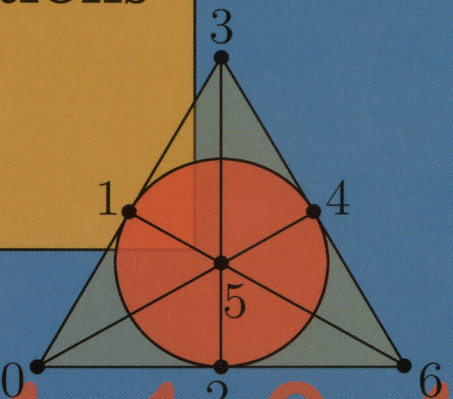
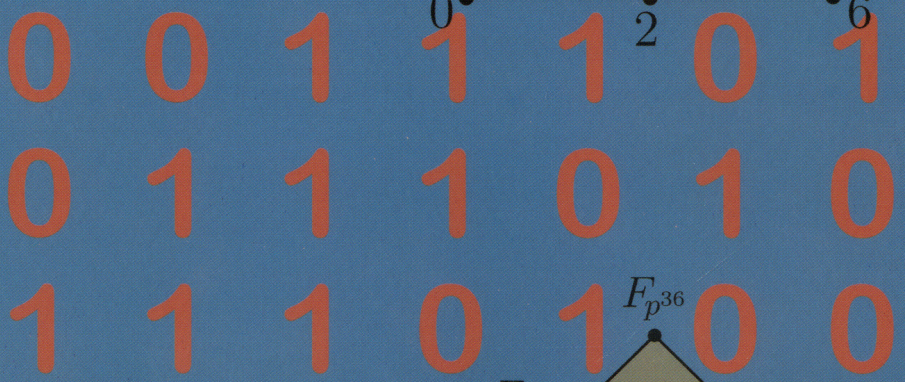
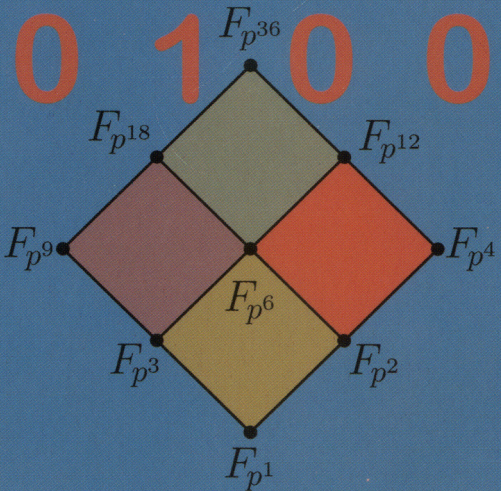
STUDENT MATHEMATICAL LIBRARY
Volume 41

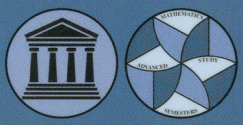
Finite Fields and Applications

Gary L. Mullen
Carl Mummert









American Mathematical Society
Mathematics Advanced Study Semesters

http://dx.doi.org/10.1090/stml/041

STUDENT MATHEMATICAL LIBRARY
Volume 41

Finite Fields and Applications

Gary L. Mullen
Carl Mummert

American Mathematical Society
Mathematics Advanced Study Semesters

2000 *Mathematics Subject Classification.* Primary 11-01, 11Txx; Secondary 11T71, 05Bxx.

For additional information and updates on this book, visit
www.ams.org/bookpages/stml-41

Library of Congress Cataloging-in-Publication Data

Mullen, Gary L.
Finite fields and applications / Gary L. Mullen, Carl Mummert.
p. cm. — (Student mathematical library, ISSN 1520-9121 ; v. 41)
Includes bibliographical references and index.
ISBN 978-0-8218-4418-2
1. Finite fields (Algebra) 2. Coding theory. 3. Cryptography. I. Mummert, Carl, 1978– II. Title.

QA247.3.M85 2007
512.7'4–dc22 2007060797

10 9 8 7 6 5 4 3 2 12 11 10 09

Contents

Foreword: MASS and REU at Penn State University

This book is part of a collection published jointly by the American Mathematical Society and the MASS (Mathematics Advanced Study Semesters) program as a part of the Student Mathematical Library series. The books in the collection are based on lecture notes for advanced undergraduate topics courses taught at the MASS and/or Penn State summer REU (Research Experiences for Undergraduates). Each book presents a self-contained exposition of a nonstandard mathematical topic, often related to current research areas, accessible to undergraduate students familiar with an equivalent of two years of standard college mathematics and suitable as a text for an upper division undergraduate course.

Started in 1996, MASS is a semester-long program for advanced undergraduate students from across the USA. The program's curriculum amounts to sixteen credit hours. It includes three core courses from the general areas of algebra/number theory, geometry/topology and analysis/dynamical systems, custom designed every year; an interdisciplinary seminar; and a special colloquium. In addition, every participant completes three research projects, one for each core course. The participants are fully immersed into mathematics, and

this, as well as intensive interaction among the students, usually leads to a dramatic increase in their mathematical enthusiasm and achievement. The program is unique for its kind in the United States.

The summer mathematical REU program is formally independent of MASS, but there is a significant interaction between the two: about half of the REU participants stay for the MASS semester in the fall. This makes it possible to offer research projects that require more than seven weeks (the length of the REU program) for completion. The summer program includes the MASS Fest, a two to three day conference at the end of the REU at which the participants present their research and that also serves as a MASS alumni reunion. A nonstandard feature of the Penn State REU is that, along with research projects, the participants are taught one or two intense topics courses.

Detailed information about the MASS and REU programs at Penn State can be found on the website `www.math.psu.edu/mass`.

Preface

The aim of this book is to provide a brief introduction to finite fields and some of their many fascinating applications. The book arose from lectures of the first author in a course entitled "Finite Fields and Their Applications," which was taught in the Department of Mathematics at The Pennsylvania State University during the Fall semester of 2004. The course was part of the department's Mathematics Advanced Study Semesters (MASS) program. The second author produced an initial online set of notes from these lectures, which have been greatly expanded into the present volume.

The most important chapter of this text is the first, which discusses a variety of properties of finite fields. Many of these properties are used in later chapters where various applications of finite fields are discussed. The chapter begins with a discussion of the basic properties of finite fields and extension fields. It then defines the important trace and norm functions and establishes some of their properties. Bases for extension fields, including dual, normal, and primitive normal bases, are then discussed. The first chapter concludes with a few results concerning polynomials over finite fields. These include a discussion of the order of a polynomial, formulas for the number and orders of irreducible polynomials, and properties of linearized polynomials and permutation polynomials over finite fields.

Chapter 2 includes some combinatorial applications of finite fields. It includes a detailed discussion of latin squares and their applications to affine and projective planes as well as more general block designs. The chapter closes with a brief discussion of Hadamard matrices which arise from an elementary finite field construction.

Chapter 3 deals with algebraic coding theory and includes a discussion of some properties of codes as well as bounds on the parameters of linear codes. Several encoding and decoding methods are also discussed. Constructions for various kinds of codes including Hamming, cyclic, BCH, and Goppa codes are given. A brief discussion of perfect codes is also included. The chapter ends with a discussion of some relations and connections between codes, latin squares, and combinatorial designs.

The final chapter covers some elementary aspects of cryptography. The discussion includes some basic properties of cryptographic systems as well as symmetric key and public key cryptography. The RSA cryptosystem and a double-round quadratic system are presented, along with key exchange systems including the Diffie-Hellman system. The discrete logarithm problem for finite fields is presented in this context. Several threshold systems for distributing secret information are presented, including one based on latin squares. The chapter ends with a brief discussion of digital signatures and several cryptosystems based on Dickson polynomials and elliptic curves over finite fields.

Appendix A provides a brief review of some basic algebraic concepts that are needed for a full understanding of some of the topics covered in the first four chapters. These concepts include topics from number theory, groups, rings and fields, homomorphisms, polynomials and splitting fields. A brief review of a few concepts from the theory of vector spaces, including dual spaces, is presented.

Each chapter, and the first appendix, concludes with a brief set of notes related to that chapter's material. These notes describe a variety of references that provide material for further reading on the topics presented here. Each chapter, and the first appendix, contains a set of exercises of varying levels of difficulty that expand upon

the ideas presented. Appendix B provides hints for many of these exercises.

The first author would like to sincerely thank Sergei Tabachnikov, Director of the MASS program at Penn State, for inviting him to teach a course in the MASS program. Both authors would like to thank Sergei for his encouragement to convert our initial class notes into this text. The first author used this text in his MASS class taught during the Fall semester of 2006. We would like to sincerely thank Charles F. Laywine for his careful reading and many excellent suggestions which greatly improved the readability of our book. A special word of thanks is owed to the 2006 class of MASS students, who provided numerous comments and helpful suggestions for improvements in addition to locating a number of typographical errors. We also thank the publishing staff of the American Mathematical Society who helped bring this book to a successful conclusion.

G. L. Mullen
C. Mummert

This page intentionally left blank

http://dx.doi.org/10.1090/stml/041/01

Chapter 1

Finite Fields

1. Introduction

A field is an algebraic structure consisting of a set of elements for which the operations of addition, subtraction, multiplication, and division satisfy certain prescribed properties. The real numbers are probably the best known example, along with the fields of rational numbers and complex numbers. These are all examples of infinite fields because each contains an infinite number of distinct elements. Certain finite sets also satisfy the field properties when assigned appropriate operations; these *finite fields* are our focus of study in this chapter.

Because some readers may not be familiar with algebraic structures such as groups, rings, fields, and vector spaces, we have included a brief introduction to them in Appendix A. The material there includes the basic definitions and background theorems required here.

2. Finite fields

We begin our exploration of finite fields by determining the possible sizes of a finite field. We will see that linear algebra plays a crucial role in the answer to this question. Recall that every field has a unique smallest subfield, called the *prime subfield*, which is the intersection of all of its subfields (see Exercise A.20).

Lemma 1.2.1. *Suppose F is a finite field with a subfield K containing q elements. Then F is a vector space over K and $|F| = q^m$, where m is the dimension of F viewed as a vector space over K.*

Proof. It is straightforward to verify that F is a vector space over K using the field operations in F; we leave this to the reader. Since F is finite, we can choose a basis $B = \{\beta_1, \ldots, \beta_m\}$ for F over K. Every element α of F can thus be written in the form $\alpha = a_1\beta_1 + \cdots + a_m\beta_m$, where $a_i \in K$ for $1 \leq i \leq m$ and the sequence $a_1, a_2, \ldots, a_m$ is uniquely determined by α. There are $|K|^m = q^m$ distinct sequences of coefficients, because there are $|K| = q$ choices for each a_i. □

The m occurring in Lemma 1.2.1, which is the dimension of F as a vector space over K, is called the *degree* of F over K. By combining the lemma with the fact that every finite field has prime characteristic (Lemma A.3.6), we obtain a characterization of the number of elements that a finite field can possess.

Theorem 1.2.2. *Let F be a finite field. The cardinality of F is p^m, where p is the characteristic of F and m is the degree of F over its prime subfield.*

We can conclude from this theorem, for example, that there is no finite field of order 36. We will prove a converse of Theorem 1.2.2 as Theorem 1.2.5 below.

We remark that for any integer $n > 2$ there is an algebraic structure of cardinality n, known as a *neofield*, which satisfies all of the field axioms except for associativity of addition. These structures are discussed by Dénes and Keedwell [11, pp. 246–249].

Lemma 1.2.3. *If F is a finite field with q elements and $a \in F$ is nonzero, then $a^{q-1} = 1$. Thus $a^q = a$ for all $a \in F$.*

Proof. The result is immediate when a is zero. If a is not zero, we know that a is a unit in F. There are $q-1$ units in F, so by Lagrange's theorem (Theorem A.2.6) the multiplicative order of a in F divides $q - 1$. Therefore $a^{q-1} = 1$ and $a^q = a$. □

An immediate consequence of the previous lemma is that the multiplicative inverse of any nonzero element a in a field of order q is a^{q-2}, because $a^{q-2} \cdot a = a^{q-1} = 1$.

The polynomial $x^q - x$ has degree q and so can have at most q roots in any field. Lemma 1.2.3 indicates that if F is a field of order q, then every element of F is a root of $x^q - x$. The next lemma follows immediately.

Lemma 1.2.4. *If F is a finite field with q elements, then $x^q - x$ factors in $F[x]$ as $\Pi_{a \in F}(x - a)$.*

Our next result precisely characterizes the orders of finite fields. We first ask the reader to review the definition of a splitting field (Definition A.5.7).

Theorem 1.2.5 (Existence and uniqueness of finite fields). *For every prime p and positive integer $n \geq 1$ there is a finite field with p^n elements. Any finite field with p^n elements is isomorphic to the splitting field of $x^{p^n} - x$ over F_p.*

Proof. We first prove the existence part of the theorem. Assume that the prime power q is of the form p^n, where p is a prime. Consider the polynomial $r(x) = x^q - x$ as a polynomial with coefficients in the field F_p. Let F be a splitting field of $r(x)$ over F_p.

Consider the set $S = \{a \in F \mid a^q - a = 0\}$. Since the derivative $r'(x)$ is identically -1, it has no roots. Thus the derivative test (Lemma A.5.2) shows that $r(x)$ has no multiple roots, so $|S| = q$. We leave it to the reader to verify that S is a subfield of the field F. This means S is a finite field with $q = p^n$ elements.

The uniqueness part of the theorem follows from the fact that if F is a finite field with p^n elements, then F must have characteristic p and so it must contain the field F_p as a subfield. Hence F is the splitting field of $x^q - x$ over F_p. By Theorem A.5.9, splitting fields are unique up to field isomorphism. □

The previous theorem shows that a finite field of a given order is unique up to field isomorphism. Thus we speak of "the" finite field of a particular order q, and we write F_q to denote this field. Another

common notation for a field of order q is $\mathrm{GF}(q)$, where G stands for Galois and F stands for field. This name is used in honor of Evariste Galois (1811–1832), who in 1830 was the first person to seriously study properties of general finite fields (fields with a prime power but not a prime number of elements). We will use the notation F_q in this book.

Remark 1.2.6. We note that when p is a prime, the field F_p is the same as (isomorphic to) the ring Z_p of integers modulo p. In Exercise 1.9, the reader is asked to show that when $m > 1$ the finite field F_{p^m} is not the same as the ring Z_{p^m} of integers modulo p^m. The reader should be sure to understand the difference between the two commutative rings in the nonprime case, when one is a field and the other is not.

Our next result gives a characterization of the subfield structure of a finite field.

Theorem 1.2.7 (Subfield structure). *Let F be a finite field with p^n elements. Every subfield of F has p^m elements for some integer m dividing n. Conversely, for any integer m dividing n there is a unique subfield of F of order p^m.*

Proof. A subfield K of the finite field F_{p^n} must have p^m distinct elements for some positive integer m with $m \leq n$. By Lemma 1.2.1, p^n must be a power of p^m, so m must divide n.

On the other hand, assume that m divides n. Then the polynomial $x^{p^m-1}-1$ divides $x^{p^n-1}-1$, and thus $x^{p^m}-x$ divides $x^{p^n}-x$. It follows that each root of $x^{p^m}-x$ is also a root of $x^{p^n}-x$. Hence the field F_{p^n} must contain a splitting field of the polynomial $x^{p^m}-x$ over F_p and this splitting field must have exactly p^m distinct elements. If the subfield was not unique, that is, if there were two such fields contained in F_{p^n}, then their union would contain more than p^m roots of the polynomial $x^{p^m}-x$ in F_{p^n}, which is impossible. □

The next theorem is one of the most important results in finite field theory. It will be central to the proofs of many later results in this book.

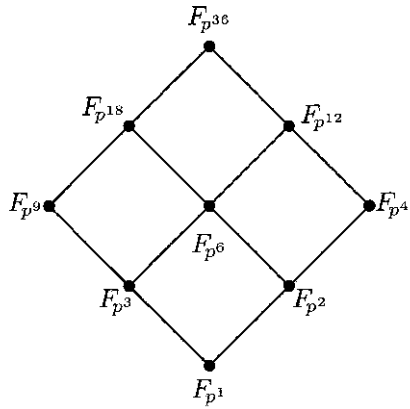


Figure 1.1. The subfields of $F_{p^{36}}$.

Theorem 1.2.8. *The multiplicative group F_q^* of all nonzero elements of the finite field F_q is cyclic.*

Proof. The case where $q = 2$ is trivial so we assume that $q \geq 3$. Let $q - 1 = h > 1$ have the prime factorization $\prod_{i=1}^{t} p_i^{r_i}$. For each i consider the polynomial $f_i(x) = x^{h/p_i} - 1$. This polynomial has degree $h/p_i < h$ and thus has at most h/p_i roots. Choose a_i, an element of F_q which is not a root of f_i, so that $a_i \neq 0$. Let $b_i = a_i^{h/p_i^{r_i}}$ for each $i \leq t$.

We will show the multiplicative order of b_i is $p_i^{r_i}$. Clearly $b_i^{p_i^{r_i}} = a_i^h = 1$. Thus the order of b_i must divide $p_i^{r_i}$ and so must be a power of p_i. Suppose $b_i^{p_i^k} = 1$ for some $k < r_i$. Then we have

$$b_i^{p_i^{r_i-1}} = b_i^{\left(p_i^k p_i^{r_i-(k+1)}\right)} = 1^{p_i^{r_i-(k+1)}} = 1.$$

This is impossible, because $b_i^{p_i^{r_i-1}} = a_i^{h/p_i}$ and $a_i^{h/p_i} \neq 1$. Thus the order of b_i is exactly $p_i^{r_i}$.

Finally, let $b = \prod_{i=1}^{t} b_i$. Then by Lemma A.2.9 the order of b is $q - 1$, because this is the least common multiple of the orders of the elements b_i for $i = 1, \ldots, t$. □

An element $\theta \in F_q$ which multiplicatively generates the group F_q^* of all nonzero elements of the field F_q is called a *primitive element*.

Remark 1.2.9. Let θ be a primitive element of a finite field F. Then every nonzero element of F can be written as a power of θ. This representation makes multiplication of field elements very easy to compute. Suppose for example that $a = \theta^t$ and $b = \theta^r$; then $ab = \theta^t\theta^r = \theta^{t+r}$. It is difficult, however, to find the power s of θ such that $\theta^t + \theta^r = \theta^s$. Conversely, as we will see later in our discussion of bases for finite fields, representations which make addition easy to compute often have a more complex multiplicative structure.

Lemma 1.2.10. *If g is a primitive element of F_q, then g^t is a primitive element of F_q if and only if $(t, q-1) = 1$.*

The reader is asked to provide a proof of this lemma in Exercise 1.19.

It follows from Lemma 1.2.10 that there are $\phi(q-1)$ primitive elements in F_q, where $\phi(n)$ denotes Euler's function from elementary number theory (see Appendix A for a discussion of this function).

3. Extension fields

In this section, we explore further the concept of adding elements to a given finite field to produce a larger finite field. We have seen one application of this method already in the existence and uniqueness theorem. Recall that the intersection of any collection of subfields of a given field F is itself a subfield of F.

Definition 1.3.1 (Adjoining elements to a field)**.** Let K be a subfield of F and let M be a subset of F. Then $K(M)$ denotes the intersection of all subfields of F containing K and M as subsets. This field is called "K adjoin M." When M is finite, say $M = \{\theta_1, \ldots, \theta_k\}$, we write $K(\theta_1, \ldots, \theta_k)$ for $K(M)$.

Definition 1.3.2. Let $K \subseteq F$, $\theta \in F$, and $p(\theta) = 0$ where $p(x)$ is a monic polynomial in $K[x]$. Then $p(x)$ is called the *minimal polynomial* of θ if θ is not a root of any nonzero polynomial in $K[x]$ of lower degree.

The following result provides a method by which one can obtain irreducible polynomials.

Proposition 1.3.3. *The minimal polynomial of any element is irreducible.*

Proof. Suppose $g(x) \in K[x]$ is a minimal polynomial which factors as $g_1(x)g_2(x)$, and suppose $g(\theta) = 0$. Then $g_1(\theta)g_2(\theta) = 0$, so either $g_1(\theta) = 0$ or $g_2(\theta) = 0$. Therefore either $g_1 = g$ and $g_2 = 1$, or else $g_1 = 1$ and $g_2 = g$, because g is the monic minimal polynomial of θ. This shows that g is irreducible. □

Definition 1.3.4. A field L is a *finite extension* of K if $K \subseteq L$ and L is a finite dimensional vector space over K. In this case we refer to the dimension m of L over K as the *degree* of the extension, and we write $[L : K] = m$.

Our next result indicates that a finite extension of a finite extension is again a finite extension.

Theorem 1.3.5 (Transitivity of degree). *Let L be a finite extension of K and let M be a finite extension of L. Then M is a finite extension of K. Moreover, we have $[M : K] = [M : L][L : K]$.*

Proof. Let $A = \{\alpha_1, \ldots, \alpha_m\}$ be a basis for L over K and $B = \{\beta_1, \ldots, \beta_n\}$ a basis for M over L. We now show that

$$\{\alpha_i\beta_j \mid 1 \leq i \leq m, 1 \leq j \leq n\}$$

is a basis for M over K. Suppose there are scalars $c_{i,j}$ such that

$$\sum_{1 \leq i \leq m, 1 \leq j \leq n} c_{i,j}\alpha_i\beta_j = 0.$$

Because B is independent and $\alpha_i \in L$, it must be the case that

$$\sum_{1 \leq i \leq m} c_{i,j}\alpha_i = 0$$

for each j. Because A is independent, we must have $c_{i,j} = 0$ for all i and j. □

Example 1.3.6. Let p be prime. In Figure 1.1, we see that F_{p^3} is an extension of F_p of degree 3, and $F_{p^{18}}$ is an extension of F_{p^3} of degree 6. Moreover, $F_{p^{18}}$ is an extension of F_p of degree $18 = 3 \cdot 6$.

Example 1.3.7. For p a prime, let $M = F_{p^6}$, $L = F_{p^2}$, and $K = F_p$. Let β_1, β_2 be a basis for L over K and $\alpha_1, \alpha_2, \alpha_3$ a basis for M over L. Then the set $\{\alpha_1\beta_1, \alpha_1\beta_2, \alpha_2\beta_1, \alpha_2\beta_2, \alpha_3\beta_1, \alpha_3\beta_2\}$ is a basis for M over K.

Definition 1.3.8. Let $K \subseteq F$ and let $\theta \in F$. Then θ is said to be *algebraic* over K if there is a nonzero polynomial $p(x) \in K[x]$ such that $p(\theta) = 0$ in $F[x]$. An extension field is called *algebraic* if every element of the extension field is algebraic over the base field.

Note that every algebraic element of an extension field has a minimal polynomial over the base field. The *degree* of an algebraic element of an extension field over a base field is defined to be the degree of its minimal polynomial over the base field.

Theorem 1.3.9. *Every finite extension of a finite field is algebraic.*

Proof. Let L be a finite extension of K. We have shown that the multiplicative group L^* of L is cyclic; let θ be a generator of L^*. It follows immediately that $L = K(\theta)$. □

Our next result summarizes some important results for the actual construction of finite fields.

Theorem 1.3.10. *Let K be a subfield of F with $\theta \in F$ algebraic of degree n over K and let $g(x)$ be the minimal polynomial of θ over K. Then:*

(1) *The field $K(\theta)$ is isomorphic to the factor ring $K[x]/(g(x))$.*

(2) *The dimension of $K(\theta)$ over K is n.*

(3) *The set $\{1, \theta, \theta^2, \ldots, \theta^{n-1}\}$ is a basis for $K(\theta)$ over K.*

(4) *Every element of $K(\theta)$ is algebraic over K with degree dividing n.*

Proof. To prove part 1, we use some ring theory. We first construct the mapping $\tau\colon K[x] \to K(\theta)$ defined by $\tau(f) = f(\theta)$. We leave it to the reader to check that τ is indeed a ring homomorphism. The reader should also verify that the kernel of τ is the set of polynomials $f \in K[x]$ such that $f(\theta) = 0$, and that $f(x)$ is in this kernel if and only if $f(x)$ is in the ideal generated by g. Let S be the range of the

mapping τ. It is clear that S is isomorphic to the factor ring $K[x]/(g)$, which is a field since $g(x)$ is irreducible. We have the sequence $K \subseteq S \subseteq K(\theta)$ of nested fields. Since $\theta \in S$, we see that $S = K(\theta)$, and the proof of part (1) is complete.

We now prove parts (2) and (3). From part (1), if $\alpha \in S = K(\theta)$, α can be written as $\alpha = f(\theta)$ for some $f \in K[x]$. We can write f as $qg + r$, where q and r are polynomials over K and the degree of r is less than the degree of g (see Exercise A.18). Let n be the degree of g. A simple calculation shows that $\alpha = f(\theta) = r(\theta)$, so α is a linear combination, with coefficients in K, of the elements $1, \theta, \ldots, \theta^{n-1}$. But if θ satisfies $a_0 + a_1\theta + \cdots + a_{n-1}\theta^{n-1} = 0$, then the polynomial $h(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}$ over K has θ as a root. Therefore $h(x)$ must be a multiple of $g(x)$, but this can happen only if $h(x) = 0$ is the zero polynomial, that is, if all of the $a_i = 0$. Hence the elements $1, \theta, \ldots, \theta^{n-1}$ must be linearly independent over the field K, and thus they form a basis of $K(\theta)$ over K.

For part (4), first note that by part (2), $K(\theta)$ is a finite extension field of K, so any $\alpha \in K(\theta)$ is algebraic over K. Moreover, $K(\alpha)$ is a subfield of $K(\theta)$. Letting d be the degree of α over K, we have

$$n = [K(\theta) : K] = [K(\theta) : K(\alpha)][K(\alpha) : K] = [K(\theta) : K(\alpha)]d,$$

and hence we see that d divides n. □

Corollary 1.3.11. *Suppose that F is an algebraic extension of K and $\alpha \in F$. Then the minimal polynomial of α over K has degree dividing $[F : K]$.*

Proof. Exercise 1.20 □

An extension obtained by adjoining a single element is called a *simple* extension (see Exercises 1.6 and 1.10 for examples of such extensions). The next theorem gives an important property of finite fields that is not shared by infinite fields (there are finite extensions of infinite fields that are not simple, as illustrated by Exercise 1.4).

Theorem 1.3.12. *Let F_q be a finite field and let F_r be a finite extension of F_q. Then F_r is a simple algebraic extension of F_q, and for any primitive element θ of F_r the relation $F_r = F_q(\theta)$ holds.*

Proof. Let θ be a primitive element of F_r. Choose $\alpha \in F_q(\theta)$, so $\alpha = a_0 + a_1\theta + \cdots + a_m\theta^m$, where m is the degree of F_r over F_q. This sum is an element (recall that $F_q \subseteq F_r$) of F_r, so $\alpha \in F_r$. Now notice that $F_q(\theta)$ contains 0, θ, and all powers of θ. Hence $F_q(\theta)$ contains F_r. □

Corollary 1.3.13. *For any prime power q and any integer $n \geq 1$ there is an irreducible polynomial of degree n over F_q.*

Proof. Let $p(x)$ be the minimal polynomial over F_q of a primitive element of F_{q^n}. □

Example 1.3.14. Consider the polynomial $p(x) = x^2+x+1$ over the field F_2. Since $p(x)$ does not have a root in F_2 (recall that $p(x)$ will have a root $a \in F_2$ if and only if $p(a) = 0$ for some $a \in F_2$ where the arithmetic is computed modulo 2), $p(x)$ is irreducible over F_2. Let θ be a root of $p(x)$ so that $\theta^2+\theta+1 = 0$, that is, $\theta^2 = -(\theta+1) = \theta+1$. The field $F_4 = F_{2^2}$ can be represented as the set $\{a\theta + b | a, b \in F_2\}$. We now give the addition and multiplication tables for the field F_{2^2}.

$+$	0	1	θ	$\theta+1$
0	0	1	θ	$\theta+1$
1	1	0	$\theta+1$	θ
θ	θ	$\theta+1$	0	1
$\theta+1$	$\theta+1$	θ	1	0

$\times$	0	1	θ	$\theta+1$
0	0	0	0	0
1	0	1	θ	$\theta+1$
θ	0	θ	$\theta+1$	1
$\theta+1$	0	$\theta+1$	1	θ

We note that (after simplification) $(\theta + 1)(\theta + 1) = \theta$. We also note that θ is a primitive element in the field F_4, so $\theta^1 = \theta$, $\theta^2 = \theta+1$, and $\theta^3 = 1$.

We now consider an example of a finite field with a larger non-prime number of elements.

Example 1.3.15. Consider the field F_9, which is a vector space of dimension 2 over F_3. Consider $p(x) = x^2 + x + 2$ in $F_3[x]$. This

$+$	0	1	2	θ	$\theta+1$	$\theta+2$	2θ	$2\theta+1$	$2\theta+2$
0	0	1	2	θ	$\theta+1$	$\theta+2$	2θ	$2\theta+1$	$2\theta+2$
1	1	2	0	$\theta+1$	$\theta+2$	θ	$2\theta+1$	$2\theta+2$	2θ
2	2	0	1	$\theta+2$	θ	$\theta+1$	$2\theta+2$	2θ	$2\theta+1$
θ	θ	$\theta+1$	$\theta+2$	2θ	$2\theta+1$	$2\theta+2$	0	1	2
$\theta+1$	$\theta+1$	$\theta+2$	θ	$2\theta+1$	$2\theta+2$	2θ	1	2	0
$\theta+2$	$\theta+2$	θ	$\theta+1$	$2\theta+2$	2θ	$2\theta+1$	2	0	1
2θ	2θ	$2\theta+1$	$2\theta+2$	0	1	2	θ	$\theta+1$	$\theta+2$
$2\theta+1$	$2\theta+1$	$2\theta+2$	2θ	1	2	0	$\theta+1$	$\theta+2$	θ
$2\theta+2$	$2\theta+2$	2θ	$2\theta+1$	2	0	1	$\theta+2$	θ	$\theta+1$

$\times$	0	1	2	θ	$\theta+1$	$\theta+2$	2θ	$2\theta+1$	$2\theta+2$
0	0	0	0	0	0	0	0	0	0
1	0	1	2	θ	$\theta+1$	$\theta+2$	2θ	$2\theta+1$	$2\theta+2$
2	0	2	1	2θ	$2\theta+2$	$2\theta+1$	θ	$\theta+2$	$\theta+1$
θ	0	θ	2θ	$2\theta+1$	1	$\theta+1$	$\theta+2$	$2\theta+2$	2
$\theta+1$	0	$\theta+1$	$2\theta+2$	1	$\theta+2$	2θ	2	θ	$2\theta+1$
$\theta+2$	0	$\theta+2$	$2\theta+1$	$\theta+1$	2θ	2	$2\theta+2$	1	θ
2θ	0	2θ	θ	$\theta+2$	2	$2\theta+2$	$2\theta+1$	$\theta+1$	1
$2\theta+1$	0	$2\theta+1$	$\theta+2$	$2\theta+2$	θ	1	$\theta+1$	2	2θ
$2\theta+2$	0	$2\theta+2$	$\theta+1$	2	$2\theta+1$	θ	1	2θ	$\theta+2$

Table 1.1. The addition and multiplication tables for F_{3^2}, where θ is a root of x^2+x+2.

polynomial has no roots in F_3 so it is irreducible over F_3. Let θ be a root of $p(x)$, so $\theta^2+\theta+2=0$. Hence $\theta^2=-\theta-2=2\theta+1$ (recall that we are working in characteristic 3).

The field F_{3^2} is isomorphic to the set $\{a\theta+b \mid a,b\in F_3\}$ with its natural operations. We can compute the addition and multiplication tables by hand. For example, $2\theta(\theta+2)=2\theta^2+4\theta=2(2\theta+1)+\theta=2\theta+2$. The complete addition and multiplication tables are shown in Table 1.1.

We can use the multiplication table to check that the multiplicative order of θ in F_9 is 8, which means that θ is a primitive element of F_9.

Example 1.3.16. Let $p(x)=x^2+1\in F_3[x]$. It is straightforward to check that $p(x)$ has no roots in F_3 and is thus irreducible over the field F_3. Let θ be a root of $p(x)$. We compute $\theta^2=-1$ and $\theta^4=1$.

Hence no root of $p(x)$ can have order 8, that is, no root of $p(x)$ can be a primitive element. Nevertheless, the splitting field of $p(x)$ over F_3 is F_9. It can be seen that $\theta + 1$ has order 8 and is thus a primitive element for F_9 over F_3.

Example 1.3.17. Consider $q = 2^{100}$. We can identify the elements of F_q with polynomials of the form $a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_{99}\alpha^{99}$, where $0 \leq a_i \leq 1$ for each i and where α is a root of an irreducible polynomial of degree 100 over the field F_2. Corollary 1.3.13 shows that such an irreducible polynomial always exists. Theorem 1.6.4 below shows that there are exactly

$$\frac{1}{100}(2^{100} - 2^{50} - 2^{20} + 2^{10})$$

irreducible polynomials of degree 100 over F_2. We pose the following question for readers with an interest in computation: how can one locate an irreducible polynomial of degree 100 over F_2?

Let $N_q(n)$ be the number of monic irreducible polynomials of degree n over F_q. In Theorem 1.6.4 below, we will prove that

$$N_q(n) = \frac{1}{n}\sum_{d|n} \mu(d) q^{n/d}. \tag{1}$$

Here μ is the Möbius function of elementary number theory defined by the rule

$$\mu(m) = \begin{cases} 1 & \text{if } m = 1, \\ (-1)^k & \text{if } m = m_1 m_2 \cdots m_k, \text{ a product of distinct primes,} \\ 0 & \text{otherwise, } i.e., \text{ if } p^2 \text{ divides } m \text{ for some prime } p. \end{cases}$$

Let us look at equation (1) above and show that it immediately implies that $N_q(n) \geq 1$ for all prime powers q and all integers $n > 1$. This follows from the following inequality:

$$N_q(n) = \frac{1}{n}\sum_{d|n} \mu(d) q^{n/d} \geq \frac{1}{n}(q^n - q^{n-1} - q^{n-2} - \cdots - q) > 0.$$

Hansen and Mullen [22] give a list of primitive (and thus irreducible) polynomials of degree n over F_p for each prime $p \leq 97$ with $p^n < 10^{50}$.

The next lemma shows that an irreducible polynomial over a finite field is the minimal polynomial for each of its roots. We leave the proof to the reader.

Lemma 1.3.18. *Let $f(x)$ be an irreducible polynomial over F_q and let α be a root of $f(x)$. Then for any $h(x) \in F_q[x]$, $h(\alpha) = 0$ if and only if $f(x)$ divides $h(x)$.*

This lemma shows that if $f(x)$ is irreducible over F_q and $f(\alpha) = 0$, then $F_q(\alpha)$ contains all the roots of $f(x)$; thus $F_q(\alpha) = F_q(\beta)$ whenever α and β are roots of the same irreducible polynomial over F_q.

Lemma 1.3.19. *Let $f(x)$ be an irreducible polynomial of degree m over F_q. Then $f(x)|(x^{q^n} - x)$ if and only if $m|n$.*

Proof. First suppose m divides n, so F_{q^m} is a subfield of F_{q^n}. Let α be a root of $f(x)$ in its splitting field, so $[F_q(\alpha) : F_q] = m$. Then because m divides n, and $F_q(\alpha) = F_{q^m}$, we have $\alpha^{q^n} - \alpha = 0$ in F_{q^m}. This shows that every root of $f(x)$ is a root of $x^{q^n} - x$, and therefore $f(x)|(x^{q^n} - x)$.

For the converse, suppose $f(x)|(x^{q^n} - x)$. Let α be a root of $f(x)$. Then we have the nested fields $F_q \subseteq F_q(\alpha) \subseteq F_{q^n}$. Now $[F_{q^n} : F_q] = n$ and $[F_q(\alpha) : F_q] = m$ so we indeed have that $m|n$. □

The next theorem describes the roots of an irreducible polynomial over a finite field.

Theorem 1.3.20. *If $f(x)$ is an irreducible polynomial of degree m over F_q, then f has a root α in F_{q^m}. Moreover, all of the roots of $f(x)$ are simple and are given by $\alpha, \alpha^q, \alpha^{q^2}, \ldots, \alpha^{q^{m-1}}$.*

Proof. Let α be a root of f in its splitting field. Since $[F_q(\alpha) : F_q] = m$, the element α is in F_{q^m}. Now write $f(x) = \sum_{i=0}^{m} a_i x^i$, where $a_i \in F_q$. Let β be any root of $f(x)$. Then $f(\beta^q) = \sum_{i=0}^{m} a_i (\beta^q)^i = \left(\sum_{i=0}^{m} a_i \beta^i\right)^q = 0$. Hence β^q is a root of f. Similarly β^{q^i} is a root for all $i > 0$.

Suppose that for $1 \leq i < j < m$ we have $\beta^{q^i} = \beta^{q^j}$. Then raising both sides to the q^{m-j} power we obtain $\beta^{q^{i+m-j}} = \beta^{q^m} = \beta$. Hence β is a root of $x^{q^{i+m-j}} - x$, so $m|(i + m - j)$. Thus i and j are congruent modulo m, a contradiction. □

Definition 1.3.21. Let $\alpha \in F_{q^m}$. Then $\alpha, \alpha^q, \alpha^{q^2}, \ldots, \alpha^{q^{m-1}}$ are the *conjugates* of α over F_q.

An element $\alpha \in F_{q^m}$ will have distinct conjugates if and only if its minimal polynomial has degree m. If the minimal polynomial of α has degree d (which must be a divisor of m), then the distinct conjugates of α will be $\alpha, \alpha^q, \ldots, \alpha^{q^{d-1}}$, each repeated exactly m/d times. The proof of the next lemma follows immediately from these observations.

Lemma 1.3.22. *Let $\alpha \in F_{q^m}$ and let the minimal polynomial of α over F_q have degree d. Consider the set $\alpha, \alpha^q, \alpha^{q^2}, \ldots, \alpha^{q^{m-1}}$ of conjugates of α. The elements of this set are distinct if $m = d$; otherwise each conjugate is repeated m/d times.*

An *automorphism* of a field F is a bijection $\phi\colon F \to F$ such that $\phi(x+y) = \phi(x)+\phi(y)$ and $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in F$. Several properties of automorphisms are described in Exercise 1.2.

Our next result describes the set of automorphisms of a finite field.

Theorem 1.3.23. *The distinct automorphisms of F_{q^m} over F_q are given by the functions $\sigma_0, \sigma_1, \ldots, \sigma_{m-1}$ where $\sigma_j\colon F_{q^m} \to F_{q^m}$ is defined by $\sigma_j(\alpha) = \alpha^{q^j}$ for any $\alpha \in F_{q^m}$.*

Proof. We have shown above that if β is a primitive element of F_q and $i \neq j \in \{0, 1, \ldots, m-1\}$, then $\beta^{q^i} \neq \beta^{q^j}$. Hence if $i \neq j$, then $\sigma_i \neq \sigma_j$.

Now let σ be any automorphism of F_{q^m} over F_q and let $f(x)$ be its minimal polynomial over F_q. Direct calculation shows that σ sends β to another root of $f(x)$ so we may assume that $\sigma(\beta) = \beta^{q^k}$. Then $\sigma = \sigma_k$, because the action of an automorphism of a finite field is completely determined by its action on a primitive element of the field. □

The set of automorphisms of F_q forms a group with the operation of function composition (recall that if f and g are functions, the *composition* $g \circ f$ is computed for a point x as $g(f(x))$). This group is called the *Galois group of F_{q^m} over F_q*. It is a cyclic group with

generator $\sigma_1 : \alpha \mapsto \alpha^q$, which is called the *Frobenius automorphism*. The conjugates of α are thus the elements which α is sent to by iterated applications of the Frobenius map.

Note that the subfields of F_{q^m} are exactly the fields of the form F_{q^n} where $n|m$. The proper subgroups of the Galois group of F_{q^m} over F_q are exactly the groups generated by σ_1^n where $n|m$. Moreover, $\sigma_1^n(\alpha) = \alpha$ if and only if $\alpha \in F_{q^n}$. Thus there is a one-to-one correspondence between the subfields of F_{q^m} and the subgroups of its Galois group.

Remark 1.3.24 (Galois theory). In general, if L is an extension of a field K, then the set of automorphisms of L that leave K fixed pointwise is called the *Galois group* of L over K. The field of *Galois theory* is the study of Galois groups, especially in the case when L is a finite extension of K. We have shown that if K is finite and L is a finite extension of K, then the Galois group is cyclic. When K is infinite, the Galois group need not be cyclic, even if L is a finite extension of K.

4. Trace and norm functions

For this section, let $K = F_q$ and $F = F_{q^m}$.

Definition 1.4.1. For $\alpha \in F$, we define the *trace* of α over K as

$$\mathrm{Tr}_{F/K}(\alpha) = \alpha + \alpha^q + \cdots + \alpha^{q^{m-1}}.$$

Equivalently, $\mathrm{Tr}_{F/K}(\alpha)$ is the sum of the conjugates of α. If K is the prime subfield of F, then the trace function is called the *absolute trace*.

Example 1.4.2. Let $K = F_2$ and $F = F_{2^4}$. Then $\mathrm{Tr}_{F/K}(\alpha) = \alpha+\alpha^2+\alpha^4+\alpha^8$. For $K = F_4$ and $F = F_{16}$ we have $\mathrm{Tr}_{F/K}(\beta) = \beta+\beta^4$.

Our first result shows that the range of the trace function is contained in the base field. We will later show that the range equals the base field. In fact, each element of the base field occurs as an element in the range of the trace function equally often.

Lemma 1.4.3. *For any* $\alpha \in F$, $\mathrm{Tr}_{F/K}(\alpha) \in K$.

Proof. Let $f(x)$ be the minimal polynomial over K of $\alpha \in F$, and assume that the degree of $f(x)$ is d. Recall $[F : K] = m$. Let $g(x) = f(x)^{m/d}$ so $g(x)$ is the *characteristic polynomial* of α (it is the m/d-th power of the minimal polynomial of α). Then $g(x)$ will have the same roots as $f(x)$; and these roots are $\{\alpha, \alpha^q, \ldots, \alpha^{q^{d-1}}\}$ each repeated m/d times. So $\mathrm{Tr}_{F/K}(\alpha)$ is exactly the sum of the roots of $g(x)$. Therefore $\mathrm{Tr}_{F/K}(\alpha)$ is the negative of the coefficient of x^{m-1} in $g(x)$, because this coefficient is the negative of the sum of the roots of the polynomial (see Exercise 1.34). □

We now give a second proof of Lemma 1.4.3. This proof illustrates an important technique: to show that an element α of an extension field F is in a subfield K, it is enough to show that α is fixed by the Frobenius automorphism of F over K. For example $\alpha \in F_q$ if and only if $\alpha^q = \alpha$ and, more generally, $\alpha \in F_{q^d}$ if and only if $\alpha^{q^d} = \alpha$.

Second proof of Lemma 1.4.3. We wish to show that $\mathrm{Tr}_{F/K}(\alpha) \in K$ for any $\alpha \in F$. We will show that $\mathrm{Tr}_{F/K}(\alpha)^q = \mathrm{Tr}_{F/K}(\alpha)$, which implies $\alpha \in K$ because K is precisely the set of all elements fixed by the Frobenius map. We compute

$$\left(\mathrm{Tr}_{F/K}(\alpha)\right)^q = \left(\sum_{i=0}^{m-1} \alpha^{q^i}\right)^q = \sum_{i=0}^{m-1} \alpha^{q^{i+1}}.$$

Because $\alpha \in F_{q^m}$, we know that $\alpha^{q^m} = \alpha$. Thus the terms in the second sum are the same as terms in the sum $\mathrm{Tr}_{F/K}(\alpha)$ (but in a different order). □

The trace function is of fundamental importance in the study of finite field theory. The next theorem summarizes some of its properties.

Theorem 1.4.4. *The trace function has the following properties:*

(1) $\mathrm{Tr}_{F/K}(\alpha + \beta) = \mathrm{Tr}_{F/K}(\alpha) + \mathrm{Tr}_{F/K}(\beta)$ *for* $\alpha, \beta \in F$;

(2) $\mathrm{Tr}_{F/K}(c\alpha) = c\,\mathrm{Tr}_{F/K}(\alpha)$ *for* $\alpha \in F$;

(3) *The trace function is a linear map from* F *onto* K;

(4) $\mathrm{Tr}_{F/K}(\alpha) = m\alpha$ *for* $\alpha \in K$;

(5) $\mathrm{Tr}_{F/K}(\alpha^q) = \mathrm{Tr}_{F/K}(\alpha)$ *for* $\alpha \in F$.

Proof. We will only prove part (3), which says that the trace function maps onto the base field K. The other properties are easily verified. Note that $\mathrm{Tr}_{F/K}(0) = 0$; we will show that $\mathrm{Tr}_{F/K}(\alpha) \neq 0$ for some $\alpha \in F$. This implies that the range of the function $\mathrm{Tr}_{F/K}$ is all of F_q, because the range of $\mathrm{Tr}_{F/K}$ is closed under scalar multiplication by elements of F_q.

To see that the range has a nonzero element, note that the kernel of the trace function is exactly the set of roots of the polynomial $\sum_{i=0}^{m-1} x^{q^i}$. This polynomial has degree q^{m-1} and so it has at most q^{m-1} distinct roots, but the field F_{q^m} contains q^m elements. Therefore some element of F_{q^m} is not in the kernel of the trace function. □

The previous result shows that every element of the base field is obtained at least once as the trace of an element in the extension field. It can be shown that the trace function maps onto each element in the base field equally often; we leave the proof of the next result to Exercise 1.24.

Lemma 1.4.5. *For any $\alpha \in K$, we have $|\{\beta \in F \mid \mathrm{Tr}_{F/K}(\beta) = \alpha\}| = q^{m-1}$.*

The next result provides an easy method to generate all of the linear transformations from the extension field F to the subfield K.

Theorem 1.4.6. *For $\beta \in K$ let L_β be the map $\alpha \mapsto \mathrm{Tr}_{F/K}(\beta\alpha)$. Then $L_\beta \neq L_\gamma$ if $\beta \neq \gamma$. Moreover, the linear transformations from F to K are exactly the maps of the form L_β as β varies over the elements of the field K.*

Proof. Because $\mathrm{Tr}_{F/K}$ is a linear map and the map $\alpha \mapsto \beta\alpha$ is linear, it is easy to check that L_β is linear. Suppose $\beta \neq \gamma$, so $\beta - \gamma \neq 0$. Choose α such that $\mathrm{Tr}_{F/K}(\alpha) \neq 0$ and let $\alpha' = (\beta - \gamma)^{-1}\alpha$. Then $\mathrm{Tr}_{F/K}((\beta - \gamma)\alpha') = \mathrm{Tr}_{F/K}(\alpha) \neq 0$, and hence $L_\beta(\alpha') \neq L_\gamma(\alpha')$.

A linear transformation is completely determined by its action on a basis. In our case the field F has a basis of m elements over the base field K. Thus there are at most q^m linear mappings from F_{q^m} to F_q. But $\{L_\beta \mid \beta \in F_{q^m}\}$ is a set of q^m distinct linear maps. Therefore these are all of the linear maps. □

Another consequence of the previous theorem is that the map $\gamma \mapsto L_\gamma$ is a vector space isomorphism from F_q to $\mathrm{Dual}(F_q)$, the set of linear functions from F_q to itself. We refer to Section 6 for a summary of the important properties of $\mathrm{Dual}(V)$ for a finite-dimensional vector space V.

We end our discussion of the trace function with the following theorem.

Theorem 1.4.7 (Transitivity of the trace function). *Suppose that $K \subseteq F \subseteq E$ are nested finite fields. Then for any $\alpha \in E$,*

$$\mathrm{Tr}_{E/K}(\alpha) = \mathrm{Tr}_{F/K}(\mathrm{Tr}_{E/F}(\alpha)).$$

Proof. Let $n = [E : F]$ and $m = [F : K]$ and let $\alpha \in E$ be fixed. The following calculation proves the theorem:

$$\begin{aligned}\mathrm{Tr}_{F/K}(\mathrm{Tr}_{E/F}(\alpha)) &= \sum_{i=0}^{m-1} \mathrm{Tr}_{E/F}(\alpha)^{q^i} = \sum_{i=0}^{m-1} \left(\sum_{j=0}^{n-1} \alpha^{q^{jm}} \right)^{q^i} \\ &= \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \alpha^{q^{jm+i}} = \sum_{k=0}^{mn-1} \alpha^{q^k} = \mathrm{Tr}_{E/K}(\alpha). \quad \square\end{aligned}$$

The trace function is a linear functional from an extension field to a subfield. We now turn our attention to a multiplicative analogue of the trace function known as the norm function.

Definition 1.4.8. The *norm* of an element $\alpha \in F$ over K is defined to be

$$\mathrm{Norm}_{F/K}(\alpha) = \alpha\alpha^q \cdots \alpha^{q^{m-1}} = \prod_{i=0}^{m-1} \alpha^{q^i} = \alpha^{(q^m-1)/(q-1)}.$$

The norm of an element α is thus calculated by taking the product of all of the conjugates of α, just as the trace of α is obtained by taking the sum of all of the conjugates of α.

Theorem 1.4.9. *The norm function has the following properties:*

(1) $\mathrm{Norm}_{F/K}(\alpha\beta) = \mathrm{Norm}_{F/K}(\alpha)\,\mathrm{Norm}_{F/K}(\beta)$ *for* $\alpha, \beta \in F$;

(2) *The norm maps F onto K and F^* onto K^*;*

(3) $\mathrm{Norm}_{F/K}(\alpha) = \alpha^m$ *if* $\alpha \in K$;

(4) *If $K \subseteq F \subseteq E$ are nested finite fields, then*

$$\mathrm{Norm}_{E/K}(\alpha) = \mathrm{Norm}_{F/K}(\mathrm{Norm}_{E/F}(\alpha)) .$$

Proof. We prove only part (4), leaving the reader to check the remaining parts. We use the same notation as in the proof of Theorem 1.4.7 for the transitivity of the trace function. We then have the following calculation:

$$\begin{aligned}\mathrm{Norm}_{F/K}(\mathrm{Norm}_{E/F}(\alpha)) &= \mathrm{Norm}_{F/K}(\alpha^{(q^{mn}-1)/(q^m-1)}) \\ &= (\alpha^{(q^{mn}-1)/(q^m-1)})^{(q^m-1)/(q-1)} \\ &= \alpha^{(q^{mn}-1)/(q-1)} = \mathrm{Norm}_{E/K}(\alpha). \qquad \square\end{aligned}$$

5. Bases

We have seen that every finite field F is a vector space over each of its subfields, and thus has a vector space basis over each of its subfields. We now discuss several different kinds of bases for finite fields, each of which facilitates certain computations. When doing computations in a finite field, there are several important operations: addition, multiplication, computing powers, and finding inverses. With some bases addition will be very easy, while multiplication will be more involved. With other bases, one can do multiplications and calculate inverses quickly, at the cost of more complicated addition.

Let θ be a root of an irreducible polynomial of degree m over F_q. We have already shown that $\{1, \theta, \theta^2, \ldots, \theta^{m-1}\}$ is a basis of the field F_{q^m} over F_q. We now give a name to these kinds of bases.

Definition 1.5.1. Suppose θ is a root of an irreducible polynomial of degree m over F_q. Then the basis $\{1, \theta, \theta^2, \ldots, \theta^{m-1}\}$ for F_{q^m} over F_q is called a *polynomial basis*.

When we use a polynomial basis for F_{q^m} we can regard field elements, which in reality are polynomials in θ of degree at most $m-1$, as vectors. We can then add vectors in the usual way by adding the corresponding coefficients. As was seen in Section 3, in the multiplication table for the field F_{3^2}, field multiplication is more complicated. This is because we must gather terms with like powers of the basis elements when we simplify a product.

Definition 1.5.2. Suppose there is a $\theta \in F_{q^m}$ such that $\{\theta^{q^i} \mid 0 \leq i < m\}$ is a basis for F_{q^m} over F_q. A basis of this form is known as a *normal basis* of F_{q^m} over F_q.

For example, let $\alpha = a_0\theta + a_1\theta^q + \cdots + a_{m-1}\theta^{q^{m-1}}$, so that α is represented by the vector $(a_0, \ldots, a_{m-1})$. Then, as the reader should check, α^q is represented by the shifted vector $(a_{m-1}, a_0, \ldots, a_{m-2})$. Thus if we represent elements using normal basis, it is extremely easy to raise a field element to the power q. Addition will of course still be easy to compute using a normal basis, while the multiplication of field elements is still complicated. In Section 5.3, we will prove that normal bases always exist for any extension field of F_q.

5.1. Tests for independence. Consider F_{q^m} as a vector space over F_q of dimension m. We know there are many bases for this vector space. Given $B = \{\alpha_1, \ldots, \alpha_m\} \subseteq F_{q^m}$, how can we tell if B is a basis for F_{q^m} over F_q? We begin with a test which determines whether a set of elements of F_{q^m} is independent over F_q. If this result is applied to a set containing m elements, it can thus be used to determine whether these elements form a basis of F_{q^m} over F_q (see Lemma A.6.4). We require the following notation.

Definition 1.5.3. Let $\{\alpha_1, \ldots, \alpha_m\}$ be a set of m elements of F viewed as a vector space over a subfield K. We define the *discriminant* $\Delta_{F/K}$ as using a determinant:

$$\Delta_{F/K}(\alpha_1, \ldots, \alpha_m) = \begin{vmatrix} \mathrm{Tr}_{F/K}(\alpha_1\alpha_1) & \cdots & \mathrm{Tr}_{F/K}(\alpha_1\alpha_m) \\ \vdots & \ddots & \vdots \\ \mathrm{Tr}_{F/K}(\alpha_m\alpha_1) & \cdots & \mathrm{Tr}_{F/K}(\alpha_m\alpha_m) \end{vmatrix}.$$

The next two results use the discriminant to provide tests that determine whether a given set of vectors forms a basis.

Theorem 1.5.4. *If $\alpha_1, \ldots, \alpha_m \in F$, then the set $\{\alpha_1, \ldots, \alpha_m\}$ is a basis for F over K if and only if $\Delta_{F/K}(\alpha_1, \ldots, \alpha_m)$ is nonzero where m is the dimension of F over K.*

Proof. Assume first that $\alpha_1, \ldots, \alpha_m$ is a basis. We will show that the discriminant is nonzero by showing that the columns of the matrix

in the definition of the discriminant are linearly independent. Let $C_1, \ldots, C_m$ denote the columns of $\Delta_{F/K}(\alpha_1, \ldots, \alpha_m)$. Assume that $c_1C_1 + \cdots + c_mC_m = 0$ so that for $1 \leq j \leq m$ we have,

$$c_1 \operatorname{Tr}_{F/K}(\alpha_1\alpha_j) + \cdots + c_m \operatorname{Tr}_{F/K}(\alpha_m\alpha_j) = 0,$$

where each c_j lies in the base field K. Let $\beta = c_1\alpha_1 + \cdots + c_m\alpha_m$ so that $\operatorname{Tr}_{F/K}(\beta\alpha) = 0$ for all α in the extension field. This can only happen if $\beta = 0$, which implies $c_1 = \cdots = c_m = 0$.

For the converse, suppose $\Delta_{F/K}(\alpha_1, \ldots, \alpha_m) \neq 0$ and suppose $c_1\alpha_1 + \cdots + c_m\alpha_m = 0$ for some $c_1, \ldots, c_m$ in the base field K. Then for each α_j, $1 \leq j \leq m$,

$$c_1\alpha_1\alpha_j + \cdots + c_m\alpha_m\alpha_j = 0$$

for $1 \leq j \leq m$. We now apply the trace function to obtain

$$c_1 \operatorname{Tr}_{F/K}(\alpha_1\alpha_j) + \cdots + c_m \operatorname{Tr}_{F/K}(\alpha_m\alpha_j) = 0$$

for $1 \leq j \leq m$. By assumption the rows $C_1, \ldots, C_m$ of the matrix in $\Delta_{F/K}(\alpha_1, \ldots, \alpha_m)$ are linearly independent; since we have shown $c_1C_1 + \cdots + c_mC_m = 0$, we must have $c_1 = \cdots = c_m = 0$. Hence the elements $\alpha_1, \ldots, \alpha_m$ are linearly independent; by Lemma A.6.4, they thus form a basis. □

The following result provides an alternative method to determine whether a given set of elements forms a basis. We note that the calculations for this method must be done in the extension field, not in the base field. Working in the extension field may have a significant computational cost. For example, if the base field is F_2 and the extension field is $F_{2^{1000000}}$, then computations in the base field may be much faster than computations in the extension field.

Corollary 1.5.5. *The set $\{\alpha_1, \ldots, \alpha_m\}$ is a basis for F over K if and only if the determinant*

$$\begin{vmatrix} \alpha_1 & \cdots & \alpha_m \\ \alpha_1^q & \cdots & \alpha_m^q \\ \vdots & \ddots & \vdots \\ \alpha_1^{q^{m-1}} & \cdots & \alpha_m^{q^{m-1}} \end{vmatrix}$$

is nonzero.

Proof. Let A be the $m \times m$ matrix whose entry in row i and column j is the element $\alpha_j^{q^{i-1}}$. Let A^T denote the transpose of the matrix A. The reader should verify that the matrix product $A^T A$, is the $m \times m$ matrix B whose entry in row i and column j is $\mathrm{Tr}_{F/K}(\alpha_i \alpha_j)$. Taking determinants we obtain

$$\Delta_{F/K}(\alpha_1, \ldots, \alpha_m) = |B| = |A|^2,$$

where $|A|$ denotes the determinant of the square matrix A. The proof of the corollary now follows from Theorem 1.5.4. □

5.2. Dual bases. In this section, we show how dual bases for finite fields fit into the general theory of dual spaces of finite dimensional vector spaces.

Definition 1.5.6. Two ordered bases $\{\alpha_1, \ldots, \alpha_k\}$ and $\{\beta_1, \ldots, \beta_k\}$ are *complementary* (or *dual*) if $\mathrm{Tr}_{F/K}(\alpha_i \beta_j) = \delta_{ij}$, where $\delta_{ij} = 0$ if $j \neq i$ and $\delta_{ij} = 1$ if $i = j$. An ordered basis is *self-dual* if it is dual with itself.

The definition of a dual basis just given becomes the same as Definition A.6.8 once we identify each $\gamma \in F$ with the linear functional (already defined above)

$$L_\gamma \colon x \mapsto \mathrm{Tr}_{F/K}(\gamma x)$$

(that is, if we apply the vector space isomorphism from F to $\mathrm{Dual}(F)$ that sends γ to L_γ). This can be seen by examining the proof of the next theorem.

Theorem 1.5.7. *Each basis of F_{q^m} has a unique dual basis.*

Proof. Let $B = \{\alpha_1, \ldots, \alpha_m\}$ be a basis for F_{q^m} over F_q. Let $\{\alpha_1^*, \ldots, \alpha_m^*\}$ be the dual basis for B, as in Definition A.6.8. Thus α_i^* is defined so that $\alpha_i^*(\alpha_j)$ is 1 if $i = j$ and 0 otherwise.

By Theorem 1.4.6, for each α_i there is a unique $\gamma_i \in F_{q^m}$ such that α_i^* and L_{γ_i} are equal as linear functionals. Thus $\{L_{\gamma_1}, \ldots, L_{\gamma_m}\}$ is a basis for $\mathrm{Dual}(F_{q^m})$, by Theorem A.6.9.

It only remains to show that $\{\gamma_1, \ldots, \gamma_m\}$ is a basis for F_q. This follows immediately from the fact that $\gamma \mapsto L_\gamma$ is an isomorphism from F_{q^m} to $\mathrm{Dual}(F_{q^m})$. □

5.3. Existence of normal bases. In this section, we prove that every finite field has a normal basis.

Theorem 1.5.8. *For any $m \geq 2$, there is a normal basis for F_{q^m} over F_q.*

Before proving Theorem 1.5.8, we review some terminology from linear algebra. Let T be a linear operator on a vector space V. We say that a polynomial $p(x)$ *annihilates* T if $p(T) = 0$. The *minimal polynomial* of T is the unique monic polynomial of minimal degree which annihilates T. The *characteristic polynomial* of T is the formal determinant of $xI - T$. The degree of the characteristic polynomial is always the dimension of the vector space V. We recall from linear algebra that the minimal polynomial always divides the characteristic polynomial.

A vector v is a *cyclic vector* for T if $\{T^k(v) \mid k \geq 0\}$ spans V. An important result in linear algebra shows that a linear operator has a cyclic vector if and only if the characteristic polynomial of the operator equals its minimal polynomial.

Proof of Theorem 1.5.8. We begin by showing that $x^m - 1$ is the minimal polynomial of the Frobenius map σ. The Frobenius automorphism is annihilated by $x^m - 1$, because $(\sigma^m - I)(a) = a^{q^m} - a = 0$ for all $a \in F_{q^m}$. Let $p(x)$ be a polynomial of degree $< m$ and consider the operator

$$p(\sigma) = a_{m-1}\sigma^{m-1} + a_{m-2}\sigma^{m-2} + \cdots + a_1\sigma + a_0.$$

Now the Frobenius automorphism and its powers form a collection of m distinct automorphisms (including the identity automorphism). By Artin's lemma (Theorem A.4.3), there is some $\alpha \in F$ such that $(p(\sigma))(\alpha)$ is nonzero. Thus $p(x)$ does not annihilate the Frobenius map. Therefore $x^m - 1$ is the minimal polynomial of the Frobenius map σ.

The characteristic polynomial of the Frobenius map is also of degree m. Therefore $x^m - 1$ is also the characteristic polynomial of the Frobenius map σ, because the minimal polynomial must divide the characteristic polynomial.

This shows that the Frobenius map σ has a cyclic vector, say α. It follows from the definitions that a cyclic vector for the Frobenius map σ generates a normal basis; that is, if $\alpha \in F_{q^m}$, then $\{\alpha, \sigma(\alpha) = \alpha^q, \ldots, \sigma^{m-1}(\alpha) = \alpha^{q^{m-1}}\}$ spans F_{q^m} and so forms a basis of F_{q^m} over F_q since F_{q^m} is a vector space of dimension m over F_q. □

Definition 1.5.9. For $f(x) \in F_q[x]$, let $\Phi_q(f)$ denote the number of polynomials over F_q which are of smaller degree than the degree of $f(x)$ and which are relatively prime to f.

A proof of the following result can be obtained along the lines used in proving the corresponding properties for the Euler function ϕ from number theory; see also Lidl and Niederreiter [36, Lemma 3.69].

Lemma 1.5.10. *The function Φ_q has the following properties:*

(1) $\Phi_q(f) = 1$ *if the degree of f is 0;*

(2) $\Phi_q(fg) = \Phi_q(f)\Phi_q(g)$ *if f and g are relatively prime;*

(3) *If f has degree $n \geq 1$, then*

$$\Phi_q(f) = q^n(1 - q^{-n_1}) \cdots (1 - q^{-n_r}),$$

where $\{n_i\}$ are the degrees of the distinct monic irreducible polynomials appearing in the unique factorization of f in $F_q[x]$.

We note that the function Φ_q is analogous to, and in fact has many of the same properties as, Euler's function ϕ from elementary number theory (see Definition A.1.2). For example, $\Phi_q(f^e) = q^{me} - q^{m(e-1)}$ if f is irreducible and has degree m over F_q. We refer to Lidl and Niederreiter [36, Theorem 3.73] for a proof of the following result which enumerates the normal bases of F_{q^m} over F_q.

Theorem 1.5.11. *The number of elements in F_{q^m} that generate normal bases over F_q is $\Phi_q(x^m - 1)$. Because each normal basis has m elements, this shows that there are exactly $\frac{1}{m}\Phi_q(x^m - 1)$ normal bases of F_{q^m} over F_q.*

Our next result illustrates one method to determine whether a particular element α in an extension field generates a normal basis

over the base field. The reader should keep in mind that the following greatest common divisor computation must be performed in the extension field F_{q^m}, not in the base field F_q. We refer to Lidl and Niederreiter [36, Theorem 2.39] for a proof of this result.

Theorem 1.5.12. *The set* $\{\alpha, \alpha^q, \ldots, \alpha^{q^{m-1}}\}$ *is a normal basis for* F_{q^m} *over* F_q *if and only if the greatest common divisor of the polynomials* $x^m - 1$ *and* $\alpha x^{m-1} + \alpha^q x^{m-2} + \cdots + \alpha^{q^{m-1}}$ *is* 1.

As the next theorem shows, not all finite fields have self-dual normal bases. We showed in Section 5.3 that every finite extension field has a normal basis over the base field; thus not all normal bases are self-dual. This question was first resolved by Lempl and Weinberger [33].

Theorem 1.5.13. *The field* F_{q^m} *has a self-dual normal basis over* F_q *if and only if* q *is even and* m *is not a multiple of* 4, *or both* q *and* m *are odd.*

5.4. Existence of primitive normal bases. It is natural to ask whether the generator for a normal basis can be taken to be a primitive element of the field.

Definition 1.5.14. A *primitive normal basis* for an extension field F_{q^m} over F_q is a basis of the form $\{\alpha, \alpha^q, \alpha^{q^2}, \ldots, \alpha^{q^{m-1}}\}$, where α is a primitive element in F_{q^m}.

The problem of proving that primitive normal bases exist is quite difficult. In 1952, Carlitz [6] showed that F_{p^m} has a primitive normal basis over F_p for large m. In 1968, Davenport [10] showed that for any $m \geq 2$ the extension field F_{p^m} has a primitive normal basis over F_p for p prime. Finally, in 1987 Lenstra and Schoof [34] showed that primitive normal bases exist for every prime power q and every positive integer $m \geq 2$. We remark that these proofs used a technique involving the estimation of character sums. These estimates are usually obtained for large values but when q and m are small, they usually require the use of some machine computation to handle the remaining cases not covered by the theoretical techniques.

Theorem 1.5.15 (Lenstra–Schoof). *For any prime power* q *and any integer* $m \geq 2$ *there is a primitive normal basis for* F_{q^m} *over* F_q.

Cohen and Huczynska [7] have given a proof of the primitive normal basis theorem which does not require any machine computation. The problem of determining the number of primitive normal bases for F_{q^m} over F_q (as a function of q and m) remains unsolved. All that is currently known is that the number of primitive normal bases is at least one!

6. Polynomials

In this section we discuss various properties related to polynomials over finite fields. Our first result is that every function defined on a finite field can be represented by a polynomial with coefficients in that field. This is an extremely important property of finite fields; in fact, it characterizes finite fields in the sense that finite fields are the only commutative rings with identity with the property that every function defined on the ring can be realized by a polynomial with coefficients in that ring. The next result tells us how to obtain a polynomial representing a given function over a field.

Theorem 1.6.1 (Lagrange Interpolation Formula). *Let $a_0, \ldots, a_n$ be $n+1$ distinct elements of a field F and let $b_0, \ldots, b_n$ be $n+1$ arbitrary elements in F. Then there is a unique $f(x) \in F[x]$ of degree at most n such that $f(a_i) = b_i$ for $0 \leq i \leq n$.*

Proof. Such a polynomial is given by

$$f(x) = \sum_{0 \leq i \leq n} \left(b_i \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - a_j}{a_i - a_j} \right). \qquad \square$$

For finite fields, we can do somewhat better.

Theorem 1.6.2. *Every function $f\colon F_q \to F_q$ can be represented by a unique polynomial over F_q of degree at most $q-1$.*

Proof. Let f be a function from F_q to itself. Define a polynomial $P_f(x)$ over F_q by

$$P_f(x) = \sum_{a \in F_q} f(a)\left[1 - (x-a)^{q-1}\right].$$

Note that $(a-b)^{q-1}$ is equal to 1 if $a \neq b$ and equal to 0 if $a = b$. A straightforward calculation shows that $P_f(a) = f(a)$ for all $a \in F_q$, so the polynomial $P_f(x)$ indeed represents the function $f(x)$. □

We leave as an exercise the proof of the following result which extends the above theorem to functions with any number of variables.

Theorem 1.6.3. *For any integer $n \geq 1$, let $f\colon F_q^n \to F_q$. Then the function f can be represented by a unique polynomial $P_f(x_1, \ldots, x_n)$ over F_q of degree at most $q-1$ in each variable. Moreover, such a polynomial is given by*

$$P_f(x_1, \ldots, x_n) = \sum_{(a_1,\ldots,a_n)\in F_q^n} \left(f(a_1, \ldots, a_n) \prod_{1\leq i\leq n} [1-(x_i-a_i)^{q-1}] \right).$$

6.1. Counting irreducible polynomials. Recall from Section 3 that $N_q(n)$ denotes the number of distinct monic irreducible polynomials of degree $n \geq 1$ over F_q. In this section, we derive a formula for $N_q(n)$ using the technique from number theory known as Möbius inversion.

Theorem 1.6.4. *For any prime power q and any $n \geq 1$ we have*

$$N_q(n) = \frac{1}{n} \sum_{d|n} \mu(d) q^{n/d},$$

where μ is the Möbius function defined in Section 1.4.

The proof of this theorem will be postponed briefly. We first prove a result about polynomials of the form $x^{q^n} - x$.

Theorem 1.6.5. *Let T_n be the set of all monic irreducible polynomials over F_q of degree dividing n. Then $x^{q^n} - x$ factors in $F_q[x]$ as $\prod_{f\in T_n} f$.*

Proof. Let $f(x)$ be a monic irreducible polynomial in $F_q[x]$. Then $f(x)$ divides $x^{q^n} - x$ if and only if the degree of f divides n. Now $x^{q^n} - x$ has no multiple roots, because its derivative is -1. Hence each monic irreducible polynomial f whose degree divides n occurs once in the factorization of f. Therefore no power of f greater than 1 divides $x^{q^n} - x$. □

We require the following theorem from number theory; a proof is given by Lidl and Niederreiter [37, Theorem 3.24].

Theorem 1.6.6 (The additive Möbius inversion formula). *Let h and H be two functions from the positive integers to an additive Abelian group G. Then*

$$H(n) = \sum_{d|n} h(d)$$

holds for every n if and only if

$$h(n) = \sum_{d|n} \mu\left(\frac{n}{d}\right) H(d)$$

holds for all n. In this case we also have

$$h(n) = \sum_{d|n} \mu(d) H\left(\frac{n}{d}\right)$$

for all n.

Proof of Theorem 1.6.4. By Theorem 1.6.5, we know that $x^{q^n} - x$ is the product of all the monic irreducible polynomials over F_q of degree dividing n. We want to enumerate the set of such polynomials. By comparing degrees, we see that for every n,

$$q^n = \sum_{d|n} dN_q(d) \,.$$

We now apply the Möbius inversion formula with $G = \mathbb{Z}$, $H(n) = q^n$, and $h(n) = nN_q(n)$. This gives

$$nN_q(n) = \sum_{d|n} \mu(d) q^{n/d},$$

and the theorem follows. □

We close this subsection by referring the reader to Exercise 1.37, which asks for a formula for the number of monic irreducibles of degree n over F_q (when $(q, n) = 1$) for which the coefficient of x^{n-1} is equal to a given value of $a \in F_q$.

6.2. Orders of polynomials. The order of an irreducible polynomial is the multiplicative order of any of its roots in the splitting field. In this section, we show the order of a polynomial is well defined. We draw several conclusions about primitive polynomials, ultimately leading us to an algorithm for determining whether a given polynomial is primitive. We begin with an easy lemma.

Lemma 1.6.7. *Let $f \in F_q[x]$ have degree $m \geq 1$ with $f(0) \neq 0$. Then there is an integer e such that $f(x)$ divides $x^e - 1$.*

Proof. The factor ring $F_q[x]/(f)$ contains $q^m - 1$ nonzero elements. Moreover, the q^m classes $x^j + (f)$, $0 \leq j \leq q^m - 1$, are nonzero. Thus we must have $x^r \equiv x^s \pmod{f(x)}$ for some $0 \leq r < s \leq q^m - 1$. It follows that $x^{s-r} \equiv 1 \pmod{f(x)}$ and the result follows. □

Definition 1.6.8. Let $f(x)$ be an arbitrary polynomial in $F_q[x]$. Let $g(x)$ be the unique polynomial such that $g(0) \neq 0$ and $f(x) = x^d g(x)$ for some d. Then the *order* of f is the smallest positive integer e such that $g(x)$ divides $x^e - 1$.

If $f(x)$ is irreducible over F and $f(0) \neq 0$, then the order of $f(x)$ can be seen to be the multiplicative order of x in $F[x]/(f(x))$. We know that if α is any root of $f(x)$, then $F(\alpha)$ contains all the roots of $f(x)$. It follows from Theorem 1.3.20 and the characterization of the automorphisms of a finite field that there is an automorphism of $F(\alpha)$ taking any root of $f(x)$ to any other root of $f(x)$; thus any two roots share the same multiplicative order. This proves the following theorem.

Theorem 1.6.9. *If $f(x) \in F_q[x]$ is irreducible of degree m with $f(0) \neq 0$, then the order of f is equal to the multiplicative order of every root of $f(x)$ in any extension field containing a root of $f(x)$.*

Our next result follows from the fact that the multiplicative group of nonzero elements of F_{q^m} has order $q^m - 1$.

Corollary 1.6.10. *If $f \in F_q[x]$ is irreducible of degree m, then the order of f divides $q^m - 1$.*

Corollary 1.6.11. *If $f(x)$ is irreducible over F_q and one root of $f(x)$ is primitive, then every root of $f(x)$ is primitive.*

Proof. This follows from the fact that all roots of an irreducible polynomial over F_q have the same order. □

In light of the previous corollary, a polynomial is defined to be *primitive* if any of its roots are primitive elements.

In our future work we will not require a formula for calculating the order of an arbitrary polynomial over the field F_q and so we refer the reader to Lidl and Niederreiter [36, Theorem 3.11] for such a result.

Definition 1.6.12. For a polynomial $f(x) = \sum_{i=0}^{n} a_i x^i \in F_q[x]$ of degree n, the *reciprocal polynomial* $f^*(x)$ of $f(x)$ is $f^*(x) = x^n f(1/x) = \sum_{i=0}^{n} a_i x^{n-i}$.

We leave to Exercise 1.35 the proof of our next result.

Proposition 1.6.13. *The reciprocal of an irreducible polynomial is again irreducible and the reciprocal of a primitive polynomial is again primitive.*

We close this section by stating two tests for determining whether a polynomial is primitive. We refer to Lidl and Niederreiter [36, Theorems 3.16 and 3.18] for proofs of these results.

Theorem 1.6.14. *A polynomial $f(x)$ of degree m over F_q is primitive if and only if f is monic, $f(0) \neq 0$, and the order of f is $q^m - 1$.*

The following theorem provides an algorithm which can be used to determine if a given polynomial is a primitive polynomial over F_q.

Theorem 1.6.15. *A monic polynomial $f(x)$ of degree m over F_q is primitive if and only if the following hold:*

(1) $(-1)^m f(0)$ *is a primitive element in* F_q.

(2) *The least r for which x^r is congruent to an element $a \in F_q$ is*

$$r = \frac{q^m - 1}{q - 1}.$$

6.3. Linearized polynomials. Let $L(x) = \sum_{i=0}^{n} \alpha_i x^{q^i}$, where $\alpha_i \in F_{q^m}$. A polynomial of this form is called a *linearized polynomial* (another name is *q-polynomial* because the exponents are all powers of q). These polynomials form an important class of polynomials over finite fields because they induce linear functions from F_{q^m} to F_q. We state this fact in the next theorem.

Theorem 1.6.16. *Let $L(x)$ be a linearized polynomial. Then for all $\alpha, \beta \in F_{q^m}$ and all $c \in F_q$:*

(1) $L(\alpha + \beta) = L(\alpha) + L(\beta)$,

(2) $L(c\alpha) = cL(\alpha)$.

Theorem 1.6.17. *Let $L(x)$ be a nonzero linearized polynomial over F_{q^m} and assume that the roots of $L(x)$ lie in the field F_{q^s}, an extension field of F_{q^m}. Then each root of $L(x)$ has the same multiplicity, which is either 1, or a power of q. Moreover, the roots form a linear subspace of the vector space F_{q^s}.*

Proof. From the previous theorem it is clear that the roots of $L(x) = \sum_{i=0}^{n} \alpha_i x^{q^i}$ form a subspace of F_{q^s}. It is easy to check that $L'(x) = \alpha_0$ so, by the derivative test, $L(x)$ only has simple roots if $\alpha_0 \neq 0$. Otherwise, for some $k \geq 1$ we would have $\alpha_0 = \cdots = \alpha_{k-1} = 0$, with $\alpha_k \neq 0$. Then a computation shows that

$$L(x) = \sum_{i=k}^{n} \alpha_i x^{q^i} = \sum_{i=k}^{n} \alpha_i^{q^{mk}} x^{q^i} = \left(\sum_{i=k}^{n} \alpha_i^{q^{(m-1)k}} x^{q^{i-k}} \right)^{q^k}.$$

This polynomial is the q^k-th power of a linearized polynomial having only simple roots. Hence each root of $L(x)$ has multiplicity q^k and the proof is complete. □

We have already seen several examples of linearized polynomials. The Frobenius automorphism $x \mapsto x^q$ is one such example, and the trace function provides another important example of a linearized polynomial over F_q. Properties of linearized polynomials can be used to obtain a formula for the number of normal bases of an extension field; in particular, in the proof of Theorem 1.5.11. They are also useful in the construction of sets of mutually orthogonal frequency squares.

6.4. Permutation polynomials. In this final section we briefly discuss permutation polynomials, a class of polynomials which are not only interesting in their own right, but which also have various applications in combinatorics and cryptography. We begin by defining a polynomial $f(x)$ over F_q to be a *permutation polynomial* if $f(x)$ induces an injective mapping on the field F_q. Recall that a mapping from a finite set to itself is injective if and only if it is surjective, so we may also say that $f(x)$ is a permutation polynomial if it induces a bijective mapping from F_q onto F_q.

In Exercise 1.43, we ask the reader to show that if $f(x)$ is a permutation polynomial, then so is $af(x+b)+c$ for all $a \neq 0, b, c \in F_q$, and these are distinct. Thus, given one permutation polynomial, we can easily generate $q^2(q-1)$ others. Given a polynomial $f(x)$ over F_q, how does one determine if $f(x)$ is actually a permutation polynomial on F_q? One could of course substitute the q field elements into the polynomial and then check if the q image values $f(a)$ are distinct. This is not efficient, however, if q is very large. In fact no efficient test, in terms of the coefficients of the polynomial, is known. We state the following test; a proof is given by Lidl and Niederreiter [37, Theorem 7.4].

Theorem 1.6.18 (Hermite–Dickson criterion). *Let $q = p^m$ where p is a prime. Then a polynomial $f(x)$ over F_q is a permutation polynomial on F_q if and only if the following two conditions hold:*

(1) *$f(x)$ has exactly one root in F_q.*

(2) *For each integer t with $1 \leq t \leq q-2$ and $t \not\equiv 0 \pmod{p}$, the reduced polynomial $(f(x))^t \pmod{x^q - x}$ has degree at most $q-2$.*

We obtain a simple corollary by applying the Hermite–Dickson criterion (with $t = (q-1)/d$) to a polynomial of degree $d > 1$ over F_q.

Corollary 1.6.19. *There is no permutation polynomial of degree $d > 1$ over F_q if d divides $q-1$.*

Theorem 1.6.20. *The polynomial x^n induces a permutation of F_q if and only if $(n, q-1) = 1$.*

Proof. The proof follows from the fact that the multiplicative group of F_q is cyclic; see Lemma A.4.2. □

In Exercise 1.48 we ask the reader to provide a different proof of this theorem.

We briefly consider another important class of permutation polynomials. If $a \in F_q$ and $n \geq 2$ is an integer, we define the *Dickson polynomial of degree n and parameter a* by

$$D_n(x,a) = \sum_{i=0}^{\lfloor n/2 \rfloor} \frac{n}{n-i} \binom{n-i}{i} (-a)^i x^{n-2i}.$$

For $n = 0$ we define $D_0(x,a) = 2$ and similarly we define $D_1(x,a) = x$. Note that $D_n(x,0) = x^n$ for $n \geq 1$, so Dickson polynomials may be viewed as generalizations of the cyclic, or power polynomial, x^n. These polynomials satisfy a second order recurrence, namely $D_{n+2}(x,a) = xD_{n+1}(x,a) - aD_n(x,a)$ for $n \geq 0$; see Lidl *et al.* [35, p. 10]. Dickson polynomials were first studied by L. E. Dickson in his PhD thesis in 1896. If one works over the field of complex numbers, Dickson polynomials are related to the classical Chebyshev polynomials, as explaned by Lidl *et al.* [35, p. 9].

Dickson polynomials satisfy a functional equation which can be obtained as follows. Let $x = y + a/y$ for some $y \in F_{q^2}$. We can always find such a y by solving the quadratic equation $y^2 - xy + a = 0$, which must have a solution in the field F_{q^2}. Using Waring's formula (see Lidl and Niederreiter [37, Theorem 1.76] for a statement and proof of the formula), we obtain

$$u_1^n + u_2^n = \sum_{i=0}^{\lfloor n/2 \rfloor} \frac{n}{n-i} \binom{n-i}{i} (-u_1 u_2)^i (u_1 + u_2)^{n-2i},$$

so $u_1^n + u_2^n = D_n(u_1 + u_2, u_1 u_2)$. If we substitute $u_1 = y$, $u_2 = a/y$, and $x = y + a/y$, we obtain the extremely useful functional equation for Dickson polynomials:

$$D_n(x,a) = y^n + \frac{a^n}{y^n}.$$

We refer to Lidl *et al.* [35] for further details and basic properties of Dickson polynomials.

The following result provides a class of permutation polynomials over F_q, where we note that, as long as $a \in F_q$ is nonzero, the value of a does not enter into the problem of determining whether the Dickson polynomial $D_n(x, a)$ permutes the field F_q. For nonzero $a \in F_q$, whether or not the polynomial $D_n(x, a)$ induces a permutation polynomial on F_q is determined only by the greatest common divisor of n and $q^2 - 1$.

Theorem 1.6.21. *For any nonzero $a \in F_q$, the Dickson polynomial $D_n(x, a)$ is a permutation polynomial on F_q if and only if $(n, q^2-1) = 1$.*

Proof. For sufficiency, assume that $D_n(b, a) = D_n(c, a)$ for some $b, c \in F_q$. Then there are $\beta, \gamma \in F_{q^2}^*$ with $\beta + a/\beta = b$ and $\gamma + a/\gamma = c$. Using the functional equation implies after some simplification, that

$$(\beta^n - \gamma^n)(\beta^n \gamma^n - a^n) = 0.$$

Since $(n, q^2 - 1) = 1$, x^n is a permutation polynomial on F_q and hence if $\beta^n = \gamma^n$, then $\beta = \gamma$ so that $b = c$. Also if $\beta^n \gamma^n = a^n$, then $\beta = a/\gamma$ and it again follows that $b = c$. Hence $D_n(x, a)$ is a permutation polynomial on F_q.

Conversely, assume that $(n, q^2 - 1) = d > 1$. If d is even, then q is odd and n is even so the Dickson polynomial contains only even powers of x. Hence for nonzero $c \in F_q$, $D_n(c, a) = D_n(-c, a)$, and thus the polynomial $D_n(x, a)$ is not a permutation polynomial of F_q. Thus we may assume that d is odd, which means that there is an odd prime r dividing d and hence also dividing n. Since r is a prime, it either divides $q - 1$ or $q + 1$.

If r divides $q - 1$, then the equation $x^r = 1$ has r solutions in F_q. Because $r \geq 3$, there is an element $b \in F_q$ with $b \neq 1$ or a with $b^r = 1$ so $b^n = 1$. The functional equation implies that

$$D_n(b + ab^{-1}, a) = 1 + a^n = D_n(1 + a, a).$$

If $b + ab^{-1} = 1 + a$, then $b = 1$ or $b = a$, which is a contradiction. Hence $D_n(x, a)$ does not permute the field F_q.

If r divides $q + 1$, choose $\gamma \in F_{q^2}^*$ so that $\gamma^{q+1} = a$. The equation $x^r = 1$ has r solutions in F_{q^2} so there is a $\beta \in F_{q^2}$ with $\beta^r = 1$ and

$\beta \neq 1$ and $\beta \neq a\gamma^{-2}$. Hence $\beta^{q+1} = 1$ and $\beta^n = 1$ so once again from the functional equation we have

$$D_n(\gamma + a\gamma^{-1}, a) = D_n(\beta\gamma + a(\beta\gamma)^{-1}, a).$$

In addition, $\gamma + a\gamma^{-1} = \gamma + \gamma^q$ and $\beta\gamma + a(\beta\gamma)^{-1} = \beta\gamma + (\beta\gamma)^q \in F_q$. If $\beta\gamma + a(\beta\gamma)^{-1} = \gamma + a\gamma^{-1}$, then $\beta = 1$ or $\beta = a\gamma^{-2}$, a contradiction. Hence $D_n(x, a)$ is not a permutation polynomial on F_q. □

Dickson polynomials are intimately tied to a famous conjecture posed by Schur [57] in 1923. We first note from Exercise 1.45 and Exercise 1.46 that when considered modulo a prime p, the polynomial x^n and the Dickson polynomial $D_n(x, a)$ permute the field F_p for infinitely many primes p. Are there other polynomials with integer coefficients which permute F_p for infinitely many primes p? The answer is yes, because polynomials of the form $ax^n + b$, where a is a nonzero integer, will be permutation polynomials for infinitely many primes p. Are there still other such integral polynomials or classes of integral polynomials?

Conjecture 1.6.22 (Schur 1923). If $f(x)$ is a polynomial with integer coefficients which is a permutation polynomial of F_p (when considered modulo p) for infinitely many primes p, then $f(x)$ must be a composition of binomials $ax^n + b$ and Dickson polynomials.

This conjecture remained open until 1970; see Lidl *et al.* [35] for a proof of the Schur Conjecture along with discussion of many other algebraic and number theoretic properties of Dickson polynomials.

We close this section by alluding to the fact that Dickson polynomials have applications in several areas of combinatorics. Exercise 2.8 illustrates how they can be used to construct complete sets of MOLS of any prime power order q. For p a prime, it is conjectured that all complete sets of $p-1$ MOLS of order p are isomorphic, and thus one could argue that all such sets of $p-1$ MOLS come from a set of $p-1$ MOLS obtained by using Dickson polynomials. A class of translation planes, called *j-planes*, and several infinite families of flocks of a cone in the projective space $PG(3, q)$, as well as certain translation planes, can be constructed from Dickson polynomials of small degrees. Dickson polynomials have also been used to construct ovals in projective

planes $PG(2,q)$, with $q = 2^e$, and e odd. Ovals can be constructed from certain power permutation polynomials x^n. We refer the reader to Lidl *et al.* [35] for more details.

We also briefly mention several applications of power and Dickson polynomials to cryptography. If one wants to quickly encipher messages, one can simply choose a value of n with $(n, q-1) = 1$, so that the polynomial x^n is a permutation polynomial over F_q. Then one can encrypt the message m by calculating m^n. This can be easily deciphered by using the inverse polynomial x^k where $kn \equiv 1 \pmod{q-1}$. One can also use a Dickson analogue of this method of encryption. There is a Dickson polynomial analogue of the RSA cryptosystem for the secure transmission of information; and there is a Dickson Diffie–Hellman scheme for the exchange of keys; see Chapter 4 for details.

7. Notes

There are numerous books dealing with theoretical as well as applied aspects of finite fields. Without question, the standard reference on the subject is by Lidl and Niederreiter [37]. This is a very comprehensive and thorough survey of the theory and applications of finite fields. This volume also contains an incredibly complete list of references through 1983. A second book by Lidl and Niederreiter [36] is a very readable and shortened textbook version of [37]. It also contains a chapter on cryptographic applications of finite fields. We also refer to Jungnickel [25] for a very readable textbook which approaches many of the finite field topics from a computational point of view. Small [59] and Wan [65] provide very readable texts dealing with various aspects, especially theoretical aspects, of finite fields. The book by Lidl, Mullen, and Turnwald [35] provides a summary of numerous algebraic and number theoretic properties of Dickson polynomials.

The following books discuss additional topics related to finite fields. Menezes [43] provides a treatment of applications of finite fields to elliptic curves and elliptic curve cryptosystems. Hachenberger [21] provides an extensive discussion of topics related to, and extensions

of, normal bases over finite fields. Finally, we mention the very comprehensive monograph by Shparlinski [58] that deals with many theoretical as well as computational topics related to finite fields. This monograph also contains a collection of 3075 references.

8. Exercises

1.1. For a commutative ring R of characteristic p, show that

$$(a_1 + \cdots + a_s)^{p^n} = a_1^{p^n} + \cdots + a_s^{p^n}$$

for every $n \geq 1$ and $a_1, a_2, \ldots, a_s \in R$.

1.2. Let F be a field (not necessarily finite) and let ϕ be an automorphism of F. Show that $\phi(0) = 0$ and $\phi(1) = 1$. Show that $\phi(x^{-1}) = (\phi(x))^{-1}$ for all $x \in F$.

1.3. Fix a field F. Let S be the set of expressions of the form $f(x,y)/g(x,y)$ where f and g are polynomials over F using the variables x and y and $g(x,y)$ is not the zero polynomial. If a suitable equivalence relation E is placed on S, then S/E will be a field under the ordinary operations of fraction addition and multiplication. State this equivalence relation and prove that S/E is a field. This field is called the *field of rational functions* in the variables x, y over F and is often denoted $F(x,y)$. Note that even if F is finite, the field of rational functions in one or more variables over F will be infinite.

1.4. Let p be a prime number and let $F_p(x,y)$ be the field of fractions in the variables x, y over F_p (see Exercise 1.3 for a definition of this field). Let $K = F_p(x^p, y^p)$ be the smallest subfield of $F_p(x,y)$ containing F_p and the elements x^p and y^p. Show that $[F_p(x,y) : K] = p^2$, and that $a^p \in K$ for any $a \in F_p(x,y)$. Use these facts to prove that $F_p(x,y)$ is a finite extension of K that is not a simple extension.

1.5. Let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be a monic polynomial of degree n over a field F. Consider the following algorithm. If $a_0 = 0$, calculate $f(x)/x$. If $a_0 \neq 0$, let j with $1 \leq j \leq n$ be the smallest value so that $a_j \neq 0$. In this case calculate

$$(x - \frac{a_0}{a_j})f(x) + \frac{a_0^2}{a_j}.$$

Show that for any polynomial $f(x)$ over F of degree $n \geq 1$, upon iteration of this algorithm, we always obtain a sequence of polynomials that ends at 1. Hint: Use induction on the degree n.

This problem is analogous to the *Collatz* $3n+1$ *problem* for the integers where one divides by 2 if the number n is even and calculates $3n+1$ if n is odd. The famous, and still unsolved, $3n+1$ conjecture postulates that starting with any positive integer n, this algorithm will produce a sequence of positive integers which always ends at 1.

Finite Fields.

1.6. Construct the addition and multiplication tables for $F_2[x]/(x^3+x^2+x)$. Determine whether or not this ring is a field.

1.7. Let $f(x) = 15x + 2x^{100}$. Evaluate $f(4)$ in F_3 and F_7.

1.8. Construct a field with 8 elements.

1.9. If p is a prime, then the field F_p is the same (isomorphic) as the ring $\mathbb{Z}_p$ of integers modulo p. Explain why the field F_4 is not the same as the ring $\mathbb{Z}_4$. In fact, it turns out that if $m > 1$, the field $F_{p^m} \neq \mathbb{Z}_{p^m}$, the ring of integers modulo p^m. Explain why.

1.10. Assume that $F_{2^6} = F_2(\alpha)$, where $\alpha^6 + \alpha^5 + 1 = 0$. Compute α^{40} in F_{2^6} as a polynomial of degree less than 6 in α. Find the unique $n < 64$ such that $\alpha^n = \alpha^4 + \alpha + 1$.

1.11. Determine the multiplicative order of each nonzero element in F_{17}.

1.12. Construct a field of order 16 and determine the multiplicative order of each nonzero element in the field.

1.13. Determine the number of functions mapping F_q to itself. How many of these functions can be represented by a polynomial over F_q? By representing a function $f(x)$ by a polynomial $P_f(x)$ over F_q, we mean, as in Theorem 1.6.2, that $P_f(a) = f(a)$ for all $a \in F_q$.

1.14. Show that any polynomial of degree two over F_q splits over F_{q^2} into a product of two linear polynomials.

1.15. Show that the sum of all elements of a finite field is 0, except for the field F_2.

1.16. Prove that $(f(x))^q = f(x^q)$ for $f(x) \in F_q[x]$. The property described in this exercise is of great use in finite field calculations.

1.17. Let F be a finite extension of $K = F_q$ and let $\alpha = \beta^q - \beta$ for some $\beta \in F$. Prove that $\alpha = \gamma^q - \gamma$ with $\gamma \in F$ if and only if $\beta - \gamma \in K$.

1.18. Construct the field F_{16} by viewing it as a vector space of dimension four over the base field F_2. Similarly, construct the same field by viewing it as a vector space of dimension two over the base field F_4. We know since up to isomorphism there is only one field of a given prime power order, that these two constructions must lead to the same field F_{16}. Can you construct an isomorphism between these two fields?

1.19. Prove Lemma 1.2.10.

1.20. Prove Corollary 1.3.11.

1.21. Let K be an extension of F and let $\alpha \in K$ have minimal polynomial $p(x) \in F[x]$. If $q(x) \in F(x)$ is such that $q(\alpha) = 0$ in F, then $p(x)|q(x)$.

Trace and Norm.

1.22. Let F be a finite extension of the finite field K of characteristic p. Prove that $\mathrm{Tr}_{F/K}(\alpha^{p^n}) = (\mathrm{Tr}_{F/K}(\alpha))^{p^n}$ for all $\alpha \in F$ and $n \geq 1$.

1.23. Prove that the dual basis of a normal basis of F_{q^m} over F_q is again a normal basis of F_{q^m} over F_q.

1.24. Prove Lemma 1.4.5 which states that the trace function maps F_{q^m} onto each element of the base field equally often. What can be said about the distribution of values for the norm function?

1.25. Let F be a finite extension of $K = F_q$. Prove that for $\alpha \in F$ we have $\mathrm{N}_{F/K}(\alpha) = 1$ if and only if $\alpha = \beta^{q-1}$ for some $\beta \in F^*$.

Bases.

1.26. Show that if one takes into account the order of the elements, then the total number of distinct bases of F_{q^m} over F_q is given by

$$(q^m - 1)(q^m - q) \cdots (q^m - q^{m-1}).$$

1.27. Construct a self-dual basis of F_{2^4} over F_2.

1.28. Prove that if $\{\alpha_1, \ldots, \alpha_m\}$ is a basis of $F = F_{q^m}$ over $K = F_q$, then $\mathrm{Tr}_{F/K}(\alpha_i) \neq 0$ for at least one i, $1 \leq i \leq m$.

1.29. Prove that there exists a normal basis $\{\alpha, \alpha^q, \ldots, \alpha^{q^{m-1}}\}$ of $F = F_{q^m}$ over $K = F_q$ with $\mathrm{Tr}_{F/K}(\alpha) = 1$.

1.30. Show that there is a self-dual normal basis of F_4 over F_2, but no self-dual normal basis of F_{16} over F_2.

Polynomials.

1.31. Let F_q be a finite field of characteristic p. Prove that $f \in F_q[x]$ satisfies $f'(x) = 0$ if and only if $f(x) = g(x^p)$ for some polynomial $g(x)$ in $F_q[x]$.

1.32. Show that $f(x) = x^4 + x + 1 \in F_2[x]$ is irreducible over F_2. Then construct the addition and multiplication tables for the simple extension $F_2(\theta)$, where θ is a root of f.

1.33. In the previous exercise the reader was asked to show that $p(x) = x^4 + x + 1$ is irreducible over F_2. Is this polynomial primitive over F_2? That is, does $p(x)$ have a root which generates the multiplicative group $F_{2^4}^*$ of all nonzero elements in the field F_{2^4}?

1.34. Let $g(x)$ be a monic polynomial with degree m over a field F such that g has m distinct roots. Show that the coefficient of x^{m-k} in $g(x)$ is the sum of all $k!\binom{m}{k}$ ordered products of k roots of g multiplied by $(-1)^k$.

1.35. Prove that the reciprocal polynomial of an irreducible polynomial f over F_q is again irreducible over F_q. Prove that the reciprocal of a primitive polynomial is also primitive.

1.36. Show that if f is a self-reciprocal polynomial of degree m ($f(x) = f^*(x) = x^m f(1/x)$) in $F_q[x]$ of degree $m > 1$, then m must be even.

1.37. If $(n, p) = 1$, show that the number of monic irreducible polynomials of degree n over F_q which have the trace coefficient equal to any fixed value of $a \in F_q$ is given by

$$\frac{1}{nq} \sum_{d|n} \mu(d) q^{n/d},$$

where p is the characteristic of F_q.

Linearized polynomials.

1.38. Prove Theorem 1.6.16, which states that linearized polynomials indeed give linear mappings.

1.39. Show that a linearized polynomial $L(x)$ is a permutation polynomial on the field F_{q^r} if and only if $L(x)$ only has the root 0 in F_{q^r}.

1.40. Let $L(x) = \sum_{s=0}^{r-1} a_s x^{q^s}$ be a linearized polynomial over the field F_{q^r}. It can be shown, as for example in Lidl and Niederreiter [37, pp. 361–362], that $L(x)$ induces a permutation on the field F_{q^r} if and only if the determinant of the matrix $A_L = (a_{i-j}^{q^j})$ where $0 \leq i, j \leq r-1$ and the subscript $i-j$ is calculated modulo r, is nonzero. In fact, the set of all such permutation polynomials forms a group under composition of functions modulo $(x^{q^r} - x)$, called the *Betti–Mathieu group*, which is isomorphic to the *general linear group* $\mathrm{GL}(r, F_q)$ of all nonsingular $r \times r$ matrices over F_q under matrix multiplication.

(i) Show that the polynomial x^q induces a permutation on the field F_{q^r} for any $r \geq 1$.

(ii) If $q = 3^e$, show that the polynomial $x^3 - ax$ is a permutation polynomial on the field F_q if and only if a is not a square, that is, if $a \neq b^2$ for any $b \in F_q$.

Permutation polynomials.

1.41. For $q = 2$, 3, and 5, construct all permutation polynomials on the field F_q of degree no higher than $q-1$.

1.42. Show that there are $q!$ permutations on the field F_q. Show that these permutations form a group under functional composition modulo $x^q - x$, and that this group is isomorphic to the symmetric group S_q.

1.43. If $f(x)$ is a permutation polynomial over the field F_q, show that for all $a \neq 0, b, c \in F_q$, the polynomial $af(x+b)+c$ is also a permutation polynomial over F_q. Show that the set $\{af(x+b)+c \mid a, b, c \in F_q, a \neq 0\}$ contains $q^2(q-1)$ polynomials.

1.44. For an odd prime p, show that the polynomial $x^{p-2} + x^{p-3} + \cdots + x^2 + 2x + 1$ represents the permutation which consists of the transposition (01) that interchanges 0 and 1 and fixes all other elements of the field F_p.

1.45. Show that the polynomial x^n permutes the field F_p for infinitely many primes p if and only if n is odd. Hint: use Dirichlet's theorem which states that if $(a, b) = 1$, then the arithmetic progression $ax + b$ contains infinitely many primes as x ranges over the positive integers. Note that the special case of Dirichlet's theorem when $a = 2$ and $b = 1$ is Euclid's theorem that there are infinitely many primes.

1.46. Show that the Dickson polynomial $D_n(x, a)$, with $a \neq 0$, permutes the field F_p for infinitely many primes p if and only if $(n, 6) = 1$. Hint: use Dirichlet's theorem from Exercise 1.45.

1.47. For a polynomial $f(x)$ over F_q, let the *value set* V_f *of* $f(x)$ be defined by $V_f = \{f(a) \mid a \in F_q\}$. Show that the value set V_{x^n} of the polynomial x^n over F_q has cardinality $|V_{x^n}| = 1 + (q-1)/(n, q-1)$, where (a, b) denotes the greatest common divisor of the integers a and b.

1.48. Use the preceding exercise to show x^n induces a permutation polynomial on F_q if and only if $(n, q-1) = 1$.

http://dx.doi.org/10.1090/stml/041/02

Chapter 2

Combinatorics

1. Introduction

Combinatorics is a very large, fascinating, and extremely active area of mathematics which, in addition, has numerous practical applications. In this chapter, we briefly discuss several area of combinatorics in which finite fields play major roles. These include sets of mutually orthogonal latin squares, affine and projective planes, balanced incomplete block designs, and Hadamard matrices. We begin with a discussion of latin squares and sets of orthogonal latin squares. We will later see that latin squares are closely related to finite geometries and algebraic codes, and have applications in cryptography.

2. Latin squares

Latin squares have fascinated individuals for centuries. Because they have so many interesting properties, useful applications, and are associated with many open questions, they are still being actively studied today.

Definition 2.2.1. A *latin square of order* n is an $n \times n$ matrix containing n distinct symbols (usually denoted by $0, 1, \ldots, n-1$) such that each row and column of the matrix contains each symbol exactly once.

Example 2.2.2. A latin square of order 6:

0	1	2	3	4	5
1	2	3	4	5	0
2	3	4	5	0	1
3	4	5	0	1	2
4	5	0	1	2	3
5	0	1	2	3	4

While the reader can no doubt convince himself or herself that there is indeed at least one latin square of order n for any positive integer n, the following gives a more formal proof of this fact.

Theorem 2.2.3. *There is a latin square of order n for each $n \geq 1$.*

Proof. The addition table of the additive group $\mathbb{Z}/n\mathbb{Z}$ of integers modulo n is always a latin square of order n. In fact, the Cayley table of any finite group of order n is a latin square of order n. □

Although the latin squares constructed in Theorem 2.2.3 are the Cayley tables of groups, not every latin square is the Cayley table of a group; see Exercise 2.13.

A central problem in the theory of latin squares is to determine how many latin squares of each size exist.

Definition 2.2.4. Let L_n denote the number of distinct latin squares of order n.

How can we determine the value of L_n? We begin by restricting our focus to reduced latin squares. A latin square of order n is said to be *reduced* if its first row and first column are in the standard order $0, 1, \ldots, n-1$.

Definition 2.2.5. Let l_n denote the number of distinct reduced latin squares of order n.

The value of l_n can be computed by hand for small values of n, and by computer for slightly larger values of n, but the value of l_n is unknown for $n \geq 12$. As of the writing of this book, no formula for l_n has been found and it seems possible that none exists. The next theorem shows that we can compute L_n from l_n, and *vice versa*.

Theorem 2.2.6. *For any $n \geq 2$, $L_n = n!\,(n-1)!\,l_n$.*

Proof. Given a latin square of order n, it is easy to see that, by interchanging (permuting) the n columns, we can obtain $n!$ latin squares of order n and each such square will be different. Similarly, by interchanging the last $n-1$ rows, we can obtain $(n-1)!$ distinct squares of order n. Moreover, each of these squares will be different from those obtained by the column permutations. Thus if we begin with a reduced latin square of order n, we can generate a total of $n!(n-1)!$ distinct latin squares of order n, exactly one of which is reduced. Since there are l_n distinct reduced latin squares of order n, we can construct $n!(n-1)!l_n$ distinct latin squares of order n. Moreover, by permuting the rows and columns of any latin square, we can obtain a unique reduced latin square of order n, and so the result follows. □

The values of l_n for $2 \leq n \leq 7$ are shown in the following table. In Exercise 2.1, we ask the reader to compute the first three values by hand, and the last three values by machine.

n	2	3	4	5	6	7
l_n	1	1	4	56	9,408	16,942,080

The values of l_{10} and l_{11} have recently been determined using large amounts of machine computation. Using these values, we obtain:

$$L_{10} = 10! \cdot 9! \cdot 7,580,721,483,160,132,811,489,280,$$
$$L_{11} = 11! \cdot 10! \cdot 5,363,937,773,277,371,298,119,673,540,771,840.$$

See McKay and Wanless [42] for a discussion of the computational effort used to determine the value l_{11} and verify the values of l_n for $n \leq 10$.

2.1. Sets of orthogonal latin squares. Given two latin squares of the same size, we can superimpose them to create a single square of ordered pairs, as shown below.

0	1	2	0	1	2	(0,0)	(1,1)	(2,2)
1	2	0	2	0	1	(1,2)	(2,0)	(0,1)
2	0	1	1	2	0	(2,1)	(0,2)	(1,0)
Square 1			*Square 2*			*Superimposed square*		

Notice that each of the nine possible ordered pairs based on three symbols appears exactly once in the superimposed square. Extending this concept to latin squares of any order gives the following definition.

Definition 2.2.7. Two latin squares of order n are *orthogonal* if when the squares are superimposed each of the n^2 ordered pairs of symbols appears exactly once. A collection $\{L_1, \ldots, L_t\}$ of $t \geq 2$ latin squares of the same order is said to be *mutually orthogonal* if every pair of distinct squares in the collection is orthogonal.

Let $N(n)$ denote the size of the largest collection of mutually orthogonal latin squares (MOLS) of order n. We remind the reader that $N(n)$ denotes the maximum number of MOLS of order n that exist, not the maximum number that we have been able to construct. The remainder of this section will present several results about the function $N(n)$.

Theorem 2.2.8. *$N(n) \leq n - 1$ for any $n \geq 2$.*

Proof. It is easy to see that the n symbols in a latin square L of order n can be renamed in any way without affecting the square's orthogonality with a second latin square, say M, also of order n. Thus in any set of orthogonal squares of order n, we may assume that the first row is in the standard order $0, 1, , \ldots, n-1$. The reader is asked to prove these statements in Exercise 2.24.

Now consider the elements that occur in the second row, first column of each of the $N(n)$ orthogonal latin squares of order n. Clearly the symbol in this cell in any of the $N(n)$ squares cannot be 0 or that square would not be latin. Since the symbols in this cell must be different in each square to preserve orthogonality, there are at most $n-1$ possibilities. Thus we have that $N(n) \leq n-1$, and our proof is complete. □

Definition 2.2.9. A set of $t \geq 2$ MOLS of order n is called a *complete set* if $t = n - 1$.

We now show that if q is a prime power, then we can use the finite field F_q to easily construct a complete set of $q-1$ MOLS of order q. The famous Indian mathematical statistician Raj Chandra Bose (1901–1987) [3] is usually credited with the following important result, although he was preceded by E. H. Moore [46].

Theorem 2.2.10 (Bose 1938). *If q is a prime power, then $N(q) = q-1$.*

Proof. Let $F_q^* = \{a_1, \ldots, a_{q-1}\}$. Label the rows and columns of a $q \times q$ matrix with the elements of F_q, listed in any order. For each $1 \leq i \leq q-1$ we construct a latin square L_i as follows. Let $f_i(x,y)$ be the linear polynomial $f_i(x,y) = a_i x + y$. In the location (x,y) in the square L_i place the field element $f_i(x,y)$. It is straightforward to verify that each polynomial $f_i(x,y)$ generates a latin square of order q, and that any two latin squares so generated are distinct.

We now show that the latin squares are mutually orthogonal, that is, that any two distinct squares in the set are orthogonal. Let $1 \leq i < j \leq q-1$, and let (b_1, b_2) be any pair of elements of F_q. Showing that this pair occurs when the squares L_i and L_j are superimposed is equivalent to showing that the system of equations

$$a_i x + y = b_1,$$
$$a_j x + y = b_2$$

has a solution (x,y) over F_q. This follows from basic linear algebra, because if $a_i \neq a_j$, then the coefficient matrix

$$\begin{pmatrix} a_i & 1 \\ a_j & 1 \end{pmatrix}$$

is invertible, and hence the pair of equations has a unique solution. □

Example 2.2.11. For $q = 3$, using the polynomials $x + y$ and $2x + y$ over F_3, we obtain the pair of MOLS of order 3 given near the beginning of this section.

We now present a larger example using the finite field F_4, which is generated over F_2 by an element α satisfying $\alpha^2 + \alpha + 1 = 0$. From the three polynomials $x + y$, $\alpha x + y$, and $\alpha^2 x + y$, we obtain the following complete set K_1, K_2, K_3 of three MOLS of order four:

$$\begin{array}{cccc} 0 & 1 & \alpha & \alpha^2 \\ 1 & 0 & \alpha^2 & \alpha \\ \alpha & \alpha^2 & 0 & 1 \\ \alpha^2 & \alpha & 1 & 0 \end{array} \qquad \begin{array}{cccc} 0 & 1 & \alpha & \alpha^2 \\ \alpha & \alpha^2 & 0 & 1 \\ \alpha^2 & \alpha & 1 & 0 \\ 1 & 0 & \alpha^2 & \alpha \end{array} \qquad \begin{array}{cccc} 0 & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & 0 \\ 1 & 0 & \alpha^2 & \alpha \\ \alpha & \alpha^2 & 0 & 1 \end{array}$$

K_1 K_2 K_3

We have shown that if n is a prime power, then $N(n) = n-1$. The problem of determining $N(n)$ for other n is much more difficult and for most n remains an open question. An important conjecture is that only prime power values of n have the property that $N(n) = n - 1$.

Conjecture 2.2.12 (The Prime Power Conjecture). For $n \geq 2$, $N(n) = n - 1$ if and only if n is a prime power.

After Fermat's Last Theorem was proved in 1994, the first author proposed [48] the Prime Power Conjecture as a candidate for the "Next Fermat Problem." We now provide some motivation and evidence for attaching such a lofty title to the Prime Power Conjecture. We begin with a famous conjecture of Leonard Euler.

Conjecture 2.2.13 (Euler 1782). If n is an odd multiple of 2, that is, if $n = 2(2k + 1)$ with $k \geq 0$, then $N(n) = 1$.

Euler's conjecture is true for $k = 0$ and $k = 1$ (that is, for $n = 2$ and $n = 6$), but is now known to be false for all other values of k; see Theorem 2.2.19. Euler was unable to prove even the case when $k = 1$, and may have been led to his conjecture through unsuccessful attempts to construct 2 MOLS of orders $6, 10, 14, \ldots$. We now provide a brief history related to Euler's conjecture. A partial positive result was obtained over one hundred years later; see Dénes and Keedwell [11, p. 140].

Theorem 2.2.14 (Tarry 1899–1900). $N(6) = 1$.

We remind the reader that in 1900, without the aid of a computer, Tarry's result was no small feat; recall that the number of latin squares of order 6 is given by $L_6 = 6! \cdot 5! \cdot 9,408 \approx 8 \cdot 10^8$. An elegant modern proof (using properties from coding and design theory) that $N(6) = 1$ is due to Stinson [61].

For latin squares of order 6, one can get very close to having a pair of MOLS. Exercise 2.3 asks you to construct two latin squares of order 6 which, when superimposed, yield a total of 34 distinct ordered pairs.

Thinking that Euler's conjecture was likely true (after all, it was true in the first two cases, and Euler had such a fabulous ability to compute) H. F. MacNeish [40] was led to generalize Euler's conjecture in the following way.

Conjecture 2.2.15 (MacNeish 1922). Suppose $n = q_1 \cdots q_r$, where the numbers q_i are powers of distinct primes and $q_1 < \cdots < q_r$. Then $N(n) = q_1 - 1$.

MacNeish's conjecture is of course true at $n = 6$ and all prime powers. It however suffered a negative fate in 1959 when Parker [52] constructed a counterexample.

Counterexample 2.2.16 (Parker 1959). $N(21) \geq 4$.

Parker used finite fields to build the necessary orthogonal latin squares. In particular, he showed that if $m > 3$ is a Mersenne prime (a prime of the form $2^n - 1$) or if $m + 1$ is a Fermat prime (a prime of the form $2^{2^n} + 1$) greater than 3, then there is a set of m MOLS of order $m^2 + m + 1$. Using the Fermat prime $m + 1 = 2^2 + 1$, Parker was able to construct 4 MOLS of order 21, thus showing that MacNeish's conjecture is false. We refer to Dénes and Keedwell [11, pp. 394–396] for further details.

The same year, Euler's conjecture suffered the same fate.

Counterexample 2.2.17 (Bose and Shrikhande 1959). $N(22) \geq 2$.

Bose and Shrikhande also used finite fields in their disproof of the Euler conjecture. In particular, they showed that if $q \equiv 3 \pmod 4$ is a prime power, then there exists a pair of MOLS of order $(3q - 1)/2$. Taking $q = 7$ yields a pair of MOLS of order 10; see Dénes and Keedwell [11, pp. 397–400] for details of this order 10 construction.

These two results of course showed that both the Euler and MacNeish conjectures were false. It turned out that Euler's conjecture was incorrect at other values as well, including $n = 10$; see Parker [53].

Counterexample 2.2.18 (Parker 1960). $N(10) \geq 2$. The following table shows two superimposed MOLS of order 10.

1,2	2,3	3,1	4,6	5,9	6,4	7,8	8,7	9,5	0,0
7,4	4,2	2,7	0,9	6,1	5,8	8,5	9,0	3,3	1,6
5,1	1,4	4,5	6,7	0,8	8,0	9,3	2,2	7,6	3,9
0,7	7,1	1,0	3,8	8,3	9,2	4,4	5,6	2,9	6,5
3,5	5,7	7,3	8,2	9,4	1,1	0,6	4,9	6,0	2,8
2,0	0,5	5,2	9,1	7,7	3,6	1,9	6,3	4,8	8,4
4,3	3,0	0,4	5,5	2,6	7,9	6,2	1,8	8,1	9,7
8,9	9,8	6,6	2,4	3,2	0,3	5,0	7,5	1,7	4,1
6,8	8,6	9,9	7,0	1,5	4,7	2,1	3,4	0,2	5,3
9,6	6,9	8,8	1,3	4,0	2,5	3,7	0,1	5,4	7,2

The following result of Bose, Shrikhande, and Parker [4] is the most fundamental result in the theory of MOLS since the 1938 proof by Bose that $N(n) = n - 1$ if n is a prime power. The authors not only proved that a pair of MOLS exists for every $n \neq 2, 6$, they, in passing, show that Euler's conjecture is false for every value of $k \geq 2$. To be fair to Euler, he of course did not have at his disposal any computer or he might have made a conjecture along the lines of the Prime Power Conjecture. In some sense, the Prime Power Conjecture is more likely to be true since one is trying to show the nonexistence of a highly complex structure (a set of $n-1$ MOLS of order n) rather than the nonexistence of a significantly simpler structure (a pair of MOLS of order n).

Theorem 2.2.19 (Bose, Shrikhande, Parker 1960). $N(n) \geq 2$ *for all* n *except* 2 *and* 6.

After $n = 6$, the next nonprime power case occurs when $n = 10$. The value of $N(10)$ remains unknown to this day; in fact, we are very far from knowing the exact value of $N(10)$. There was a lot of interest in whether $N(10) = 9$, and a lot of mathematical effort was put into this problem using various ideas from coding theory (see Chapter 3) and projective planes (see Section 3 of this chapter). Progress on the value of $N(10)$ remained very elusive until a result by Lam, Thiel, and Swiercz [29] in 1989.

Theorem 2.2.20 (Lam, Thiel, Swiercz 1989). $N(10) < 9$.

This result required sophisticated mathematical tools from algebraic coding theory together with over 2000 hours of computation on a Cray supercomputer! Nevertheless, it is still not known today whether there exist three MOLS of order 10. It is known that if $n > 4$ and $N(n) < n - 1$, then $N(n) \leq n - 4$ (see Dénes and Keedwell [11, p. 385]). It follows that $2 \leq N(10) \leq 6$.

To prove Theorem 2.2.24 below, MacNeish invented a method of combining latin squares of sizes n_1 and n_2 into a larger latin square of order $n_1 n_2$. We now describe this construction, which is closely related to the Kronecker product of matrices, in detail. Let $H = (h_{ij})$ be a latin square of order n_1 and let $K = (k_{rs})$ be a latin square of order n_2. We will form an $n_1 n_2 \times n_1 n_2$ matrix, denoted $H \otimes K$. We replace each element h_{ij} of H with the $n_2 \times n_2$ matrix whose entries $\{a_{rs}\}$ are ordered pairs: $a_{rs} = (h_{ij}, k_{rs})$.

As an illustration of this construction with $n_1 = 2$ and $n_2 = 3$, let H and K be the following squares.

$$\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \qquad \begin{array}{ccc} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 2 & 0 & 1 \end{array}$$

H $\qquad\qquad$ K

The Kronecker product construction yields the following 6×6 square $H \otimes K$ whose elements are ordered pairs (for simplicity, we have omitted the parentheses and commas).

$$\begin{array}{cccccc} 00 & 01 & 02 & 10 & 11 & 12 \\ 01 & 02 & 00 & 11 & 12 & 10 \\ 02 & 00 & 01 & 12 & 10 & 11 \\ 10 & 11 & 12 & 00 & 01 & 02 \\ 11 & 12 & 10 & 01 & 02 & 00 \\ 12 & 10 & 11 & 02 & 00 & 01 \end{array}$$

$H \otimes K$

Of course one can easily replace the ordered pairs $00, 01, 02, 10, 11, 12$ by the integers $0, 1, 2, 3, 4, 5$ to obtain a latin square of order 6 whose elements are the usual symbols.

Example 2.2.21. Figure 2.1 below illustrates the Kronecker product construction for a pair $H_1 \otimes K_1, H_2 \otimes K_2$ of MOLS of order 12 constructed from the pair of MOLS of order 3 given at the beginning of this section, and the first two MOLS K_1 and K_2 of order 4 given above. We replace the symbols 0, 1, α, α^2 in the MOLS of order 4 by 0, 1, 2, 3, respectively.

The next two lemmas show that the construction we have just demonstrated can be used to construct sets of MOLS. The proofs are left as exercises for the reader.

Lemma 2.2.22. *If H and K are latin squares of orders n_1 and n_2, then $H \otimes K$ is a latin square of order $n_1 n_2$.*

Proof. Exercise 2.25. □

Lemma 2.2.23. *If H_1 and H_2 are orthogonal latin squares of order n_1 and K_1 and K_2 are orthogonal latin squares of order n_2, then $H_1 \otimes K_1$ and $H_2 \otimes K_2$ are orthogonal latin squares of order $n_1 n_2$.*

Proof. Exercise 2.26. □

Except for positive integers of the form $n = 2(2k+1)$, we can use this construction to yield a pair of MOLS of order n. (The problem with $n = 2(2k + 1)$ is that $N(2) = 1$.) It thus follows that for 3/4 of the positive integers n, we have $N(n) \geq 2$.

Theorem 2.2.24 (MacNeish 1922). *Let $n = q_1 \cdots q_r$, where q_i are distinct prime powers and $q_1 < \cdots < q_r$. Then $N(n) \geq q_1 - 1$.*

Proof. If q is a prime power greater than 2, then $N(q) \geq 2$. By using the construction outlined above, we can construct a set of MOLS of order n from collections of MOLS of orders q_i for $1 \leq i \leq r$. □

If n is a prime power, then the lower bound in MacNeish's theorem is exact. It was known to MacNeish that $N(6) = 2 - 1 = 1$, so the bound is optimal in that case as well.

As indicated earlier, $N(21) \geq 2$. A more recent conjecture is that Conjecture 2.2.15 is almost completely wrong. This rather daring conjecture was raised by Laywine, Mullen, and Whittle [32]. It says

0	1	2
1	2	0
2	0	1

H_1

0	1	2
2	0	1
1	2	0

H_2

0	1	2	3
1	0	3	2
2	3	0	1
3	2	1	0

K_1

0	1	2	3
2	3	0	1
3	2	1	0
1	0	3	2

K_2

00	01	02	03	10	11	12	13	20	21	22	23
01	00	03	02	11	10	13	12	21	20	23	22
02	03	00	01	12	13	10	11	22	23	20	21
03	02	01	00	13	12	11	10	23	22	21	20
10	11	12	13	20	21	22	23	00	01	02	03
11	10	13	12	21	20	23	22	01	00	03	02
12	13	10	11	22	23	20	21	02	03	00	01
13	12	11	10	23	22	21	20	03	02	01	00
20	21	22	23	00	01	02	03	10	11	12	13
21	20	23	22	01	00	03	02	11	10	13	12
22	23	20	21	02	03	00	01	12	13	10	11
23	22	21	20	03	02	01	00	13	12	11	10

$H_1 \otimes K_1$

00	01	02	03	10	11	12	13	20	21	22	23
01	00	03	02	11	10	13	12	21	20	23	22
02	03	00	01	12	13	10	11	22	23	20	21
03	02	01	00	13	12	11	10	23	22	21	20
20	21	22	23	00	01	02	03	10	11	12	13
21	20	23	22	01	00	03	02	11	10	13	12
22	23	20	21	02	03	00	01	12	13	10	11
23	22	21	20	03	02	01	00	13	12	11	10
10	11	12	13	20	21	22	23	00	01	02	03
11	10	13	12	21	20	23	22	01	00	03	02
12	13	10	11	22	23	20	21	02	03	00	01
13	12	11	10	23	22	21	20	03	02	01	00

$H_2 \otimes K_2$

Figure 2.1. Kronecker products of two pairs of MOLS of orders 3 and 4

	0	1	2	3	4	5	6	7	8	9
00	-	-	1	2	3	4	1	6	7	8
10	2	10	5	12	3	4	15	16	3	18
20	4	5	3	22	7	24	4	26	5	28
30	4	30	31	5	4	5	8	36	4	5
40	7	40	5	42	5	6	4	46	8	48
50	6	5	5	52	5	6	7	7	5	58
60	4	60	5	6	63	7	5	66	5	6
70	6	70	7	72	5	7	6	6	6	78
80	9	80	8	82	6	6	6	6	7	88
90	6	7	6	6	6	6	7	96	6	8

Table 2.1. Known lower bounds on $N(n)$

that, like the Euler conjecture, the MacNeish conjecture is always wrong except at $n = 6$ and prime powers.

Conjecture 2.2.25 (Laywine, Mullen, Whittle 1995)**.** Except for prime powers n and $n = 6$, the MacNeish conjecture is always false. That is, if $n = q_1 \cdots q_r$, where $q_1 < \cdots < q_r$ are prime powers of distinct primes, then $N(n) > q_1 - 1$.

The first unresolved case of MacNeish's conjecture occurs for the value $n = 63$.

Theorem 2.2.26 (Bruck and Ryser 1949)**.** *For infinitely many n, $N(n) < n - 1$. In particular, if n is congruent to 1 or 2 modulo 4 and the squarefree part of n contains a prime of the form $4k+3$, then $N(n) < n - 1$.*

The proof uses Lagrange's four squares theorem, which states that every positive integer n can be written in the form $n = a_1^2 + a_2^2 + a_3^2 + a_4^2$ for integers a_i. It is interesting to see how Lagrange's number theoretic result is so useful in the result of Bruck and Ryser [5] concerning sets of MOLS.

Table 2.1 gives lower bounds for $N(n)$ for $n \leq 100$. To read the table, locate the desired tens digit on the left hand side and the desired ones digit on the top.

For much more latin square information, we refer the reader to the monographs by Colbourn and Dinitz [8], Dénes and Keedwell [11,12], and Laywine and Mullen [31]. In particular, we refer to Colbourn and Dinitz [8] for a table of lower bounds for $N(n)$ for $n \leq 10,000$.

2.2. Sudoku squares. There has been a tremendous amount of recent worldwide activity related to Sudoku puzzles, which are, in our terminology, partial latin squares of order 9. The goal of the puzzle is to complete the partial latin square of order 9 to a full latin square of order 9 with the additional property that each of the nine 3×3 subsquares contains each of the nine numbers exactly once. Such squares are called *Sudoku squares*, which translates to "the number that is alone." Sudoku puzzles are published in many newspapers and the squares are being sold commercially. The following partial latin square is an example of a Sudoku puzzle:

	6		1		4		5	
		8	3		5	6		
2								1
8			4		7			6
		6				3		
7			9		1			4
5								2
		7	2		6	9		
	4		5		8		7	

We say that a latin square of order q^2 is a *Sudoku square* if each of the $q \times q$ subsquares contains each number exactly once. An ordinary Sudoku puzzle is thus a partial latin square of order 9 with the property that it has a unique extension to a full Sudoku square of order 9.

We now provide a simple construction for Sudoku squares of order q^2, where q is any prime power. In addition, our construction yields a Sudoku square with the additional property that each of the rows

and each of the columns, of each of the q^2, $q \times q$ subsquares, has the same constant sum. We might call such a square a *row/column magic Sudoku square.*

As indicated in Exercise 2.11, if one has a pair L_1 and L_2 of orthogonal diagonal latin squares of order n, then the square $M = nL_1 + L_2$ is a magic square of order n, where the square M is computed using arithmetic modulo n^2. Here a *magic square of order* n is an $n \times n$ array based on the numbers $0, 1, \ldots, n-1$ with the property that each row, each column, and both of the main diagonals, have a constant sum, called the *magic sum.*

For q a prime power, consider a pair of polynomials $a_1x + y$ and $a_2x + y$ where $a_1 \neq a_2 \in F_q$ and neither a_1 nor a_2 is zero. These polynomials yield a pair of orthogonal latin squares of order q, and in fact they will both be diagonal if neither a_1 or a_2 is 1 or -1. Now form a row/column magic square of order q as illustrated above.

Given a row/column magic square M of order n, one can construct a Sudoku magic square of order n^2 by the following construction. First note that any rearrangement of the rows or columns of a row/column magic square yields another row/column magic square. Place the magic square M in the top left corner of the big square and then cyclically shift the rows down by one, and place the new row/column magic square next to the right, then shift again and place a third row/column magic square, continuing until one has placed $n - 1$ shifted row/column magic squares across the top of the big square. Similarly shift the columns of M to fill in the left side of the square, and then shift each of those squares across to the right as above.

As an illustration consider the case where $q = 3$ and the resulting orthogonal latin squares of order 3 are as at the beginning of Section 2.1. The resulting nonmagic square of order 3 is shown below.

$$\begin{array}{ccc} 0 & 4 & 8 \\ 5 & 6 & 1 \\ 7 & 2 & 3 \end{array}$$

Using the above construction, we obtain the following Sudoku latin square of order 9 with the property that each of the nine $3 \times$

3 subsquares is a row/column magic square of order 3 with magic sum 12.

0	4	8	7	2	3	5	6	1
5	6	1	0	4	8	7	2	3
7	2	3	5	6	1	0	4	8
8	0	4	3	7	2	1	5	6
1	5	6	8	0	4	3	7	2
3	7	2	1	5	6	8	0	4
4	8	0	2	3	7	6	1	5
6	1	5	4	8	0	2	3	7
2	3	7	6	1	5	4	8	0

A Sudoku puzzle can be obtained by removing a small number of elements from such a square. But how many elements can be removed? Little research seems to have been conducted on this interesting question. It is known that there are Sudoku puzzles with 77 cells filled but which do not have a unique extension to a latin square of order 9, and a Sudoku puzzle with 17 cells filled in that does have a unique extension, but unknown whether there is a Sudoku puzzle with only 16 cells filled that has exactly one extension to a Sudoku square. The reader may enjoy trying to construct examples of such squares.

2.3. Generalizations of latin squares. We close this section with a brief discussion of several possible generalizations of latin squares. First, one could retain two dimensions and allow repetitions of elements in each row and column. Such squares are known as frequency squares, and will shortly be defined more formally. Proceeding in a different direction, one could keep the number of symbols fixed at n and extend the dimension from two (as in a square) to an arbitrary dimension $d \geq 2$. This leads to the study of sets of orthogonal (latin) hypercubes of order n and dimension d. Finally, to generalize further,

one could study d-dimensional frequency hypercubes, or even hyper-rectangles, based on m symbols, as in Suchower [62]. We will not go into those details here, however.

Definition 2.2.27. Let $n = \lambda m$. An $F(n;\lambda)$ *frequency square* is an $n \times n$ matrix consisting of m distinct symbols such that each symbol appears exactly λ times in each row and column.

Example 2.2.28. An $F(4;2)$ frequency square:

$$\begin{array}{cccc} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}$$

Definition 2.2.29. Two $F(n;\lambda)$ frequency squares are *orthogonal* if when they are superimposed each of the m^2 pairs appears exactly λ^2 times, and a set of $F(n;\lambda)$ frequency squares is *orthogonal* if every pair of distinct squares in the set is orthogonal.

We note in passing that an $F(n;1)$ frequency square is simply a latin square of order n, and a set of $F(n;1)$ mutually orthogonal frequency squares (MOFS) is a set of MOLS of order n.

The following result provides an upper bound on the maximum possible number of MOFS; a proof is given by Laywine and Mullen [31, Theorem 4.1]. We note that when $n = m$ (so the frequency squares are latin squares of order n), this upper bound reduces to $n-1$, the maximum possible number of MOLS of order n.

Theorem 2.2.30. *Let $n = \lambda m$. The size of any orthogonal set of $F(n;\lambda)$ frequency squares is less than or equal to $(n-1)^2/(m-1)$.*

In Exercise 2.14, we ask the reader to construct a complete set of $F(q^i;q^{i-1})$ MOFS for a prime power q and positive integer i; see Mullen [47] for such a construction using finite fields.

Definition 2.2.31. A *d-dimensional hypercube of order n* is an $n \times \cdots \times n$ array containing n^d cells with the property that when any of the d coordinates is fixed, each of the n symbols occurs exactly n^{d-2} times in the resulting subarray.

We note that when $d = 2$, d-dimensional hypercubes are simply latin squares, and sets of such orthogonal hypercubes are sets of MOLS. Laywine, Mullen, and Whittle [32] show that the maximum number of such mutually orthogonal hypercubes of dimension $d \geq 2$ and order n is bounded above by $(n^d - 1)/(n - 1) - d$; when $d = 2$ this specializes to $(n^2 - 1)/(n - 1) - 2$ which yields the bound $n - 1$ for the cardinality of a maximal (complete) set of MOLS of order n.

In Exercise 2.15, we ask the reader to construct a complete set of d-dimensional hypercubes of order q, where $d \geq 2$ and q is a prime power. We also mention that a complete set of d-dimensional hypercubes can be used in the next section to construct the affine geometry $\mathrm{AG}(d, q)$ of dimension d over the field F_q.

We close this section by mentioning that Laywine *et al.* [32] give a recursive method for constructing sets of orthogonal cubes or, more generally, sets of orthogonal hypercubes, from sets of MOLS.

3. Affine and projective planes

In this section, we briefly develop the theory of some finite analogues of Euclidean geometry. Not only are these objects interesting in their own right, they also provide very natural ways to construct the block designs that will be studied in the next section.

As geometric objects, projective and affine planes have been studied for many years. Higher dimensional projective and affine geometries over finite fields have also been studied widely. As of today, no such finite objects are known to exist except in the case of prime powers q when of course we also have a finite field F_q containing q elements.

We will define these objects shortly, but throughout this section, and as is standard, we will denote a projective plane and affine plane defined over the field F_q, respectively, by $\mathrm{PG}(2, F_q)$ and $\mathrm{AG}(2, F_q)$. More generally, $\mathrm{PG}(m, F_q)$ will denote a projective space over F_q of dimension $m \geq 2$, and $\mathrm{AG}(m, F_q)$ will denote an affine space over F_q, also of dimension $m \geq 2$.

Definition 2.3.1. A *projective plane* consists of a set of points, a set of lines, and an incidence relation that determines which points lie

on which lines. The incidence relation needs to satisfy the following three axioms:

(1) Every two lines intersect in exactly one point.

(2) Every two points have a unique line incident with them.

(3) There are at least four points of which no three are on the same line.

Note that the definition of a projective plane has a duality between lines and points; if the words lines and points are exchanged in the definition, then the first two axioms are preserved. Moreover, the dual version of the third axiom is a consequence of the original three axioms. Thus any proof about lines can be turned into a proof about points, and *vice versa.*

Theorem 2.3.2. *Let Π be a finite projective plane. There is an integer $m \geq 2$ such that every point (line) is incident with $m+1$ lines (points, respectively). Moreover, Π has exactly $m^2 + m + 1$ points (lines, respectively).*

We omit a proof of the above result, which can be located in many combinatorics books. The important point for us is that the positive integer m in the theorem is called the *order* of the projective plane. As alluded to above, there is no known case where the order m of a projective plane is not a prime power, and as a result of Conjecture 2.2.12 and Theorem 2.3.8 below, it is conjectured that nonprime power order projective planes do not exist.

Definition 2.3.3. An *affine plane* consists of a set $\mathcal{P}$ of points, a set $\mathcal{L}$ of lines, and an incidence relation that determines which points lie on which lines. The incidence relation must satisfy the following three axioms:

(1) Every pair of points determines a unique line.

(2) For every point p not on a line l there is a unique line m incident with p and not intersecting l.

(3) There are four points, no three of which are on the same line.

The next result illustrates how to construct an affine plane over any field K. We leave the proof to the reader as Exercise 2.27.

Theorem 2.3.4. *Let K be a field. Let $\mathcal{P} = \{(x, y) \mid x, y \in K\}$. Let $\mathcal{L}$ be the set of graphs of linear functions over K: each $L \in \mathcal{L}$ is of the form $L = \{(x, y) \mid ax + by + c = 0\}$ for some $a, b, c \in K$. Then the pair $(\mathcal{P}, \mathcal{L})$, with its natural incidence structure, is an affine plane, denoted by* $\mathrm{AG}(2, K)$.

We now illustrate a method to extend an affine plane to a projective plane. To do this, we will add a "line at infinity," L_∞. Begin by renaming each point (x, y) in $\mathrm{AG}(2, K)$ as $(x, y, 1)$. Hence, for $a \neq 0$ and $b \neq 0$, view $ax + by + cz = 0$ as the equation of a line. Put $L_\infty = \{(1, 0, 0)\} \cup \{(x, 1, 0) \mid x \in K\}$ so that we can view this line as having $z = 0$. Let $\mathcal{P}'$ be the union of $\mathcal{P}$ and the points on L_∞. Let $\mathcal{L}' = \mathcal{L} \cup \{L_\infty\}$.

Theorem 2.3.5. *If $(\mathcal{P}, \mathcal{L})$ is an affine plane over a field K, then $(\mathcal{P}', \mathcal{L}')$ is a projective plane over K.*

Corollary 2.3.6. *If q is a prime power, then* $\mathrm{PG}(2, F_q)$ *and* $\mathrm{AG}(2, F_q)$ *both exist.*

Example 2.3.7. We will now explicitly construct $\mathrm{AG}(2, F_2)$. We have the set of equations $\{ax + by = c \mid a, b, c \in F_2, (a, b) \neq (0, 0)\}$ along with the following incidence relation:

Line	Equation	Points	
L_1	$x + y = 1$	101	011
L_2	$x = 1$	101	111
L_3	$y = 1$	011	111
L_4	$x + y = 0$	001	111
L_5	$x = 0$	011	001
L_6	$y = 0$	101	001

Here, as usual, we omit the parentheses and commas in ordered tuples. For example, 101 denotes $(1, 0, 1)$. It is easy to check that this is a model of the affine plane $\mathrm{AG}(2, F_2)$ axioms (see Exercise 2.18).

We now add the points 100, 010, and 110 to form a projective plane $\mathrm{PG}(2, F_2)$, often called the *Fano plane*, illustrated in Figure 2.3.

Line	Equation	Points		
L_1	$x + y + z = 0$	101	011	110
L_2	$x + z = 0$	101	111	010
L_3	$y + z = 0$	011	111	100
L_4	$x + y = 0$	001	111	110
L_5	$x = 0$	011	001	010
L_6	$y = 0$	101	001	100
L_7	$z = 0$	100	010	110

The following result is one of the most fundamental results in combinatorics, providing an equivalence between affine planes and complete sets of MOLS.

Theorem 2.3.8 (Bose 1938). *There are $n-1$ mutually orthogonal latin squares of order n if and only if there is an affine plane* $\mathrm{AG}(2,n)$ *(or, equivalently, there is a projective plane* $\mathrm{PG}(2,n)$*).*

Proof. We first assume that we have a complete set $M_1, \ldots, M_{n-1}$ of $n-1$ MOLS of order n, and construct the affine plane $\mathrm{AG}(2,n)$. Let P denote the set of all pairs (x,y) with $0 \leq x, y \leq n-1$. Thus the points of $\mathrm{AG}(2,n)$ will consist of ordered pairs.

For each d with $1 \leq d \leq n-1$, and for each c with $0 \leq c \leq n-1$, form the line

$$L_c^{(d)} = \{(x,y) | M_d(x,y) = c\},$$

where $M_d(x,y) = c$ means that symbol c occurs in position (x,y) of the dth latin square M_d in the set of MOLS. This gives n lines for each of the $n-1$ latin squares, and hence it gives a total of $(n-1)n = n^2 - n$ lines.

We now form $2n$ more lines from the canonical row and column squares (which are not latin squares). The row square R consisting of rows with constant values $x = c$, gives n lines: for $0 \leq c \leq n-1$ the line R_c consists of all pairs with first coordinate c. Similarly consider the column square C defined by $y = c$; this generates an additional n lines C_c, each consisting of those points with second coordinate c, where $0 \leq c \leq n-1$.

We have now formed a total of $n^2 - n + 2n = n^2 + n$ lines. We leave it to the reader to verify that the set

$$\{L_c^{(d)}, R_0, \ldots, R_{n-1}, C_0, \ldots, C_{n-1} \mid 1 \leq d \leq n-1, 0 \leq c \leq n-1\}$$

forms an affine plane $\mathrm{AG}(2, n)$.

Conversely, given the affine plane $\mathrm{AG}(2, n)$, we now construct $n-1$ MOLS of order n. We first label the $n+1$ parallel classes of the affine plane as $0, 1, \ldots, n$, and then label the n lines in each class as $0, 1, \ldots, n-1$. The next step is to use two parallel classes to set up a correspondence between the elements in the affine plane, and the numbers $0, 1, \ldots, n-1$ which will be used to build the set of $n-1$ MOLS of order n. We can obtain such a correspondence as follows: assign the ordered pair (i, j) to the unique point of intersection of line i of class 0 and line j of class n.

In order to construct the $n-1$ squares of order n, we place symbol s in cell (i, j) of the square L_e if line s of class e contains the ordered pair (i, j). The latin property of the squares follows since any line in one of the classes $1, 2, \ldots, n-1$ intersects any line from classes 0 or n in exactly one point. Similarly, any line from class e with $e = 1, 2, \ldots, n-1$ intersects any line from class f, $f \neq e, 0, n$ in exactly one point. Hence the latin squares L_e and L_f are indeed orthogonal. This completes the proof. □

We now illustrate the above connection between sets of MOLS and affine planes in the case when $q = 2$. Consider the squares

$$M_1 = \begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}, \quad R = \begin{matrix} 0 & 0 \\ 1 & 1 \end{matrix}, \quad C = \begin{matrix} 0 & 1 \\ 0 & 1 \end{matrix},$$

so that M_1 is a latin square of order 2, and R and C are the canonical row and column squares of order 2. Form the lines

$$L_0^1 = 00\ 11 \qquad R_0 = 00\ 01 \qquad C_0 = 00\ 10$$
$$L_1^1 = 01\ 10 \qquad R_1 = 10\ 11 \qquad C_1 = 01\ 00$$

It is easily checked that $\{L_0^1, L_1^1, R_0, R_1, C_0, C_1\}$ gives an affine plane $\mathrm{AG}(2, 2)$.

Conversely, consider the following affine plane $\mathrm{AG}(2, 3)$.

Class	Line	Points	Class	Line	Points
0	0	0 5 7	2	0	0 1 2
0	1	1 3 8	2	1	3 4 5
0	2	2 4 6	2	2	6 7 8
1	0	0 4 8	3	0	0 3 6
1	1	1 5 6	3	1	1 4 7
1	2	2 3 7	3	2	2 5 8

The correspondence between the numbers $0, 1, \ldots, 8$ in the affine plane and the nine ordered pairs (where we have omitted the parentheses) is given by

$$\begin{array}{lll} 0 \mapsto 00 & 3 \mapsto 10 & 6 \mapsto 20 \\ 7 \mapsto 01 & 1 \mapsto 11 & 4 \mapsto 21 \\ 5 \mapsto 02 & 8 \mapsto 12 & 2 \mapsto 22 \end{array}$$

Finally, using this correspondence, we obtain the following pair of MOLS of order 3:

$$\begin{array}{ccc} 0 & 1 & 2 \\ 2 & 1 & 0 \\ 1 & 0 & 2 \end{array} \qquad \begin{array}{ccc} 0 & 2 & 1 \\ 1 & 0 & 2 \\ 2 & 1 & 0 \end{array}$$

$$L_1 \qquad\qquad L_2$$

From our earlier Prime Power Conjecture for MOLS (Conjecture 2.2.12), we obtain the following conjecture.

Conjecture 2.3.9 (The prime power conjecture for affine and projective planes)**.** There are $n-1$ mutually orthogonal latin squares of order n if and only if n is a prime power. Equivalently, there is an affine plane of order n, or equivalently a projective plane of order n, if and only if n is a prime power.

We digress from our combinatorial constructions for a moment to discuss the notion of "different," or nonisomorphic, planes. It is known that there is only one plane of each of the orders $n = 2, 3, 4, 5, 7, 8$. By this we mean that any two planes of the same order $n \leq 8$ are isomorphic. It is known that there are exactly four nonisomorphic projective planes of order $n = 9$; see Dénes and Keedwell [11, Section 8.4] and Lam, Kolesova, and Thiel [28].

A plane is called *Desarguesian* if the theorem of Desargues concerning perspective triangles is universally valid. All projective planes constructed in the manner of Example 2.3.7, that is, with linear polynomials over finite fields, are Desarguesian. A famous conjecture is that all planes of prime order are Desarguesian (see Dénes and Keedwell [11, p. 276]) but this has only been validated for the primes 2, 3, 5, and 7.

Definition 2.3.10. Let m be a natural number, $m \geq 2$. A *triangle* is a set of three lines (the sides) such that any two of the three lines intersect.

A point is a 0-*space*, also called a 0-*flat*. A line is a 1-*space*, also called a 1-*flat*. A k-*flat*, $k > 1$, will be determined by $k+1$ points which do not lie in any $(k-1)$-flat. An m-*space* is a set of points and lines such that:

(1) There is a unique line through any two distinct points.

(2) A line which intersects two sides of a triangle must intersect the third side.

(3) Every line contains at least three points.

(4) There is no $(m+1)$-flat.

Normally m-spaces are constructed over fields, and since our focus is on finite fields, we will now illustrate the construction of an m-space over the finite field F_q. As is customary, this space will be called a projective geometry $\mathrm{PG}(m, F_q)$ of dimension $m \geq 2$ over the field F_q.

A point will be an $(m+1)$-tuple of elements of F_q not all of which are zero. We identify two points $(a_0, \ldots, a_m)$ and $(b_0, \ldots, b_m)$ as being equivalent if there is a $c \in F_q^*$ such that $(b_0, \ldots, b_m) = (ca_0, \ldots, ca_m)$. It is easy to check that there are $(q^{m+1}-1)/(q-1)$ equivalence classes of points.

A k-*flat* in $\mathrm{PG}(m, F_q)$ is the set of all solutions to a system of $m-k$ independent linear homogeneous equations, that is, a system of equations of the form

$$\begin{pmatrix} a_{10} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m-k,0} & \cdots & a_{m-k,m} \end{pmatrix} \begin{pmatrix} x_0 \\ \vdots \\ x_m \end{pmatrix} = 0\,,$$

where the matrix (a_{ij}) is nonsingular and so has rank $m-k$. Each k-flat has $(q^{k+1}-1)/(q-1)$ points on it (recall that points are equivalence classes of solutions).

Definition 2.3.11. A *hyperplane* in $\mathrm{PG}(m, F_q)$ is an $(m-1)$-flat.

From a projective geometry of dimension $m \geq 2$, we now construct an affine geometry $\mathrm{AG}(m, F_q)$ of dimension $m \geq 2$ by deleting a hyperplane. We begin with $\mathrm{PG}(m, F_q)$, and remove any hyperplane and the points that lie on it. The resulting space has

$$\frac{q^{m+1}-1}{q-1} - \frac{q^m-1}{q-1} = q^m$$

points. Suppose we remove the hyperplane corresponding to the equation $x_m = 0$. The remaining points can be rescaled to have $x_m = 1$. Therefore the remaining k-flats can be viewed as solution sets to systems of linear equations of the form

$$\begin{pmatrix} a_{10} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m-k,0} & \cdots & a_{m-k,m} \end{pmatrix} \begin{pmatrix} x_0 \\ \vdots \\ x_{m-1} \\ 1 \end{pmatrix} = 0\,,$$

where (a_{ij}) is nonsingular and so has rank $m-k$. This yields the affine geometry $\mathrm{AG}(m, F_q)$ of dimension $m \geq 2$ over the field F_q. In Exercise 2.20, the reader is asked to construct the affine geometry $\mathrm{AG}(3, F_2)$.

4. Block designs

Informally, a combinatorial or block design is a collection of subsets from a finite set. Invariably certain specified conditions are imposed on the selection of the subsets.

Definition 2.4.1. A *tactical configuration* is a set of v symbols arranged in b sets, called blocks, such that each block is of size k and each symbol occurs in exactly r blocks. If $v = b$, the configuration is said to be *symmetric*. Note that $vr = bk$ in any tactical configuration.

Definition 2.4.2. A (v, k, λ) *balanced incomplete block* (*BIB*) *design* is a tactical configuration with v symbols and blocks of size k such that each pair of distinct symbols occurs in exactly λ blocks.

Hereafter we will omit the term BIB and simply say design.

Proposition 2.4.3. *In a* (v, k, λ) *design, the following relations hold among the five parameters* v, r, b, k, λ:

$$vr = bk,$$
$$r(k-1) = \lambda(v-1).$$

Proof. To prove that $vr = bk$ we simply note that vr is the total number of times, counting multiplicities, that the v distinct elements occur in the design. Since the elements are arranged in b blocks each containing exactly k elements, we must have $vr = bk$. For the second equality, each symbol a occurs in r blocks together with $k-1$ other symbols in each block. Hence a is a member of $r(k-1)$ ordered pairs. Also element a must occur λ times with each of the $v-1$ other symbols, so the second equality follows. □

Hence given (v, k, λ), we can always determine b and r. In fact, given any three of the parameters, we can uniquely determine the remaining two.

Example 2.4.4. The rows shown in Figure 2.2 form the blocks of a $(7, 3, 1)$ design. Notice that the rows are developed cyclically from the first by successively adding 1 modulo $v = 7$ to each element. Moreover, we will soon see that each row has a special property: if we take the 6 possible differences modulo 7 between ordered pairs in that row, we obtain each nonzero residue modulo 7 exactly once. The rows are also the lines in the projective plane $\mathrm{PG}(2, F_2)$ with points $\{0, 1, 2, 3, 4, 5, 6\}$, as shown in Figure 2.3.

The following is a fundamental, and still unresolved, problem in the theory of designs.

Question 2.4.5. Characterize the values of (v, k, λ) for which there is a (v, k, λ) design. If integer values satisfy the conditions for a block design and if such a design exists, how does one construct it?

0	1	3
1	2	4
2	3	5
3	4	6
4	5	0
5	6	1
6	0	2

Figure 2.2. A (7,3,1) block design

It is known that the conditions in Proposition 2.4.3 are not enough to guarantee the existence of a design. The values $v = 22$, $r = 7$, $k = 7$, and $\lambda = 2$ satisfy $vr = bk$ and $r(k-1) = \lambda(v-1)$ but it can be shown that no $(22, 7, 2)$ design exists. This proof uses a more general form of the Bruck/Ryser theorem (2.2.26); see Colbourn and Dinitz [8, p. 76].

Open Problem 2.4.6. Is there a $(22, 8, 4)$ design? In such a design, $v = 22$, $k = 8$, $\lambda = 4$, $b = 33$, and $r = 12$. This is the smallest unresolved case of Question 2.4.5.

Definition 2.4.7. A set $D = \{d_1, \ldots . d_k\}$ of residues modulo v is a (v, k, λ) *difference set* if every nonzero residue modulo v occurs exactly λ times as a difference $d_i - d_j$.

Example 2.4.8. The set $\{0, 1, 3\}$ from Example 2.4.4 forms a $(7, 3, 1)$ difference set.

Theorem 2.4.9. *Let $\{d_1, \ldots, d_k\}$ be a (v, k, λ) difference set. Define*

$$B_t = \{d_i + t \pmod{v} \mid 1 \leq i \leq k\}$$

for $0 \leq t \leq v-1$. The collection $\{B_t \mid 0 \leq t \leq v-1\}$ forms a (v, k, λ) design.

Proof. Let a be a nonzero residue modulo v. We can see that a is in B_r for each r which can be obtained as $a - d_i$ for some i. So a is in exactly k blocks. Consider the pair (a, c). If $a = d_i + t$ and $c = d_j + t$, that is, a and c are both in block B_t, then $a - c = d_i - d_j$. Since D is a difference set, this occurs exactly λ times. □

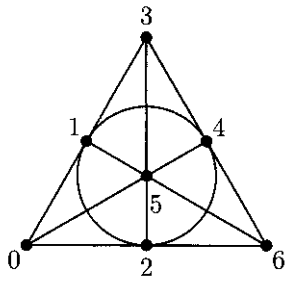


Figure 2.3. The projective plane $\mathrm{PG}(2, F_2)$ has seven lines, each containing three points.

We leave it to the reader to check that for $k \geq 3$, the existence of a symmetric $(v, k, 1)$ design is equivalent to the existence of a projective plane of order v.

Corollary 2.4.10. *If* $\{d_1, \ldots, d_k\}$ *is a* $(v, k, 1)$ *difference set with* $k \geq 3$, *then the collection* $\{B_t\}$ *forms a finite projective plane* $\mathrm{PG}(2, k-1)$.

Finite fields can be used to construct many different kinds and types of designs. We summarize, without proof, the following results.

Let q be a prime power, let $m \geq 2$ be a positive integer, and let t be a positive integer with $1 \leq t \leq m$. We now form designs by using the elements of $\mathrm{PG}(m, F_q)$ or $\mathrm{AG}(m, F_q)$ as points of the design and the t-flats as blocks. In this way one can build an infinite number of (v, k, λ) designs whose parameters in the projective case are shown below:

$$v = \frac{q^{m+1}-1}{q-1}, \qquad b = \prod_{i=1}^{t+1} \frac{q^{m-t+i}-1}{q^i-1}, \qquad r = \prod_{i=1}^{t} \frac{q^{m-t+i}-1}{q^i-1},$$

$$k = \frac{q^{t+1}-1}{q-1}, \qquad \lambda = \prod_{i=1}^{t-1} \frac{q^{m-t+i}-1}{q^i-1}.$$

We note that these designs are symmetric if $t = m-1$, the case when the t-flats are hyperplanes in $\mathrm{PG}(m, F_q)$. If $t = 1$, then the product for λ is interpreted to be 1.

In the affine case, the designs are not symmetric; their parameters are shown below:

$$v = q^m, \qquad b = q^{m-t}\prod_{i=1}^{t}\frac{q^{m-t+i}-1}{q^i-1}, \qquad r = \prod_{i=1}^{t}\frac{q^{m-t+i}-1}{q^i-1},$$

$$k = q^t, \qquad \lambda = \prod_{i=1}^{t-1}\frac{q^{m-t+i}-1}{q^i-1}.$$

As in the projective case, if $t = 1$, then the product for λ is interpreted to be 1.

We now illustrate how a projective space over the field F_q can be used to construct difference sets.

Theorem 2.4.11. *The points in any hyperplane of* $\mathrm{PG}(m, F_q)$ *determine a* (v, k, λ) *difference set with* $v = (q^{m+1}-1)/(q-1)$, $k = (q^m-1)/(q-1)$, *and* $\lambda = (q^{m-1}-1)/(q-1)$.

Proof. Assume α is a primitive element of $F_{q^{m+1}}$ so that the multiplicative group $F^*_{q^{m+1}}$ is generated by α. The set $\{1, \alpha, \alpha^2, \ldots, \alpha^m\}$ forms a polynomial basis for $F_{q^{m+1}}$ over F_q so every element α^i of $F_{q^{m+1}}$ can now be represented as $\alpha^i = a_0 + a_1\alpha + \cdots + a_m\alpha^m$, with $a_j \in F_q$ for each $j = 0, 1, \ldots, m$. Thus we can associate the element α^i with the $(m+1)$-tuple $(a_0, a_1, \ldots, a_m)$. In this way we can identify the points of $\mathrm{PG}(m, F_q)$ with powers of α.

Let $H = \{\alpha^{d_1}, \ldots, \alpha^{d_k}\}$ be any hyperplane of $\mathrm{PG}(2, F_q)$. Then any other hyperplane can be put in the form $H_e = \alpha^e H$, where $0 \le e \le v-1$. Note that $\mathrm{PG}(m, F_q)$ has $v = (q^{m+1}-1)/(q-1)$ distinct hyperplanes, each of which contains $k = (q^m-1)/(q-1)$ points. The next table gives a list of the points in each of the hyperplanes $H_1, H_2, \ldots, H_{v-1}$ in terms of $H = H_0$. In listing the elements of the hyperplane H, we list only the exponents of α in the representation above.

Hyperplane	Points (exponents of α)			
H_0	d_1	d_2	$\ldots$	d_k
H_1	d_1+1	d_2+1	$\ldots$	d_k+1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
H_{v-1}	d_1+v-1	d_2+v-1	$\ldots$	d_k+v-1

In the above table, consider only the rows which contain the exponent 0 (actually any fixed value will work, but notationally it is easier to work with 0). These are the rows of the k hyperplanes containing the point α^0. Hence the exponents of α in the points for these k hyperplanes may be listed as:

$$\begin{array}{cccc} d_1 - d_1 & d_2 - d_1 & \dots & d_k - d_1 \\ d_1 - d_2 & d_2 - d_2 & \dots & d_k - d_2 \\ \vdots & \vdots & \ddots & \vdots \\ d_1 - d_k & d_2 - d_k & \dots & d_k - d_k \end{array}.$$

Note that a point $\alpha^i \neq \alpha^0$ appears in as many rows of these k hyperplanes as there are hyperplanes containing any two fixed points, which is $\lambda = (q^{m-1} - 1)/(q - 1)$. Hence the off-diagonal entries repeat each nonzero residue modulo v exactly λ times. We thus have a (v, k, λ) difference set. □

5. Hadamard matrices

Hadamard's inequality states that if $H = (h_{ij})$ is a real square matrix of size n such that $|h_{ij}| \leq 1$ for all $1 \leq i, j \leq n$, then $|\det(H)| \leq n^{n/2}$. Furthermore, these will be equal if and only if $HH^T = nI$, where I denotes the identity matrix of size n, and $|e|$ denotes the absolute value of e. Matrices that achieve this upper bound are called Hadamard matrices; constructing such matrices is a classical combinatorics problem which we will investigate briefly here. For more information, see Wallis [64, Chapters 8 and 9].

Definition 2.5.1. A *Hadamard matrix* H_n is an $n \times n$ matrix, each of whose entries is either -1 or 1, such that $H_n H_n^T = nI$, where I is the identity matrix of size n.

We note that in a Hadamard matrix, any two distinct rows, or columns, are orthogonal as vectors. This is equivalent to their inner product (dot product) being zero.

Examples 2.5.2. Some Hadamard matrices of orders 1, 2, and 4:

$$(1) \qquad \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Our next result limits the possible orders for which Hadamard matrices exist.

Theorem 2.5.3. *If H_n is a Hadamard matrix with $n > 2$, then n is divisible by 4.*

Proof. By multiplying the rows and columns of H_n by -1 as necessary, we can assume that the first row and column of H_n consist of only 1s.

Now consider the sum

$$\begin{aligned} &\sum_{j=1}^{n} (a_{1j} + a_{2j})(a_{1j} + a_{3j}) \\ &\quad = \sum_{j=1}^{n} a_{1j}^2 + \sum_{j=1}^{n} a_{1j}a_{3j} + \sum_{j=1}^{n} a_{2j}a_{1j} + \sum_{j=1}^{n} a_{2j}a_{3j} \\ &\quad = \sum_{j=1}^{n} a_{1j}^2 = n\,. \end{aligned}$$

We note that $(a_{1j} + a_{2j})(a_{1j} + a_{3j})$ is either 0 or 4, and thus n is divisible by 4. □

Question 2.5.4. Is there a Hadamard matrix of size $4k$ for each $k \geq 1$?

It is believed that the answer is affirmative, but this is still not proved; the first unresolved case occurs when $4k = 668$.

Let G be a finite Abelian group. A (multiplicative) *character of G* is a homomorphism from G into the multiplicative group of complex numbers of absolute value 1. Hence if χ is a character of G, then $\chi(ab) = \chi(a)\chi(b)$ for all $a, b \in G$.

We refer to Lidl and Niederreiter [36, Chapter 5][37] for a discussion and various properties of characters defined over a finite field F_q. Proofs of the following results can be found there.

Proposition 2.5.5. *For any nontrivial multiplicative character χ we have*

$$\sum_{c \in F_q^*} \chi(c) = 0.$$

For odd q, let ξ be defined on F_q^* by

$$\xi(c) = \begin{cases} 1 & c = b^2 \text{ is a square,} \\ -1 & \text{otherwise.} \end{cases}$$

The reader should verify that ξ is a character on F_q^*, called the *quadratic character*. Moreover, it can be shown that $\xi(-1) = -1$ if and only if $q \equiv 3 \pmod 4$.

The following theorem will be in the construction of Hadamard matrices. A proof is given by Lidl and Niederreiter [37, Theorem 5.18].

Theorem 2.5.6. *If q is odd, let $f(x) = a_2x^2 + a_1x + a_0$ be a polynomial over F_q with $a_2 \neq 0$, and let $d = a_1^2 - 4a_0a_2$ be the discriminant of $f(x)$. Then*

$$\sum_{c \in F_q} \xi(f(c)) = \begin{cases} -\xi(a_2) & \text{if } d \neq 0, \\ (q-1)\xi(a_2) & \text{if } d = 0. \end{cases}$$

We are now ready to prove our main result concerning the use of finite fields in the construction of Hadamard matrices.

Theorem 2.5.7. *Let $q \equiv 3 \pmod 4$ be a prime power, and let $F_q = \{a_1, \ldots, a_q\}$. Define a matrix*

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & -1 & b_{12} & b_{13} & \cdots & b_{1q} \\ 1 & b_{21} & -1 & b_{23} & \cdots & b_{2q} \\ 1 & b_{31} & b_{32} & -1 & \cdots & b_{3q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & b_{q1} & b_{q2} & b_{q3} & \cdots & -1 \end{pmatrix}$$

with $b_{ij} = \xi(a_j - a_i)$ *where* $1 \leq i, j \leq q$ *for* $i \neq j$ *(here* ξ *denotes the quadratic character). The matrix* H *is a Hadamard matrix of order* $n = q + 1$.

Proof. We need to verify that the rows of the matrix are orthogonal, that is, verify that the inner products of the distinct rows of the matrix are all zero. The proof consists of a calculation which will use some properties of characters referred to above.

We first calculate the inner product of the first row with some other row $i + 1$, where $1 \leq i \leq q$. We obtain

$$1 - 1 + \sum_{j \neq i} b_{ij} = \sum_{j \neq i} \xi(a_j - a_i) = \sum_{c \in F_q^*} \xi(c),$$

which by one of the above properties of characters, is 0.

We now calculate the inner product of rows $i+1$ and $k+1$, where $k \neq i$, with $1 \leq i < k \leq q$. In this case we obtain

$$\begin{aligned} 1 - b_{ki} - b_{ik} + \sum_{j \neq i,k} b_{ij} b_{kj} \\ = 1 - \xi(a_i - a_k) - \xi(a_k - a_i) + \sum_{j \neq i,k} \xi(a_j - a_i)\xi(a_j - a_k), \end{aligned}$$

which, after some simplification, can be rewritten as

$$1 - [1 + \xi(-1)]\xi(a_i - a_k) + \sum_{c \in F_q} \xi(c^2 - (a_i + a_k)c + a_i a_k).$$

Since $q \equiv 3 \pmod 4$, $\xi(-1) = -1$. By Proposition 2.5.5, this reduces to $1 - \xi(1) = 1 - 1 = 0$, and hence row $i + 1$ is indeed orthogonal to row $k + 1$, and the proof is complete. □

Once we have a Hadamard matrix H_n, we can obtain others as follows.

Proposition 2.5.8. *Let* H_n *is a Hadamard matrix of order* n. *Then the matrix*

$$\begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix}$$

is a Hadamard matrix of order $2n$.

In particular, if $m = 2^k$, $k \geq 0$, then we can obtain a Hadamard matrix of size m by starting with the trivial Hadamard matrix H_1 and repeating this construction.

6. Notes

For a comprehensive treatment of the theory of latin squares as well as a discussion of some applications, we refer to Dénes and Keedwell [11]. See also Dénes and Keedwell [12] for further theory of latin squares. The book by Laywine and Mullen [31] is a textbook at the undergraduate level which discusses various topics in discrete mathematics, each topic being motivated by some connection to latin squares. We refer to Section III, Chapter 1 of Colbourn and Dinitz [8] for a discussion of latin squares and to Chapter 3 of the same section for a table of the best values for the number of MOLS of order n for $n \leq 10,000$.

The *Handbook of Combinatorial Designs* [8], edited by Colbourn and Dinitz, contains a huge amount of material related to the properties and constructions of many different kinds of combinatorial objects, many of which can be constructed using various properties of finite fields and polynomials over finite fields.

Further reading related to affine and projective planes may be found in Dénes and Keedwell [11, Chapter 8]. Colbourn and Dinitz [8] provide a magnificent survey of many structures in the area of combinatorics known as design theory; a detailed discussion for the construction of block designs occurs in Section II, and a survey of results related to Hadamard matrices and their construction is in Section V, Chapter 1.

7. Exercises

Latin squares.

2.1. Compute by hand the values of l_n for $n = 2, 3, 4$; and then calculate by machine the values of l_n for $n = 5, 6, 7$.

2.2. For each $n \geq 2$, construct a *partial latin square of order* n with n cells filled which cannot be completed to a latin square of order n.

A partial latin square of order n is an $n \times n$ square with some cells filled but filled in such a way that no row or column contains any symbol twice.

2.3. Construct a pair of latin squares of order 6 with the property that when the two squares are superimposed there are 34 distinct ordered pairs. Explain why there is no pair of latin squares of order 6 with exactly 35 distinct pairs.

2.4. Construct 3 MOLS of each of the orders 5, 8, and 9.

2.5. Construct a pair of MOLS of order 21.

2.6. Construct a latin square L of order 4 which does not have an orthogonal mate, that is, for which there does not exist a latin square of order 4 orthogonal to L. Prove that your latin square does not have an orthogonal mate.

2.7. For q an odd prime power, show that the latin square L of order $q-1$ which represents the multiplication table of F_q^* does not have an orthogonal mate (that is, there is not another latin square of order $q-1$ orthogonal to the square L).

2.8. Let q be a prime power and let $n \geq 1$ be a positive integer with $(n, q^2-1) = 1$. Let $a \neq 0 \in F_q$. Show that as b varies over the nonzero elements of F_q, the Dickson polynomials $D_n(x,a) + bD_n(y,a)$ give a complete set of $q-1$ mutually orthogonal latin squares of order q. (For the definition of Dickson polynomials, see Section 6.4.)

2.9. If $q \geq 5$ is an odd prime power, construct a set of $q-3$ diagonal MOLS of order q, that is, MOLS for which each square in the set has distinct elements on each of the two main diagonals.

2.10. Construct $q^2 - q$ mutually orthogonal Sudoku latin squares of order q^2, where q is an arbitrary prime power.

2.11. Show that if one has a pair of orthogonal diagonal latin squares L_1, L_2 of order n, then one can construct a *magic square* based on the symbols $0, 1, \ldots, n^2 - 1$ (the sum of the elements in each row, each column, and each of the two main diagonals is the same) of order n via the construction $nL_1 + L_2$ where the arithmetic is performed modulo n^2. Construct a magic square of order 5.

2.12. Let $q \neq 2, 3$ be a prime power. Construct a latin square of order q which is orthogonal to its transpose. Such a square is said to be *self-orthogonal.*

2.13. Give an example of a latin square which is not the Cayley table of any group, although the first row and first column of the square are in the standard order.

Frequency squares and latin hypercubes.

2.14. For q a prime power and an integer $i \geq 1$, construct a complete set of $(q^i - 1)^2/(q - 1)$ MOFS of type $F(q^i; q^{i-1})$.

2.15. For q a prime power and an integer $d \geq 2$, construct a complete set of $(q^d - 1)/(q - 1) - d$ mutually orthogonal hypercubes of dimension d and order q.

2.16. For q a prime power and $d \geq 2$ an integer, construct a set of

$$\frac{1}{q-1} \sum_{k=j+1}^{d} \binom{d}{k} (q-1)^k$$

mutually orthogonal hypercubes of order q, type j with $1 \leq j \leq d-1$ and dimension d. By having *type* j, we mean that when any j of the d coordinates are fixed, each of the q symbols occurs exactly q^{d-j-1} times in that subarray consisting of q^{d-j} cells. Also note that a hypercube of type j automatically has type i with $0 \leq i \leq j$.

2.17. Let $i \geq 1$ be a positive integer. Show that for $a \in F_{q^i}$ and $b \in F_{q^i}$ but with no two of the b values being F_q^* multiples of each other, the squares of order q^i constructed with the element $\mathrm{Tr}_{F/K}(b(ax + y))$ being placed at the intersection of row x and column y gives a complete set of $F(q^i; q^{i-1})$ MOFS.

Finite geometries.

2.18. Verify that the finite geometries constructed in Example 2.3.7 are actually an affine plane and a projective plane.

2.19. Construct an affine plane AG(2, 3) using a complete set of MOLS of order 3, and show this plane is the same as that constructed using Theorem 2.3.4.

2.20. Construct the affine geometry $\mathrm{AG}(3, F_2)$ and the projective geometry $\mathrm{PG}(3, F_2)$.

2.21. Using the projective geometry constructed in the above exercise, construct a $(15, 7, 3)$ difference set arising from Theorem 2.4.11.

Hadamard matrices.

2.22. Construct Hadamard matrices of orders 8 and 12.

2.23. Explain why the construction from Theorem 2.5.7 cannot be used to construct Hadamard matrices when $q \equiv 1 \pmod 4$.

Proofs left to the reader.

2.24. Assume that two $n \times n$ latin squares L_1 and L_2 are orthogonal. Prove that if the symbols of L_1 are permuted, then the resulting square is still orthogonal to L_2. Prove that if $L_1, L_2, \ldots, L_i$ form a set of MOLS of order n, then there is a set $L_1', L_2', \ldots, L_i'$ of i MOLS of size n such that the first row of each square L_j' is in the standard order.

2.25. Prove Lemma 2.2.22.

2.26. Prove Lemma 2.2.23.

2.27. Prove Theorem 2.3.4.

http://dx.doi.org/10.1090/stml/041/03

Chapter 3

Algebraic Coding Theory

1. Introduction

In today's world of modern communications, we often want to send messages or files from one place to another, or from one computer to another. Scientific progress in areas such as deep-sea and deep-space research requires the ability to transmit information through difficult environments.

When we send a message, we hope the person on the receiving end can obtain the original message without errors which may have been introduced during the transmission process. The difficulty is that when one transmits information via a communication line, noise and other environmental factors often introduce errors so that the received message is not the same as the original message. For example, radiation from the sun interferes with transmissions between the earth and communication satellites.

In order to analyze situations such as this, we first assume that there is a fixed set M of messages that we might wish to send, and that this list is known to both the sender and the receiver.

In common applications, a message might be a word, a single letter, or a fixed finite word on a fixed finite alphabet. Thus a "message"

A	·–	B	–···	C	–·–·
D	–··	E	·	F	··–·
G	––·	H	····	I	··
J	·–––	K	–·–	L	·–··
M	––	N	–·	O	–––
P	·––·	Q	––·–	R	·–·
S	···	T	–	U	··–
V	···–	W	·––	X	–··–
Y	–·––	Z	––··	1	·–––––
2	··–––	3	···–––	4	····–
5	·····	6	–····	7	––···
8	–––··	9	––––·	0	–––––

Table 3.1. The codewords of Morse code

in the ordinary sense may be a sequence of what we call messages. A *code* is an injection from a set of messages to a set of words on a fixed finite alphabet (the words in the range of this function are called *codewords*). We require a code to be injective so that we can decode the sequence that is received.

We will use Morse Code as a simple example of what we mean by a code. This code can be viewed as a system for translating the messages $\{A, B, \ldots, Z, 1, 2, \ldots, 9, 0\}$ to sequences of letters from the alphabet $\{\cdot, -\}$ (usually $\cdot$ is called *dot* and $-$ is called *dash*). The assignment function for this code is shown below. To send a word by Morse code, each letter in the word is translated and sent in sequence.

One of the goals of coding theory is to make it possible for the person receiving a message to detect and correct errors that have arisen during the transmission process. The detection of errors is accomplished by noticing that the received sequence is not a codeword. For example, if a Morse Code operator receives $\cdot - \cdot -$ then the operator will immediately know that an error has occurred, because this sequence is not assigned to any letter or number. The operator has no way of telling which message was intended, and thus cannot correct the error.

For some codes, it is possible for the receiver to determine, with high probability, the intended message when the received sequence is not a codeword. Such codes are often called *error-correcting codes.* Because errors during transmission are inevitable in many situations, error-correcting codes are essential for transmitting data efficiently. These codes are often called *algebraic codes* because they are usually constructed using some algebraic system, very often a finite field.

It is important to note that the codes we discuss in this chapter are not meant to keep information secret. For example, anyone who knows Morse Code and has suitable radio equipment can snoop on other people's conversations. The goal of error-correcting codes is only to allow the intended recipient to correct errors that have occurred during transmission. The field of *cryptography*, which we will discuss in Chapter 4, is devoted to keeping information private.

In this chapter we first develop a few basic properties of codes, discuss some bounds on code parameters, and discuss some encoding and decoding techniques. We then illustrate several constructions for classes of codes using finite fields, briefly discuss perfect codes, as well as several connections between codes and combinatorial designs. We conclude the chapter with some connections between sets of orthogonal latin squares and codes.

2. Basic properties of codes

In order to use algebraic properties and methods for the construction of our codes, we will restrict our attention to the theory of linear codes. In this way we will be able to view our codes as subspaces of a vector space over the finite field F_q.

Let F_q^n denote the set of all n-tuples over the field F_q:

$$F_q^n = \{(a_1, \ldots, a_n) \mid a_i \in F_q, i = 1, \ldots, n\}.$$

The reader should recall that F_q^n is a vector space over the field F_q. It has dimension n, with a standard basis consisting of the n unit vectors of length n.

In this chapter, we will assume that the messages we wish to transmit are elements of F_q^k for some $k \geq 1$. We may thus view these messages as words of length k on the alphabet which consists

of the q elements of F_q. In particular, if $q = 2$, then our messages are binary strings of length k. Under the assumption that the messages are the elements of F_q^k, we have q^k distinct messages that can be sent. To apply our theory to the real world, we first take the true set of messages we wish to send and assign each one to an element of F_q^k. This assignment function (which must be injective) is shared with the desired recipient. Because the creation and use of assignment functions is routine, we will ignore it for the rest of this chapter.

When we assume our messages come from the set F_q^k, we will always assume our codewords come from the set F_q^n for some $n \geq k$. A code thus gives an injective function from F_q^k to F_q^n; the codewords are the range of this function. We are particularly interested in those codes for which the range is a subspace of F_q^n, for then we can use results of linear algebra to analyze the code.

Definition 3.2.1. A *linear code* C is a subspace of the vector space F_q^n. Such a code is called a *q-ary* code; the code is *binary* if $q = 2$ and *ternary* if $q = 3$. The number n is the *length* of the code.

Since a linear code C is a subspace of F_q^n, it will contain q^k distinct codewords for some k with $0 \leq k \leq n$. The integer k is called the *dimension* of the linear code C. We can also recognize k as the length of each uncoded message, for our messages will be elements from the set F_q^k. We will denote such a code C as an $[n, k]$ linear code.

Examples 3.2.2. We briefly discuss several elementary examples of linear codes. We can define a code, called a *q-ary repetition code*, which acts by repeating the message $a \in F_q$ that is to be encoded a total of n times: $a \to a \ldots a$. Clearly this is a linear code of length n and dimension 1.

A binary *parity-check code* over F_2 can be constructed with the map $(a_1, \ldots, a_n) \mapsto (a_1, \ldots, a_n, \sum_{i=1}^{n} a_i)$. This code is also linear, and has a large dimension $n - 1$, but (as will be seen later) no error-correcting ability.

There are various ways to encode messages, but we will focus our efforts on two matrix methods. One uses a parity-check matrix and

the other uses a generator matrix. We begin with the parity-check method.

Assume for the moment that the messages we wish to encode are of the form $(c_1, \ldots, c_k)$ where each $c_i \in F_q$. We encode each message by appending an additional $n-k$ digits, called *parity-check digits* $(c_{k+1}, \ldots, c_n)$, with each c_i also in F_q. We will then have a codeword $\mathbf{c} = (c_1, \ldots, c_n) \in F_q^n$ where the first k digits are information digits and the last $n-k$ digits are parity-check digits. We now show how to find the parity-check digits.

Let H be an $(n-k) \times n$ matrix over F_q of rank $n-k$. Given such a matrix H, we can construct a code C by letting $\mathbf{c}$ be a codeword if and only if $H\mathbf{c}^T = 0$, where T denotes the transpose. The matrix H is called a *parity-check matrix* for the code C.

Lemma 3.2.3. *Let H be a $(n-k) \times n$ matrix over F_q of rank $n-k$. Then $C = \{\mathbf{c} \in F_q^n \mid H\mathbf{c}^T = \mathbf{0}\}$ is a linear $[n, k]$ code.*

Proof. The code C will indeed be linear because (from linear algebra) if $\mathbf{c}_1, \mathbf{c}_2 \in C$, then $H(a\mathbf{c}_1 + b\mathbf{c}_2)^T = \mathbf{0}$. Since the rank of H is $n-k$, the code C has dimension k. □

The equation $H\mathbf{c}^T = 0$ leads to a system of linear equations over the field F_q which can be used to determine $c_{k+1}, \ldots, c_n$ given $c_1, \ldots, c_k$. These equations are often called the *parity-check equations* for the linear code C. If H is of the form $H = (A|I_{n-k})$, where I_{n-k} is the identity matrix of order $n-k$, then the code C is said to be in *systematic form*. Note that if H is in systematic form, then we can compute the parity-check digits with particular ease.

We now turn from parity-check matrices to consider a dual way of forming a linear $[n, k]$ code. To this end, let G be a $k \times n$ matrix (with no zero columns) of rank k over F_q. We form a code C by putting $\mathbf{c} \in C$ if $\mathbf{c}$ is in the row space of the matrix G. Thus C will consist of all vectors of the form $\mathbf{c} = \mathbf{a}G$ where $\mathbf{a}$ runs through all messages of length k, that is, all $1 \times k$ vectors $(a_1, \ldots, a_k)$ over F_q. The matrix G is called a *generator matrix* for the code C.

Lemma 3.2.4. *Let G be a $k \times n$ matrix over F_q. The set $C = \{\mathbf{a}G \mid \mathbf{a} \in F_q^k\}$ is a linear code. The dimension of C is the rank of G.*

The proof of Lemma 3.2.4 is left as Exercise 3.15.

If G is of the form $G = (I_k | -A^T)$, where A is of size $(n-k) \times n$ and I_k denotes the $k \times k$ identity matrix, then G is in *systematic form*. In this case, let H be the matrix $(A|I_{n-k})$. Then the matrices G and H are related by the matrix equation $GH^T = 0$; in this case, the code generated by H is the same as the code for which G is the generator matrix (see Exercise 3.9). Thus given either a generator matrix G or a parity-check matrix H, we can obtain the other. We now consider a small example to illustrate these concepts.

Example 3.2.5. We work over the binary field F_2. Let

$$H = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix} \qquad A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

so that the parity-check matrix H is in systematic form with $n = 7$ and $k = 4$. We thus have a binary $[7, 4]$ linear code.

Recall that a vector $\mathbf{c} \in F_2^7$ is a codeword if $H\mathbf{c}^T = 0$, which is equivalent to satisfying the following system of linear equations.

$$\begin{aligned} c_1 + c_3 + c_4 + c_5 &= 0, \\ c_1 + c_2 + c_3 + c_6 &= 0, \\ c_2 + c_3 + c_4 + c_7 &= 0. \end{aligned}$$

These can be rewritten as the system

$$\begin{aligned} c_5 &= c_1 + c_3 + c_4, \\ c_6 &= c_1 + c_2 + c_3, \\ c_7 &= c_2 + c_3 + c_4, \end{aligned}$$

which gives an efficient scheme for encoding the message (c_1, c_2, c_3, c_4): one can encode a message (c_1, c_2, c_3, c_4) using the formula

$$(c_1, c_2, c_3, c_4) \mapsto (c_1, c_2, c_3, c_4, c_1 + c_3 + c_4, c_1 + c_2 + c_3, c_2 + c_3 + c_4).$$

Using the same example a generator matrix G is given by

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

so that we have a $[7, 4]$ linear binary code C over the field F_2. Since C has dimension 4, it will contain $2^4 = 16$ binary codewords each of length 7. The actual codewords will be given in Example 3.5.1.

From the above discussion, it is clear that we can construct many linear codes by simply writing down matrices to use as generator or parity-check matrices. As will be seen later, the difficulty is to determine if the codes are useful for detecting and correcting errors. In order to answer these questions, we will need the concept of the weight of a vector and the distance between two vectors.

Definition 3.2.6. The *Hamming distance* $d(\mathbf{x}, \mathbf{y})$ between two vectors $\mathbf{x}$ and $\mathbf{y}$ in F_q^n is the number of coordinates where the vectors differ. The *Hamming weight* $\mathrm{wt}(\mathbf{x})$ of a vector $\mathbf{x}$ is the number of coordinates where the vector is nonzero.

In a linear code, the Hamming distance $d(\mathbf{x}, \mathbf{y})$ is always equal to the Hamming weight of the vector $\mathbf{x} - \mathbf{y}$. In particular, we have $d(\mathbf{x}, \mathbf{0}) = \mathrm{wt}(\mathbf{x})$.

Proposition 3.2.7. *The Hamming distance function is a metric. That is, for all vectors* $\mathbf{u}$, $\mathbf{v}$, *and* $\mathbf{w}$:

(1) $d(\mathbf{u}, \mathbf{v}) \geq 0$.

(2) $d(\mathbf{u}, \mathbf{v}) = 0$ *if and only if* $\mathbf{u} = \mathbf{v}$.

(3) $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$.

(4) $d(\mathbf{u}, \mathbf{w}) \leq d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w})$.

Definition 3.2.8. If C is a linear code, then the *minimum distance* d_C of C is defined as

$$d_C = \min\{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x}, \mathbf{y} \in C, \mathbf{x} \neq \mathbf{y}\} = \min\{\mathrm{wt}(\mathbf{x}) \mid \mathbf{x} \in C, \mathbf{x} \neq \mathbf{0}\}.$$

Let $\mathbf{x} \in F_q^n$ be a vector of length n over F_q. The *sphere of radius* t *about* $\mathbf{x}$ is the set of all vectors $\mathbf{y} \in F_q^n$ for which $d(\mathbf{x}, \mathbf{y}) \leq t$. Exercise 3.1 asks the reader to determine the number of distinct vectors in a sphere of radius t.

Let C be an $[n, k]$ code. We say that C can *correct* t *errors* if whenever no more than t coordinates of a codeword are changed the original codeword can be effectively recovered from the changed

codeword. For $s > 0$, we say that C can *detect s errors* if whenever no more than s coordinates of a codeword are changed the resulting word is not in C.

Definition 3.2.9. A code C is said to be *t-error correcting* if for every vector $\mathbf{x} \in F_q^n$, there is at most one codeword $\mathbf{c} \in C$ within distance t of $\mathbf{x}$, that is, with $d(\mathbf{x}, \mathbf{c}) \leq t$.

If C is a linear code over F_q of length n, dimension k, and minimum distance d, we say that C is a q-ary linear $[n, k, d]$ code. As will be seen in our next result, the minimum distance of a code is what determines its error correcting ability.

Theorem 3.2.10. *Let C be a code.*

(1) *C can correct t errors if and only if $d_C \geq 2t + 1$.*

(2) *C can detect s errors if and only if $d_C \geq s + 1$.*

Proof. For part (1), if a codeword $\mathbf{c}$ is transmitted and the received vector $\mathbf{r}$ contains t or fewer errors, that is, differs from $\mathbf{c}$ in t or fewer coordinates, then $d_C(\mathbf{c}, \mathbf{r}) \leq t$. If $\mathbf{c}' \in C$ is any other codeword, then $d_C(\mathbf{c}', \mathbf{r}) \geq t + 1$, for if not, we would have

$$d_C(\mathbf{c}, \mathbf{c}') \leq d_C(\mathbf{c}, \mathbf{r}) + d_C(\mathbf{r}, \mathbf{c}') \leq 2t,$$

a contradiction.

For part (2), assume the received vector is obtained after the introduction of s or fewer errors. Then it cannot be mistaken for a codeword say $\mathbf{c}'$, because then $d_C(\mathbf{c}, \mathbf{c}') \leq s$, which is a contradiction. □

The following result provides an efficient way to obtain a lower bound on the minimum distance of a code.

Theorem 3.2.11. *A linear code C over F_q with parity check matrix H has minimum distance $d_C \geq s + 1$ if and only if any s columns of H are linearly independent over F_q.*

Proof. For the necessity, if we have s linearly dependent columns, then for some $\mathbf{c} \in C$, with $\mathbf{c} \neq 0$, $H\mathbf{c}^T = 0$ and so $\mathrm{wt}(\mathbf{c}) \leq s$. For the converse, if any s columns are linearly independent over F_q, then there is no $\mathbf{c} \in C$ with $\mathrm{wt}(\mathbf{c}) \leq s$, and so $d_C \geq s + 1$. □

We note that $d_C = s + 1$ if any s columns of H are linearly independent and there are $s + 1$ columns of H which are linearly dependent over F_q.

Suppose one wants to construct a code C to be an $[n, k, d]$ linear code. In order to maximize the utility of our code one would like:

(1) d to be large in order to be able to correct, or detect, large numbers of errors.

(2) n to be small in order to only have to transmit short messages.

(3) The *rate* of a code is defined to be k/n. One would like to have the rate be as close to 1 as possible so that most of the coordinates in the transmitted message are information digits (for which there are k coordinates), not parity-check digits, which are added to allow for error-correction and detection.

As will be seen, these goals are somewhat contradictory. As one improves the code in one of the properties, one usually must weaken the given code relative to some of the other properties. Thus we are in a trade-off situation. In order to find the best solution to the coding problem, we would first need to determine which parameters can be obtained by a code. This is not an easy task, however.

Open Problem 3.2.12. For which n, k, d and prime powers q is there an $[n, k, d]$ linear code over F_q?

We end this section with a probabilistic argument that a code that corrects relatively few errors may correct "enough" errors in practice. If $n < m$ are nonnegative positive integers, a reasonable assumption may be that it is less likely for a received vector to have m errors than it is to have n errors. The reasonability of this assumption depends on the precise communication channel causing the errors. Many probabilistic models have been developed for specific communications channels.

Let us assume that each coordinate of a transmitted vector will be correctly received with probability $p > 0$, independent of the other coordinates. This assumption is a reasonable model of the errors that

occur, for example, when reading from a computer hard disk drive. It is not a reasonable model for radio transmission, where errors tend to come in bursts.

Under the assumption that errors occur independently with a fixed probability $1-p$, the probability $P(i)$ of having i errors in a received vector of length n is given by the formula

$$P(i) = \binom{n}{i}(1-p)^i p^{n-i}.$$

The following table shows these probabilites when $n = 5$ and $p = 0.9$.

i	$P(i)$
0	0.59049
1	0.32805
2	0.07290
3	0.00810
4	0.00045
5	0.00001

In this case there is less than a 9/1000 chance that a received vector will arrive with 3 or more errors. Thus a code with the ability to correct only 2 errors will be able to correctly decode almost all received messages.

3. Bounds for parameters of codes

In this section we provide several bounds on the possible values of the parameters n, k, and d for a linear $[n,k,d]$ code over F_q.

Theorem 3.3.1 (Hamming bound). *Suppose C is a linear $[n,k]$, t-error correcting code over F_q of length n. Then*

$$q^k\left(1+\binom{n}{1}(q-1)+\cdots+\binom{n}{t}(q-1)^t\right) \leq q^n.$$

Proof. We first notice that the term q^n on the right-hand side simply counts the total number of vectors of length n over F_q that are available. Similarly, q^k is the number of codewords in the code C since C has dimension k. Also note that

$$1+\binom{n}{1}(q-1)+\cdots+\binom{n}{t}(q-1)^t$$

is the number of distinct vectors in a sphere of radius t around any codeword; see Exercise 3.1, from which the result follows. □

Codes whose parameters yield an equality in the Hamming bound, sometimes also called the *sphere-packing bound*, form a very special class of codes.

Definition 3.3.2. A t-error correcting code C over F_q of length n is *perfect* if F_q^n is the disjoint union of the spheres of radius t around the codewords of C.

We will provide a brief discussion and classification (without proof) for all q-ary linear perfect codes in Section 5.4.

Theorem 3.3.3 (Plotkin bound). *For a linear code C over F_q with parameters $[n, k, d_C]$, we have*

$$d_C \leq \frac{nq^{k-1}(q-1)}{q^k - 1}.$$

Proof. We show that $nq^{k-1}(q-1)$ is no less than the total weight of all the codewords. Fix an integer $1 \leq i \leq n$. Let D be the set of codewords with a 0 in coordinate i. Then D is a subspace of C containing $|C|/|D| = |C/D| = q^{k-1}$ codewords. Therefore the total weight of all codewords in D is less than or equal to $q^{k-1}(q-1)$, and thus the total weight of all codewords in C is no more than $nq^{k-1}(q-1)$. □

Another very simple, but useful bound, is contained in the next result.

Theorem 3.3.4 (Singleton bound). *If C is a linear $[n, k, d]$ code over F_q, then*

$$q^k \leq q^{n-d+1}, \tag{2}$$

which is equivalent to the inequality $d \leq n - k + 1$.

Proof. Let C be an $[n, k, d]$ linear code over F_q. Delete the last $d-1$ coordinates of each code word of C. This leaves $n-(d-1)$ coordinates. Distinct elements of C cannot agree on each of the first $n-(d-1)$ coordinates, because the code has distance d. Therefore

there can in total be no more than $q^{n-(d-1)} = q^{n-d+1}$ codewords, from which the result follows. □

We note that the Singleton bound can also be obtained as a corollary to Theorem 3.2.11 because the rank of the parity-check matrix for C is $n-k$. If equality holds in (2), then the code is called a *maximum distance separable (MDS) code.* In the last section of this chapter, we will see that sets of orthogonal latin squares are useful in the construction of some MDS codes.

The Hamming, Plotkin, and Singleton bounds give only necessary conditions for a linear $[n,k,d]$ code over F_q to exist. The next result provides a sufficient condition for the existence of such a code.

Theorem 3.3.5 (Gilbert–Varshamov bound). *There is a linear $[n,k]$ code over F_q with minimum distance no less than d if*

$$q^{n-k} > \sum_{i=0}^{d-2} \binom{n-1}{i} (q-1)^i .$$

Proof. The proof is obtained by forming a parity-check matrix of rank $n-k$ whose columns have the property that any $d-1$ columns are linearly independent vectors over F_q. To this end, we use any nonzero vector as the first column. The second column can be taken to be any nonzero vector which is not dependent with the first, that is, which is not a nonzero scalar multiple of the first column. Then the third column can be taken to be any nonzero vector which is not a linear combination of the first two columns. We thus continue to add new columns as long as new vectors are available.

We can add another vector of length $n-k$ as a column of H as long as one is available. Assume that we have chosen a total of m vectors of length $n-k$ over F_q for which any $d-1$ columns are linearly independent over F_q. The collection of vectors which are not available is the set of all linear combinations of these m vectors. The number of such linear combinations is seen to be t, where

$$t = \binom{m}{1}(q-1) + \cdots + \binom{m}{d-2}(q-1)^{(d-2)}.$$

If $t \leq q^{n-k} - 1$, we can choose another vector of length $n-k$. In particular, if $m = n - 1$, we will be able to choose an n-th vector as a column of our parity check matrix H. The code thus constructed will have length n and minimum distance at least d. Because H has $n - k$ rows, its rank is at most $n - k$, and thus the dimension of the code will be at least k. □

4. Decoding methods

As we have seen earlier on, given a linear code C it is very easy to encode a message. One can either use a parity-check matrix H or generator matrix G to encode a message, say m, into a codeword c. For example, $\mathbf{c} = \mathbf{m}G$ or $H\mathbf{c}^T = 0$ using the parity-check equations.

On the receiving end, how can one decode a received vector $\mathbf{r}$ to obtain the original codeword $\mathbf{c}$? Given a particular code, there may be special methods which can be used to efficiently decode received vectors. We will discuss only one general method which can be used to decode any linear code.

We begin with the assumption that the receiver, seeing a received vector $\mathbf{y}$ with errors will decode $\mathbf{y}$ to the codeword $\mathbf{c} \in C$ which is nearest, in Hamming distance, to $\mathbf{y}$. This method of decoding is known as *nearest neighbor decoding*. It is clear that if C is t-error correcting and $\mathbf{y}$ has no more than t errors, then the nearest neighbor decoding method produces the correct result.

Let C be a linear $[n, k]$ code so that $|C| = q^k$. We construct the collection of cosets $\{\mathbf{a} + C \mid a \in F_q^n\}$. (Recall C is a subgroup of F_q^n under addition, so this definition makes sense.) There are q^{n-k} cosets which partition F_q^n.

We next describe an array for a linear code C. Let the cosets of C be $\{\mathbf{a_i} + C\}$, where each $\mathbf{a_i}$ is chosen to be an element of minimum Hamming weight in its coset. The elements $\mathbf{a_i}$ are known as *coset leaders*. While in theory these coset leaders are very useful, one must point out that in a large code, it will be far from easy to determine all of the coset leaders.

We use the coset leaders to write the elements of F_q^n in an array, called the *standard array* for the linear code C:

$$\begin{array}{lcccccc} \mathbf{0}+C & = & \mathbf{0}+\mathbf{c}_1 & \mathbf{0}+\mathbf{c}_2 & \cdots & \mathbf{0}+\mathbf{c}_{q^k} \\ \mathbf{a}_1+C & = & \mathbf{a}_1+\mathbf{c}_1 & \mathbf{a}_1+\mathbf{c}_2 & \cdots & \mathbf{a}_1+\mathbf{c}_{q^k} \\ & \vdots & & & & \\ \mathbf{a}_s+C & = & \mathbf{a}_s+\mathbf{c}_1 & \mathbf{a}_s+\mathbf{c}_2 & \cdots & \mathbf{a}_s+\mathbf{c}_{q^k} \end{array}$$

where $s = q^{n-k}$. The element in position (i, j) of the array is thus the vector $\mathbf{a}_i + \mathbf{c}_j$.

Now given a received vector $\mathbf{a} \in F_q^n$, if we find the coset leader $\mathbf{a}'$ of the coset containing $\mathbf{a}$, then we assume that the correct codeword for $\mathbf{a}$ is $\mathbf{a} - \mathbf{a}'$. Thus we assume that the error introduced during transmission is the coset leader $\mathbf{a}_i$. Since each coset leader is assumed to have minimal weight in the given coset containing it, this method of decoding is often called *nearest neighbor decoding.*

One problem with using the standard array for decoding an $[n, k]$ linear code C is its large size. Recall that the standard array for C will contain q^k columns, the number of codewords in C, and the number of rows will be q^{n-k}.

The following notion will be very helpful in our decoding efforts. If $\mathbf{y} \in F_q^n$, then the *syndrome* $S(\mathbf{y})$ is $H\mathbf{y}^T$, where H is the parity-check matrix for C.

Lemma 3.4.1. *Let* $\mathbf{y}$ *be a vector in* F_q^n. *Then*

(1) $S(\mathbf{y}) = 0$ *if and only if* $\mathbf{y} \in C$.

(2) $S(\mathbf{y}) = S(\mathbf{z})$ *if and only if the equality of cosets* $\mathbf{y}+C = \mathbf{z}+C$ *holds.*

Proof. Part (1) is just a restatement of the condition involving a parity-check matrix for a vector $\mathbf{y}$ to be in the code C. Part (2) follows from a simple calculation, namely that $S(\mathbf{y}) = S(\mathbf{z})$ if and only if $H(\mathbf{y} - \mathbf{z})^T = \mathbf{0}$. The details are left to Exercise 3.11. □

Thus any two elements in the same row of the standard array have the same syndrome, while any two elements from different rows have different syndromes.

To implement the syndrome decoding method, we compute the syndrome of each coset leader and store these syndromes for future reference. We then compute the syndrome of each incoming message and compare it with the precomputed syndromes to find the coset leader. Subtracting the coset leader from the received message gives us the codeword.

We end this section with a few words regarding a drawback of the coset leader decoding algorithm. If we have a linear code over F_q with a large number of cosets (recall in an $[n, k]$ linear code there will be q^{n-k} distinct cosets), it will be extremely difficult to calculate all of the coset leaders and store all of their syndromes. In such cases, one sometimes uses what is known as *incomplete decoding*. Here one calculates and stores the syndromes of some large subset of the coset leaders. When a vector $\mathbf{r}$ is received, one calculates the syndrome $S(\mathbf{r})$ and searches through the stored set of syndromes. If that syndrome is found, then we decode $\mathbf{r}$ to be the codeword $\mathbf{c} = \mathbf{r} - \mathbf{a}_i$, where $\mathbf{a}_i$ is the coset leader such that $S(\mathbf{a}_i) = S(\mathbf{r})$, as usual. If in our search of the stored syndromes we do not locate the syndrome $S(\mathbf{r})$ of the received vector $\mathbf{r}$, then we ask the sender to resend the message. The term *incomplete decoding* is used because we are not able to decode all possible received vectors.

As an illustration of syndrome decoding, using the binary Hamming code $H(2, 3)$ from Example 3.5.1, one calculates the set of syndromes to obtain the following:

Coset leader	Syndrome
0000001	001
0000010	010
0000100	100
0001000	101
0010000	111
0100000	011
1000000	110

Thus if the vector $\mathbf{r}$ is received, the receiver calculates $S(\mathbf{r}) = 110$ which is seen to correspond to the coset leader $\mathbf{a}_7 = 1000000$, which is the error which occurred in transmission. Thus the received vector $\mathbf{r}$ is decoded to the codeword $\mathbf{c} = \mathbf{r} - \mathbf{a}_7 = 0111001$.

5. Code constructions

We now provide methods for the construction of several important classes of linear codes over finite fields.

5.1. Hamming codes. We begin by constructing the binary Hamming codes $H(2,m)$, where $m \geq 2$ is a positive integer. We form a parity-check matrix H of size $m \times (2^m - 1)$ whose columns are the binary representations of the numbers $1, 2, \ldots, 2^m - 1$. The reader should check that the matrix H will have rank m. This matrix produces a *binary Hamming code* $H(2,m)$ whose parameters are $n = 2^m - 1$, $k = 2^m - 1 - m$, $d = 3$. Any two columns of H are linearly independent over F_2 but many collections of three columns are dependent. Therefore, by Theorem 3.2.11, this code has minimum distance 3 and is thus a binary 1-error correcting code.

Example 3.5.1. For $m = 3$, a parity-check matrix H for the binary Hamming code $H(2,3)$ is given by the matrix

$$H = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

We can perform a similar construction over F_q for q an arbitrary prime power. For a positive integer $m \geq 2$, such a code will have an $m \times (q^m - 1)/(q-1)$ parity-check matrix whose columns are the nonzero vectors in F_q^m whose first nonzero entry is 1. There will be $(q^m - 1)/(q-1)$ such vectors over F_q, and thus the *q-ary Hamming code* $H(q,m)$ will be a linear code with parameters

$$\left[\frac{q^m - 1}{q-1}, \frac{q^m - 1}{q-1} - m, 3\right].$$

As in the binary case, any two columns of the parity-check matrix H will be linearly independent over F_q and some three columns will be dependent. Thus the general Hamming codes $H(q,m)$ will have minimum distance $d = 3$ and will thus also be 1-error correcting and 2-error detecting.

5.2. Cyclic codes. A very important class of codes is the class of cyclic codes, so named because of the following property.

Definition 3.5.2. A code C of length n is *cyclic* if for any codeword $(c_1, c_2, \ldots, c_n) \in C$, the shifted vector $(c_n, c_1, \ldots, c_{n-1})$ is also in C.

One of the reasons cyclic codes are so nice to work with is that there is a convenient algebraic way to associate vectors with polynomials in a certain factor ring. Consider the factor ring $F_q[x]/(x^n - 1)$ which consists of all polynomials over F_q of degree less than n. We know this ring is isomorphic to F_q^n as a vector space over F_q. We can therefore view a code over F_q as a subset of $F_q[x]/(x^n - 1)$. To this end, we will associate a vector $(a_0, a_1, \ldots, a_{n-1}) \in F_q^n$ with the polynomial $a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1} \in F_q[x]/(x^n - 1)$.

We note that since $x^n = 1$ in this factor ring, the shifted vector $(a_{n-1}, a_0, a_1, \ldots, a_{n-2})$ corresponds to the polynomial

$$x(a_0 + a_1 x + \cdots + a_{n-1} x^{n-1}) = a_{n-1} + a_0 x + a_1 x^2 + \cdots + a_{n-2} x^{n-1}.$$

More generally, we note that the cyclic shift maps on F_q^n become the maps $\{p(x) \mapsto x^j p(x) \mid 0 \leq j < n - 1\}$ on $F_q[x]/(x^n - 1)$.

Definition 3.5.3. An *ideal* I of a ring R is a subset of R such that I is a subring of R and I is closed under multiplication by elements of R.

For example, in the ring $\mathbb{Z}$ of integers, if I denotes the set of even integers, then I is an ideal of $\mathbb{Z}$ since I is a subring of $\mathbb{Z}$ and the product of any integer with an even integer always produces an even integer in I.

Our next result provides a fundamental connection relating cyclic codes of length n over F_q to ideals in the factor ring $F_q[x]/(x^n - 1)$.

Theorem 3.5.4. *A linear code $C \subseteq F_q[x]/(x^n - 1)$ is cyclic if and only if C is an ideal of $F_q[x]/(x^n - 1)$.*

Proof. If C is an ideal, then C is closed under the cyclic shift maps $p(x) \mapsto x^j p(x)$, so C is a cyclic code. Conversely, the map $\phi_{r(x)}$ sending $p(x)$ to $r(x)p(x)$, where $r(x) \in F_q[x]$ is fixed, can be decomposed into a sum of scalar multiples of cyclic shift maps. Therefore C is closed under $\phi_{r(x)}$ for all $r(x) \in F_q[x]$, so C is an ideal. □

We now construct generator and parity-check matrices for cyclic linear codes directly from polynomials over F_q. Let $g(x) = g_0 +$

$g_1x + \cdots + g_{n-k}x^{n-k}$, be a polynomial over F_q of degree $n-k$, where $g(x)|(x^n-1)$. Construct a $k \times n$ matrix G from the coefficients of the polynomial $g(x)$:

$$G = \begin{pmatrix} g_0 & g_1 & \cdots & g_{n-k} & 0 & \cdots & 0 \\ 0 & g_0 & \cdots & \cdots & g_{n-k} & \cdots & 0 \\ \cdots & \cdots & \cdots & & & & \\ 0 & \cdots & 0 & \cdots & g_0 & \cdots & g_{n-k} \end{pmatrix}.$$

Since $g_{n-k} \neq 0$, the rank of G is k, so G generates an $[n,k]$ linear code C.

Let $h(x) = (x^n-1)/g(x) = h_0 + h_1x + \cdots + h_kx^k$ be of degree k. We define an $(n-k) \times n$ matrix H as follows:

$$H = \begin{pmatrix} 0 & \cdots & 0 & h_k & h_{k-1} & \cdots & h_0 \\ 0 & \cdots & h_k & h_{k-1} & \cdots & h_0 & 0 \\ \vdots & & & \vdots & & & \vdots \\ h_k & \cdots & h_0 & 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

Note that $GH^T = 0$, so H is a parity-check matrix for the $[n,k]$ linear code generated by the matrix G.

Example 3.5.5. Let $g(x) = 1 + x + x^3 \in F_2[x]$. In $F_2[x]$, we have $x^7 - 1 = (x-1)(x^3+x+1)(x^3+x^2+1)$, so if we let $h(x) = x^4+x^2+1$, then $g(x)h(x) = x^7 - 1$. Following the previous construction, we obtain the matrices:

$$H = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

These matrices generate a binary linear $[7,4,3]$ code, which the alert reader will realize is the binary linear Hamming code $H(2,3)$.

Definition 3.5.6. Two codes are said to be *equivalent* if the columns of the parity-check matrix for one of the codes can be permuted and multiplied by nonzero scalars to obtain the columns of the parity-check matrix of the other code.

In the above example, $g(x)$ is a primitive polynomial over F_2 and the code obtained is equivalent to a Hamming code. This is not a

coincidence, as the following theorem shows. A proof is given by Hill [23, Theorem 12.24].

Theorem 3.5.7. *Let C be the cyclic linear code whose generator polynomial is primitive of degree $m \geq 2$ over F_2. Then C is equivalent to the binary Hamming code* $[2^{m-1}, 2^{m-1} - 1 - m, 3]$.

We mention the following q-ary results. First, not all Hamming codes are equivalent to cyclic codes. For example, the reader should check that the ternary Hamming code $H(3,2)$ is not equivalent to a cyclic code. More generally, the Hamming code $H(q,m)$ is equivalent to a cyclic code if $(m, q-1) = 1$; see Hill [23, p. 161].

5.3. BCH and Goppa codes. We briefly illustrate how to construct an important class of cyclic codes, called *BCH codes*, named for R. C. Bose, D. K. Ray-Chaudhuri, and A. Hocquenghem, who first studied various special cases of such codes.

Let $b \geq 0$ be an integer and let $\alpha \in F_{q^m}$ be a primitive nth root of unity. Assume that m is the multiplicative order of q modulo n so that m is the smallest positive integer so that $q^m \equiv 1 \pmod{n}$.

The BCH code will be constructed from the set of roots $\{\alpha^{b+i} \mid 0 \leq i \leq d-2\}$ of a generator polynomial. Here d is called the *designed distance* of the BCH code and is chosen so that $2 \leq d \leq n$. For each i with $0 \leq i \leq d-2$, let $M^{(i)}(x)$ be the minimum polynomial of α^i. Let $g(x)$ be the least common multiple of the polynomials $M^{(i)}(x)$. The polynomial $g(x)$ will be the generator polynomial for the cyclic BCH code.

Codes obtained for certain values of these parameters are known by various special names. When $b = 1$, the code is known as a *narrow-sense code*; when $n = q^m - 1$, the code is known as a *primitive code*; and when $n = q - 1$, the code is called a *Reed–Solomon code*. These special cases were studied before the general construction of BCH codes was completely understood.

One important feature of BCH codes is that codes of arbitrarily large minimum distance can be obtained. This property, which is important for error correction, is seen in the following result.

Theorem 3.5.8. *The minimum distance of a BCH code is no less than the designed distance.*

See Lidl and Niederreiter [37, Theorem 8.4.5] for a proof of this result.

Another important class of codes are the *Goppa codes*, which were developed by V. D. Goppa after the BCH codes. We now illustrate one method to construct these codes. Let $g(x) \in F_{q^m}$ be of degree t, where $1 \leq t \leq m$. Let $L = \{\gamma_0, \ldots, \gamma_{n-1}\}$ be distinct elements in F_{q^m} which are not roots of $g(x)$. The *Goppa code*, usually denoted $\Gamma(L, g)$, is the set of all vectors $(c_0, \ldots, c_{n-1}) \in F_q^n$ for which

$$\sum_{i=0}^{n-1} c_i g(\gamma_i) \frac{g(x) - g(\gamma_i)}{x - \gamma_i} = 0\,,$$

where the equality is in the polynomial ring $F_{q^m}[x]$.

As with BCH codes, Goppa codes of arbitrarily large minimum distance can be constructed. This will be illustrated in the following result; see Lidl and Niederreiter [37, Theorem 8.56] for a proof.

Theorem 3.5.9. *The dimension of the Goppa code $C = \Gamma(L, g)$ is at least $n - mt$ and the minimum distance of C is at least $t + 1$.*

Note that the dimension of a Goppa code decreases quite quickly as the minimum distance increases.

5.4. Perfect codes. We now provide a brief discussion of perfect codes, that is, codes over some F_q whose parameters n, k, and d yield an equality in the Hamming bound of Theorem 3.3.1. This bound states that if C is a linear $[n, k]$ code over F_q that is t-error correcting, then

$$q^k \left(1 + \binom{n}{1}(q-1) + \cdots + \binom{n}{t}(q-1)^t\right) \leq q^n.$$

Such a code is *perfect* if it achieves equality in the Hamming bound, which means that F_q^n is the disjoint union of the spheres of radius t around the codewords of C.

Theorem 3.5.10. *For a positive integer $m \geq 2$, the q-ary Hamming code $H(q, m)$ is perfect.*

Proof. For the q-ary Hamming code C, there are $(q^m - 1)/(q - 1)$ codewords. Let $S_1(\mathbf{c})$ be the sphere of radius 1 around a codeword $\mathbf{c}$. We have $|S_1(\mathbf{c})| = 1 + \binom{n}{1}(q-1)$ for all $\mathbf{c} \in C$. A straightforward computation shows that $|C| \cdot [1 + \binom{n}{1}(q-1)] = q^n$, showing that the code C is indeed perfect. □

How many perfect linear codes are there? We now provide a short list of perfect linear codes, with a brief explanation to follow the list as to why each of the listed codes is perfect.

(1) $\{(0, \ldots, 0)\}$.

(2) F_q^n.

(3) A binary repetition code of odd length.

(4) For $m \geq 2$, the q-ary Hamming codes $H(q, m)$, which have parameters $[(q^m - 1)/(q - 1), (q^m - 1)/(q - 1) - m, 3]$.

(5) The Golay code over F_3 which has parameters $[11, 6, 5]$.

(6) The Golay code over F_2 which has parameters $[23, 12, 7]$.

It is straightforward to verify that the codes listed here are all perfect. If C has only one codeword $(0, \ldots, 0)$, the code is perfect for trivial reasons. Similarly, if every vector in F_q^n is a codeword, then the code is perfect. The proof that binary repetition codes of odd length are perfect is left to the reader in Exercise 3.10. We have already shown that the Hamming codes $H(q, m)$ are perfect. The perfectness of the binary and ternary Golay codes can be verified easily using their parameters; see Exercise 3.12.

Rather than proving any details regarding the two Golay codes, we will simply describe their generator matrices. The ternary linear Golay code with parameters $[11, 6, 5]$ has a generator matrix of the form $G = (I_6|A)$, where

$$A = \begin{matrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 & 2 \\ 1 & 1 & 0 & 1 & 2 \\ 1 & 2 & 1 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 \\ 1 & 1 & 2 & 2 & 1 \end{matrix}.$$

The binary linear Golay code with parameters $[23, 12, 7]$ has a generator matrix of the form $G = (I_{12}|B)$, where

$$B = \begin{matrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{matrix}.$$

The first three codes, while perfect, are considered to be *trivial codes*. The following result is a culmination of a tremendous volume of research in algebraic coding theory, and represents one of the most fundamental results in the field; see Tietäväinen [63] or van Lint [38] for details.

Theorem 3.5.11 (van Lint and Tietäväinen). *A nontrivial perfect t-error correcting code must have parameters of the Hamming codes or the Golay codes, whose parameters are $[11, 6, 5]$ over F_3 and $[23, 12, 7]$ over F_2.*

In light of this theorem, our earlier list of perfect linear codes is complete.

It is known that any linear code with the same parameters as the q-ary Hamming codes is equivalent to a Hamming code; see Hill [23]. It is also known that for all prime powers q there are (nonlinear) codes that have the same parameters as the Hamming codes but are not equivalent to a Hamming code.

6. Codes and combinatorial designs

There are many connections between codes, in particular with codewords of particular weights in codes, and block designs. In this section we give a brief survey of some of these relationships.

Let C be a linear $[n, k]$ code over F_q. The *dual code* $C^\perp$ is the set of all vectors which are orthogonal to every vector of C:

$$C^\perp = \{\mathbf{u} \in F_q^n \mid \mathbf{u} \cdot \mathbf{c} = 0, \text{ for all } \mathbf{c} \in C\}.$$

Hence $C^\perp$ is a subspace complementary to C, justifying the notation. If C has generator matrix G and parity-check matrix H, then $C^\perp$ has generator matrix H and parity-check matrix G.

For $i = 0, 1, \ldots, n$, let A_i be the number of codewords in C with weight i. We know $A_0 = 1$ and $A_i = 0$ when $0 < i < d_C$. Let $A(x, y) = \sum_{i=0}^{n} A_i x^i y^{n-i}$. This polynomial is known as the *weight enumerator polynomial* for the code C. We let $A^\perp(x, y)$ denote the weight enumerator polynomial for $C^\perp$.

The MacWilliams identity is an extremely fundamental and important result in algebraic coding theory because it relates, in a simple way, the weight enumerator polynomial $A(x, y)$ of a linear code C to the weight enumerator polynomial $A^\perp(x, y)$ of its dual code $C^\perp$. While a proof of the following result is not beyond the scope of our text, we omit the proof and instead refer the reader to Lidl and Niederreiter [37].

Theorem 3.6.1 (MacWilliams identity). *Let C be a linear $[n, k]$ code over F_q. Let $C^\perp$ be the dual code of C, and let $A(x, y)$ and $A^\perp(x, y)$ be the weight enumerator polynomials of C and $C^\perp$, respectively. Then*

$$A^\perp(x, y) = q^{-k} A(y - x, y + (q - 1)x).$$

We will now briefly illustrate how codes can be used to construct various kinds of combinatorial designs. In Chapter 2 we discussed (v, k, λ) designs. We now extend that definition as follows:

Definition 3.6.2. A t-(v, k, λ) *design* is a set of v points and b blocks such that each point occurs in r blocks, each block has k points, and every t-tuple of points occurs in exactly λ blocks. A (v, k, λ) block design is thus a 2-(v, k, λ) design.

From our earlier discussion in Section 3 of Chapter 2, a 2-$(v, k, 1)$ design is a projective plane. If $\lambda = 1$, the design is called a *Steiner system*, and is denoted as $S(t, k, v)$. In this terminology, a projective plane is an $S(2, n+1, n^2+n+1)$ Steiner system.

We say that a codeword $\mathbf{c}$ in a code C *holds a block* B if B is the set of indices of the nonzero coordinates of the codeword $\mathbf{c}$.

Our next result illustrates how certain codes can be used to obtain combinatorial designs.

Theorem 3.6.3. *If C is a perfect binary $[n, k, d]$ code, then the codewords in C of weight d hold a Steiner system $S(t+1, d, n)$, where $t = (d-1)/2$.*

Example 3.6.4. Consider the $[7, 4, 3]$ binary Hamming code, with parity-check matrix

$$H = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

and generator matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

Recall that C is the span of the rows of G. Table 3.2 gives a complete list of the codewords and their weights.

The weight enumerator polynomial for the code C is thus seen to be

$$A(x, y) = y^7 + 7x^3y^4 + 7x^4y^3 + x^7$$

and the dual enumerator polynomial is

$$A^{\perp}(x, y) = 2^{-4}A(y - x, y + x) = y^7 + 7x^4y^3 .$$

The code $C^{\perp}$ is known as a *constant-weight code*, because every nonzero codeword has the same weight. We can, of course, determine this from the coefficients of $A^{\perp}$.

Rows	1	2	3	4	5	6	7	Weight	
	0	0	0	0	0	0	0	0	
R_1	1	0	0	0	1	1	0	3	*
R_2	0	1	0	0	0	1	1	3	*
R_3	0	0	1	0	1	1	1	4	
R_4	0	0	0	1	1	0	1	3	*
$R_1 + R_2$	1	1	0	0	1	0	1	4	
$R_1 + R_3$	1	0	1	0	0	0	1	3	*
$R_1 + R_4$	1	0	0	1	0	1	1	4	
$R_2 + R_3$	0	1	1	0	1	0	0	3	*
$R_2 + R_4$	0	1	0	1	1	1	0	4	
$R_3 + R_4$	0	0	1	1	0	1	0	3	*
$R_1 + R_2 + R_3$	1	1	1	0	0	1	0	4	
$R_1 + R_2 + R_4$	1	1	0	1	0	0	0	3	*
$R_1 + R_3 + R_4$	1	0	1	1	1	0	0	4	
$R_2 + R_3 + R_4$	0	1	1	1	0	0	1	4	
$R_1 + R_2 + R_3 + R_4$	1	1	1	1	1	1	1	7	

Table 3.2. Codewords in the [7,4,3] binary Hamming code.

Now consider the nonzero codewords of C of minimal weight (which is $d = 3$) which are noted in Table 3.2 with an *. The nonzero coordinates of these vectors determine the following system:

$$
\begin{array}{l}
1\,5\,6 \\
2\,6\,7 \\
4\,5\,7 \\
1\,3\,7 \\
2\,3\,5 \\
3\,4\,6 \\
1\,2\,4\,.
\end{array}
$$

As indicated in Example 2.4.4, this is a cyclic block design generated by $\{1, 5, 6\}$.

Corollary 3.6.5. *For $m \geq 3$, the vectors of weight 3 in the binary Hamming code $H(2, m)$ with parameters $[2^m - 2, 2^m - 1 - m, 3]$ hold a Steiner system $S(2, 3, 2^m - 1)$.*

Corollary 3.6.6. *The vectors of weight 7 in the binary Golay code* $[23, 12, 7]$ *hold a Steiner system* $S(4, 7, 23)$.

We now state a fundamental result of Assmus and Mattson [2] that can be used to locate combinatorial designs in certain codes. A proof is given by Pless [54].

Theorem 3.6.7 (Assmus and Mattson). *Let* C *be a binary linear* $[n, k, d]$ *code with* $0 < t < d$. *Let* B_i *be the number of codewords in* $C^{\perp}$ *of weight* i. *Let* $s = |\{i \mid B_i \neq 0 \text{ and } 0 < \cdots < n - t\}|$. *If* $s \leq d - t$, *then the codewords of weight* d *hold a* t*-design and the vectors of any weight* $i \leq n - t$ *such that* $B_i \neq 0$ *hold a* t*-design.*

7. Codes and latin squares

In this section we demonstrate some connections between codes and latin squares. We will show that certain optimal codes may be constructed from sets of orthogonal latin squares.

Some of the codes we consider may be nonlinear. We will denote a code C of length n with M codewords and minimum distance d by saying that C is an (n, M, d) code based upon q symbols, where now q is not required to be a prime power.

Let $A_q(n, d)$ be the maximum value of M for which there exists an (n, M, d) q-ary code (not necessarily linear) with M codewords. For linear codes, the maximum is $M = q^k$ in which case we would have a linear $[n, k, d]$ code over the field F_q.

Theorem 3.7.1. *For all* n *and* q (*not necessarily a prime power*) *there is a* q*-ary* $(n, q^2, n-1)$ *code if and only if there are* $n-2$ *MOLS of order* q.

Proof. Let $\{A^{(k)} \mid 1 \leq k \leq n-2\}$ be a set of $n-2$ MOLS of order q, and let $a_{i,j}^{(k)}$ denote the element in row i and column j of square $A^{(k)}$. Define C to be the set of all tuples $(i, j, a_{i,j}^{(1)}, \ldots, a_{i,j}^{(n-2)})$ where $0 \leq i, j < q$. The set C will be a code with the desired properties. Conversely, given a code C, we can obtain a set of MOLS from C in an analogous fashion. The details are left to Exercise 3.16. □

Example 3.7.2. We have the following set of 2 MOLS of order 3:

$$\begin{array}{ccc} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 2 & 0 & 1 \end{array} \qquad \begin{array}{ccc} 0 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & 2 & 0 \end{array}$$

with the resulting code C shown in the following table:

i	j	a_{ij}^1	a_{ij}^2
0	0	0	0
0	1	1	1
0	2	2	2
1	0	1	2
1	1	2	0
1	2	0	1
2	0	2	1
2	1	0	2
2	2	1	0

When q is a prime power we are able to construct a complete set of $q-1$ MOLS of order q, so we are able to state the following result.

Corollary 3.7.3. *Let q be a prime power and $n \leq q+1$. Then $A_q(n, n-1) = q^2$. Moreover, there is a q-ary MDS code $(n, q^2, n-1)$.*

We note that these codes yield an equality in the Singleton bound, and so the codes are optimal in that they have the largest possible minimum distance for any codes with the same length and number of codewords; that is, these codes are maximum distance separating MDS codes.

In Exercise 2.14, the reader was asked to construct a complete set of mutually orthogonal $F(q^i; q^{i-1})$ frequency squares. Using linear polynomials over F_q to construct those squares, we can also obtain a linear code, whose proof we leave as Exercise 3.13. For more details, see Dénes, Mullen, and Suchower [13].

Theorem 3.7.4. *For an integer $i \geq 1$ and any prime power q, there is a linear code over F_q with parameters $[(q^i-1)^2/(q-1), 2i, q^{2i-1} - 2q^{i-1}]$.*

Similarly, by using linear polynomials to construct a complete set of orthogonal hypercubes of dimension $d \geq 2$ and order q as in Exercise 2.15, we can obtain another class of linear codes over F_q. The proof is left to Exercise 3.14.

Theorem 3.7.5. *For any integer $d \geq 2$ and prime power q, there is a linear code over F_q with parameters $[(q^d-1)/(q-1)-d, d, q^{d-1}-d]$.*

For certain sets of parameters, the number of MDS codes is the same as the number of permutation hypercubes. In particular, the number of q-ary MDS $(n, q^{n-1}, 2)$ codes is the same as the number of $n-1$ dimensional hypercubes of order q and type $n-2$ (see Laywine and Mullen [31, p. 223]).

8. Notes

The book by MacWilliams and Sloane [41] is, without question, the primary reference on research related to the theory of error-correcting codes. Pless [54] is an upper division coding theory textbook which does a beautiful job of tying many coding theory ideas into the theory and construction of combinatorial designs. Chapter 8 of Lidl and Niederreiter [37] provides a nice introduction to algebraic coding theory over finite fields, while Chapter 13 of the book by Laywine and Mullen [31] provides some coding theory results and constructions which are developed via the use of sets of orthogonal latin squares and hypercubes. We also refer the reader to Hill [23] for a very readable treatment of algebraic coding theory which includes connections to combinatorial designs.

9. Exercises

3.1. Show that the number of vectors of length n over F_q in a sphere of radius t about a vector is

$$1 + \binom{n}{1}(q-1) + \cdots + \binom{n}{t}(q-1)^t.$$

Constructing codes.

3.2. Show that every linear $[n, k]$ code arises from some parity-check matrix.

3.3. Construct the parity-check and generator matrices for a linear ternary Hamming code $H(3,3)$.

3.4. Construct a 4-ary code of length $n = 4$ and distance $d = 3$ containing 16 codewords.

3.5. Construct a 5-ary code of length $n = 4$ and distance $d = 3$ containing 25 codewords.

Properties of linear codes.

3.6. If C is a linear code over F_q and $\mathbf{c}_1, \mathbf{c}_2 \in C$, show that $d(\mathbf{c}_1, \mathbf{c}_2) = \mathrm{wt}(\mathbf{c}_1 - \mathbf{c}_2)$ and thus the minimum distance d_C of C can be calculated as $d_C = \min\{\mathrm{wt}(\mathbf{c}) \mid \mathbf{c} \in C, \mathbf{c} \neq 0\}$. Explain how this can make the calculation of the linear code's minimum distance much simpler than just using the definition of the minimum distance d_C.

3.7. Show that if C is an $[n,k]$ linear code over F_q, then $C^{\perp}$ is an $[n, n-k]$ linear code over F_q.

3.8. How many binary cyclic codes of length $n = 15$ are there? Find generator polynomials for each of these codes and for each of their dual codes.

3.9. Working over F_q, let G be of the form $(I_k| - A^T)$, where A is of size $(n-k) \times n$. Let H be the matrix $(A|I_{n-k})$. Show that the code generated by H is the same as the code for which G is the parity-check matrix.

Perfect codes.

3.10. Show that a binary repetition code of odd length is perfect. Explain why a binary repetition code of even length is not perfect.

3.11. For a linear code C, show that the syndromes $S(\mathbf{y})$ and $S(\mathbf{z})$ are equal if and only if we have equality of the cosets $\mathbf{y} + C = \mathbf{z} + C$. If C has a parity-check matrix H, show that $S(\mathbf{y}) = S(\mathbf{z})$ if and only if $H(\mathbf{y} - \mathbf{z})^T = \mathbf{0}$.

3.12. Show that the 3-ary Golay $[11, 6, 5]$ linear code and the binary Golay $[23, 12, 7]$ linear codes are perfect.

Proofs left to the reader.

3.13. Complete the proof that the code constructed from the vectors that produce a complete set of mutually orthogonal $F(q^i; q^{i-1})$ frequency squares in Theorem 3.7.4 has parameters

$$\left[\frac{q^i - 1^2}{q-1}, 2i, q^{2i-1} - 2q^{i-1}\right].$$

3.14. Complete the proof that the code constructed from the vectors that produce a complete set of mutually orthogonal hypercubes in Theorem 3.7.5 has the parameters $[(q^d - 1)/(q-1) - d, d, q^{d-1} - d]$.

3.15. Prove Lemma 3.2.4.

3.16. Give a complete proof of Theorem 3.7.1.

http://dx.doi.org/10.1090/stml/041/04

Chapter 4

Cryptography

In the first part of this chapter we discuss *cryptography*, the field of mathematics and computer science concerned with preventing anyone except the desired recipient from understanding a transmitted message. As in coding theory, the person sending the message replaces the true message with an encoded form. Unlike coding theory, the correspondence between original and encoded messages is only known to the sender and recipient of the message.

The last part of this chapter discusses *threshold schemes*, which are used to split a piece of secret information among several individuals so that a predetermined number of those individuals must collude in order to reconstruct the secret information. These schemes can be used to store corporate secrets, launch codes for nuclear arms, or other information which is thought to be too important to trust to a single individual.

We present cryptography from a purely mathematical point of view. When cryptography is used in the real world, nonmathematical implementation issues can arise. We do not consider such issues here, but refer the reader to the book by Ferguson and Schneier [16] that discusses these issues thoroughly.

It is important to remember that the aim of cryptography is not to send messages without errors. We assume, in fact, that no errors at all occur during transmission. Thus we assume that the transmitted

message is available to both the desired recipient and to those trying to break the system. In order to prevent or detect errors during transmission, the methods of coding theory (as described in Chapter 3) must be used.

1. Introduction to cryptography

The problem underlying cryptography is to transmit a message so that only the intended recipient is able to understand it. The general idea is that the true message (the *plaintext*) is replaced with another message (the corresponding *ciphertext*) before transmission. The desired recipient will be able to transform the ciphertext back to its original plaintext form. These transformations often require another parameter, known as a *key*. This technique allows the same general transformation method to be used repeatedly, with different keys, without compromising security if one key is discovered.

Example 4.1.1. A very simple cryptosystem for English words is the *substitution cipher*. We begin with a permutation of the letters $A, B, C, D, \ldots, Z$; this permutation serves as the key. In this example, we will use the following permutation.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
F G H I J K L M N O P Q R S T U V W X Y Z A B C D E

To encrypt a word, replace each letter using the permutation. To decrypt, use the inverse permutation. Thus the encrypted version of GALOIS is LFQTNX and the encrypted message KNJQI decrypts to FIELD.

A simple substitution cipher such as the one just described provides little security, but this is not because it is easy to try all the possible permutations. There are $26! \approx 4 \cdot 10^{26}$ permutations of the letters A–Z, and thus it would take over $3 \cdot 10^{12}$ years to test all the permutations at a rate of one million per second. On average, a search like this is expected to test about half of the permutations, which gives the search an expected running time of approximately $1.5 \cdot 10^{12}$ years. This is about 100 times the age of the Earth, which is estimated to be under $5 \cdot 10^9$ years. Thus a blind search for the

correct permutation is unlikely to succeed in a reasonable amount of time.

The weakness of the substitution cipher is that an analysis of the encrypted message based on the frequency with which each letter occurs, and the frequency with which each pattern occurs, allows the set of likely permutations to be reduced until an exhaustive search is possible. Such an analysis may be done by hand. There is, in fact, a series of puzzle books consisting entirely of famous quotations encrypted by substitution ciphers; the goal of each puzzle is to break the code to find the quotation. We provide an example of such a puzzle in Exercise 4.1. This weakness in the substitution cipher shows that the question of whether a cryptosystem is secure depends on much more than just the number of possible keys.

The substitution cipher in Example 4.1.1 is a basic example of a cryptosystem. In order to be more precise, we give the following definition.

Definition 4.1.2. A *cryptosystem* is a pair of functions E, D, each of which takes two arguments: a message M and a key K. For each key K_E there must be a key K_D such that $D(E(M, K_E), K_D) = M$ for every message M. The key K_E is the *encryption key* while K_D is the corresponding *decryption key*.

Although this definition is intentionally abstract, in practice it will always be clear what the messages are, what the keys are, and what the encryption and decryption functions are.

A cryptosystem D, E is said to be *secure* if it is not computationally feasible to determine the true message M from an encrypted message $E(M, K_E)$. This is not a formal definition; a cryptosystem may be secure only because of the lack of computational power in contemporary computers. Indeed, several cryptosystems which were considered secure in the 1970s are now thought to be insecure. One such system is the Data Encryption Standard (DES), which was developed in the 1970s and is now obsolete because an exhaustive search of the all possible keys has become feasible. Landau [30] gives a full description of DES and the methods that were developed to try to break it.

2. Symmetric key cryptography

A cryptosystem is called a *symmetric key* system if the keys K_D and K_E are always equal. A substitution cipher based on a self-inverse permutation is an example of such a system. There are many well-known symmetric key cryptosystems. These are used, for example, to encrypt transmission by automatic teller machines over public phone lines. Symmetric key cryptosystems tend to be designed for implementation in custom hardware with applications in which the ability to encrypt and decrypt information at a high rate is crucial.

The drawback of symmetric key cryptosystems is that the key must somehow be known to both parties before communication begins. This is reasonable for an automated teller machine which can be programmed with an encryption key when it is installed, but it is not reasonable for individuals who cannot meet but wish to communicate over a long distance. We will not study symmetric key cryptosystems in depth in this book. In Section 3.3 we will show how public key cryptography can be used to securely transmit a shared key over a public channel.

We present two examples of symmetric key cryptosystems. The first example, known as the Advanced Encryption Standard (AES) cryptosystem, uses finite fields. The second example is a cryptosystem based on latin squares.

Example 4.2.1. The Advanced Encryption Standard (AES) cryptosystem is widely used in contemporary applications. It was the winner of a competition sponsored by the government of the United States to produce a strong symmetric key cryptosystem. This cryptosystem is also known as Rijndael; it received the name AES after winning the competition. The AES cryptosystem and its development have been fully documented by its designers Daemen and Rijmen [9].

Like many symmetric key cryptosystems, the AES cryptosystem operates on blocks, which are fixed length sequences of binary digits (bits). The cryptosystem acts by breaking the message into blocks and replacing each input block by an output block of the same length. Thus the cryptosystem can be viewed as a permutation of the set of blocks, parameterized by a key. The overall permutation is obtained

by taking a simpler permutation and iterating it several times. Each of these iterations is called a round; each round has a different key, and each of these round keys is computed from the overall encryption key.

We now give an extremely brief description of how the AES cryptosystem works, while refering the reader to Daemen and Rijmen [9] for details. The blocks in AES are 128 bits long; these blocks may be viewed as vectors of length 128 over F_2. Each round of AES consists of three steps: a key application step, a nonlinear step and a linear step. Each of these steps replaces the block with a new block. Each AES round key is 128 bits long, and the key application step in each round consists of adding the block and the round key as vectors over F_2 (this is often called an XOR operation). The nonlinear step in each round proceeds by breaking the 128 bit block into sixteen bytes (8 bit units), considering each byte as an element of F_{2^8}, and replacing each element with its multiplicative inverse. The linear step in each round uses a fixed bijective linear map from F_2^{128} to itself. The number of rounds in AES is determined by the length of the overall key; with longer keys, more rounds can be performed, which is believed to strengthen the security of the cryptosystem. It follows from our description that an algebraic formula for AES (and a fixed key length) can be obtained; an explicit formula is given by Ferguson, Schroeppel, and Whiting [17].

Example 4.2.2. We describe a simple symmetric key cryptosystem using pairs of orthogonal latin squares. The security of this cryptosystem has, apparently, not been investigated.

Let L_1 and L_2 be two MOLS of order n based on $0, 1, \ldots, n-1$. We will identify our messages with pairs (i, j) with $0 \leq i, j < n$. To encrypt a message (i, j), find the unique location (k, l) where the pair (i, j) occurs when L_1 and L_2 are superimposed. The encrypted message is (k, l). To decrypt (k, l), superimpose the squares and read the pair (i, j) that occurs at location (k, l) in the superimposed square.

An interesting property of this cryptosystem is that the key can be canonically split into two halves, L_1 and L_2. Possession of just one half does not allow the message to be decrypted (although knowledge of L_1, for example, might make an exhaustive search for L_2 feasible).

We will discuss this property in the section on threshold schemes below.

3. Public key cryptography

Definition 4.3.1. A *public key cryptosystem* is a cryptosystem in which a recipient computes a pair of keys K_E, K_D and announces the key K_E (which is called the *public key*) to all potential senders. The *private key* K_D is kept secret. Thus anyone who possesses the public key can encrypt a message for the recipient, but (if the system is secure) only someone who possesses the private key can decrypt the message.

Public key cryptosystems are in wide use at the beginning of the twenty-first century. One important use is in secure communication on the internet.

Note that in order for a public key cryptosystem to be secure, it must not be computationally feasible to compute the private key K_D from the public key K_E; doing so is called a *direct attack* on the cryptosystem. It must also not be feasible to compute M from $E(M, K_E)$ without computing K_D; this is called an *indirect attack.*

We now present two public key cryptosystems: the RSA cryptosystem and the double round quadratic cipher. Each of these cryptosystems is notable for its use of algebraic methods. We will also discuss public key cryptosystems based on elliptic curves and on Dickson polynomials later in the chapter.

3.1. The RSA cryptosystem. RSA is perhaps the most well studied and widely used public key cryptosystem in the early twenty-first century. It was developed in 1978 and named for its inventors: Rivest, Shamir, and Adleman [55].

We will define the RSA cryptosystem by describing how it works. We identify the set of possible messages with integers $0, 1, 2, \ldots, N-1$, so that there are N distinct messages. Let $n = pq$ be a fixed integer, where p and q are large primes. The number n will be announced publicly but the prime factors p, q will be kept private.

Now choose $k < n$ such that $(k, \phi(n)) = 1$, where ϕ denotes Euler's function. Note that $\phi(n) = \phi(pq) = \phi(p)\phi(q) = (p-1)(q-1)$ is easy to compute with knowledge of p and q but believed to be difficult to compute without a factorization of n. Now choose d such that $kd \equiv 1 \pmod{\phi(n)}$. This inverse will exist because k and $\phi(n)$ are relatively prime.

The public key for this cryptosystem is the pair (n, k) and the encryption function is

$$E(M, (n, k)) \equiv M^k \pmod{n},$$

where M denotes a message. The private key is the pair (n, d) and the decryption map is

$$D(C, (n, d)) \equiv C^d \pmod{n},$$

where C denotes a received ciphertext. We leave it as Exercise 4.4 to show that

$$(M^k)^d \equiv M^{kd} \equiv M^1 = M \pmod{n},$$

and thus the decryption function is indeed the inverse of the encryption function.

The security of this cryptosystem rests on the difficulty of the factoring problem: "Given an integer n of the form pq, where p and q are distinct primes, find p and q." The ability to solve this problem efficiently would allow an attacker to determine a private key (n, d) from a public key (n, k). In particular, potential advances in quantum computing could make factoring extremely efficient, which could make the RSA cryptosystem obsolete.

We note that our description of the RSA cryptosystem, while formally correct, ignores many practical issues that would be faced in an implementation of the algorithm. These practical issues include: how to turn a computer document into a sequence of message units, how to choose the number $n = pq$, and how to choose the public key k. A poor implementation might make insecure choices that allow the implementation to be broken.

One implementation detail crucial to the RSA cryptosystem is known as *padding*. The design of the RSA algorithm is such that the messages represented by 0 and 1 will always be enciphered as

themselves; this allows the attacker to recover partial information about the plaintext message from the ciphertext. Also, if k is small and M is small, then it is possible that $M^k < n$, in which case the modular arithmetic gives the same result as nonmodular arithmetic. To avoid these problems, secure implementations of RSA use only a subset of the possible values to encode messages. These values are then *padded* with random bits so that no small message numbers are ever passed into the encryption algorithm. The decryption algorithm is modified to discard the padding, leaving the correct message.

An example that illustrates the RSA cryptosystem is given in Exercise 4.5.

3.2. Double-round quadratic enciphering. We now present a public key cryptosystem, known as the *double-round quadratic cipher*, which is based on linear algebra and finite fields. This cryptosystem is presented in detail by Koblitz [27]. The public key consists of a finite field F_{q^n}, a basis for this field over F_q, and n quartic polynomials in n variables over F_q (here, *quartic* means that the sum of the exponents of the variables in any term will be no more than 4). These public polynomials allow the sender to compute the encryption function $E\colon F_q^n \to F_q^n$. The private key is, as we will see, a formula closely related to the inverse of this function.

To form the cryptosystem, choose a prime power q such that $q \equiv 3 \pmod 4$, and choose an odd integer $n \geq 1$. We will see later why these restrictions on q and n are necessary. Let $\{b_1, \ldots, b_n\}$ be a basis for F_{q^n} over F_q; this basis will be made public. Then choose three secret invertible $n \times n$ matrices A, B, and C over F_q. These determine linear maps f_A, f_B, and f_C from F_{q^n} to itself, using the vector space isomorphism $F_{q^n} \cong F_q^n$ induced by the basis B. Let g be the function $x \mapsto x^2$ from F_{q^n} to itself.

The encryption map is the composite function

$$E = f_C \circ g \circ f_B \circ g \circ f_A.$$

The public key consists of a certain description of this function. By viewing E as a nonlinear map from F_q^n to itself, we can decompose E coordinatewise relative to the basis B. Each coordinate function of E will be given by a fourth-degree polynomial with coefficients in

F_q. The public key consists of these coordinate polynomials. This decomposition is explicitly demonstrated in Example 4.3.2 below.

We are now ready to define the set of messages for this cryptosystem. We will consider $x, y \in F_{q^n}$ to be equivalent if $x = \pm y$; the set of messages is the set of equivalence classes of F_{q^n}. Because $q \equiv 3 \pmod 4$ and n is odd, for every $x \in F_{q^n}$, exactly one of x or $-x$ is in the range of the squaring map g (see Exercise 4.3). Thus g may be regarded as a bijection from M to M, for $(-x)^2 = (x^2)$. Moreover, the linear maps f_A, f_B, f_C all respect the equivalence relation (because, for example, $f_A(-x) = -f_A(x)$). Thus we may view E as a well-defined map on the set of messages.

To decrypt a message $E(M)$, we use the fact that the map

$$g^{-1}\colon x \mapsto x^{(q^n+1)/4}$$

is (up to a factor of ± 1) the inverse of the squaring map g. This is because

$$(x^2)^{(q^n+1)/4} = x^{(q^n+1/2)} = x^{(q^n-1)/2} \cdot x = \pm 1 \cdot x.$$

Note that n must be odd so that $q^n + 1$ is divisible by 4.

The decryption function is

$$D = f_A^{-1} \circ g^{-1} \circ f_B^{-1} \circ g^{-1} \circ f_C^{-1}.$$

It is straightforward to compute the inverse maps f_A^{-1}, f_B^{-1}, and f_C^{-1} from A, B, and C, respectively.

Example 4.3.2. We give a simple example of the double-round quadratic cipher. Let $q = 3$ and $n = 3$; we construct F_{3^3} as the splitting field of the irreducible polynomial $x^3 + 2x^2 + 1$ over F_3. Let α be a root of this polynomial in F_{3^3}, so that $B = \{1, \alpha, \alpha^2\}$ is a polynomial basis of F_{3^3} over F_3. We view each element $x \in F_{3^3}$ as a vector (x_0, x_1, x_2) in F_3^3 using the basis B.

We choose the following matrices for A, B, and C:

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 2 \\ 0 & 1 & 0 \\ 2 & 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This gives us the formulas

$$\begin{aligned} f_A(x_0, x_1, x_2) &= (x_1, x_0, x_2), \\ f_B(x_0, x_1, x_2) &= (2x_2, x_1, 2x_0), \\ f_C(x_0, x_1, x_2) &= (x_1 + x_2, x_0, x_2). \end{aligned}$$

We compute the formula for the squaring map $g\colon x \mapsto x^2$ relative to our basis:

$$\begin{aligned} g(x_0, x_1, x_2) &= x_0^2 + 2x_0x_1\alpha + (x_1^2 + 2x_0x_2)\alpha^2 + 2x_1x_2\alpha^3 + x_2^2\alpha^4 \\ &= (x_0^2 + x_1x_2 + 2x_2^2) + (2x_0x_1 - x_2^2)\alpha \\ &\quad + (x_1^2 + 2x_0x_2 + x_1x_2 + x_2^2)\alpha^2. \end{aligned}$$

Recall that $E = f_C \circ g \circ f_B \circ g \circ f_A$. Write $E(x_0, x_1, x_2) = (y_0, y_1, y_2)$. A simple but long calculation in F_3^3 gives

$$\begin{aligned} y_0 &= 2x_0^3x_1 + x_0x_1^3 + 2x_0^3x_2 + x_1^3x_2 + 2x_0^2x_2^2 + x_0x_1x_2^2, \\ y_1 &= x_0^4 + x_0x_1^3 + 2x_1^4 + 2x_1^3x_2 + 2x_0^2x_1x_2 \\ &\quad + 2x_0^2x_2^2 + x_1^2x_2^2 + 2x_0x_2^3 + x_1x_2^3 + 2x_2^4, \\ y_2 &= x_0x_1^3 + x_1^4 + 2x_0^3x_2 + x_0^2x_1x_2 + x_1^3x_2 + \\ &\quad x_0^2x_2^2 + 2x_xx_1x_2^2 + x_1^2x_2^2 + 2x_0x_2^3 + 2x_1x_2^3 + 2x_2^4. \end{aligned}$$

The public key consists of these three polynomials. To encrypt a message (x_0, x_1, x_2), we evaluate the polynomials at (x_0, x_1, x_2) to obtain the ciphertext (y_0, y_1, y_2). Exercise 4.7 asks for the formulas of the corresponding decryption function.

In order to break the double-round quadratic cipher, it would be sufficient to determine A, B, and C from the public polynomials. This seems to involve decomposing the multivariate fourth-degree polynomials as a composition of polynomials of lower degree. There is no known method to accomplish this decomposition efficiently, and the double-round quadratic cipher is believed to be secure. One reason that the RSA cipher is more widely used is because its security is better established.

A single-round quadratic cipher could be defined in which the encryption map is of the form $f_B \circ g \circ f_A$. Such a cryptosystem is known to be insecure; see Koblitz [27].

3.3. Key exchange, the Diffie-Hellman system, and discrete logarithms. When only one small message needs to be transmitted, the rate at which data is sent is often of secondary importance. If many large messages must be exchanged, however, the speed of transmission becomes crucial. Public key cryptosystems tend to be computationally inefficient because they perform modular arithmetic with very large moduli. For example, the moduli used in the RSA cryptosystem may be of magnitude 10^{600} or larger in contemporary applications. Such numbers are too large to be directly manipulated by computer processors, so specialized software is used to perform these computations. Because symmetric key systems use numbers that can be directly manipulated by computer processors, they are often much faster than public key cryptosystems. In some cases, symmetric key encryption is performed using hardware specifically designed for the cryptosystem employed, which further increases the encryption speed. One of the criteria used in selecting AES from other possible cryptosystems was the existence of efficient implementations for desktop machines, custom encryption processors, and embedded processors on smart cards.

The disadvantage of symmetric key cryptosystems is that the key must be known to both parties before the cryptosystem can be used. It is obvious that the key cannot be transmitted unprotected from one party to another. The method of *key exchange* allows two parties to use public key cryptography to obtain a shared key which is then used for a symmetric key cryptosystem.

Example 4.3.3. We describe a simple key exchange system that can be implemented with any public key cryptosystem. In order to communicate with user B, user A chooses a random encryption key K. User A then encrypts K with user B's public key and sends this encrypted message to user B, who decrypts K with his private key. Now users A and B may use a symmetric key cryptosystem for further communication. The key K has only been sent in encrypted form; as long as the public key cryptosystem used to transmit K is secure, and the symmetric key cryptosystem is secure, the overall communication will be secure.

Note that in this "exchange" system the key was sent in only one direction. Thus although user A can be confident that only user B can decrypt the key, user B cannot be sure that it was user A who sent it. The Diffie–Hellman key exchange overcomes this problem by having each user send an encrypted key to the other.

Diffie and Hellman [14] proposed a method for key exchange using finite fields. To set up the system, users A and B first agree on a prime power q and a primitive element g of F_q, which they assume are publicly known. Users A and B choose secret numbers a and b, respectively. User A transmits g^a to user B, who computes $(g^a)^b = g^{ab}$. User B transmits g^b to user A, who computes $(g^b)^a = g^{ba}$. The users thus agree on the value of $g^{ab} = g^{ba}$, and they can use it as a common key for another cryptosystem. Note that user A cannot easily compute b from g^b, user B cannot easily compute a from g^a, and a third party cannot easily compute a or b.

The Diffie–Hellman method can also be used as a public key cryptosystem, as we now explain. Fix a prime power q and let messages be elements of F_q. To set up the cryptosystem, user A chooses a primitive element g of F_q and some secret value $2 \leq a \leq q-2$. User A's public key is the tuple (q, g, g^a); the private key is the value of a. To send a message m to user A, user B chooses a secret value $2 \leq b \leq q-2$ and sends the pair $(g^{ab}m, g^b)$ to user A. User B can compute $g^{ab} = (g^a)^b$ without knowing a. User A, who knows the value of a, computes $g^{ab} = (g^b)^a$ and then computes $m = (g^{ab}m)/g^{ab}$.

A potential attacker of the Diffie–Hellman key exchange will know q, g, g^a, and g^b. An attacker of the Diffie–Hellman cryptosystem has similar information. It is clear that if the attacker could compute a from the pair (g, g^a) then the cryptosystems would both be vulnerable. No efficient means of performing this computation are known.

Definition 4.3.4 (The discrete logarithm). Let g be a primitive element of F_q. The *discrete logarithm function* is the unique function $\log_g : F_q^* \to \{0, 1, \ldots, q-2\}$ which makes the equation

$$a = g^{\log_g(a)}$$

hold for every $a \in F_q^*$.

The function $\log_g$ is thought to be very difficult to compute, although its inverse $a \mapsto g^a$ is very easy to compute. If an efficient method for computing discrete logarithms were discovered, the Diffie–Hellman key exchange (and the Diffie–Hellman cryptosystem) would no longer be secure. It is known that computing discrete logarithms in F_q is of about the same level of computational difficulty as factoring an RSA modulus $n = pr$ when q and n are of comparable size.

Several formulas for $\log_g$ have been established, but they are not computationally feasible. We state one such formula in the next theorem, which is proved in Mullen and White [49].

Theorem 4.3.5. *Let p be a prime and let g be a primitive element of F_p. The following formula for the discrete* $\log_g$ *in F_p holds:*

$$\log_g(a) = -1 + \sum_{j=1}^{p-2} \frac{a^j}{g^{-j} - 1}. \tag{3}$$

3.4. Elliptic curves and elliptic curve cryptography. In this section we briefly discuss elliptic curves over finite fields and describe their use in modern cryptosystems. An *elliptic curve* over a finite field is defined using an equation of the form

$$E\colon y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6,$$

where $a_1, a_2, a_3, a_4, a_6 \in F_q$. The elliptic curve consists of all the points (x, y) in F_q^2 satisfying the equation together with another point $\mathcal{O}$ not in F_q^2 called the *point at infinity*. The solutions in F_q^2 are called the *F_q-rational points*, or just *rational points* when the field is clear from the context. In order to define an elliptic curve, the equation must be *smooth*, which means that it is impossible to find values for x and y in an algebraic extension of F_q such that (x, y) is a point on the curve where the partial derivatives

$$2y + a_1x + a_3,$$
$$x^3 + 2a_2x + a_4 - a_1y,$$

are simultaneously zero.

To simplify the theory, it is common to simplify the defining equation of an elliptic curve by making variable substitutions on the

variables. If the characteristic of the field is not 2, then we can reduce E to the curve E' defined by

$$y^2 = x^3 + b_2x^2 + b_4x + b_6$$

via the substitution $y \mapsto y - a_1x/2 - a_3/2$ (we leave the verification to the reader). Furthermore, if the characteristic of F_q is not 2 or 3, then by using the substitutions $x \mapsto (x - 3b_2)/36, y \mapsto y/216$ we can further reduce the curve to the form

$$y^2 = x^3 + ax + b, \tag{4}$$

where $a, b \in F_q$. From here forward we will assume that our elliptic curve is of this form and that it is defined over a field F_q whose characteristic is not 2 or 3. Elliptic curves can be studied over fields of characteristic 2 or 3, but the theory is more complicated because substitutions such as those above cannot be performed. In the cases we are interested in, when the characteristic is greater than 3, an elliptic curve in the form of (4) is smooth if and only if its *discriminant* $-16(4a^3 + 27b^2)$ is nonzero.

We now define a notion of addition for the points on an elliptic curve.

Definition 4.3.6. Assume that $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ are points on an elliptic curve in form (4). Then we define $P+Q = (x_3, y_3)$ where

$$x_3 = \lambda^2 - x_1 - x_2, y_3 = \lambda(x_1 - x_3) - y_1,$$

with

$$\lambda = \frac{y_2 - y_1}{x_2 - x_1} \quad \text{if } P \neq Q,$$

$$\lambda = \frac{3x_1^2 + a}{2y_1} \quad \text{if } P = Q.$$

We leave it as a challenging exercise to show that, with the operation just defined, the set of points on an elliptic curve forms an Abelian group (see Exercise 4.16). Note that $\mathcal{O} = (\infty, \infty)$ serves as the identity of the group, and for each point P we have $-P = (x_1, -y_1)$.

Example 4.3.7. Consider the elliptic curve E: $y^2 = x^3+x+6$ defined over the field F_{11}. Then $(2,4) + (3,5) = (7,2)$, $(2,4) + (2,4) = (5,9)$, and $(2,4) + (2,7) = \mathcal{O}$.

Since the group of rational points of an elliptic curve is a finite group under the addition operation defined above, each point in the group must also have finite order. Let the order of the group be n. The *elliptic curve discrete logarithm problem* is the following: Given two points P, Q in the group of rational points, determine the unique integer i with $0 \leq i \leq n-1$ such that

$$Q = iP = \underbrace{P + \cdots + P}_{i \text{ times}},$$

provided that such an integer exists.

We now turn to the cryptographic uses of elliptic curves. These have practical appeal because the addition operation for rational points of an elliptic curve involves only a few arithmetical operations in the underlying field F_q and so is easy to compute. Moreover, as indicated by Menezes [44, p. 13], the elliptic curve discrete logarithm problem is believed to be more difficult to solve than the discrete logarithm problem in finite fields of approximately the same size. This provides excellent motivation for the use of elliptic curve cryptosystems over Diffie–Hellman type cryptosystems.

Example 4.3.8. One elliptic-curve cryptosystem can be defined in a manner similar to the Diffie–Hellman system. A recipient chooses a finite field F_q, an elliptic curve E over the finite field, and a rational point P on the curve. The messages are the rational points of the elliptic curve.

To make a key pair, the recipient chooses a secret value a between 0 and the order of P. The recipient's public key contains the information about the field and curve together with values of P and aP. The private key is the value of a.

To send a message M to this recipient, a sender chooses a value b and computes bP and $b(aP) + M$. These values are transmitted to the recipient, who computes $a(bP) = b(aP)$ and then

$$M = (b(aP) + M) - (b(aP)).$$

Another advantage of elliptic curve systems is that each user may use a different elliptic curve while all users use the same underlying field. Thus each user has the ability to choose a new elliptic curve from

time to time without changing the field; this allows implementations to be optimized for working with a particular field.

3.5. Digital signatures. Electronic contracts are far more vulnerable to manipulation after adoption than printed contracts. This situation is underlined by the fact that there is no physical signature proving that an individual or company accepted the contract. In a legal context, it is desirable to have proof that an individual or corporation truly did accept a particular electronic contract and that the wording of this contract has not changed since it was accepted. Such proof can be constructed via digital signature schemes.

Definition 4.3.9. A *digital signature scheme* consists of two functions S and C. The signature function S takes a message M and returns another message $S(M)$, called the *signature*. The check function C takes a message M and a purported signature s. If $s = S(M)$, the check function returns 1, and otherwise the check function returns 0. The check function is made public, but the signature function is kept private. In order for the system to be secure, it must be infeasible to compute signatures with knowledge of the check function alone.

We now give an example of how signature schemes are used on the Internet. Before sending personal information, a user wishes to verify that a website is authentic (not an imposter). To facilitate this process, the website publishes a check function C that is known to the user's web browser. (In reality, a fixed check function with extra parameters is used for all websites, and a particular website only publishes values of the extra parameters.) To perform the verification, the user's web browser chooses a message M and transmits it to the website. The website computes $S(M)$ and sends the signature back to the user's browser. The browser receives a purported signature s and checks whether $C(M, s) = 1$. If the equation holds, then the browser will tell the user the website is authenticated. In order to impersonate the website, it would be necessary to know the website's (private) signature function.

Many public key cryptosystems can be used "backwards" to implement digital signature schemes. First, a user creates keys K_E and

K_D for a public key cryptosystem with encryption function E and decryption function D. The user publicizes the decryption key K_D and keeps the encryption key K_E secret. The signature scheme is defined by letting $S(M) = E(M, K_E)$ and letting $C(M, s) = 1$ if and only if $D(s, K_D) = M$. This scheme will be secure if it is not possible to obtain the encryption key from knowledge of the decryption key. The RSA cryptosystem, in particular, can be used for this purpose (see Exercise 4.8).

Another digital signature scheme can be obtained using finite fields. We identify our messages with elements of F_p, for p a fixed prime. Let g be a primitive element of F_p and let $h < p$ be a fixed secret integer. The signature for a message m is a tuple (g^h, r, s) such that

$$g^m \equiv (g^h)^r r^s \pmod{p}.$$

An individual will publish the prime p, a primitive element g of F_p, and another element $c = g^h$ for some secret number h. The signature assigned to m is a pair (r, s) such that $g^m = c^r r^s \pmod{p}$, where $r = g^k$ with $(k, p-1) = 1$. To see that s can be found, note that $g^m = (g^h)^r (g^k)^s = g^{hr} g^{ks} = g^{hr+ks}$, so $m = hr + ks \pmod{p-1}$, and thus $s = k^{-1}(m - hr) \pmod{p-1}$.

3.6. Dickson cryptosystems. Our final example of a cryptosystem and signature scheme is based on Dickson polynomials. In Section 6.4 of Chapter 1, we discussed a few properties of these polynomials, which are defined for a parameter $a \in F_q$ by

$$D_n(x, a) = \sum_{i=0}^{\lfloor n/2 \rfloor} \frac{n}{n-i} \binom{n-i}{i} (-a)^i x^{n-2i}.$$

In particular, we showed that the polynomial $D_n(x, a)$ induces a permutation of F_q if and only if $(n, q^2 - 1) = 1$. In this section we briefly discuss a few cryptographic applications of Dickson polynomials. We will not discuss the security of these systems. We will also omit the proofs; the reader may refer to Lidl, Mullen, and Turnwald [35, Section 7.1] for further details.

We first describe a Dickson polynomial analogue of the RSA cryptosystem. It is known (see Lidl, Mullen, and Turnwald [35, p.

156]) that the Dickson polynomial $D_n(x,1)$ induces a permutation of the ring $\mathbb{Z}_n$ with $n = p_1^{e_1} \cdots p_r^{e_r}$ if and only if $(n, v(n)) = 1$ where $v(n) = \operatorname{lcm}\{p_i^{e_i-1}(p_i^2 - 1)\}$. Moreover, the inverse permutation is given by the Dickson polynomial $D_m(x,1)$ where $nm \equiv 1 \pmod{v(n)}$. Thus if $n = pq$ with p and q both large primes, we have a system analogous to the original RSA system. One may also replace the value $a = 1$ by $a = -1$ with similar results.

Dickson polynomials can also be used to obtain signatures. The map $x \to x^k$ in the RSA system is replaced by the map $x \to D_k(x,a)$ where $a = \pm 1$. For example if user A wants to sign a message m, then user A computes $D_{d_A}(m,a) \equiv s \pmod{n}$ which is sent to user B. Then user B computes $D_{e_A}(s,a) \equiv D_{e_A}(D_{d_A}(m,a),a) \equiv m \pmod{n}$. Thus the receiver knows that the message came from user A, because only user A knows the private key d_A. In addition, user A could compute $D_{e_B}(s,a)$ and send this to user B, who then calculates $D_{e_A}(s,a) \equiv m \pmod{n}$.

We now illustrate how Dickson polynomials can be used for key exchange. Let q be a prime power which is made public, and choose another public element $x_0 = \gamma^{q-1} + \gamma^{-(q-1)}$ where γ is a primitive element in the field F_{q^2}. Users A and B choose secret integers a and b. They then publish the public values $D_a(x_0,1)$ and $D_b(x_0,1)$. User A computes $D_a(D_b(x_0,1),1) = D_{ab}(x_0,1)$ and similarly B computes $D_b(D_a(x_0,1),1) = D_{ba}(x_0,1)$. Users A and B have thus established the common key $D_{ab}(x_0,1) = D_{ba}(x_0,1)$.

As a final example of the use of Dickson polynomials in cryptography using a finite field F_q, we illustrate the construction of a symmetric key cryptosystem. Let $(k, q^2 - 1) = 1$. One enciphers a message m as $D_k(m,a) = c$; the deciphering is accomplished by calculating $D_l(c,a)$ where $kl \equiv 1 \pmod{q^2 - 1}$.

Lidl, Mullen, and Turnwald [35, pp. 160–161] discuss possible attacks on Dickson cryptosystems.

4. Threshold schemes

Suppose there is a secret piece of information s. A (k,t)-*threshold scheme* (or (k,t) *secret-sharing scheme*) is a system of k objects,

called *shares*, and a numeric threshold $t \leq k$ such that any collection of t shares allows s to be computed, but no collection of fewer than t shares allows s to be computed. In many applications, we would like to know that any collection of fewer than t shares gives absolutely no information about s.

Example 4.4.1. We present a threshold scheme based on the Lagrange interpolation formula (Theorem 1.6.1). Let $k(x) = b_0 + b_1 x + \cdots + b_{t-1}x^{t-1}$ be a polynomial in $F_q[x]$ of degree at most $t-1$. The secret s is the coefficient b_0. Choose nonzero elements $c_1, \ldots, c_n \in F_q$. The shares are pairs of the form $(c_i, k(c_i))$.

Given t pairs of the form $(c_i, k(c_i))$, with each $c_i \neq 0$, there is a unique polynomial function of degree at most $t-1$ which contains all these pairs; this polynomial can be obtained by Lagrange interpolation. But given $t-1$ pairs, for every pair $(0, c)$ there is a polynomial function containing all $t-1$ pairs and containing $(0, c)$. Hence no collection of $t-1$ shares gives any information about the secret.

We now present a second secret sharing scheme that uses latin squares.

Definition 4.4.2. A *critical set* for a latin square of order n is a collection C of triples $\{(i, j, n_{ij})\} \subseteq \{1, 2, \ldots, n\}^3$ such that (1) there is a unique latin square S of order n such that $S(i, j) = n_{ij}$ for all (i, j, n_{ij}) in the collection C, and (2) no proper subset of C has property (1).

To construct the secret sharing scheme, we identify the secret information with a fixed latin square L. We then find a critical set C for L, and give each user one element of the critical set.

Example 4.4.3. Let L be the following latin square.

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{array}$$

The set $C = \{(1,1,1), (1,2,2), (2,4,3), (3,2,4), (4,3,2)\}$ is a critical set for L, as the reader may verify.

Example 4.4.4. Let M be the latin square

$$\begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{array}$$

and let $S = \{(2,1,2),(3,2,1),(1,3,3)\}$. We claim that if M is the secret, then S is a $(3,2)$ secret-sharing scheme. Direct computation can be used to show that any two-element subset of S extends uniquely to the latin square M. Clearly, a one-element subset cannot determine a latin square.

The secret-sharing scheme just described leads us to the question of which partial latin squares of order n can be extended to latin squares of order n. Not every partial latin square extends; for example, the following partial 2×2 square cannot be completed to a latin square of order 2.

$$\begin{array}{cc} 0 & \\ & 1 \end{array}$$

In 1960, Evans [15] conjectured that any partial $n \times n$ latin square with no more than $n-1$ entries filled can be completed to a latin square of order n. This conjecture was verified by Smetaniuk [60] in 1981.

Theorem 4.4.5 (Smetaniuk). *Any partial latin square of order n with at most $n-1$ cells filled can be extended to a latin square of order n.*

The value $n-1$ in Smetaniuk's theorem is optimal; see Exercise 4.9.

A second threshold scheme with two shares can be constructed using latin squares as described in Example 4.2.2.

5. Notes

Schneier [56] gives a thorough account of modern cryptographic methods used in real-world applications. A popularization by Kahn [26] gives a thorough account of the history and development of cryptography. General information about cryptography may be found in the *Handbook of Applied Cryptography* [45].

Lidl and Niederreiter [36] describe a class of symmetric key cryptosystems known as stream ciphers. They also give a thorough introduction to the theory of discrete logarithms. More information on cryptosystems and threshold schemes using latin squares may be found in Laywine and Mullen [31].

Koblitz [27] gives an introduction to several cryptosystems based on finite fields, including the double-round quadratic cipher, as well as cryptosystems based on elliptic curves. The elliptic curve systems are of practical interest because although subexponential algorithms are known for factoring integers and computing discrete logarithms in finite fields, there are no known subexponential algorithms for computing discrete logarithms in general elliptic curves (which have a group structure but not a field structure).

Ferguson and Schneier [16] give a detailed introduction to the practice of implementing cryptographic methods securely. They emphasize the difficulties involved in implementing secure systems in real-world applications. These implementations must involve both secure cryptosystems and secure practices by the users of the cryptosystems.

An understanding of the subject of computational complexity, as described by Papadimitriou [51], is essential for assessing the security of cryptosystems.

6. Exercises

4.1. The following text has been encrypted by a simple substitution cipher using the letters A–Z. The punctuation and spacing are unchanged. Determine the original message.

> "PEI BGPEIBGPDLDGV OCIY VCP YPQOZ
> HQKI BGPEIBGPDLY AILGQYI DP DY QYI-
> SQN; EI YPQODIY DP AILGQYI EI OINDFEPY
> DV DP GVO EI OINDFEPY DV DP AILGQYI
> DP DY AIGQPDSQN." – EIVKD HCDVLGKI

4.2. Let q be a power of an odd prime and let $x \in F_q^*$ be fixed. Show that there is a $y \in F_q^*$ with $y^2 = x$ if and only if $x^{(q-1)/2} = 1$ in F_q.

4.3. Let q be a prime power such that $q \equiv 3 \pmod 4$ and let n be odd. Use Exercise 4.2 to show that -1 is not a perfect square in F_{q^n}. Use this to show that for any $x \in F_{q^n}$, exactly one of x and $-x$ is a perfect square.

Public key cryptosystems.

4.4. Show that if $n = pq$ is a product of distict primes, M is arbitrary, and $kd \equiv 1 \pmod{\phi(n)}$, then $M^{kd} \equiv 1 \pmod n$. This verifies the decryption step of the RSA cryptosystem.

4.5. The RSA cryptosystem is used with the public key $n = 517537$ and $k = 17$. Find the corresponding decryption key d any way you can. Then decrypt the following sequence of message units, each of which represents one letter or space. To convert a decoded message unit M to a letter, first reduce M modulo 27. Then $A = 1$, $B = 2$, and so on to $Z = 26$. If $M \equiv 0 \pmod{27}$, then M represents a space.

301985	260072	280987	329845	378568	391456	376789
311874	229335	419880	32739	20292	192273	70755
280987	301985	144317	280987	507536	378568	391456
301985	144317	357744	192273	126852	491968	475436
350585	378935	285376	2046	280987		

4.6. Verify the formulas for y_0, y_1, and y_2 in Example 4.3.2.

4.7. Compute the formulas for the decryption function in Example 4.3.2.

4.8. Show that the RSA cryptosystem can be used as a digital signature scheme. To do this, show that an attack on the digital signature scheme formed from the RSA cryptosystem can be turned into an attack on the RSA cryptosystem itself. Thus if the RSA cryptosystem is secure, then so is the signature scheme formed from it.

Latin squares.

4.9. Prove that for each $n \geq 2$ there is a partial latin square L of order n with exactly n cells filled such that L cannot be extended to a latin square of order n.

4.10. Show that any $(n-1) \times n$ latin rectangle, where $n \geq 2$, can be completed to a latin square of order n in exactly one way.

Discrete logarithms.

4.11. Make a table showing the discrete logarithm of each nonzero element in F_{17} with respect to the primitive element $g = 3$.

4.12. Let $\log_g$ denote the discrete logarithm, where g is a primitive element of the finite field F_q. Prove that the following identities hold for all $a, b \in F_q^*$ and $n \in \mathbb{N}$:

$$\begin{aligned} \log_g(ab) &\equiv \log_g(a) + \log_g(b) \pmod{q-1}, \\ \log_g(ab^{-1}) &\equiv \log_g(a) - \log_g(b) \pmod{q-1}, \\ \log_g(a^n) &\equiv n \log_g(a) \pmod{q-1}. \end{aligned}$$

Elliptic curves.

4.13. Determine all 13 of the rational points of the elliptic curve

$$E\colon y^2 = x^3 + x + 6$$

defined over the field F_{11}.

4.14. Make an addition table for the group operation on the rational points of the elliptic curve $E\colon y^2 = x^3 + x + 6$ defined over the field F_{11}, which were found in the previous exercise.

4.15. Consider the elliptic curve $E : y^2 = x^3 + 7x$ defined over the field F_{13}. It can be shown that there are 18 rational points on E. Find each of these rational points and determine its order. Make an addition table for the group operation on the set of rational points of this curve.

4.16. Prove that the addition operation for points on elliptic curves is commutative and has $\mathcal{O}$ as an identity element. Then prove that every point has an inverse. (Note that to show that the points form a group, it is also necessary to verify that the operation is associative, which is a much more difficult result.)

4.17. Let E be an elliptic curve over a field F_q with characteristic greater than 3 and assume that the curve contains three rational points P_1, P_2, and P_3 with different x-values. Show that the sum of these three points is $\mathcal{O}$. Use this fact to give a geometrical interpretation of the addition operation for rational points.

This page intentionally left blank

http://dx.doi.org/10.1090/stml/041/05

Appendix A

Background in Number Theory and Abstract Algebra

In this appendix, we present some basic definitions and results of number theory and abstract algebra. Our goal is to cover the background material required for other parts of this book, and thus we do not discuss these subjects in detail. Although the exposition here is self contained, we encourage the reader unfamiliar with this material to consult a more thorough text. The books by Andrews [1], Grosswald [20], and Niven and Zuckerman [50] cover the fundamentals of number theory, and the books by Fraleigh [18], Gallian [19], and Hungerford [24] cover abstract algebra. Further material related to vector spaces, matrices, and linear algebra is given by Lipschutz [39].

1. Number theory

At its most basic level, number theory involves the study of the additive and multiplicative structure of the set $\mathbb{Z}$ of integers, with an emphasis on divisibility and primality. We use the standard notation (a, b) to denote the greatest (positive) common integer divisor of the integers a and b. If $(a, b) = 1$, then the integers a and b are said to be *relatively prime*.

Lemma A.1.1 (Euclid). *For any two integers a, b there are integers n, m such that $na + mb = (a, b)$. Moreover, (a, b) is the least positive integer that can be written as a sum of integer multiples of a and b.*

Proof. We prove the first statement by induction on the larger of a and b. The result is trivial when $a = b$, and we may assume without loss of generality that a and b are both positive and $a < b$. Consider $c = b - a$. Clearly $0 < c < b$. It is immediate that $(a, c) = (a, b)$ because any divisor of a and c divides b while any divisor of a and b divides c. By induction, there are integers n' and m' such that $n'a + m'c = (a, c)$. Hence $(n' - m')a + m'b = (a, b)$. We let $n = n' - m'$ and $m = m'$. This completes the induction.

To prove the second statement, we note that if $r|a$ and $r|b$, then $r|(na + mb)$ for all integers n and m. In particular, $(a, b)|(na + mb)$ for all n and m. Thus no positive integer smaller than (a, b) can be of the form $na + mb$. □

It should be noted that the proof of the previous lemma gives an algorithm to find (a, b): replace the larger of a and b with the (positive) difference of the numbers, and continue doing this until the numbers are equal. A more efficient algorithm for computing (a, b) is given in Exercise A.7.

Definition A.1.2. Let ϕ denote *Euler's function*, defined so that $\phi(n)$ counts the number of integers less than n and relatively prime to n (including 1). For example: $\phi(2) = 1$, $\phi(3) = 2$, $\phi(6) = 2$, and $\phi(8) = 4$. This function is also known as the *totient function.*

The next lemma gives several important properties of the ϕ function. These properties allow $\phi(n)$ to be computed quickly once n has been factored into prime powers. If n is large, it is believed to be difficult to calculate $\phi(n)$ without having the prime factorization of n. (As explained in Chapter 4, this difficulty is the basis for the generally believed security of the RSA cryptosystem.) Other properties of ϕ are explored in the exercises.

Lemma A.1.3. *The ϕ function has the following properties:*

(1) *If a and b are relatively prime, then $\phi(ab) = \phi(a)\phi(b)$.*

(2) *If p is a prime and $k \geq 1$ is an integer, then*

$$\phi(p^k) = p^k - p^{k-1}.$$

Proof. Exercise A.1. □

2. Groups

We now consider the fundamental notion of an abstract group. We will see that finite fields contain two groups, one under addition and another under multiplication. Recall that a binary operation $\cdot$ on a set G is *closed* if $a \cdot b$ lies in the set G for all elements $a, b \in G$.

Definition A.2.1. A *group* is a nonempty set G with a closed binary operation $\cdot$ such that the following properties hold:

(1) Associativity: $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all a, b, and c in G.

(2) Identity: There is an element $e \in G$ such that $e \cdot a = a \cdot e = a$ for all $a \in G$.

(3) Inverses: For each $a \in G$ there is an element $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$.

A group G is *Abelian* (or *commutative*) if $a \cdot b = b \cdot a$ for all $a, b \in G$. For simplicity of notation, the group operation may sometimes be denoted by juxtaposition of group elements: ab denotes $a \cdot b$.

Example A.2.2. The following structures are groups:

(1) The integers under addition.

(2) The nonzero rational numbers under multiplication.

(3) The set of all 2×2 matrices with real entries whose determinant is nonzero, with the operation of matrix multiplication.

Definition A.2.3. If G is a group, then a nonempty subset H of G is a *subgroup* of G if H is itself a group under the same operation as in G.

The next lemma gives a convenient criterion for a subset of a group to be a subgroup.

Lemma A.2.4. *Let G be a group and let H be a nonempty subset of G. Then H is a subgroup of G if and only if hk^{-1} is in H for all $h, k \in H$. If H is finite, then H is a subgroup if and only if H is closed under the group operation.*

Proof. Exercise A.11. □

The *order* of a finite group G is the cardinality of G, which we denote $|G|$. We may also speak of the order of an element of a group, as the next definition shows.

Let a be an element of a group G. If n is a positive integer, a^n denotes the multiplication of a with itself n times. If n is negative, then $a^n = (a^{-1})^{-n}$, and a^0 is e, the identity of G.

Definition A.2.5. The *order* of an element a in a group G is the least positive integer n such that $a^n = e$, if such an n exists.

Every element of a finite group has some order. It is not difficult to see that if $a \in G$ and the order of a is n, then the set $\{a^i \mid 1 \leq i \leq n\}$ is a subgroup of G, and the order of this subgroup is n, the same as the order of a.

Every group has exactly one element of order 1, which is the identity element e. The next theorem characterizes which other orders may occur in a finite group.

Theorem A.2.6 (Lagrange). *If G is a finite group, the order of any subgroup of G divides the order of G. In particular, the order of any element in a finite group divides the order of the group.*

Proof. Let $h \in G$. Define g_1 and g_2 to be equivalent if there is an integer n such that $g_1 = h^n g_2$. The reader should verify that this is an equivalence relation and that the cardinality of each equivalence class is the order of h. Hence the order of h divides the order of the group. □

Definition A.2.7. A group G is said to be *cyclic* if there is an element $g \in G$ such that every element h of G can be written as an integral power $h = g^n$ of g. Such an element g is called a *generator* of G. If g is an element of a group G we write $\langle g \rangle$ for the set $\{g^n \mid$

$n \in \mathbb{Z}\}$. The set $\langle g \rangle$ forms a subgroup of G called the *subgroup of G generated by g*.

Several facts follow immediately from the definitions. A group G is cyclic if and only if there is some $g \in G$ with $G = \langle g \rangle$. A finite group G of order n is cyclic if and only if there is an element $g \in G$ of order n. In this case, $G = \langle g \rangle = \{g^1, g^2, \ldots, g^{n-1}, g^n = e\}$.

Lemma A.2.8. *Every subgroup of a cyclic group is cyclic.*

Proof. Let g generate a cyclic group G and let H be a subgroup of G. If $H = \langle e \rangle$, then H is trivially cyclic. Otherwise, let $N = \{n \in \mathbb{Z} \mid g^n \in H\}$. The set N is nonempty and closed under negation (because H is closed under inverses). Therefore there is a least positive $n \in N$. We claim that $H = \langle g^n \rangle$. If not, then there is a positive integer m such that $g^m \in H$ and n does not divide m. Write $l = (n, m)$ and choose a, b such that $an + bm = l$ (using Lemma A.1.1). Then $g^l = g^{an+bm} = g^{an} g^{bm} \in H$. But $l < n$, which is a contradiction. □

The following result is useful for the calculation of the orders of elements in a finite group.

Lemma A.2.9. *Let G be a finite commutative group. If the order of $a \in G$ is relatively prime to the order of $b \in G$, then the order of ab is the product of the orders of a and b. More generally, the order of ab is the least common multiple of the orders of a and b.*

Proof. Exercise A.9. □

3. Rings and fields

Definition A.3.1. A *ring* is a set R with two closed binary operations $+$ and $\cdot$ satisfying the following properties:

(1) R is an Abelian group under the operation $+$.

(2) The operation $\cdot$ is associative.

(3) For all $a, b, c \in R$ the *distributive laws* hold:

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

If the multiplication operation is commutative, then the ring is called *commutative* (note that, by definition, the addition operation in a ring is always commutative).

Example A.3.2. The following structures form rings:

(1) The integers under the usual operations of addition and multiplication.

(2) The rational numbers with the usual operations of addition and multiplication.

(3) The set of all 2×2 real matrices with the usual operations of matrix addition and multiplication.

(4) The set of all continuous functions on the real line, with pointwise addition and multiplication.

Definition A.3.3. A nonzero element y of a commutative ring is called a *zero divisor* if there is a nonzero element z in the ring with the property that $y \cdot z = 0$.

The ring of integers with the usual multiplication operation has no zero divisors, but some other rings do have them. No field has zero divisors (Exercise A.16). The ring of 2×2 real matrices has many zero divisors, as does the ring of continuous functions on the real line with pointwise addition and multiplication.

Definition A.3.4. A *field* is a commutative ring with a multiplicative identity with the property that every nonzero element has a multiplicative inverse.

Definition A.3.5. The *characteristic* of a ring is the least positive integer n such that $n \cdot 1 = 1 + 1 + \cdots + 1$ (the sum of 1 with itself n times) equals 0; if no such n exists, then the characteristic is declared to be 0.

Lemma A.3.6. *If the characteristic of a field is nonzero, then the characteristic is prime.*

Proof. Suppose the characteristic n of a field F factors as $n_1 n_2$; thus $n_1 n_2 \cdot 1 = 0$. Since there are no zero divisors in F (see Exercise A.16), either $n_1 \cdot 1$ or $n_2 \cdot 1$ is zero. Therefore there is a prime divisor p of n

such that $p \cdot 1 = 0$. A similar argument shows that the characteristic is exactly p. □

The field of rational numbers, the field of real numbers, and the field of complex numbers all have characteristic zero. It follows from Lagrange's theorem that a finite field of order $q = p^n$ has prime characteristic p.

Definition A.3.7. An *integral domain* is a commutative ring with a multiplicative identity and no zero divisors.

Thus in an integral domain R, if $a \cdot b = 0$ for some $a, b \in R$, then either $a = 0$ or $b = 0$.

Lemma A.3.8. *A finite integral domain is a field.*

Proof. The only property which must be verified is the existence of multiplicative inverses. Let R be a finite integral domain and choose a nonzero element $a \in R$. Consider the set $aR = \{ar \mid r \in R\}$. For $r_1 \neq r_2$ we must have $ar_1 \neq ar_2$, for if $ar_1 = ar_2$, then $a(r_1 - r_2) = 0$. Since a is nonzero, we see that $r_1 = r_2$. Therefore $|aR| = |R|$, so the multiplicative identity 1_R is an element of aR. Thus there is a $b \in R$ such that $ab = 1_R$. Clearly b is the inverse of a. □

The hypothesis of finiteness in the previous lemma is necessary; the ring of integers is an integral domain but not a field.

The ring of integers $\mathbb{Z}$ is the fundamental example of an integral domain. We can form many interesting finite rings from it, as we now explain. Let n, a positive integer, be fixed. To form the ring $\mathbb{Z}_n$, we define a and b to be *equivalent modulo* n if their difference is a multiple of n. Thus n, $2n$, and $-5n$ are equivalent modulo n, as are $n - 2$, $19n - 2$, and $14n - 2$. To perform addition or multiplication modulo n, we perform the operations as usual in the ring $\mathbb{Z}$ of integers, but then replace the result with the smallest equivalent nonnegative integer. Thus, working modulo 5, we have:

$$3 + 4 \cdot 2 \equiv 3 + 8 \equiv 3 + 3 \equiv 6 \equiv 1 \pmod 5.$$

We write $\equiv$ instead of $=$ in order to distinguish this arithmetic from the usual integer arithmetic with its equality of integers, and write

(mod 5) to remind ourselves of the modulus $n = 5$. The integers $0, 1, 2, \ldots, n-1$ are pairwise inequivalent modulo n, and every integer is equivalent to one of them. These numbers form a commutative ring with the operations of addition and multiplication modulo n. We denote this ring by $\mathbb{Z}_n$ and call it the *ring of integers modulo n*.

Example A.3.9. The following tables show the addition and multiplication operations on $\mathbb{Z}_5$, the ring of integers modulo 5.

$+$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\cdot$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Note that the nonzero elements of $\mathbb{Z}_5$ form a group under multiplication with identity 1. The following lemma explains why this has occurred.

Lemma A.3.10. *A nonzero element k in $\mathbb{Z}_n$ has a multiplicative inverse if and only if $(k, n) = 1$. In particular, if p is prime, then the commutative ring $\mathbb{Z}_p$ is a finite field.*

Proof. Suppose that $(k, n) = 1$. By Euclid's theorem, there are r, s such that $rk + sn = 1$; that is, $rk \equiv 1 \pmod{n}$. Now assume p is prime. Since the multiplication on $\mathbb{Z}_p$ is associative, commutative, and has an identity, the existence of a multiplicative inverse for every element is enough to make the nonzero elements of $\mathbb{Z}_p$ into a group and $\mathbb{Z}_p$ itself into a field. □

If n is a not prime, the ring $\mathbb{Z}_n$ has zero divisors; for example, 2 and 3 are zero divisors in the ring $\mathbb{Z}_6$. More generally, an element $a \in \mathbb{Z}_n$ is a zero divisor if and only if $(a, n) > 1$.

4. Homomorphisms

Definition A.4.1. A *group homomorphism* is a map ψ from one group G with operation $\cdot$ to another group G' with operation $\circ$ such that $\psi(g_1 \cdot g_2) = \psi(g_1) \circ \psi(g_2)$ for all $g_1, g_2 \in G$. A *group isomorphism*

is a group homomorphism from a group G to another group G' that is both injective (one-to-one) and surjective (onto). If there is an isomorphism between two groups, we call the groups *isomorphic.*

A *ring homomorphism* is a map ϕ from one ring R with operations $+_R$ and $\cdot_R$ to another ring S with operations $+_S$ and $\cdot_S$ so that $\phi(r_1 +_R r_2) = \phi(r_2) +_S \phi(r_2)$ and $\phi(r_1 \cdot_R r_2) = \phi(r_1) \cdot_S \phi(r_2)$ for all $r_1, r_2 \in R$. A *ring isomorphism* is a bijective ring homomorphism. If there is an isomorphism between two rings, we call the rings *isomorphic.*

Lemma A.4.2. *Let G be a finite Abelian group of order m, and let n be an integer with $(n, m) = 1$. Then the map $f: a \mapsto a^n$ is an isomorphism from G to itself (an automorphism).*

Proof. We first show that f is bijective. Suppose $a^n = b^n$; then $(ab^{-1})^n = e$ (see Exercise A.12). This shows that the order of ab^{-1} divides n. But the order of ab^{-1} divides m, so the order divides $(n, m) = 1$. We conclude $ab^{-1} = e$; that is, $a = b$. This shows that f is injective. Clearly, since G is finite, f is also surjective.

Since G is Abelian, we may prove by induction on n that $(ab)^n = a^n b^n$ for all n. Thus f is a group isomorphism. □

The following result will be useful in Chapter 1, when we prove that every extension field has a normal basis over the base field. This result can be proved by induction on m, the number of homomorphisms; a proof is provided by Lidl and Niederreiter [36, Lemma 2.33].

Theorem A.4.3 (Artin's Lemma). *Let $\psi_1, \ldots, \psi_m$ be a set of distinct nonzero homomorphisms from a group G into the multiplicative group of a field F. Let $a_1, \ldots, a_m$ be elements of F not all of which are zero. Then there is a $g \in G$ such that $\alpha_1\psi_1(g) + \cdots + \alpha_m\psi_m(g) \neq 0$.*

5. Polynomials and splitting fields

Given a ring R with identity, we let $R[x]$ denote the set whose elements are of the form $r_0 + r_1x + r_2x^2 + \cdots + r_nx^n$, where each $r_i \in R$. The variable x is treated as a new element distinct from every element of R. The elements of $R[x]$ are called *polynomials* with coefficients in R

or sometimes just called *polynomials over R*. We add and multiply elements of $R[x]$ using naive polynomial addition and multiplication. For example, we have:

$$\begin{aligned}(r_0+r_1x+r_2x^2)(s_0+s_1x)&\\ &= r_0s_0 + r_1s_0x + r_2s_0x^2 + r_0s_1x + r_1s_1x^2 + r_2s_1x^3\\ &= r_0s_0 + (r_1s_0 + r_0s_1)x + (r_2s_0 + r_1s_1)x^2 + r_2s_1x^3.\end{aligned}$$

The reader should check that with these operations the set $R[x]$ forms a ring; this ring is called the *ring of polynomials over R*. We also note that if R is commutative, then so is $R[x]$.

For any polynomial $p(x)$ over R we may form the *factor ring* $R[x]/(p(x))$ which is the set of all polynomials over R of degree less than the degree of $p(x)$. To form this ring, we consider two polynomials $r(x), r'(x)$ to be equivalent if their difference is a (polynomial) multiple of $p(x)$. This is entirely analogous to the way the ring $\mathbb{Z}_n$ is formed from the ring $\mathbb{Z}$.

A *root* of a polynomial $p(x) = r_0 + r_1x + \cdots + r_nx^n \in R[x]$ is an element r such that $p(r) = r_0 + r_1r + r_2r^2 + \cdots + r_nr^n = 0$ in R.

Lemma A.5.1. *Let F be a field. An element $r \in F$ is a root of a polynomial $p(x) \in F[x]$ if and only if the polynomial $x - r$ divides the polynomial $p(x)$ in the ring $F[x]$.*

Proof. Exercise A.22. □

If r is a root of $p(x) \in F[x]$, where F is a field, we define the *multiplicity* of r to be the largest n such that $(x - r)^n$ divides $p(x)$.

Lemma A.5.2 (The derivative test). *Let $p(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ be an element of $F[x]$, where F is a field. Define the* formal derivative *of $p(x)$ to be $p'(x)$, where*

$$p'(x) = na_nx^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + 2a_2x + a_1.$$

A root r of $p(x)$ has multiplicity greater than 1 if and only if $p'(r) = 0$.

Proof. Suppose that the multiplicity of r is $k \geq 1$. Write $p(x) = (x - r)^kq(x)$ (here $q(x)$ may be 1). It can be shown that the formal derivative obeys the same algebraic formulas as the familiar derivative

from elementary calculus (see Exercise A.24). In particular, $p'(x) = (k)(x-r)^{k-1}q(x) + (x-r)^k q'(x)$. It is clear that if $k > 1$, then $p'(r) = 0$. If $k = 1$, then $p'(x) = q(x) + (x-r)q'(x)$. We have assumed $q(r) \neq 0$, so if $k = 1$, we have $p'(r) = q(r) \neq 0$. □

Lemma A.5.3. *A polynomial of degree n over a field has at most n roots (counting multiplicities).*

Proof. The proof follows by repeatedly applying Lemma A.5.1. We leave the details to the reader in Exercise A.26. □

In the previous lemma, we did not make the false claim that a polynomial over a field must have exactly as many roots (counting multiplicity) as its degree. For example, the polynomial $x^2 + 1$ has no roots at all over the field of real numbers.

We say that a polynomial $p(x)$ over a ring R is *irreducible* if there are no polynomials $q(x), r(x) \in R[x]$ of positive degree such that $p(x) = q(x)r(x)$. It can be seen that a quadratic or cubic polynomial over a field is irreducible if and only if it has no roots, but this is not true for polynomials of degree 4 or higher (Exercise A.23).

Lemma A.5.4. *If F is a field, then $F[x]/(p(x))$ is a field if and only if $p(x)$ is irreducible in $F[x]$. Moreover, there is a field isomorphism from F to a proper subfield of $F[x]/(p(x))$.*

Proof. Exercise A.27. □

We note that the construction in the previous lemma is used in Chapter 1 to construct finite fields. In particular, to construct the finite field F_{p^n} one uses an irreducible polynomial $p(x)$ of degree n over F_p. Such an irreducible exists for every prime p and every positive integer $n \geq 2$ by Corollary 1.3.13.

The key fact about $F[x]/(p(x))$ is that the equivalence class of x in the factor ring is a root of $p(x)$ in the extension field. The next lemma shows that the construction in Lemma A.5.4 is unique up to isomorphism. We require some notation for field extensions.

Definition A.5.5. Let F be a subfield of a field K and let $g_1, \ldots, g_k \in K$. We write $F(g_1, \ldots, g_k)$ for the smallest subfield H of K such that

$F \subseteq H$ and $g_i \in H$ for each $i \leq k$. We call $F(g_1, \ldots, g_k)$ the *extension field* of F generated by $g_1, \ldots, g_k$.

Lemma A.5.6. *Suppose that $p(x)$ is irreducible over a field F. Suppose there is a field $E = F(r)$, obtained by adjoining r to F, in which r is a root of $p(x)$. Then E is isomorphic to $F[x]/(p(x))$.*

Proof. It is easy to show that the evaluation map $\phi\colon F[x] \to E$ which sends each $q(x)$ to $q(r)$ is a ring homomorphism. Moreover, the kernel of ϕ includes $p(x)$, so ϕ induces a homomorphism $\hat{\phi}$ from $F[x]/(p(x))$ to E. That is, if $q(x)$ and $r(x)$ are equivalent modulo $p(x)$, then $\phi(q(x))$ and $\phi(r(x))$ are equal in E, and so $\phi(r(x))$ is uniquely determined by the equivalence class of $r(x)$ in $F[x]/(p(x))$.

Because $p(x)$ is irreducible, Lemma A.5.4 tells us that $F[x]/(p(x))$ is a field. Since $\hat{\phi}$ does not send every element to 0, $\hat{\phi}$ is injective (see Exercise A.25). Now the range of $\hat{\phi}$ is a subfield of E that includes F and includes r, so the range includes $E = F(r)$. Thus $\hat{\phi}$ is surjective. We have shown $\hat{\phi}$ is an isomorphism between $F[x]/(p(x))$ and E. □

Definition A.5.7. Suppose F is a field and $p(x) \in F[X]$. A field K is called a *splitting field* of $p(x)$ if the following conditions hold.

(1) The monic polynomial $p(x)$ factors into linear factors in $K[x]$. This means

$$p(x) = (x - a_1)(x - a_2) \cdots (x - a_k)$$

for some $a_1, \ldots, a_k \in K$.

(2) $K = F(a_1, \ldots, a_k)$, that is, the roots of $p(x)$ generate K over F.

We will show in Theorem A.5.9 that every polynomial has a splitting field and that such a field is unique up to isomorphism. This result is crucial for the classification of finite fields in Chapter 1. We require the following lemma.

Lemma A.5.8. *Suppose that K is the splitting field of $p(x)$ over a field F. Suppose that M is any field containing F in which $p(x)$ splits into linear factors. Then there is an injective homomorphism from K to M which is the identity on F.*

Proof. The proof is by induction on the degree of $p(x)$. If the degree is 1, then the result is trivial. We assume the degree is greater than 1 and write $p(x) = q_1(x) \cdots q_k(x)$, where each $q_i(x)$ is irreducible and of degree greater than 1.

Let $a \in K$ be any root of $q_1(x)$. By Lemma A.5.6, $F(a)$ is isomorphic to $F[x]/(q_1(x))$. There is a root b of $q_1(x)$ in M, and $F(b)$ is also isomorphic to $F[x]/(p(x))$. Thus $F(a) \subseteq K$ is isomorphic to $F(b) \subseteq M$. Let $r(x) = p(x)/(x - a)$ in $F[x]$. The degree of $r(x)$ is strictly less than the degree of $p(x)$, so we may apply our inductive hypothesis, replacing F with $F(a)$ and identifying $F(a)$ with $F(b) \subseteq M$. By induction, there is an isomorphism from K to M which is the identity on $F(a)$ assuming we have identified $F(a)$ with $F(b)$. We already have an isomorphism which identifies $F(a)$ with $F(b)$ and is the identity on F. By putting these together, we obtain a homomorphism from K to M which is the identity on F. To see that the homomorphism is injective, apply the result of Exercise A.25. □

Theorem A.5.9 (Existence and uniqueness of splitting fields). *Let $p(x)$ be a polynomial over a field F. There is a splitting field for $p(x)$ over F, and it is unique in the following sense. If E and E' are splitting fields for $p(x)$ over F, then there is an isomorphism between E and E' which is the identity on F.*

Proof. We first prove that $p(x)$ has a splitting field over F. We prove the result for all fields simultaneously by induction on the degree of $p(x)$. If the degree is 1, then F is trivially a splitting field for $p(x)$. Now assume the degree of $p(x)$ is larger than 1, and write $p(x)$ as a product $q_1(x) \cdots q_k(x)$ of irreducible factors. Apply Lemma A.5.4 to construct a field E such that q_1 has a root α in E and $E = F(\alpha)$. We now view $p(x)$ as a polynomial in $E[x]$, and note that, by Lemma A.5.1, $x - \alpha$ divides $p(x)$ in the extension field. Let $q(x)$ be $p(x)/(x - \alpha)$ in $E[x]$. Then the degree of $q(x)$ is less than the degree of $p(x)$, so by induction there is a splitting field $K = E(\beta_1, \ldots, \beta_r)$ for $q(x)$ over E, where $\beta_1, \ldots, \beta_r$ are the roots of $q(x)$. Now $K = F(\alpha, \beta_1, \ldots, \beta_k)$ is a splitting field for $p(x)$ over F.

We now turn to the uniqueness of the splitting field. Let E and E' be splitting fields for $p(x)$ over F. We apply Lemma A.5.8 to obtain

an injective homomorphism ϕ from E to E' which is the identity on F. It is not difficult to see that each root of $p(x)$ in E maps to a root of $p(x)$ in E' (see Exercise A.21). Thus, since ϕ is a injective and $p(x)$ has only finitely many roots, every root of $p(x)$ in E' is in the range of ϕ. Since E' is generated by these roots, E' is a subset of the range of ϕ, so ϕ is surjective. The result follows immediately. □

6. Vector spaces

Vector spaces are another important class of algebraic structures. They are essential to the study of finite fields in Chapter 1 and to the construction of error-correcting codes in Chapter 3.

Definition A.6.1. Let F be a field. A *vector space* over F consists of a nonempty set V, an Abelian group operation $+$ on V, and a *scalar multiplication* function that takes an $a \in F$ and a $\mathbf{v} \in V$ and returns $a\mathbf{v} \in V$. In addition to the Abelian group axioms for $+$, the following are required to hold for all $a, b \in F$ and all $\mathbf{u}, \mathbf{v} \in V$:

(1) $a(b\mathbf{u}) = (ab)\mathbf{u}$,

(2) $(a + b)\mathbf{u} = a\mathbf{u} + b\mathbf{u}$,

(3) $a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$.

A *subspace* of a vector space V is a nonempty subset which is itself a vector space under the same operations as in V.

We now give several examples of vector spaces.

Examples A.6.2. (1) Let $V = \mathbb{R}^2$ and let $F = \mathbb{R}$, where $\mathbb{R}$ denotes the field of real numbers. Addition is the usual addition of ordered pairs defined by $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$ and scalar multiplication is defined in the usual way as $a(x_1, y_1) = (ax_1, ay_1)$. With these operations, V is nothing more than the usual Euclidean plane. Similarly, the set $\mathbb{R}^3$ with the ordinary operations is a vector space over $\mathbb{R}$ which is just the usual three-dimensional Euclidean space.

(2) For $n \geq 1$ let $V = \mathbb{R}^n$ and let $F = \mathbb{R}$. Define addition as

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$$

and scalar multiplication by

$$a(x_1, \ldots, x_n) = (ax_1, \ldots, ax_n).$$

As in (1) above, V is a vector space over $\mathbb{R}$.

(3) In (1) and (2) above, we may replace the field $\mathbb{R}$ of real numbers by any field F and $V = F^n$ is still a vector space. In particular, if $F = F_q$, the finite field with q elements, then $V = F_q^n$ is a vector space over F_q which contains exactly q^n distinct elements.

(4) It follows from (3) that any field F is a vector space over itself, using the field addition for vector addition and the field multiplication for scalar multiplication. It can be seen that if F is a subfield of a field K, then K is a vector space over F, using the field operations in K.

(5) Let $m \geq 1, n \geq 1$ be integers and let $M_{m,n}$ denote the set of all $m \times n$ matrices over a field F. Then $M_{m,n}$ becomes a vector space over the field F with operations defined as follows. Let $A = (a_{ij}), B = (b_{ij}) \in M_{m,n}$ and let $c \in F$ be a scalar. Then $A + B = (a_{ij} + b_{ij})$ and $cA = (ca_{ij})$. The reader should also note that with this definition of addition and the usual matrix multiplication, $M_{m,n}$ is a ring but the ring is not commutative, that is, there are $A, B \in M_{m,n}$ for which $AB \neq BA$.

Definition A.6.3. A set $\{\mathbf{v}_2, \ldots, \mathbf{v}_k\}$ of vectors in a vector space V is said to be *linearly independent* over the field F if the only solution to the vector equation

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v}_k = \mathbf{0}, \tag{5}$$

where each $a_i \in F$, is given by $a_1 = a_2 = \cdots = a_k = 0$. If there is a solution to Equation (5) in which at least one $a_i \neq 0$, then the set of vectors is said to be *linearly dependent* over F.

We say that a vector $\mathbf{v} \in V$ is a *linear combination* of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ if there are elements $c_1, \ldots, c_k \in F$, so that

$$\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k.$$

A set of vectors $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is said to *span* V if every vector $\mathbf{v} \in V$ can be written as a linear combination of B. In addition, we

say that B forms a *basis* for V over F if B is linearly independent and spans V.

A vector space V is *finite dimensional* if it has a finite basis. One of the fundamental results in the theory of finite dimensional vector spaces is that any two bases of a fixed finite dimensional space V must contain the same number of vectors. This common size of all bases is called the *dimension* of V.

The following lemma gives a convenient criterion for determining whether a set of vectors forms a basis.

Lemma A.6.4. *Let V be a vector space of finite dimension m and let $v_1, \ldots, v_m$ be a set of linearly independent vectors in V. Then this set forms a basis for V.*

Proof. Exercise A.30. □

Examples A.6.5.

(1) Let $\mathbb{R}^2$ be the vector space of all ordered pairs of real numbers, with addition and scalar multiplication defined coordinatewise. Let $\mathbf{v}_1 = (1, 0)$ and $\mathbf{v}_2 = (0, 1)$. Then $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis for V, so the dimension of $\mathbb{R}^2$ is 2. Similarly, for $n \geq 1$, let $\mathbb{R}^n$ be the set of all n-tuples or real numbers. Then R^n is a vector space of dimension n over the field of real numbers.

(2) Let $M_{m,n}$ denote the set of all $m \times n$ matrices over a field F, where addition and scalar multiplication are defined in the usual way. That is if $A = (a_{ij}), B = (b_{ij})$, then $A + B = (a_{ij} + b_{ij})$ and $cA = (ca_{ij})$. The dimension of this vector space is mn, as the reader should verify by constructing a basis. An additional problem in this vein is stated as Exercise A.28.

(3) Because the rational numbers are a subfield of the real numbers, the real numbers form a vector space over the field of rational numbers. This vector space is not finite dimensional.

Natural functions to study in the context of vector spaces are those that preserve both the vector addition and scalar multiplication operations.

Definition A.6.6. Suppose V and W are vector spaces over the same field F. A map $f\colon V \to W$ is called *linear* if the following hold for all $\mathbf{u}, \mathbf{v} \in V$ and $a \in F$:

$$f(\mathbf{u}+\mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v}),$$
$$f(a\mathbf{u}) = af(\mathbf{u}),$$

where operations on the left side of each equation are performed in V and operations on the right side of each equation are performed in W.

Several important properties of linear functions are sketched in Exercise A.29.

We now turn to the subject of dual spaces, which are an essential tool in Chapter 1. We point out that the following definitions and theorem are valid for vector spaces over an arbitrary (possibly finite) field.

Definition A.6.7. Let V be a vector space over a field F. A *linear functional* on V is a linear function from V to F. We let $\mathrm{Dual}(V)$ denote the set of all linear functionals on V; this is called the *dual space* of V. This space is a vector space over F with pointwise addition and pointwise scalar multiplication of functions. It is common to denote the dual space of V by V^*, but we reserve the notation F^* to denote the set of nonzero, and thus invertible, elements in a field F.

Definition A.6.8. Suppose that V is a finite dimensional vector space over a field F and $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is an ordered basis for V (that is, the elements of B have been assigned numbers $1, \ldots, k$). For each $i \leq k$ let $\mathbf{v}_i^*$ be the unique linear functional defined by the rule

$$\mathbf{v}_i^*(\mathbf{v}_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

(The definition of $\mathbf{v}_i^*$ for each i depends on the entire basis. We leave it to the reader to verify that this is a valid definition; see Exercise A.32.) The ordered set $B^* = \{\mathbf{v}_1^*, \ldots, \mathbf{v}_k^*\}$ is the *dual basis* of the ordered basis B.

Theorem A.6.9. *Let V be a finite dimensional vector space over a field F and let B be a finite basis for V. The space* $\mathrm{Dual}(V)$ *is also*

a vector space over F with the canonical scalar multiplication. The dual basis B^* is a basis for $\mathrm{Dual}(V)$. *In particular, the dimension of* $\mathrm{Dual}(V)$ *is the same as the dimension of* V.

Proof. Exercise A.32. □

We caution the reader that there is a different definition of "dual space" which is more appropriate for use with infinite dimensional vector spaces that arise in the field of functional analysis. In the case that a vector space is finite dimensional, the other definition of a dual space coincides precisely with the definition we have given here.

7. Notes

There are numerous excellent texts dealing with elementary number theory and abstract algebra. For number theory we refer to books by Andrews [1], Grosswald [20], and Niven and Zuckerman [50]. For abstract algebra we refer the reader to books by Fraleigh [18], Gallian [19], and Hungerford [24]. A discussion of basic properties of vector spaces, matrices, and linear algebra is given by Lipschutz [39].

8. Exercises

Number Theory.

A.1. Prove Lemma A.1.3.

A.2 (The Chinese Remainder Theorem). Assume that $a_1, a_2, \ldots, a_k$ are natural numbers that are pairwise relatively prime and $n_1, \ldots, n_k$ are arbitrary natural numbers. Show that there is a natural number s such that $s \equiv n_i \pmod{a_1}$ for each $i \leq k$. Moreover, show that if s and s' are two solutions to the system of congruences, then $s \equiv s' \pmod{a_1 a_2 \cdots a_k}$.

A.3. Show that for any positive integer n the following equation holds:

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

where the product is taken over all prime divisors of n.

A.4. Show that for any positive integer n the following equation holds:

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

where the sum is taken over all divisors of n (including 1 and n).

A.5 (Fermat's Little Theorem). Let p be a prime number. For all integers a not divisible by p, show that $a^{p-1} \equiv 1 \pmod{p}$.

A.6 (Euler's Theorem). If $(a, n) = 1$, show that $a^{\phi(n)} \equiv 1 \pmod{n}$.

A.7 (The Euclidean Algorithm). Consider the following algorithm.

> *Input:* two positive integers a and b.
> *Procedure:* If $a = b$, stop and return a. Otherwise, swap the numbers if needed so that $a < b$. Replace b by the remainder r when b is divided by a. If this remainder is zero, return a. Otherwise, repeat from the beginning using a and r instead of a and b.

Show that for any positive integers a and b this algorithm will terminate and return the greatest common divisor (a, b) of the inputs.

A.8. This exercise assumes some knowledge of computational complexity. Prove the algorithm presented in Exercise A.1 has an upper bound on the number of iterations that is a linear function of the sum of the lengths of a and b in binary digits. This can be used to show that the greatest common divisor of a pair of integers can be computed in polynomial (in fact, quadratic) time.

Groups.

A.9. Prove Lemma A.2.9

A.10. Let a be an element of finite order k in a multiplicative group G. Show that for $m \in \mathbb{Z}$ we have $a^m = e$ if and only if k divides m.

A.11. Prove Lemma A.2.4.

A.12. Let G be an Abelian group and suppose $c = b^n$ in G. Show that $cb^{-n} = 1$. This elementary fact was used in the proof of Lemma A.4.2.

Rings and Fields.

A.13. Show that each of those structures in Example A.3.2 is a ring. Determine which of the structures are fields.

A.14. Show that the ring $\mathbb{Z}_n$ is a field if and only if n is prime.

A.15. Note that if p is a prime, then the field F_p is the same as the ring $\mathbb{Z}_p$ of integers modulo p. Explain why the field F_4 is not the same as the ring $\mathbb{Z}_4$. In fact, it turns out that if $m > 1$, then the field $F_{p^m} \neq \mathbb{Z}_{p^m}$, the ring of integers modulo p^m. Explain why.

A.16. Show that a field cannot have zero divisors. That is, if F is a field and $a, b \in F$ satisfy $ab = 0$, then either $a = 0$ or $b = 0$. Use this to give a complete proof of Lemma A.3.6.

A.17. Let F be a field. Show that the polynomial ring $F[x]$ has no zero divisors.

A.18. Let F be a field and let $p(x)$ and $q(x)$ be polynomials over F. Show that there are unique polynomials $r(x)$ and $s(x)$ over F such that the degree of $r(x)$ is strictly less than the degree of $q(x)$ and $p(x) = q(x)s(x) + r(x)$.

A.19. Let F be a field, let $p(x) \in F[x]$ be irreducible, and fix $q(x) \in F[x]$ such that $q(x)$ is not a multiple of $p(x)$. Show that there are polynomials $a(x)$ and $b(x)$ in $F[x]$ such that $p(x)a(x) + q(x)b(x) = 1$. This is a special case of general properties of the rings called *unique factorization domains*, which include all rings of the form $F[x]$.

A.20. Let R be a ring and let S be the intersection of all subrings of R. Show that S is a subring of R, and that if R is a field, then S is a field as well.

A.21. Suppose $p(x) \in R[x]$, where R is a ring. Suppose S, S' are two rings containing R and ϕ is a ring homomorphism from S to S'. Then we have $\phi(p(s)) = p(\phi(s))$ for every $s \in S$.

A.22. Prove Lemma A.5.1.

A.23. Show that a quadratic or cubic polynomial over a field is irreducible if and only if it has no roots in the field. Construct an example of a quartic polynomial over the real numbers which is not irreducible but has no real roots.

A.24. Let F be a field and let $p'(x)$ denote the formal derivative of $p(x) \in F[x]$. Show that the formal derivative satisfies the following familiar identities:

$$(p(x) + q(x))' = p'(x) + q'(x),$$
$$(p(x)q(x))' = p'(x)q(x) + p(x)q'(x),$$
$$(p(x)^k)' = kp(x)^{k-1}p'(x).$$

Here k must be a positive integer and $(p(x))^0$ is taken to equal 1.

A.25. Suppose ϕ is a homomorphism from a field F to a field E. Show that either $\phi(a) = 0$ for every $a \in F$ or else ϕ is injective.

A.26. Prove Lemma A.5.3.

A.27. Prove Lemma A.5.4

Vector spaces.

A.28. Show that the collection of all polynomials of degree n or less, with real coefficients, forms a vector space over the reals, where vector addition and scalar multiplication are given by the usual operations on polynomials. Construct a basis for this vector space. What is its dimension?

A.29. Let U, V, and W be vector spaces over the same field F. Let $f: U \to V$ and $g: V \to W$ be linear maps. Show that:

(1) The composition $g \circ f: U \to W$ is linear.

(2) The function f is completely determined by its values on an arbitrary basis of U.

(3) The function f is injective if and only if the only vector $\mathbf{u} \in U$ with $f(\mathbf{u}) = \mathbf{0}$ is $\mathbf{u} = \mathbf{0}$.

(4) Suppose $U = V$ and U is finite dimensional. Then f is injective if and only if f is surjective. Show that this may not be true if $U = V$ and U is infinite dimensional.

A.30. Prove Lemma A.6.4.

A.31. Find a basis for the vector space V of all $m \times n$ matrices over a field F. What is the dimension of V?

A.32. Let V be a finite dimensional vector space over a field F. Let B be a basis for V. Prove that the dual space $\mathrm{Dual}(V)$ is a vector space over F and that the dual basis B^* of B is well defined. Prove Theorem A.6.9.

http://dx.doi.org/10.1090/stml/041/06

Appendix B

Hints for Selected Exercises

Chapter 1

1.1 For $s = 2$ use the binomial theorem; then use induction on s.

1.2 Use properties of an automorphism.

1.3 Two fractions a/b and c/d should be equivalent if $ad = bc$. Recall that the ring of polynomials over a field has no zero divisors (Exercise A.17).

1.4 To show that $a^p \in K$, use Exercise 1.1. To finish the proof, use Theorem 1.2.8.

1.5 Assume $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$. If $a_0 = 0$, divide by x and apply the induction hypothesis. If $a_0 = 1$, define j with $1 \leq j \leq n$ to be the smallest value so that $a_j = 1$. Hence $f(x) = x^n + \cdots + x^j + 1$. After two iterations, we obtain $x^n + \cdots + x^{j-1} + 1$. Continuing, we see that with each pair of iterations, the smallest exponent on the first nonzero term decreases by one.

After $2(j-1)$ iterations, we reach a polynomial of the form $x^n + \cdots + x + 1$. After three more iterations we are left with a polynomial of degree $n-1$, to which we apply the induction hypothesis.

1.6 Follow the construction given in Section 1.3 for constructing factor groups. This ring is not a field since the polynomial x^3+x^2+x is not irreducible over F_2; see Lemma A.5.4.

1.7 2 and 5.

1.8 Follow the ideas used in Example 1.3.14.

1.9 Use an argument involving the characteristic of a ring.

1.10 $\alpha^{40} = \alpha^3 + \alpha^2 + \alpha$; $n = 22$.

1.11 There will be 1 element of order 1, 1 of order 2, 2 of order 4, 4 of order 8, and 8 of order 15. In general, there will be $\phi(d)$ elements of order d dividing $q-1$.

1.12 To construct the field, follow the ideas used in Example 1.3.14. There will be 1 element of order 1, 2 elements of order 3, 4 of order 5, and 8 of order 15.

1.13 There are q^q functions mapping F_q to itself, each of which can be represented by a polynomial over F_q.

1.14 Use the fact that any two finite fields of the same order are isomorphic.

1.15 Let $a \in F_q$ be a primitive element. Then $1 + a + a^2 + \cdots + a^{q-2} = (a^{q-1} - 1)/(a - 1)$.

1.16 Write out both sides of the equation, and proceed by induction on the degree of the polynomial.

1.17 Use properties of the trace function.

1.18 Follow the ideas of Example 1.3.14.

1.19 Use the fact that the multiplicative group F_q^* of F_q is cyclic.

1.20 Use part (4) of Theorem 1.3.10.

1.21 Use Exercise A.19 to draw a contradiction if $p(x)$ does not divide $q(x)$.

1.22 Use Exercise 1.1.

1.23 Let α be a normal element in F_{q^m}, so α generates a normal basis. Let $\{\beta_1, \ldots, \beta_m\}$ be the dual basis and let

$$A = \begin{pmatrix} \alpha & \alpha^q & \cdots & \alpha^{q^{m-1}} \\ \alpha^q & \alpha^{q^2} & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{q^{m-1}} & \alpha & \cdots & \alpha^{q^{m-2}} \end{pmatrix}$$

and

$$B = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_m \\ \beta_1^q & \beta_2^q & \cdots & \beta_m^q \\ \vdots & \vdots & \ddots & \vdots \\ \beta_1^{q^{m-1}} & \beta_2^{q^{m-1}} & \cdots & \beta_m^{q^{m-1}} \end{pmatrix}.$$

Show that $AB = I_m$ and thus, since A is a symmetric matrix, $BA = I_m$. Since S is a symmetric matrix, show that $(AB)^T = B^T A^T = B^T A = I_m$ and conclude that $B = B^T$. Then it follows that $\beta_i = \beta_1^{q^{i-1}}$, so that the dual basis is also normal.

1.24 Show that the trace function maps onto each element of F_q exactly q^{m-1} times. The norm function maps onto each element of F_q^* exactly $(q^m - 1)/(q - 1)$ times.

1.25 If $\alpha = \beta^{q-1}$, calculate the norm of α.

1.26 Any nonzero vector of length m can be used as a basis element; then we can add any vector which is not a multiple of the first; then add a third vector which is not a linear combination of the first two vectors, etc.

1.28 Show that if $\mathrm{Tr}(\alpha_i) = 0$ for each i, then the set of elements consisting of the α_i cannot form a basis.

1.29 Use the fact that the range of the trace function is all of F_q.

1.30 For the second part, assume there is a self-dual normal basis and obtain a contradiction.

1.32 Follow the ideas of Exercise 1.6.

1.33 Yes.

1.34 First show this for polynomials of small degrees, such as 2, 3, and 4.

1.35 Recall that the reciprocal polynomial $f^*(x) = x^n f(1/x)$ if $f(x)$ is of degree n. Proceed by contradiction.

1.37 If $f(x)$ is irreducible over F_q, show that $f(x+e)$ is also irreducible over F_q for any $e \in F_q$. Now consider a special value of e which maps a given trace coefficient say b, to another trace coefficient, say c.

1.39 First show that the function $L : b \to L(b), b \in F_q$ is a linear operator on the vector space F_{q^r} over F_q. Then use part (3) of Exercise A.29.

1.40 Use the matrix A_L to show that in both cases (i) and (ii), the determinant of A_L is nonzero.

1.41 Use Corollary 1.6.19.

1.42 Recall that the composition of two permutations is another permutation.

1.44 If $f(x)$ is the given polynomial, show that $f(0) = 1, f(1) = 0$, and $f(c) = c$ if $c \neq 0, 1 \in F_p$.

1.45 Use the hint and Theorem 1.6.20.

1.46 Use the hint and Theorem 1.6.21.

1.47 Show that the mapping x^n from F_q to itself sends 0 to 0, and the range of x^n on F_q^* consists of $(q-1)/(q-1,n)$ distinct values, each repeated exactly $(q-1,n)$ times.

Chapter 2

2.1 For $n = 2, \ldots, 7$, $l_n = 1; 1; 4; 56; 9,408; 16,942,080$.

2.2 Across the first row put the values $1, 2, \ldots, n-1$ in the first $n-1$ cells. Then in the last cell of the second row put the value n.

2.3 This can be done by hand, or by using a computer. See Laywine and Mullen [31, pp. 6–7] for an example. For the second part, assume you have a pair of latin squares of order 6 with 35 distinct ordered pairs, and obtain a contradiction.

2.4 First construct fields of orders 5, 8, and 9, and then use Theorem 2.2.10.

2.5 Use fields of orders 3 and 7 to construct pairs of MOLS of orders 3 and 7; then use Lemmas 2.2.22 and 2.2.23.

2.6 The following latin square is one example.

0	1	2	3
1	3	0	2
2	0	3	1
3	2	1	0

2.7 Use the fact that a latin square of order n has an orthogonal mate if and only if the square consists of n disjoint transversals. Here a *transversal* is a set of n cells, one in each row and one in each column, which contain n distinct symbols.

2.8 Since for $a \neq 0 \in F_q$ the Dickson polynomial $D_n(x, a)$ gives a permutation of the field F_q if and only if $(n, q^2 - 1) = 1$, the proof is similar to that given in Theorem 2.2.10 for the polynomials $ax + y$.

2.9 Consider the polynomials $ax + y$ with $a \neq 0, 1, -1 \in F_q$.

2.10 Consider the polynomials $ax + y$ with $a \in F_{q^2}$ and $a \notin F_q$.

2.11 Use the orthogonality of the latin squares and the definition of a magic square.

2.12 Consider the polynomials $ax+y$ and $x+ay$ with $a \neq 0, 1 \in F_q$.

2.13 Consider the following latin square. Show that this cannot be the Cayley table of a group by considering the order of the group element corresponding to the fifth row.

a	b	c	d	e
b	d	a	e	c
c	e	d	b	a
d	a	e	c	b
e	c	b	a	d

2.14 Consider the polynomials $a_1x_1 + \cdots + a_{2i}x_{2i}$ over F_q where

(1) $(a_1, \ldots, a_i) \neq (0, \ldots, 0)$,

(2) $(a_{i+1}, \ldots, a_{2i}) \neq (0, \ldots, 0)$, and

(3) $(a'_1, \ldots, a'_{2i}) \neq e(a_1, \ldots, a_{2i})$ for any $e \in F_q$.

2.15 Use the same construction as in Exercise 2.14 except use d variables.

2.16 Use linear polynomials over F_q in d variables with at least $j+1$ nonzero coefficients.

2.17 Use properties of the trace function.

2.19 Use Theorem 2.3.8.

2.20 For the projective geometry $\mathrm{PG}(3, F_2)$, use the construction following Definition 2.3.10; for the affine geometry $\mathrm{AG}(3, F_2)$, use the construction following Definition 2.3.11.

2.22 For order 8, use Theorem 2.5.7 with $q = 7$. For order 12, use $q = 11$.

2.23 Since $q \equiv 1 \pmod 4$, we have that $\psi(-1) = 1$, not -1, and hence near the end of the proof of Theorem 2.5.7, we will not be able to show that row $i+1$ is orthogonal to row $k+1$.

2.24 Show that if two squares are orthogonal, then they remain orthogonal after permutations of the elements of either square.

2.25 Write out the ordered pairs in the new square and convince yourself that each row and each column has $n_1 n_2$ distinct ordered pairs.

2.26 Convince yourself that upon superposition of the resulting two new squares, no ordered pair occurs more than once.

Chapter 3

3.1 For each $i = 0, 1, \ldots, t$, count the number of vectors at a distance i from the given vector.

3.2 Let B be a basis for a linear code C. By using the vectors of B as the rows of a matrix, one can construct a generating matrix G for C and then construct from G a parity-check matrix H.

3.3 For the partiy check matrix, use as columns, all nonzero vectors over F_3 of length 3 whose first nonzero entry is 1.

3.4 Use Theorem 3.7.1 with $n = q = 4$.

3.5 Similar to the solution to the previous exercise.

3.6 If C is linear of dimension k, one only needs to calculate the weights of the $q^k - 1$ nonzero codewords; whereas if C is nonlinear,

then one must calculate the distance between all pairs of nonzero codewords which requires roughly $(q^k - 1)^2/2$ calculations.

3.7 This follows from a vector space argument involving the dimension of the vector space and its dual space.

3.8 The number of distinct binary cyclic codes is the number of divisors of the polynomial $x^{15} - 1$ over F_2. To find this number, factor the polynomial over the field F_2.

3.10 For a binary repetition code of odd length say $2k + 1$, use the fact that

$$\binom{2k+1}{0} + \binom{2k+1}{1} + \cdots + \binom{2k+1}{k} = 2^{2k}.$$

The sphere packing bound does not yield an equality for a binary repetition code of even length.

3.11 Show that $S(\mathbf{y}) = S(\mathbf{z})$ if and only if $H(\mathbf{y} - \mathbf{z})^T = 0$, and this happens if and only if $\mathbf{y} - \mathbf{z} \in C$.

3.12 Show that both of these codes yield equalities in the sphere packing bound.

3.13 Construct a $2i \times (q^i - 1)^2/(q - 1)$ generator matrix over F_q of rank $2i$ by using the vectors that give the linear polynomials generating the frequency squares.

3.14 Construct a $d \times (q^d - 1)/(q - 1) - d$ generator matrix over F_q of rank d.

3.15 Use properties of matrix algebra.

Chapter 4

4.1 The letter G stands for A, the letter I stands for E, and the letter Q stands for U.

4.2 If there is a y with $y^2 = x$, then the order of x can be calculated from the order of y. Conversely, assuming $x^{(q-1)/2} = 1$, let $x = g^\alpha$ for a primitive element g and prove that α must be even.

4.4 Since $kd \equiv 1 \pmod{(p-1)(q-1)}$, we have $kd \equiv 1 \pmod{p-1}$ and $kd \equiv 1 \pmod{q-1}$. Apply Fermat's little theorem (Exercise A.5) and the Chinese remainder theorem (Exercise A.2) to finish the proof.

4.5 You may need to use a computer program to search for the decryption key.

4.6 and 4.7 These calculations are best carried out using a computer algebra program.

4.8 The proof is based on the symmetry between the encryption and decryption aspects of the RSA cryptosystem.

4.9 Arrange the partial square so that every possible choice is eliminated for one particular space in the square.

4.12 These identities of the logarithm follow from properties of exponentiation in finite fields.

4.17 If $P_1 + P_2 + P_3 = \mathcal{O}$, then $P_1 + P_2 = -P_3$. Consider a line drawn through P_1 and P_2 on the curve, and a characterization of additive inverses.

Appendix A

A.1 For the first part, given n with $(n, ab) = 1$, let n_1 be the value of n modulo a and n_2 the value of n modulo b.

A.2 Consider $N = n_1 n_2 \cdots n_k$. For each i, n_i and N/n_i are relatively prime, so by the Euclidean algorithm there are a and b with $an_i + bN/n_i = 1$. Note that $bN/n_i \equiv 1 \pmod{n_i}$ and $bN/n_i \equiv 0 \pmod{n_j}$ for all $j \leq k$ with $j \neq i$.

A.3 Use the second part of Lemma A.1.3.

A.4 Given $m < n$, let d be the order of m in $\mathbb{Z}_n$. Show that each such value d occurs for $\phi(d)$ values of m.

A.5 Consider the order of a in $\mathbb{Z}_p$.

A.6 Consider the multiplicative order of a in the ring $\mathbb{Z}_n$.

A.7 Prove that the algorithm always terminates and then use induction on the number of iterations of the algorithm. Note that if $n = ma_1 + b_1$ and $m = b_1 a_2 + b_2$, then $m(1 + a_1 a_2) - na_2 = b_2$.

A.8 Establish an upper bound on the number of iterations of the algorithm, using the fact that the remainder when b is divided by a is no greater than a. Recall that the division operation runs in $O(n)$ time.

A.9 First show that the order of ab divides the product k of the order of a and the order of b. Now, assuming the order of ab is r, consider the order of the element $c = a^r = b^{-r}$.

A.10 Show that if $a^k = e$ and $a^m = e$, then for $n = (k, m)$ we have $a^n = e$ as well.

A.11 First show that the order l of ab divides the least common multiple of the orders of a and b. Then show by contradiction that l cannot be a proper divisor.

A.12 First prove by induction that $(b^n)^{-1} = b^{-n}$.

A.13 For the first part, first verify that H is closed under taking inverses. For the second part, under the assumption that H is finite, consider what must happen if an element is repeatedly multiplied by itself, and use the fact that G is a group.

A.14 and A.15 Consider zero divisors.

A.16 Use multiplicative inverses.

A.17 Consider the terms of highest degree in each of the two factors of a polynomial product.

A.18 Suppose that $q(x)s_1(x) + r_1(x) = q(x)s_2(x) + r_2(x)$. Then $q(x)(s_1(x) - s_2(x)) = r_2(x) - r_1(x)$. Use a degree argument to show that $r_2(x) - r_1(x) = 0$ and then use Exercise A.17 to show that $s_1(x) - s_2(x) = 0$.

A.19 The inductive proof uses Exercise A.18 repeatedly. As a preliminary step, if the degree of $q(x)$ is not smaller than that of $p(x)$, replace $q(x)$ by the remainder obtained when dividing $q(x)$ by $p(x)$. Now, since the degree of $p(x)$ is greater than that of $q(x)$, one can write $p(x) = q(x)s(x) + r(x)$, where $s(x)$ is nonconstant and therefore the remainder $r(x)$ is of smaller degree than $p(x)$. Moreover, $r(x)$ cannot be 0 because $p(x)$ is irreducible. Replace $q(x)$ by $r(x)$ and repeat until the degree of the new remainder $r(x)$ is zero, as in Exercise A.7, and thus $r(x) = c$ for some $c \in F^*$. By tracing backwards, show that there are polynomials $a(x)$ and $b(x)$ such that $p(x)a(x) + q(x)b(x) = c$. Finally, multiply by c^{-1}.

A.21 Decompose $p(x)$ into monomials.

A.22 The divisibility result can be proved by induction on the degree of $p(x)$. Then, assuming $p(r) = 0$, apply the result with $q(x) = x - r$ to show that $x - r$ divides $p(x)$.

A.23 Show that $p(a) = 0$ if and only if $x - a$ divides $p(x)$.

A.24 For the second equality, decompose $q(x)$ into monomials and use the first equality. For the third equality, use induction on k.

A.25 Use the multiplicative property of field isomorphisms.

A.26 Use Lemma A.5.1 and proceed by induction on the degree of the polynomial.

A.27 The embedding of the field F uses constant polynomials. If $p(x)$ is reducible, then $F[x]/(p(x))$ will have zero divisors, and is thus not a field by Exercise A.16. For the converse, assume $p(x)$ is irreducible and apply Exercise A.19 to show that all nonzero elements of $F[x]/(p(x))$ have multiplicative inverses.

A.28 The dimension is $n + 1$. One basis is $\{1, x, x^2, \ldots, x^n\}$.

A.29 For part (3), use the linearity of the map. For part (4), show that if f is not surjective, then the dimension of the range of f is less than the dimension of U, and if f is injective, then the dimensions must be the same.

A.30 Show that if independent vectors $v_1, \ldots, v_m$ do not span V, then the dimension of V must be greater than m.

A.31 Consider all of the mn matrices containing all 0s except for one 1, which is moved throughout the mn cells of the matrices.

A.32 To show that the dual basis is well defined, use the fact that a linear map is determined by its values on a basis. To show that the dual basis is a basis, assuming the dimension of V is k, consider a linear combination of the form

$$a_1\mathbf{v_1}^* + a_2\mathbf{v_2}^* + \cdots + a_k\mathbf{v_k}^* = \mathbf{0}.$$

Since this is an equation relating functions, it must hold for all values in the domain of the functions. In particular, for each vector $\mathbf{v}_i$ in the original basis we have

$$a_1\mathbf{v}_1^*(\mathbf{v}_i) + a_2\mathbf{v}_2^*(\mathbf{v}_i) + \cdots + a_k\mathbf{v}_k^*(\mathbf{v}_i) = \mathbf{0}$$

from which it follows that $a_i = 0$.

References

[1] G. E. Andrews, *Number Theory*, Saunders, Philadelphia, 1971. Reissued by Dover, 1995.

[2] E. F. Assmus Jr. and H. F. Mattson Jr., *Coding and combinatorics*, SIAM Review **16** (1974), 349–388.

[3] R. C. Bose, *On the application of the properties of Galois fields to the construction of hyper-Graeco-Latin squares*, Sankhyā **3** (1938), 323–338.

[4] R. C. Bose, S. S. Shrikhande, and E. T. Parker, *Further results on the construction of mutually orthogonal latin squares and the falsity of Euler's conjecture*, Canad. J. Math. **12** (1960), 189–203.

[5] R. H. Bruck and H. J. Ryser, *The non-existence of certain projective planes*, Canad. J. Math. **1** (1949), 149–168.

[6] L. Carlitz, *Primitive roots in a finite field*, Trans. Amer. Math. Soc. **73** (1952), 373–382.

[7] S. D. Cohen and S. Huczynska, *The primitive normal basis theorem — without a computer*, J. London Math. Soc. **67** (2003), 41–56.

[8] C. J. Colbourn and J. H. Dinitz (eds.), *Handbook of Combinatorial Designs*, second ed., Discrete Mathematics and its Applications, Chapman and Hall/CRC, Boca Raton, FL, 2007.

[9] J. Daemen and V. Rijmen, *The Design of Rijndael*, Springer, New York, 2001.

[10] H. Davenport, *Bases for finite fields*, J. London Math. Soc. **43** (1968), 21–39.

[11] J. Dénes and A. D. Keedwell, *Latin Squares and Their Applications*, Academic Press, New York, 1974.

[12] ———, *Latin Squares*, Ann. Disc. Math., vol. 46, North-Holland, Amsterdam, 1991.

[13] J. Dénes, G. L. Mullen, and S. Suchower, *Another generalized Golomb–Posner code*, IEEE Trans. Information Theory **IT-36** (1990), 408–411.

[14] W. Diffie and M. E. Hellman, *New directions in cryptography*, IEEE Trans. Information Theory **IT-22** (1976), no. 6, 644–654.

[15] T. Evans, *Embedding incomplete latin squares*, Amer. Math. Monthly **67** (1960), 958–961.

[16] N. Ferguson and B. Schneier, *Practical Cryptography*, John Wiley & Sons, 2003.

[17] N. Ferguson, R. Schroeppel, and D. Whiting, *A simple algebraic representation of Rijndael*, Selected areas in cryptography, Lecture Notes in Comput. Sci., vol. 2259, Springer, Berlin, 2001, pp. 103–111.

[18] J. B. Fraleigh, *A First Course in Abstract Algebra*, Addison-Wesley, Reading, MA, 1982.

[19] J. A. Gallian, *Contemporary Abstract Algebra*, D. C. Heath, Lexington, MA, 1994.

[20] E. Grosswald, *Topics from the Theory of Numbers*, second ed., Birkhäuser, Boston, 1984.

[21] D. Hachenberger, *Finite Fields*, Kluwer Academic Publishers, Boston, MA, 1997.

[22] T. Hansen and G. L. Mullen, *Primitive polynomials over finite fields*, Math. Comp. **59** (1992), 639–643; Supplement S47–S50.

[23] R. Hill, *A First Course in Coding Theory*, Oxford Appl. Math. and Comp. Sci. Ser., Clarendon Press, Oxford, 1986.

[24] T. Hungerford, *Abstract Algebra*, Saunders, Philadelphia, 1990.

[25] D. Jungnickel, *Finite Fields*, Bibliographisches Institut, Mannheim, Germany, 1993.

[26] D. Kahn, *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet*, Scribner, New York, 1996.

[27] N. Koblitz, *Algebraic Aspects of Cryptography*, Algorithms and Computation in Mathematics, vol. 3, Springer-Verlag, Berlin, 1998.

[28] C. W. H. Lam, G. Kolesova, and L. Thiel, *A computer search for finite projective planes of order* 9, Disc. Math. **92** (1991), 187–195.

[29] C. W. H. Lam, L. Thiel, and S. Swiercz, *The non-existence of finite projective planes of order* 10, Canad. J. Math. **41** (1989), 1117–1123.

[30] S. Landau, *Standing the test of time: the data encryption standard*, Notices Amer. Math. Soc. **47** (2000), no. 3, 341–349.

[31] C. F. Laywine and G. L. Mullen, *Discrete Mathematics Using Latin Squares*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc., New York, 1998.

[32] C. F. Laywine, G. L. Mullen, and G. Whittle, *D-dimensional hypercubes and the Euler and MacNeish conjectures*, Monatsh. Math. **119** (1995), 223–238.

[33] A. Lempel and M. J. Weinberger, *Self-complementary normal bases in finite fields*, SIAM J. Disc. Math. **1** (1988), 193–198.

[34] H. W. Lenstra Jr. and R. J. Schoof, *Primitive normal bases for finite fields*, Math. Comp. **48** (1987), no. 177, 217–231.

[35] R. Lidl, G. L. Mullen, and G. Turnwald, *Dickson Polynomials*, Longman Scientific and Technical, Essex, United Kingdom, 1993.

[36] R. Lidl and H. Niederreiter, *Introduction to Finite Fields and Their Applications*, revised ed., Cambridge University Press, Cambridge, 1994.

[37] ———, *Finite Fields*, revised ed., Cambridge University Press, Cambridge, 1997.

[38] J. H. van Lint, *A survey of perfect codes*, Rocky Mountain J. Math. **5** (1975), 199–224.

[39] S. Lipschutz, *Linear Algebra*, second ed., McGraw-Hill, New York, 1991. Schaum's Outline Ser.

[40] H. F. MacNeish, *Euler squares*, Ann. of Math. **23** (1922), 221–227.

[41] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, Vol. 16, North-Holland, Amsterdam, 1977. Eleventh impression, 1993.

[42] B. D. McKay and I. M. Wanless, *On the number of latin squares*, Ann. Combin. **9** (2005), 335–344.

[43] A. J. Menezes (ed.), *Applications of Finite Fields*, Kluwer Academic Publishers, Boston, MA, 1993.

[44] Alfred Menezes, *Elliptic curve public key cryptosystems*, The Kluwer International Series in Engineering and Computer Science, 234, Kluwer Academic Publishers, Boston, MA, 1993.

[45] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press Series on Discrete Mathematics and its Applications, CRC Press, Boca Raton, FL, 1997.

[46] E. H. Moore, *Tactical memoranda I-III*, Amer. J. Math. **18** (1896), 264–303.

[47] G. L. Mullen, *Polynomial representation of complete sets of mutually orthogonal frequency squares of prime power order*, Discrete Math. **69** (1988), 79–84.

[48] ———, *A candidate for the next Fermat problem*, Math. Intelligencer **17** (1995), 18–22.

[49] G. L. Mullen and D. White, *A polynomial representation for logarithms in* GF(q), Acta Arithmetica **47** (1986), 255–261.

[50] I. Niven and H. S. Zuckerman, *An Introduction to the Theory of Numbers*, Wiley, New York, 1980. Fourth Ed.

[51] C. C. H. Papadimitriou, *Computational Complexity*, Addison-Wesley Publishing Company, Reading, MA, 1994.

[52] E. T. Parker, *Construction of some sets of mutually orthogonal latin squares*, Proc. Amer. Math. Soc. **10** (1959), 946–949.

[53] ———, *Orthogonal latin squares*, Proc. Nat. Acad. Sci. U. S. A. **21** (1960), 859–862.

[54] V. Pless, *Introduction to the Theory of Error-Correcting Codes*, Wiley-Interscience Series in Discrete Mathematics, North-Holland, New York, 1982.

[55] R. L. Rivest, A. Shamir, and L. Adleman, *A method for obtaining digital signatures and public-key cryptosystems*, Comm. ACM **21** (1978), no. 2, 120–126.

[56] Bruce Schneier, *Applied Cryptography: Protocols, Algorithms, and Source Code in C*, second ed., John Wiley & Sons, New York, 1996.

[57] I. Schur, *Über den Zusammenhang zwischen einem Problem der Zahlentheorie und einem Satz über algebraische Funktionen*, Sitzungsber. Preuss. Akad. Wiss. Berlin (1923), 123–134.

[58] I. E. Shparlinski, *Finite Fields: Theory and Computation*, Kluwer Academic Publishers, Boston, MA, 1999.

[59] C. Small, *Arithmetic of Finite Fields*, Monographs and Textbooks in Pure and Applied Mathematics, Dekker, New York, 1991.

[60] B. Smetaniuk, *A new construction of latin squares* I*: A proof of the Evans conjecture*, Ars Combinatoria **11** (1981), 155–172.

[61] D. R. Stinson, *A short proof of the nonexistence of a pair of orthogonal latin squares of order six*, J. Combin. Theory, Ser. A **36** (1984), 373–376.

[62] S. J. Suchower, *Polynomial representations of complete sets of frequency hyperrectangles with prime power dimensions*, J. Combin. Theory, Ser. A **62** (1993), 46–65.

[63] A. Tietäväinen, *On the nonexistence of perfect codes over finite fields*, SIAM J. Appl. Math. **24** (1973), 88–96.

[64] W. D. Wallis, *Combinatorial Designs*, Marcel Dekker, Inc., New York, 1988.

[65] Z.-X. Wan, *Lectures on Finite Fields and Galois Rings*, World Scientific, Singapore, 2003.

This page intentionally left blank

Index

This page intentionally left blank

Titles in This Series

41 **Gary L. Mullen and Carl Mummert,** Finite fields and applications, 2007

40 **Deguang Han, Keri Kornelson, David Larson, and Eric Weber,** Frames for undergraduates, 2007

39 **Alex Iosevich,** A view from the top: Analysis, combinatorics and number theory, 2007

38 **B. Fristedt, N. Jain, and N. Krylov,** Filtering and prediction: A primer, 2007

37 **Svetlana Katok,** p-adic analysis compared with real, 2007

36 **Mara D. Neusel,** Invariant theory, 2007

35 **Jörg Bewersdorff,** Galois theory for beginners: A historical perspective, 2006

34 **Bruce C. Berndt,** Number theory in the spirit of Ramanujan, 2006

33 **Rekha R. Thomas,** Lectures in geometric combinatorics, 2006

32 **Sheldon Katz,** Enumerative geometry and string theory, 2006

31 **John McCleary,** A first course in topology: Continuity and dimension, 2006

30 **Serge Tabachnikov,** Geometry and billiards, 2005

29 **Kristopher Tapp,** Matrix groups for undergraduates, 2005

28 **Emmanuel Lesigne,** Heads or tails: An introduction to limit theorems in probability, 2005

27 **Reinhard Illner, C. Sean Bohun, Samantha McCollum, and Thea van Roode,** Mathematical modelling: A case studies approach, 2005

26 **Robert Hardt, Editor,** Six themes on variation, 2004

25 **S. V. Duzhin and B. D. Chebotarevsky,** Transformation groups for beginners, 2004

24 **Bruce M. Landman and Aaron Robertson,** Ramsey theory on the integers, 2004

23 **S. K. Lando,** Lectures on generating functions, 2003

22 **Andreas Arvanitoyeorgos,** An introduction to Lie groups and the geometry of homogeneous spaces, 2003

21 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis III: Integration, 2003

20 **Klaus Hulek,** Elementary algebraic geometry, 2003

19 **A. Shen and N. K. Vereshchagin,** Computable functions, 2003

18 **V. V. Yaschenko, Editor,** Cryptography: An introduction, 2002

17 **A. Shen and N. K. Vereshchagin,** Basic set theory, 2002

16 **Wolfgang Kühnel,** Differential geometry: curves – surfaces – manifolds, second edition, 2006

15 **Gerd Fischer,** Plane algebraic curves, 2001

14 **V. A. Vassiliev,** Introduction to topology, 2001

TITLES IN THIS SERIES

13 **Frederick J. Almgren, Jr.,** Plateau's problem: An invitation to varifold geometry, 2001

12 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis II: Continuity and differentiation, 2001

11 **Michael Mesterton-Gibbons,** An introduction to game-theoretic modelling, 2000

10 **John Oprea,** The mathematics of soap films: Explorations with Maple®, 2000

9 **David E. Blair,** Inversion theory and conformal mapping, 2000

8 **Edward B. Burger,** Exploring the number jungle: A journey into diophantine analysis, 2000

7 **Judy L. Walker,** Codes and curves, 2000

6 **Gérald Tenenbaum and Michel Mendès France,** The prime numbers and their distribution, 2000

5 **Alexander Mehlmann,** The game's afoot! Game theory in myth and paradox, 2000

4 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis I: Real numbers, sequences and series, 2000

3 **Roger Knobel,** An introduction to the mathematical theory of waves, 2000

2 **Gregory F. Lawler and Lester N. Coyle,** Lectures on contemporary probability, 1999

1 **Charles Radin,** Miles of tiles, 1999

This book provides a brief and accessible introduction to the theory of finite fields and to some of their many fascinating and practical applications.

The first chapter is devoted to the theory of finite fields. After covering their construction and elementary properties, the authors discuss the trace and norm functions, bases for finite fields, and properties of polynomials over finite fields.

Each of the remaining chapters details applications. Chapter 2 deals with combinatorial topics such as the construction of sets of orthogonal latin squares, affine and projective planes, block designs, and Hadamard matrices. Chapters 3 and 4 provide a number of constructions and basic properties of error-correcting codes and cryptographic systems using finite fields.

Each chapter includes a set of exercises of varying levels of difficulty which help to further explain and motivate the material. Appendix A provides a brief review of the basic number theory and abstract algebra used in the text, as well as exercises related to this material. Appendix B provides hints and partial solutions for many of the exercises in each chapter. A list of 65 references to further reading and to additional topics related to the book's material is also included.

Intended for advanced undergraduate students, it is suitable both for classroom use and for individual study.

For additional information and updates on this book, visit

www.ams.org/bookpages/stml-41

ISBN 978-0-8218-4418-2

9 780821 844182

STML/41