, 1] of measure greater than 1 − ε that contains only irrational numbers.

Exercise 8.3. a) Show that countable sets have Lebesgue measure zero. (Hint: use the Definition 8.4 and Corollary 8.5 (4).)
   b) What is the Lebesgue measure of the following sets: the rationals in [0, 1], the algebraic numbers in [0, 1], the transcendental numbers in [0, 1], and the irrational numbers in [0, 1]?

Exercise 8.4. Show that any open set O in R is a union of disjoint open intervals. (Hint: for every x ∈ O there is an open interval (a, b) ⊆ O that contains x. Now let α = inf{a : (a, b) ⊆ O , x ∈ (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.)

In the next exercise, we prove the following Lemma.

Lemma 8.17. 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 8.5. a) Show that the empty set has measure zero. (Hint: X and ∅ are disjoint. Use criterion (4) in Section 8.1.)
   b) Prove the 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 8.2.)
   c) Prove part (ii) of the lemma. (Hint: use equation 8.1.)

Exercise 8.6. Let X = [0, 1], E the set of irrational numbers in X, and µ the Lebesgue measure.
   a) Use exercise 8.3 to show that \( \int_E d\mu = 1 \). (Hint: approximate the Lebesgue integral as in Section 8.1.)
   b) Show that the Riemann integral \( \int_E dx \) is undefined. (Hint: look up the exact definition of Riemann integral)

Exercise 8.7. Construct the middle third set Cantor set \( C \subseteq [0, 1] \) in the following way (Figure 25). 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 1.11.
   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 1.20.)
Exercise 8.8. Construct the set \( C \subseteq [0, 1] \) in the same way as in exercise 8.7, but now at stage \( n \), take out (open intervals of) an arbitrary fraction \( m_n \in (0, 1) \) 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 \( e^{-\alpha}/(1-\alpha) \).)

We remark that Cantor sets with positive measure such as those in exercise 8.8 are sometimes called fat Cantor sets.

Exercise 8.9. 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 8.7) is a Borel set.

c) Show that the Cantor sets of exercise 8.8 are Borel sets.

d) Show the sets in (a), (b), and (c) are measurable.

e) Show that the complements of the sets in (d) are measurable.

Exercise 8.10. Construct the Cantor function \( c : [0, 1] \to [0, 1] \), also called Devil’s staircase as follows. See also exercise 8.7.

a) Start with stage 0: \( c(0) = 0 \) and \( c(1) = 1 \). At stage 1, set \( c(x) = \frac{1}{2} \) if \( x \in \left[ \frac{1}{3}, \frac{2}{3} \right] \).

b) At stage 2, set \( c(x) = \frac{1}{4} \) if \( x \in \left[ \frac{1}{9}, \frac{2}{9} \right] \) and \( c(x) = \frac{3}{4} \) if \( x \in \left[ \frac{7}{9}, \frac{8}{9} \right] \).

c) Use a computer program to draw 5 or more stages. \( c(x) \) is the continuous function that is the limit of this process.

Exercise 8.11. See exercise 8.10 for the definition of the Cantor function, \( c(x) \).

a) Use exercise 8.7 (a) to show that for \( x \) in the Cantor set

\[ x = \sum_{i=1}^{\infty} a_i 3^{-i} \implies c(x) = \sum_{i=1}^{\infty} a_i 2^{-i}. \]

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.)
8.6. Exercises

Since $c$ is increasing, we can interpret it as a cumulative distribution function. The measure $\mu$ 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 8.12.** Find the Lebesgue decomposition (Theorem 8.15) of $c$ in exercise 8.11 interpreted as a measure. Explain!

**Exercise 8.13.**

a) Show that the derivative $c'$ of $c$ of exercise 8.10 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 8.2.*)

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 8.13 holds in the case where the function $c$ admits a derivative everywhere.

**Exercise 8.14.** 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 $t$.


c) Use (b) to show that if an invariant set contains an interval, then it equals $[0,1]$.

d) Show that the frequency with $t^n(x)$ visits the interval $I$ equals the frequency with which 358 occurs (if that average exists).

e) Assuming ergodicity, 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 8.15.** 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 8.16.  a) Show that there exist \( x \) in whose decimal expansion the word “358” occurs more often than in almost all other numbers (see exercise 8.14 (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 measure of the set of numbers referred to in (a) and (b). (Hint: use Birkhoff’s theorem and its corollary.)

Exercise 8.17.  a) Fix \( b > 1 \) and let \( w \) be any finite word in \([0, 1, \ldots, b - 1]^n\) 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: follow the reasoning in exercise 8.14.)
b) The measure of the set of \( x \) for which that frequency is not \( b^{-n} \) is zero.

Definition 8.18.  Let \( b > 2 \) an integer. A real number in \([0, 1]\) is called normal in base \( b \) 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 absolutely normal if the property holds for every integer \( b > 2 \).

Exercise 8.18.  Use exercise 8.17, Corollary 1.21, and Lemma 8.17 to show that the set of words not normal in base \( b \) has measure 0.

Exercise 8.19.  Show that the set of absolutely normal numbers has full measure. (Hint: follow the reasoning of exercise 8.18.)

Exercise 8.20.  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 8.7.)
b) Repeat (a), but now for base 2. (Hint: rewrite in base 4 with digits 00, 01, 10, and 11; follow (a).)

Exercise 8.21.  a) Show that the set of absolutely normal numbers is dense. (Hint: follows from exercise 8.19.)
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: pick any number and approximate it.)

Exercise 8.22.  Show that a rational number is non-normal in any base. (Hint: generalize proposition 5.6 to show that the expansion of a rational number in base \( b \) is eventually periodic.)
8.6. Exercises

Exercise 8.23. a) In base 2, construct a number whose expansion is the list of all finite words. Start with all length 1 words: “01”. Then obtain all length 2 words by concatenating first a “0”, then a “1”, so you get “010011011”. And so forth.

b) Show that the number 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 $2^n$ in every “level” at least $n$.)

Definition 8.19. A sequence $\{a_i\}_{i=1}^\infty$ is equidistributed modulo 1 if for each subinterval $[a, b]$ of $[0, 1]$

$$\lim_{n \to \infty} \frac{|\{a_1, a_2, \ldots, a_n\} \cap [a, b]|}{n} = b - a.$$ 

Exercise 8.24. Show that $x$ is normal in base $b > 2$ in $\mathbb{N}$ if and only if the sequence $a_n = \{xb^n\}$ is equidistributed modulo 1, where $\{\cdot\}$ means fractional part.

As with so many issues in number theory, for any of the numbers we care about — such as $e$, $\pi$, $\sqrt{2}$, et cetera — it is not known whether they are normal in any base.

Exercise 8.25. Show that a rotation on $\mathbb{R}/\mathbb{Z}$ preserves the Lebesgue measure. (Hint: Corollary 8.5 (iii).)
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 \(\mathbb{R}/\mathbb{Z}\)).

9.1. 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. In most cases, and certainly in this text, we are interested in invariant measures \(\nu\) that are absolutely continuous with respect to the Lebesgue measure (see Section 8.3). Thus \(d\nu = \rho(x)dx\). It is often easier to compute with densities than it is with measures.

Recall the pushforward of a measure (Definition 8.6). We need to reformulate it for densities.
Lemma 9.1. The pushforward \( \tilde{\rho} \) by \( T \) of a density \( \rho \) is given by

\[
\tilde{\rho}(y) = \sum_{T(x) = y} \frac{\rho(x)}{|T'(x)|}.
\]

This is called the Perron-Frobenius operator.

Proof. The measure of the pushforward \( \tilde{\rho} \) contained in the small interval \( dy \) is \( \tilde{\rho}(y)dy \). By Definition 8.6, it is equal to \( \sum_{T(x) = y} dx \) where \( dx \) is the length of the interval \( T^{-1}(dy) \) (see Figure 26). Now the length of \( T^{-1}(dy) \)

\[
\frac{d}{dy} T^{-1}(y) \bigg| dy = \frac{dy}{|T'(x)|},
\]

the result follows.

Thus \( T \) preserves an absolute continuous measure with density \( \rho \) if and only if

\[
(9.1) \quad \rho(y) = \sum_{T(x) = y} \frac{\rho(x)}{|T'(x)|}.
\]

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 9.1 also yields 1 for its pushforward \( \tilde{\rho} \), and thus equation (9.1) is satisfied. If instead \( T \) is defined as \( x \rightarrow \tau x \) modulo 1, where \( \tau \) is any integer other than \( \pm 1 \) or 0, the situation is different, but still not very complicated. We will call these transformation angle multiplications for short. Now each \( y \) has
9.1. Invariant Measures

|τ| inverse images \{x_1, \cdots, x_{|\tau|}\} and \(T'(x) = \frac{1}{x}\). So if \(\rho(x) = 1\), Lemma 9.1 yields \(\rho(x) = 1\) for the pushforward.

The situation is slightly more complicated for the Gauss map of Definition 6.1.

**Proposition 9.2.**

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

\[ \nu([0, x]) = \frac{1}{\ln 2} \int_0^x \frac{1}{1 + x} \, ds = \frac{1}{\ln 2} \ln(1 + x), \]

so \(\nu([0, 1]) = 1\) and \(\nu\) is as probability measure. It is easy to check that (see also Figure 9) that the inverse image under \(T\) of \([0, x]\) is the union of the intervals \([\frac{1}{a+1}, \frac{1}{a}]\), and so

\[ \nu(T^{-1}(\{0, x\})) = \nu(\bigcup_{a=1}^{\infty} \left[\frac{1}{a+1}, \frac{1}{a}\right]) \]

\[ = \frac{1}{\ln 2} \sum_{a=1}^{\infty} \left\{ \ln \left(\frac{a+1}{a}\right) - \ln \left(\frac{a+1+x}{a+x}\right) \right\} \]

\[ = \frac{1}{\ln 2} \sum_{a=1}^{\infty} \left\{ \ln \left(\frac{a+1}{a}\right) - \ln \left(\frac{a+1+x}{a+x}\right) \right\} \]

\[ = \frac{1}{\ln 2} \ln(1 + x). \]

The last equality follows because the sum telescopes.

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 on any interval \((x, y)\) is invariant. Therefore, the same is true for any open set (see 8.4). Thus it is true any Borel set (Definition 8.1). Since measurable can be approximated by Borel sets (Proposition 8.3), the result follows. \(\blacksquare\)

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 (9.1). But to do that, you first need to know a tricky sum, see exercises 9.6 and 9.7.

With the invariant measures in hand, we can now turn to proving the ergodicity of the three maps starring in this chapter.

9.2. The Lebesgue Density Theorem

**Proposition 9.3.** Given a measurable set $E \subseteq [0,1]$ with $\mu(E) > 0$, 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 $A$ in $I$ is greater than $1 - \varepsilon$.

**Proof.** By Proposition 8.3, there are open sets $O_n$ containing $E$ such that $\mu(O_n \setminus E) = \delta_n$, where $\delta_n$ tends to 0 as $n$ tends to infinity. Using property (4) of Corollary 8.5, we see that

$$\mu(O_n) = \mu(O_n \setminus E) + \mu(E) = \mu(E) + \delta_n.$$ 

According to exercise 8.4, for each $n$, there is a collection of disjoint open intervals $\{I_{n,i}\}$ such that

$$O_n = \cup_i I_{n,i}.$$ 

Now 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

$$\mu(E \cap O_n) = \mu(E \cap (\cup_i I_{n,i})) = \sum_i \mu(E \cap I_{n,i}) \leq \sum_i (1 - \varepsilon)\mu(I_{n,i}).$$

The middle equality follows again from property (4) of Corollary 8.5. Notice that the left hand equals $\mu(E)$, since $O_n$ contains $E$, and the right hand equals $(1 - \varepsilon)\mu(O_n)$ by definition of the intervals $I_{n,i}$. Together with equation (9.2), this gives

$$\mu(E) = (1 - \varepsilon)\mu(O_n) = (1 - \varepsilon)(\mu(E) + \delta_n).$$

If $n$ tends to infinity, $\delta_n$ tends to 0, and thus $\mu(E)$ must be 0. \hfill \square

This is a weak version of a much better theorem. We will 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 [30].
Theorem 9.4 (Lebesgue Density Theorem). If $E$ is a measurable set in $\mathbb{R}^n$ with $\mu(E) > 0$, then for almost all $x \in E$

$$\lim_{\varepsilon \to 0} \frac{\mu(E \cap B_\varepsilon(x))}{\mu(B_\varepsilon(x))} = 1.$$ 

That is: this holds for all $x$ in $E$, except possibly for a set of $\mu$ measure 0.

9.3. Rotations and Multiplications on $\mathbb{R}/\mathbb{Z}$

In this section, we will use a strategy involving 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 $\omega$ modulo 1 on $\mathbb{R}/\mathbb{Z}$. We denote the Lebesgue measure by $\mu$.

Lemma 9.5. Every orbit of an irrational rotation $R_\omega$ 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 $\omega$ (of Definition 6.4). By Lemma 6.12

$$\lim_{n \to \infty} x + q_n\omega - p_n = x.$$ 

Fix $n$ be big enough enough, so that the distance (on the circle) between $x$ and $x + q_n\omega - p_n$ is less than $\delta$. Then the points $x_i := x + iq_n\omega$ modulo 1 advance (or recede) by less than $\delta$. And thus at least one must land in the stipulated interval (see Figure 27).

![Figure 27](image)

Figure 27. $r$ 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 9.6. Irrational rotations modulo 1 are ergodic with respect to the Lebesgue measure.
Proof. By Proposition 9.2, 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 9.3, for every $\varepsilon$ 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 9.5, 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 28). In all cases, this 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 (see exercise 9.4).

In the proof of the next theorem, we employ the same strategy as in the proof of Proposition 8.13 and Corollary 8.14. But this time, the Lebesgue density theorem helps us get a much stronger result.

**Theorem 9.7.** Multiplication by $\tau \in \mathbb{Z}$ with $|\tau| > 1$ modulo 1 is ergodic.

**Proof.** By Proposition 9.2, the Lebesgue measure is invariant.

Suppose that the set $A$ is invariant and has positive measure. For any $\varepsilon > 0$, we can find an interval $J$ in which $A$ has density at least $1 - \frac{\varepsilon}{2}$. We now cover $J$ by intervals of the form $\left[ \frac{k}{n}, \frac{k+1}{n} \right]$. These intervals form a larger interval $C$: $J \subseteq C$. If we take $n$ large enough, $\mu(C \setminus J)$ will be very small, and so the density of $A$ in $C$ will be at least $1 - \varepsilon$.

There will be at least one interval $I = \left[ \frac{k}{n}, \frac{k+1}{n} \right]$ where the density of $A$ is at least equal to the average, $1 - \varepsilon$. But this interval is an inverse image of
9.3. Rotations and Multiplications on $\mathbb{R}/\mathbb{Z}$

$[0, 1]$ under an affine branch of $T^n$. Thus the density of $A$ in $[0, 1]$ is at least $1 - \varepsilon$. Since $\varepsilon > 0$ is arbitrary, it follows that $A$ must have full measure. ■

These theorems have interesting consequences. The most important one for rotations is the following.

**Corollary 9.8.** For $\omega$ irrational, the sequence $\{R_\omega^i(x)\}_{i=1}^\infty$ is equidistributed (see Definition 8.19) modulo 1 for every $x$.

**Proof.** Define $f : [0, 1) \to [0, \infty)$ as $f(x) = 1$ if $x$ is in the interval $[a, b]$ and 0 else. Note that $R_\omega$ is ergodic by Theorem 9.6 with respect to the Lebesgue measure. By Corollary 8.9 to $f$, for almost all $x$

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x)) = \int_X f(x) \, dx = b - a.$$  

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. ■

The principal consequence of the ergodicity of multiplication is the absolute normality (see Definition 8.18) of almost all numbers. This was discussed at length in the exercises of Chapter 8.

There is an extension of Theorem 9.7 that will be useful in the next section.

**Corollary 9.9.** Let $\{I_i\}$ be a finite or countable partition of $[0, 1]$ of intervals of positive length $\ell_i$ 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 $[0, 1]$. Let $T = \cup_i f_i$. Then $T$ preserves the Lebesgue measure and is ergodic with respect to that measure.

**Proof.** By hypothesis we have $\sum_i |f_i|^{-1} = \sum_i \ell_i = 1$ and so the Perron-Frobenius equation (9.1) immediately implies that the Lebesgue measure is preserved.

Note that $T^n$ is piecewise affine and, since $|f_i'| \geq c > 1$, each branch of $T^n$ maps an interval of size less than $c^{-n}$ onto $[0, 1]$. Denote these intervals by the $n$th level intervals. As in the proof of Theorem 9.7, assume there is an invariant set $A$ of positive measure. We can then again construct for any positive $\varepsilon$ an $n$th level interval $I$ such that $A$ has density at least $1 - \varepsilon$. Since
that is an affine inverse image of $[0,1]$, $A$ must have density at least $1 - \varepsilon$ on all of $[0,1]$.

9.4. 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 9.2, we know the invariant measure. It might seem that Corollary 9.9 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 $A$ is an invariant set of positive measure. Just like before, for every $\varepsilon > 0$ we can take $n$ big enough so that there an $n$th level interval $I$ where $A$ has density $1 - \varepsilon$. This interval $I$ is of course an inverse image of $[0,1]$ under $T^n$. Thus there is a branch of $T^n$ that maps $I$ to $[0,1]$ just as before. What is the problem? That branch is not affine. It might have bigger derivative in $A^c \cap I$ than it does in $A \cap I$. That could distort the image under $T^n$ in such a way that it changes dramatically the proportion between the measure of $A^c$ and $A$ in $[0,1]$. There is no way to tell, because we do not even know what $A$ is. The solution lies in controlling that distortion. If we can prove that for that particular branch $\left| \frac{\partial T^n(s_0)}{\partial x_{[0,1]}} \right|$ is bounded independent of $n$ by, say, $K$, then the argument of the proof Proposition 9.2 and gives that a small interval with the density of $A$ being greater than $1 - \varepsilon$ must map to a large interval with density at least $1 - K\varepsilon$. Since we can let $\varepsilon$ as small as we want, the set $A \cap [0,1]$ must have measure 1.
9.4. The Return of the Gauss Map

The exposition in the remainder of this section and the next closely follows [36].

**Definition 9.10.** Let \( I_0 \) be an interval. The distortion \( D \) of \( T^n \) on that interval is defined as

\[
D := \sup_{x_0, y_0 \in I_0} \left| \ln \frac{\partial T^n(x_0)}{\partial T^n(y_0)} \right|.
\]

Here, \( \partial \) stands for the derivative with respect to \( x \).

**Proposition 9.11.** Let \( T \) be the Gauss map. The distortion of \( T^n \) on any \( n \)th 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

\[
D \leq \sum_{i=0}^{n-1} \sup_{x_i, y_i \in I_i} |\ln |\partial T(x_i)| - \ln |\partial T(y_i)||.
\]

By the mean value theorem, there are \( z_i \in I_i \) such that the right hand 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 \( |\frac{\partial^2 T}{\partial T}| \). Furthermore, the mean value theorem (once again) gives \( |I_i| = \frac{1}{\partial T(u_i)} \) for some \( u_i \in I_i \). Substituting this into the last equation, we get

\[
(9.3) \quad D \leq \sum_{i=0}^{n-1} \sup_{z_i, u_i \in I_i} \left| \frac{\partial^2 T(z_i)}{\partial T(z_i) \partial T(u_i)} \right| \cdot |I_{i+1}|.
\]

We need to estimate the two expressions in the right hand. 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 \( \left( \frac{1}{a_i+1}, \frac{1}{a_i} \right) \) depicted in figure 9, where \( a_i \) the continued fraction coefficient associated with that particular branch. For that branch, we have

\[
\left| \frac{\partial^2 T(z_i)}{\partial T(z_i)} \right| \leq 2(a_i + 1) \quad \text{and} \quad \left| \frac{1}{\partial T(u_i)} \right| \leq \frac{1}{a_i^2}.
\]
Next we estimate the length on $n$th level interval $|I|$. In figure 9, 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 (9.3) gives

$$D \leq \sum_{i=0}^{n-1} \frac{2(a_i + 1)}{a_i^2} Kd^{(n-i-1)/2}.$$  

Since $a_i \in \mathbb{N}$, this tells us that the expression in (9.3) is uniformly bounded in $n$. 

As explained in the introduction to this section, our main result follows immediately.

**Corollary 9.12.** The Gauss map is ergodic with respect to $d\nu = \frac{1}{\ln 2} \frac{dx}{1+e^x}$.

### 9.5. 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 8 (Definition 8.18). But the real pay-off is in understanding some basic properties of the continued fraction of "typical" real numbers. This is what we do in this section.

In this section, $T$ denotes the the Gauss transformation and $\nu$ its invariant measure (see Proposition 9.2) while $\mu$ will denote the Lebesgue measure. Note that a set has $\mu$ measure zero if and only if it has $\nu$ measure zero (exercise 9.1). For the continued fraction coefficients $a_n$ and the continued fraction convergents $p_n/q_n$, see Definition 6.4.

We start with a remarkable result that says that the arithmetic (usual) mean of the continued fraction coefficients diverges for almost all numbers, but their geometric mean is almost always well-defined.

**Theorem 9.13.** For almost all numbers $x$, the continued fraction coefficients $a_n = a_n(x)$ satisfy:

i) $\lim_{n \to \infty} \frac{a_1 + \ldots + a_n}{n} = \infty$ and

ii) $\lim_{n \to \infty} (a_1 \cdot \ldots \cdot a_n)^{1/n} = \prod_{a=1}^{\infty} \left(1 - \frac{1}{(1+a)^2}\right)^{-\log_2 a} < \infty$. 

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

For $a \in \{1, \cdots, k\}$ : \quad $f_k(x) = a$ if $x \in \left(\frac{1}{a+1}, \frac{1}{a}\right]$

\[ f_k(x) = 0 \text{ if } x \in \left[0, \frac{1}{k+1}\right]. \]

Denote the pointwise limit by $f_\infty$. We really want to use Corollary 8.9 to show that the “time average”

\[ \lim_{n \to \infty} \left( \frac{a_1 + \cdots + a_n}{n} \right) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} f_\infty(T^i(x)) \]

is unbounded. But $f_\infty$ is not integrable and so cannot be used. However this sum is bounded from below by the right hand if we replace $f_\infty$ by $f_k$ (which is integrable). Proposition 9.2 and Corollary 8.9 say that the time average of $f_k$ equals

(9.4) \[ \frac{1}{\ln 2} \int_0^1 f_k(x) \frac{1}{1+x} dx = \frac{1}{\ln 2} \sum_{a=1}^k \int_{\frac{1}{a+1}}^{\frac{1}{a}} \frac{1}{1+x} \, dx \]

The integral of $1/(1+x)$ is of course $\ln(1+x)$ and so the above gives

(9.5) \[ \frac{1}{\ln 2} \sum_{a=1}^k a \left( \ln \left( \frac{a+1}{a} \right) - \ln \left( \frac{a+2}{a+1} \right) \right). \]

This sum telescopes and the student should verify (see exercise 9.14) that this gives

(9.6) \[ \frac{1}{\ln 2} \left( \ln(k+1) - k \ln \left( 1 + \frac{1}{k+1} \right) \right), \]

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”

\[ \lim_{n \to \infty} \left( \frac{\ln a_1 + \cdots + \ln a_n}{n} \right). \]

The exponential of this will give us the result we need. So this time, we define

(9.7) \[ g_\infty(x) = \ln a \text{ if } x \in \left(\frac{1}{a+1}, \frac{1}{a}\right], \]

For $a \in \mathbb{N}$. 

This time around, $g_\infty$ is $\nu$-integrable (as we will see below) and we get
\[(9.8) \quad \frac{1}{\ln 2} \int_0^1 \frac{g_\infty(x)}{1 + x} \, dx = \sum_{a=1}^{\infty} \frac{\ln a}{\ln 2} \left( \ln \left( \frac{a+1}{a} \right) - \ln \left( \frac{a+2}{a+1} \right) \right). \]
(Note that $\frac{\ln a}{\ln 2} = \log_2 a$.) Since we can write
\[(9.9) \quad \ln \left( \frac{a+1}{a} \right) - \ln \left( \frac{a+2}{a+1} \right) = -\ln \left( 1 - \frac{1}{(a+1)^2} \right), \]
we finally get the result (as well as the assertion that $g_\infty$ is $\nu$-integrable) by taking the exponential of the sum in (9.8).

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 $n = 2^k$, we set $a_{2^k} = 2^{2k}$. For $n = 2^k$, we have
\[
\frac{a_1 + \ldots + a_n}{n} = \frac{a_n}{n} = 2^{2k} - 2k,
\]
which clearly diverges as $k \to \infty$. Meanwhile, the geometric average at that point is (after taking the logarithm):
\[
\frac{\ln a_1 + \ldots + \ln a_n}{n} = \frac{\sum_{j=1}^{k} \ln 2}{2^k} = \frac{2^{k+1} - 1}{2^k} \ln 2.
\]
The latter converges to 0, which makes the geometric average 1.

**Theorem 9.14.** For almost all numbers $x$, the denominators $q_n = q_n(x)$ of the convergents satisfy $\lim_{n \to \infty} \frac{\ln q_n}{n} = \frac{\pi^2}{12 \ln 2}$. The constant $\frac{\pi^2}{12 \ln 2} \approx 1.1866 \ldots$ is called Lévy’s constant.

**Proof.** To simplify notation in this proof, we will write $x_i := T^i(x_0)$ where $T$ is the Gauss map. For the $n$th 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
\[
p_n(x_0) = \frac{1}{a_1(x_0) + p_{n-1}(x_1)/q_{n-1}(x_1)} = \frac{q_{n-1}(x_1)}{a_1(x_0)q_{n-1}(x_1) + p_{n-1}(x_1)}.
\]
See also exercise 9.2 (a). By Corollary 6.8 (ii), $\gcd(p_n, q_n) = 1$, and so from exercise 9.2 (b) we see that $p_n(x_0)$ equals $q_{n-1}(x_1)$. More generally, we have by the same reasoning
\[
p_n(x_j) = q_{n-1}(x_{j-1}).
\]
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
\[
-\frac{1}{n} \ln q_n(x_0) = \frac{1}{n} \sum_{i=0}^{n-1} \ln x_i - \frac{1}{n} \sum_{i=0}^{n-1} \left( \ln x_i - \ln \frac{p_{n-i}(x_i)}{q_{n-i}(x_i)} \right).
\]

Two more steps are required. The first is showing that the last sum is finite. This not difficult, because
\[
\frac{1}{n} \sum_{i=0}^{n-1} \ln q_{n-i}(x_i) = \frac{1}{n} \sum_{i=0}^{n-1} \ln \left( 1 + \frac{(q_{n-i}(x_i)x_i - p_{n-i}(x_i))}{p_{n-i}(x_i)} \right).
\]

Corollary 6.7 or, more precisely, exercise 6.16 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 9.12).

Since the above sum is bounded and \( x_i = T^i(x_0) \), we now divide by \( n \) and 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 of this expression. Naturally, the ergodicity of the Gauss map invites us to employ Birkhoff’s theorem in the guise of Corollary 8.9 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 9.19.

9.6. Exercises

Exercise 9.1. Show that for a set \( A: \mu(A) = 0 \) (Lebesgue measure) if and only if \( \nu(A) = 0 \) (invariant measure of the Gauss map). (Hint: write both equalities in terms of Lebesgue integrals.)
9. Three Maps and the Real Numbers

Exercise 9.2. To reacquaint ourselves with continued fractions, consider

\[
x_0 = \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}} \overset{\text{def}}{=} [a_1, a_2, a_3, \ldots].
\]

a) Show that \(\lfloor x_0^{-1} \rfloor = a_1\) and that

\[
T(x_0) = x_0^{-1} - a_1 = \frac{1}{a_2 + \frac{1}{a_3 + \cdots}} \overset{\text{def}}{=} [a_2, a_3, \ldots].
\]

b) Show that if \(\gcd(p, q) = 1\), then \(\gcd(p + aq, q) = 1\). (Hint: use Lemma 2.6.)

Exercise 9.3. a) Show that every probability density \(\rho\) 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 8.5 — 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 \(x \to nx \text{ modulo } 1\) (where \(n \in \mathbb{N}\)).

Exercise 9.4. At the end of the proof of Theorem 9.6, assume that \(|I| = |J|\) and complete the proof in that case.

a) Show that for every \(\varepsilon > 0\), there is \(i\) such that \(R_{\omega}^i(I)\) falls in an \(\varepsilon\)-neighborhood of \(J\).

b) Estimate the fraction of \(J\) that must be in \(A\).

c) Show that this gives a contradiction.

Exercise 9.5. Let \(R_0\) be identity on \(\mathbb{R}/\mathbb{Z}\).

a) Show that for any \(x\), the delta measure \(\delta_x\) is an invariant measure for \(R_0\), and that \(I\) is ergodic with respect to that measure.

b) Show that for any of the invariant measure in exercise 9.3 (a), \(R_0\) is not ergodic.

c) Show that \(R_{\omega}\) is not ergodic with respect to any of the measures of exercise 9.3 (c).

Exercise 9.6. Show that \(\sum_{i=1}^{\infty} \frac{1}{(x+i)(x+i+1)} = \frac{1}{x+1}\) for all \(x \in \mathbb{R}\) except the negative integers. (Hint: use partial fractions, then note that the resulting sum telescopes.)

Exercise 9.7. Prove that the Gauss map preserves the measure of Proposition 9.2 via equation (9.1). Do not use the computation in the proof of that proposition. (Hint: use exercise 9.6.)
9.6. Exercises

Exercise 9.8. Show that \( \rho(x) = 1 \) is the only continuous invariant density of an irrational rotation \( R \). (Hint: if \( \rho \) is invariant under \( R \), it must be invariant under \( R^i \) for all positive \( i \). Use Lemma 9.5.)

Exercise 9.9. a) Show that \( \rho(x) = 1 \) is the only continuous invariant density for the angle doubling map. (Hint: use a reasoning similar to that of Proposition 8.13.)
b) Check that the same is true for the map \( x \rightarrow \tau x \mod 1 \) where \( \tau \in \mathbb{Z} \) and \( \tau > 1 \).

Exercise 9.10. The orbit of any irrational rotations is uniformly distributed. So why do we encounter specifically the golden mean in phyllotaxis — the placement of leaves? Research this and add illustrations.

Exercises 9.11 and 9.12 discuss some very useful properties for reference. In fact, they are useful in a much wider context than discussed here. For instance, exercise 9.11 comes up in any discussion of entropy [11] or in deciding the stability of Lotka-Volterra dynamical systems [32]. Exercise 9.12 is important for deciding the convergence of products of the form \( \prod (1 + x_i) \).

Exercise 9.11. a) Show that if \( x > -1, \) then \( \ln(1 + x) \leq x \) with equality iff \( x = 0 \). (Hint: draw the graphs of \( \ln(1 + x) \) and \( x \).)  
b) Let \( p_i \) and \( q_i \) positive and \( \sum p_i = \sum q_i \). Use (a) to show that \( \sum p_i \ln p_i \geq \sum q_i \ln q_i \). (Hint: \( -\sum p_i (\ln p_i - \ln q_i) = \sum p_i \ln \frac{p_i}{q_i} \leq -\sum q_i - p_i \) by (a)).
c) Let \( S_n \) be the open \( n \)-dimensional simplex \( p_i > 0 \) and \( \sum p_i = K \). Show that \( h : S_n \rightarrow \mathbb{R} \) given by \( h(p) = -\sum p_i \ln p_i \) has a single extremum at \( p_i = \frac{1}{n} \). (Hint: The constraint is \( C = \sum p_i \) must be equal to 1. Deduce that at the maximum, the gradients of \( h \) and \( C \) must be parallel.)
d) Show that this extremum is a maximum. (Hint: set \( f(x) := -x \ln x \) and show that \( f''(x) < 0 \). As a consequence, if \( w_i \) are positive weights such that \( \sum w_i = 1 \), we have Jensen's inequality or \( f(\sum w_i p_i) \geq \sum w_i f(p_i) \). See Figure 30.)

In the next exercise, we show this lemma.

Lemma 9.15. Suppose that \( x_n > -1 \) and \( \lim_{n \to 0} x_n = 0 \). Then \( \sum x_n \ln(1 + x_n) \) converges absolutely if and only if \( \sum x_n \) converges absolutely. Also \( \sum x_n \ln(1 + x_n) \) diverges absolutely if and only if \( \sum x_n \) diverges absolutely.
Exercise 9.12.  
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 \( \exists a > 0 \) such that for all \( n \) large enough \( |\ln(1+x_n)| \leq b|x_n| \). (Hint: use the direct comparison text.)

c) From (a), conclude that if \( x_n > -1 \) and \( \lim_{n \to 0} x_n = 0 \), then \( \exists b > 0 \) such that for all \( n \) large enough \( |x_n| \leq b|\ln(1+x_n)| \).

d) Show that (b) and (c) imply Lemma 9.15.

The next four exercises provide some computational details of the proof of Theorem 9.13.

Exercise 9.13. Compute the frequency with which \( a_n(x) = a \) occurs. (Hint: set \( f(x) = 1 \) on \((1/(1+a), 1/a]\). Then use Birkhoff.)

a) Show that the right hand of (9.4) gives (9.5). 
b) Show that (9.5) gives (9.6). (Hint: write out the first few terms explicitly.)
c) Use exercise 9.11 (a) to bound the second term of (9.6).
d) Conclude that (9.6) is unbounded.

Exercise 9.15.  
a) Show the equality in (9.8) holds. 
b) Show the equality in (9.9) holds. 
c) Show that (9.9) implies part (ii) of Theorem 9.13.
Exercise 9.16. a) Show that instead of (9.9), 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 9.11 (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 (9.8) 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 \(g_\infty\) 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 5.)

Exercise 9.17. Use exercise 9.16 (a) to show that Khinchin’s constant equals \(\prod_{a=1}^\infty \left(1 + \frac{1}{(1 + a)^2}\right) \log_2 a\).

Exercise 9.18. Let \(\nu\) be absolutely continuous with respect to the Lebesgue measure \(\mu\). Show that if a set has full \(\mu\) measure then it has full \(\nu\) measure.

Figure 31. Plot of the function \(\ln(x) \ln(1 + x)\).
Exercise 9.19.  a) Show that \( \lim_{x \to 0} \ln(x) \ln(1 + x) = 0 \) (Figure 31). \( \text{(Hint: for the limit as } x \to 0, \text{ substitute } x = e^y, \text{ then use L'Hopital.)} \)

b) Use (a) to show that \( I := \int_0^1 \frac{\ln x}{1 + x} \, dx = -\int_0^1 \frac{\ln(1 + x)}{x} \, dx. \) \( \text{(Hint: integration by parts.)} \)

c) Show that \( \ln(1 + x) = \sum_{n=1}^{\infty} \frac{(-1)^n x^n}{n} \).

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 9.14. \( \text{(Observation: we sure took the cowardly way out in this last step; to really work out that last sum from first principles is elementary but very laborious. The interested student should look this up on the web.)} \)

In exercise 9.19, note the curious fact that \( \sum_{n=1}^{\infty} (-1)^n n^{-2} = \frac{\pi^2}{12} \) while from exercise 2.25 we have that \( \zeta(2) = \sum_{n=1}^{\infty} n^{-2} = \frac{\pi^2}{6} \).

Exercise 9.20.  a) Use exercise 6.16 and Theorem 9.14 to show that for almost all \( \omega \in [0, 1] \)

\[
\lim_{n \to \infty} \frac{1}{n} \ln |\omega - p_n/q_n| = \frac{\pi^2}{6 \ln 2}.
\]

b) What do you in (a) get if \( \omega \) is rational? Is that a problem?

Definition 9.16.  Given a one dimensional smooth map \( T : [0, 1] \to [0, 1] \), the Lyapunov exponent \( \lambda(x) \) at a point \( x \) is given by

\[
\lambda(x) := \lim_{n \to \infty} \frac{1}{n} \ln |D(T^n)(x)|,
\]

assuming that the limit exists. \( \text{(There is a natural extension of this notion for systems in dimension greater than or equal to 2, but we do not need it here.)} \)

Exercise 9.21.  What does Definition 9.16 tell you about how fast \( T^n x \) and \( T^n y \) separate if \( x \) is a typical point and \( y \) is very close to \( x \)?
9.6. Exercises

Exercise 9.22. Let $T$ be the Gauss map and $\mu$ its invariant measure. Show that the Lyapunov exponent at $x$ satisfies

$$\lambda(x) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} \ln |DT(J^j(x))| .$$

(Hint: think chain rule.)

b) Show that Birkhoff’s theorem (Corollary 8.9) implies that for almost all $x \in [0,1]$

$$\lambda(x) = \int_0^1 \frac{-2\ln x}{\ln 2(1+x)} dx .$$

(Hint: refer to exercise 9.18.)

c) Use the last part of the proof of Theorem 9.14 and exercise 9.19 to show that for almost all $x$, the Lyapunov exponent equals $\frac{\pi^2}{6\ln 2}$.

Exercise 9.23. a) See exercise 9.22. Let $T$ be the Gauss map and $x = [n,n,\cdots]$. Determine the Lyapunov exponent at $x$. (Hint: see also exercise 6.3.)

b) Why are these exponents different from the one computed in exercise 9.22?

Exercise 9.24. Let $T$ be the map given in Corollary 9.9. 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 9.22.)

b) Show that the answer in (a) is greater than 1.

c) Show if the map has $n$ branches, then the Lyapunov exponent is extremal if all branches have the same slope. (Hint: exercise 9.11 (c).)

d) Show that this extremum is a maximum. (Hint: exercise 9.11 (d).)

Exercise 9.25. 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: Prime factorization.)

b) From now on, suppose $k \neq 10^r$. Show $T : x \to x + \log_{10} k$ modulo 1 is ergodic.

c) Let $f(x) = 1$ when $x \in [\log_{10} 7, \log_{10} 8]$ and 0 elsewhere. Compute $\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x))$.

d) Explain how often 1 through 9 occur in $\{k^i\}_{i=0}^\infty$ 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 9.25. 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 [7].
Chapter 10

The Cauchy Integral Formula

Overview. Again, we need to venture very far, apparently, from number theory to make progress. In the mid-nineteenth 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 Riemann zeta function to the complex plane. In this chapter, we develop the necessary complex analysis tools — essential the Cauchy integral formula — to study the convergence of a certain Laplace transform, which is the key to the proof of the prime number theorem.

10.1

10.2. Exercises
Chapter 11

The Prime Number Theorem

Overview. In 1850, it seemed that Chebyshev was awfully close to proving the prime number theorem (Theorem 2.22). 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 [28]. Even so, 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 [25]. We made heavy use of Zagier’s rendition of this proof [38] and of [31].

11.1. Introduction

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 2.20 and Proposition 2.21. In this chapter, we will frequently encounter sums of the form \( \sum_p \). For example see Definition 11.1 below. Such sums will always be understood to be over positive primes only. A similar convention holds for products.
11. The Prime Number Theorem

We now define a couple of new functions. It is not really necessary to know the names of all these functions to understand the arguments in this chapter. However, since the names are commonly used in the literature, it is convenient to know them.

**Definition 11.1.** The first Chebyshev function is given by

\[ \theta(x) := \sum_{p \leq x} \ln p. \]

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

\[ \psi(x) := \sum_{n \leq x} \Lambda(n). \]

In what follows, we need to integrate expressions like

\[ I(x) := \int_2^x f(t) d\theta(t). \]

If we partition the interval \([2, x]\) by \(2 = x_0 < x_1 \cdots x_n = x\), then \(I(x)\) can be approximated as

\[ I(x) = \sum_{i=1}^{n} f(c_i)(\theta(x_{i+1}) - \theta(x_i)), \]

where \(c_i \in (x_i, x_{i+1})\) and then the appropriate limit (assuming it exists) can be taken. This is well-defined as a Riemann integral (see [26] for Riemann integrals). Now, \(\theta(t)\) is constant except at the values \(t = p\) (a prime) where it has a jump discontinuity of size \(\ln p\). Thus

\[ (11.1) \quad I(x) = \sum_{p \geq x} f(p) \ln(p). \]

**Lemma 11.2.** We have for \(x \geq 2\)

\[ \pi(x) = 1 + \frac{\theta(x)}{\ln x} + \int_e^x \frac{\theta(t)}{(\ln t)^2} \, dt. \]
11.2. Chebyshev’s Theorem

Proof. First note that since 2 is the smallest prime, by equation (11.1), we have 
\( \pi(x) = \int_x^y \frac{d\theta(t)}{\ln t} \) for any \( y \in (1, 2) \). Since \( e \in (2, 3) \), we get 
\( \pi(x) = 1 + \int_2^e \frac{d\theta(t)}{\ln t} \).

Apply integration by parts to obtain 
\( \pi(x) = 1 + \frac{\theta(x)}{\ln x} - \int_2^x \frac{\theta(t)}{\ln t} \frac{1}{t} \) dt.

Working out the last term yields the lemma.

11.2. Chebyshev’s Theorem

We prove Theorem 11.6, an approximate version of the prime number theorem (Theorem 2.22). 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 5.20).

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 \( a^k \parallel b \) but not \( a^{k+1} \parallel b \).

Lemma 11.3. Let \( 0 < m < n \) and \( p \) prime. Suppose that \( p^k \parallel n! \). Then we have \( p^k \leq n \).

Proof. Let \( p \) prime and suppose that \( p^k \parallel n! \). We want to find \( k \). Any multiple \( ap \leq n \) contributes one factor to \( p^k \). The number of multiples \( ap \) less than \( n \) equals \( \lfloor \frac{n}{p} \rfloor \). So these contribute \( \lfloor \frac{n}{p} \rfloor \) 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 \( \lfloor \frac{n}{p^2} \rfloor \). Continuing like that, we find

\[ p^k \parallel n! \implies k = \sum_{j=1}^{\infty} \left\lfloor \frac{n}{p^j} \right\rfloor. \]

As a consequence, we have

\[ (11.2) \quad p^k \parallel \binom{n}{m} \implies k = \sum_{j=1}^{\infty} \left( \left\lfloor \frac{n}{p^j} \right\rfloor - \left\lfloor \frac{m}{p^j} \right\rfloor - \left\lfloor \frac{n-m}{p^j} \right\rfloor \right). \]

Consider the expression \( E = |x_1 + x_2| - |x_1| - |x_2| \). 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 (11.2), we
see that if \( p^j > n \), then the contribution is always zero. Thus 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 11.4.** For \( n \geq 2 \), we have \( \frac{2^n}{n+1} < \left( \frac{n}{\lceil \frac{n}{2} \rceil} \right) < 2^n \).

**Proof.** We prove the right hand first. From the binomial theorem (Theorem 5.20), we see that
\[
\left( \frac{n}{\lceil \frac{n}{2} \rceil} \right) < \sum_{i=0}^{n} \binom{n}{i} = 2^n.
\]
For the left hand, we note that \( \left( \frac{n}{\lceil \frac{n}{2} \rceil} \right) \) is the largest of the \( n+1 \) numbers \( \binom{n}{i} \) and so
\[
2^n = \sum_{i=0}^{n} \binom{n}{i} < (n+1) \left( \frac{n}{\lceil \frac{n}{2} \rceil} \right).
\]

**Lemma 11.5.** For \( n \geq 2 \) we have \( e^{\theta(n) - \theta(n/2)} \leq \left( \frac{n}{\lceil \frac{n}{2} \rceil} \right) \leq n^{\pi(n)} \), where the product is over primes only.

**Proof.** We start by noticing that any prime \( p \) in the interval \( \left( \frac{n}{2}, n \right] \) is a divisor of \( n! \) but not of the denominator of \( \left( \frac{n}{\lceil \frac{n}{2} \rceil} \right) \). Therefore \( p \) divides \( \left( \frac{n}{\lceil \frac{n}{2} \rceil} \right) \).
This implies that \( \prod_{\frac{n}{2} < p < n} p < \left( \frac{n}{\lceil \frac{n}{2} \rceil} \right) \). Noting that \( p = e^{\ln p} \) and inserting the definition of \( \theta(x) \) (Definition 11.1) yields the left hand of the formula.

For the right hand, use unique factorization (Theorem 2.15) and the definition of \( \pi(n) \) to write
\[
\left( \frac{n}{\lceil \frac{n}{2} \rceil} \right) = \prod_{i=1}^{\pi(n)} p_i^{k_i}.
\]
By Lemma 11.3, \( p_i^{k_i} \leq n \). Thus \( \prod_{i=1}^{\pi(n)} p_i^{k_i} < n^{\pi(n)} \), which yields the right hand.

**Theorem 11.6 (Chebyshev’s Theorem).** There are constants \( a < 1 \) and \( b > 1 \) such that for large enough \( K \)
\[
\forall x \geq K : \frac{\pi(x)}{x/\ln x} \in [a, b].
\]
11.3. Exercises

Proof. Putting Lemmas 11.4 and 11.5 together gives
\[ \frac{2^n}{n+1} < n^{\pi(n)} \quad \text{and} \quad e^{\theta(n) - \theta\left(\frac{n}{2}\right)} < 2^n. \]
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 \( x \in [n,n+1) \) we have
\[ e^{\theta(x) - \theta\left(\frac{x}{2}\right)} = e^{\theta(n) - \theta\left(\frac{x}{2}\right)}. \]
so the second inequality still holds if we replace \( n \) by a real number \( x \geq 2. \)
Taking logarithms again, we obtain
\[ \theta(x) - \theta\left(\frac{x}{2}\right) \leq n \ln 2 \quad \text{and} \quad \theta\left(\frac{x}{2}\right) - \theta\left(\frac{x}{4}\right) \leq \frac{n}{2} \ln 2 \quad \text{and} \quad \cdots \]
and so on. Thus \( \theta(n) \leq 2n \ln 2. \) Again, this inequality holds if we replace \( n \)
by \( x. \) Substituting this into Lemma 11.2 gives that
\[ (11.3) \quad \pi(x) \leq \frac{1}{\ln x} + 2 \ln 2 \cdot \frac{x}{\ln x} + 2 \ln 2 \int_{e}^{x} (\ln t)^{-2} \, dt. \]
L’Hôpital’s rule implies that
\[ \lim_{x \to \infty} \frac{\int_{e}^{x} (\ln t)^{-2} \, dt}{x (\ln x)^{-2}} = 1. \]
Thus the integral in (11.3) can be replaced by \( x (\ln x)^{-2}. \) The dominant term
of the right hand of that equation is the middle one. Thus for any \( b > 2 \ln 2, \)
we have for \( x \) large enough that \( \pi(x) < b \frac{\ln x}{\ln 2}. \)  

11.3. Exercises

Exercise 11.1. Consider \( E(x_1, x_2) := |x_1 + x_2| - |x_1| - |x_2| \) as in the proof
of Lemma 11.3 and show that \( E \in \{0,1\}. \)

Exercise 11.2. In Theorem 11.6, give an explicit estimate for \( K, a, \) and \( b \)
that works for integers \( n \geq 2. \) (Hint: no need to find the best possible
estimate.)

Exercise 11.3. Suppose we had an “perfect” estimate for Lemma 11.4
of the form \( \left( \frac{3}{2} \right) = c \frac{2^n}{\pi^m} \) for some \( c > 0. \) Can you improve Theorem
11.6? (Hint: no. Conclusion: we need a different method to make further
progress.)
In the following exercise, we prove the equivalence Theorem 2.22 (a) and (b).

**Exercise 11.4.**

a) Compute the derivative of \( \frac{x}{\ln x} \).

b) Use (a) to prove that \( \lim_{x \to \infty} \frac{f'(\ln t) \cdot dt}{x/\ln x} = 1 \).

c) Use (b) to show that parts (a) and (b) of Theorem 2.22 are equivalent.

In the next two problems we prove the following result.

**Proposition 11.7.** Let \( p_n \) denote the \( n \)th prime. The prime number theorem is equivalent to

\[
\lim_{n \to \infty} \frac{p_n}{n \ln n} = 1.
\]

**Exercise 11.5.**

a) Assume that \( \lim_{x \to \infty} \frac{y}{x/\ln x} = 1 \) and show that \( \lim_{x \to \infty} y = \infty \).

b) Suppose \( \lim_{x \to \infty} f_1(x) = \infty \) and \( \lim_{x \to \infty} \frac{f_1(x)}{f_2(x)} = 1 \). Show that \( \lim_{x \to \infty} \frac{f_1(x)}{f_2(x)} = 1 \) and \( \lim_{x \to \infty} \frac{\ln f_1(x)}{\ln f_2(x)} = 1 \). (Hint: use L'Hôpital in each case.)

c) Show that \( \lim_{x \to \infty} \frac{\ln\ln x}{\ln x} = 1 \). (Hint: substitute \( x = e^t \).

d) Assume that \( \lim_{x \to \infty} \frac{x}{\ln x} = 1 \). Use (a) and (b) and (c) to show that

\[
\lim_{x \to \infty} \frac{x}{y \ln y} = \lim_{x \to \infty} \frac{x}{y \ln y} \frac{y}{x/\ln x} \frac{\ln y}{\ln x - \ln\ln x} = \lim_{x \to \infty} \frac{1}{1 - \ln\ln x} = 1.
\]

d) Show that (d) implies one way of Proposition 11.7.

e) Whereabouts is the \( n \)th prime located?

**Exercise 11.6.**

a) Assume that \( \lim_{y \to \infty} \frac{x}{y \ln y} = 1 \) and show that \( \lim_{y \to \infty} x = \infty \).

b) Assume that \( \lim_{y \to \infty} \frac{x}{y \ln y} = 1 \). Use (a) exercise 11.5 (b) and (c) to show that

\[
\lim_{x \to \infty} \frac{y}{x/\ln x} = \lim_{x \to \infty} \frac{x}{y \ln y} \frac{y}{x/\ln x} \frac{\ln y}{\ln x - \ln\ln x} = \lim_{x \to \infty} \frac{1}{1 - \ln\ln x} = 1.
\]

d) Show that (c) implies the other direction of Proposition 11.7.
11.3. Exercises

Exercise 11.7. In this exercise, we fix any \( K > 1 \) and \( x_i \) means a sequence such that \( \lim_{i \to \infty} x_i = \infty \). We also set \( x' = Kx \) for notational ease.

a) Show that if \( \pi(Kx_i) = \pi(x_i) \) 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 \( (x, x') \).

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 [2].

Proposition 11.8 (Bertrand’s Postulate). For all \( n \geq 2 \) there is a prime in the interval \( [n, 2n) \).

The same reference [2] also mentions an open (in 2018) problem in this direction: Is there always a prime between \( n^2 \) and \( (n + 1)^2 \)?

Exercise 11.8. a) Show for every \( m \in \mathbb{N} \), then \( \{m! + 2, \ldots, m! + m\} \) contains no primes. (Hint: for \( 2 \leq j \leq m \) we have \( j | (m! + j) \).)

b) Use Lemma 11.7 and exercise 11.3 to decide whether the gap in (a) between successive primes is greater or smaller than expected.
Bibliography

Index

\( I(n) \), 62
\( C[x] \), 7
\( \Lambda(n) \), 172
\( \Omega(n) \), 58
\( Q[x] \), 7
\( R[x] \), 7
\( Z[x] \), 7
\( \varepsilon(n) \), 58, 62
\( \lambda(n) \), 70
\( \mu(n) \), 57
\( \omega(n) \), 58, 69
\( \varphi(n) \), 58
\( \pi(x) \), 29
\( \psi(x) \), 172
\( \sigma_k(n) \), 56
\( \tau(n) \), 57
\( \theta(x) \), 172
\( \{1, 2, \cdots, b - 1\} \), 11
\( 1(n) \), 62

absolutely continuous measure, 141
absolutely normal, 146
additive function, 58
additive order, 74
affine cipher, 86
algebraic integer, 116
algebraic number, 7
aliquot sequence, 67
amicable numbers, 66, 67
angle multiplications, 150
arithmetic function, 55
aspiring numbers, 67
associative, 69, 111
axiom of choice, 131
Bézout’s Lemma, 20
Benford’s law, 168
Bertrand’s Postulate, 177
Birkhoff Ergodic Theorem, 135
Borel sets, 132
can of worms, 113
Cancellation Theorem, 22, 114
Cantor function, 144
cardinality, 13
cardinality of a finite set, 13
Carmichael number, 78
Catalan number, 123
Catalan-Dickson conjecture, 67
ceiling function, 19
Chebyshev function, 172
Chebyshev’s Theorem, 174
Chinese remainder theorem, 52
Chinese remainder theorem, generalized, 53
closed set, 131
closest return, 99
commutative, 69
commutative ring, 111
complete set of least residues, 83
<table>
<thead>
<tr>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete set of residues</td>
<td>74</td>
</tr>
<tr>
<td>completely additive function</td>
<td>58</td>
</tr>
<tr>
<td>completely multiplicative function</td>
<td>55</td>
</tr>
<tr>
<td>composite</td>
<td>4, 23</td>
</tr>
<tr>
<td>congruent</td>
<td>5</td>
</tr>
<tr>
<td>continued fraction approximants</td>
<td>92</td>
</tr>
<tr>
<td>continued fraction coefficients</td>
<td>92</td>
</tr>
<tr>
<td>continued fraction convergents</td>
<td>92</td>
</tr>
<tr>
<td>continued fraction expansion</td>
<td>91</td>
</tr>
<tr>
<td>cool</td>
<td>62</td>
</tr>
<tr>
<td>countable</td>
<td>9</td>
</tr>
<tr>
<td>degree</td>
<td>7</td>
</tr>
<tr>
<td>dense</td>
<td>6</td>
</tr>
<tr>
<td>density of a measure</td>
<td>141</td>
</tr>
<tr>
<td>density of a set</td>
<td>152</td>
</tr>
<tr>
<td>Devil's staircase</td>
<td>144</td>
</tr>
<tr>
<td>diagonal argument</td>
<td>10</td>
</tr>
<tr>
<td>Diophantine equation</td>
<td>39</td>
</tr>
<tr>
<td>Dirichlet convolution</td>
<td>62</td>
</tr>
<tr>
<td>Dirichlet series</td>
<td>63</td>
</tr>
<tr>
<td>Dirichlet's Theorem</td>
<td>107</td>
</tr>
<tr>
<td>discrete measure</td>
<td>142</td>
</tr>
<tr>
<td>distortion</td>
<td>157</td>
</tr>
<tr>
<td>distributive</td>
<td>69, 111</td>
</tr>
<tr>
<td>division algorithm</td>
<td>20</td>
</tr>
<tr>
<td>divisor</td>
<td>4, 113</td>
</tr>
<tr>
<td>dynamical system</td>
<td>91</td>
</tr>
<tr>
<td>elementary</td>
<td>31</td>
</tr>
<tr>
<td>emphasis</td>
<td>44</td>
</tr>
<tr>
<td>encryption</td>
<td>85</td>
</tr>
<tr>
<td>equidistributed</td>
<td>147</td>
</tr>
<tr>
<td>equivalence relation</td>
<td>13</td>
</tr>
<tr>
<td>Eratosthenes' sieve</td>
<td>5</td>
</tr>
<tr>
<td>ergodic transformation</td>
<td>136</td>
</tr>
<tr>
<td>Euclid numbers</td>
<td>35</td>
</tr>
<tr>
<td>Euclid's division lemma</td>
<td>20</td>
</tr>
<tr>
<td>Euclid's Lemma</td>
<td>21</td>
</tr>
<tr>
<td>Euclidean algorithm</td>
<td>40</td>
</tr>
<tr>
<td>Euclidean domain</td>
<td>118</td>
</tr>
<tr>
<td>Euclidean function</td>
<td>118</td>
</tr>
<tr>
<td>Euclidean ring</td>
<td>118</td>
</tr>
<tr>
<td>Euler's phi function</td>
<td>58</td>
</tr>
<tr>
<td>Euler's Product Formula</td>
<td>27</td>
</tr>
<tr>
<td>Euler's Theorem</td>
<td>80</td>
</tr>
<tr>
<td>Euler's totient function</td>
<td>58</td>
</tr>
<tr>
<td>fat Cantor set</td>
<td>144</td>
</tr>
<tr>
<td>Fermat prime</td>
<td>79</td>
</tr>
<tr>
<td>Fermat's little theorem</td>
<td>76</td>
</tr>
<tr>
<td>Fibonacci numbers</td>
<td>51, 92</td>
</tr>
<tr>
<td>field</td>
<td>81, 112</td>
</tr>
<tr>
<td>field norm</td>
<td>118</td>
</tr>
<tr>
<td>first Chebyshev function</td>
<td>172</td>
</tr>
<tr>
<td>floor function</td>
<td>19</td>
</tr>
<tr>
<td>formal expression</td>
<td>91</td>
</tr>
<tr>
<td>fractional part</td>
<td>19</td>
</tr>
<tr>
<td>Fundamental Theorem of Algebra</td>
<td>51</td>
</tr>
<tr>
<td>fundamental theorem of arithmetic</td>
<td>23</td>
</tr>
<tr>
<td>Gauss map</td>
<td>89</td>
</tr>
<tr>
<td>Gaussian integers</td>
<td>119</td>
</tr>
<tr>
<td>general solution</td>
<td>44</td>
</tr>
<tr>
<td>golden mean</td>
<td>92</td>
</tr>
<tr>
<td>greatest common divisor</td>
<td>4, 36, 44, 119</td>
</tr>
<tr>
<td>greatest common divisor for polynomials</td>
<td>50</td>
</tr>
<tr>
<td>greedy algorithm</td>
<td>48</td>
</tr>
<tr>
<td>homogeneous</td>
<td>44</td>
</tr>
<tr>
<td>inclusion-exclusion</td>
<td>68</td>
</tr>
<tr>
<td>Infinity of Primes</td>
<td>25</td>
</tr>
<tr>
<td>inhomogeneous</td>
<td>44</td>
</tr>
<tr>
<td>integer</td>
<td>4</td>
</tr>
<tr>
<td>integrable function</td>
<td>134</td>
</tr>
<tr>
<td>integral domain</td>
<td>114</td>
</tr>
<tr>
<td>invariant set</td>
<td>136</td>
</tr>
<tr>
<td>invertible element</td>
<td>113</td>
</tr>
<tr>
<td>irrational number</td>
<td>6</td>
</tr>
<tr>
<td>irreducible</td>
<td>113</td>
</tr>
<tr>
<td>Jensen’s inequality</td>
<td>163</td>
</tr>
<tr>
<td>key</td>
<td>85</td>
</tr>
<tr>
<td>Khinchin’s constant</td>
<td>159</td>
</tr>
<tr>
<td>Lévy’s constant</td>
<td>160</td>
</tr>
<tr>
<td>Lamé’s theorem</td>
<td>51</td>
</tr>
<tr>
<td>Lambert series</td>
<td>63</td>
</tr>
<tr>
<td>least common multiple</td>
<td>4, 44, 119</td>
</tr>
<tr>
<td>least common multiplier</td>
<td>36</td>
</tr>
<tr>
<td>least residue</td>
<td>83</td>
</tr>
<tr>
<td>Lebesgue Decomposition</td>
<td>141</td>
</tr>
<tr>
<td>Lebesgue density theorem</td>
<td>153</td>
</tr>
<tr>
<td>Lebesgue integral</td>
<td>133</td>
</tr>
<tr>
<td>Lebesgue measurable sets</td>
<td>132</td>
</tr>
<tr>
<td>Lebesgue measure</td>
<td>133</td>
</tr>
</tbody>
</table>
Index

left module, 117
Liouville \(\lambda\)-function, 70
Liouville number, 8
Liouville’s Theorem, 8
Lyapunov exponent, 166
Möbius inversion, 59
Möbius function, 57
measurable function, 133
measurable sets, 131
measurable space, 133
measure, 133
measure space, 133
Mersenne prime, 79
middle third Cantor set, 15
middle third set Cantor set, 143
minimal polynomial, 8
modular arithmetic, 5, 73
modular exponentiation, 77
module, 117
modulo, 5
monic, 116
multiple, 4
multiplicative cipher, 85
multiplicative function, 55
multiplicative order, 74
mutually singular measures, 137
non-units, 23
norm, 20, 118
normal in base \(b\), 146
number theoretic function, 55
open set, 131
order, 73
order of \(a\) modulo \(b\), 74
particular solution, 44
perfect number, 66, 79
Perron-Frobenius operator, 150
plain text, 85
Pointwise Ergodic Theorem, 135
power set, 16
prime, 4, 23, 113
Prime Number Theorem, 29
prime omega functions, 58
primitive root, 75
probability measure, 133
pseudo prime to the base \(a\), 78
pushforward of a measure, 135
Radon-Nikodym, 141
rational number, 6
reduced set of residues, 74
reducible, 113
relatively prime, 5
residue, 5
Riemann Hypothesis, 30
Riemann zeta function, 27
right module, 117
ring, 81, 111
Roth’s Theorem, 9
second Chebyshev function, 172
sequences, 55
singular continuous measure, 142
Skewes number, 30
sociable numbers, 67
square free, 35, 57
square full, 35
sub-additivity, 133
sum of divisors, 56
time average equals spatial average, 136
transcendental number, 7
twin primes, 13
uncountable, 9
underline, 132
unique factorization theorem, 23
uniquely ergodic, 137
unit, 113
units, 23
untouchable numbers, 67
Veerman, 3
von Mangoldt function, 172
well-ordering principle, 5
Wilson’s theorem, 82
word, 11