Undergraduate Texts in Mathematics

Editors
S. Axler
K.A. Ribet

Undergraduate Texts in Mathematics

Abbott: Understanding Analysis.
Anglin: Mathematics: A Concise History and Philosophy.
Readings in Mathematics.
Anglin/Lambek: The Heritage of Thales.
Readings in Mathematics.
Apostol: Introduction to Analytic Number Theory. Second edition.
Armstrong: Basic Topology.
Armstrong: Groups and Symmetry.
Axler: Linear Algebra Done Right. Second edition.
Beardon: Limits: A New Approach to Real Analysis.
Bak/Newman: Complex Analysis. Second edition.
Banchoff/Wermer: Linear Algebra Through Geometry. Second edition.
Beck/Robins: Computing the Continuous Discretely
Berberian: A First Course in Real Analysis.
Bix: Conics and Cubics: A Concrete Introduction to Algebraic Curves. Second edition.
Brèmaud: An Introduction to Probabilistic Modeling.
Bressoud: Factorization and Primality Testing.
Bressoud: Second Year Calculus.
Readings in Mathematics.
Brickman: Mathematical Introduction to Linear Programming and Game Theory.
Browder: Mathematical Analysis: An Introduction.
Buchmann: Introduction to Cryptography. Second Edition.
Buskes/van Rooij: Topological Spaces: From Distance to Neighborhood.
Callahan: The Geometry of Spacetime: An Introduction to Special and General Relavitity.
Carter/van Brunt: The Lebesgue– Stieltjes Integral: A Practical Introduction.
Cederberg: A Course in Modern Geometries. Second edition.
Chambert-Loir: A Field Guide to Algebra
Childs: A Concrete Introduction to Higher Algebra. Second edition.
Chung/AitSahlia: Elementary Probability Theory: With Stochastic Processes and an Introduction to Mathematical Finance. Fourth edition.
Cox/Little/O'Shea: Ideals, Varieties, and Algorithms. Second edition.
Croom: Basic Concepts of Algebraic Topology.
Cull/Flahive/Robson: Difference Equations. From Rabbits to Chaos
Curtis: Linear Algebra: An Introductory Approach. Fourth edition.
Daepp/Gorkin: Reading, Writing, and Proving: A Closer Look at Mathematics.
Devlin: The Joy of Sets: Fundamentals of-Contemporary Set Theory. Second edition.
Dixmier: General Topology.
Driver: Why Math?
Ebbinghaus/Flum/Thomas: Mathematical Logic. Second edition.
Edgar: Measure, Topology, and Fractal Geometry. Second edition.
Elaydi: An Introduction to Difference Equations. Third edition.
Erdős/Surányi: Topics in the Theory of Numbers.
Estep: Practical Analysis on One Variable.
Exner: An Accompaniment to Higher Mathematics.
Exner: Inside Calculus.
Fine/Rosenberger: The Fundamental Theory of Algebra.
Fischer: Intermediate Real Analysis.
Flanigan/Kazdan: Calculus Two: Linear and Nonlinear Functions. Second edition.
Fleming: Functions of Several Variables. Second edition.
Foulds: Combinatorial Optimization for Undergraduates.
Foulds: Optimization Techniques: An Introduction.
Franklin: Methods of Mathematical Economics.
Frazier: An Introduction to Wavelets Through Linear Algebra.
Gamelin: Complex Analysis.
Ghorpade/Limaye: A Course in Calculus and Real Analysis
Gordon: Discrete Probability.
Hairer/Wanner: Analysis by Its History.
Readings in Mathematics.
Halmos: Finite-Dimensional Vector Spaces. Second edition.
Halmos: Naive Set Theory.
Hämmerlin/Hoffmann: Numerical Mathematics.
Readings in Mathematics.
Harris/Hirst/Mossinghoff: Combinatorics and Graph Theory.
Hartshorne: Geometry: Euclid and Beyond.
Hijab: Introduction to Calculus and Classical Analysis. Second edition.
Hilton/Holton/Pedersen: Mathematical Reflections: In a Room with Many Mirrors.
Hilton/Holton/Pedersen: Mathematical Vistas: From a Room with Many Windows.
Hoffstein/Pipher/Silverman: An Introduction to Mathematical Cryptography.
Iooss/Joseph: Elementary Stability and Bifurcation Theory. Second Edition.

(continued after index)

Jeffrey Hoffstein
Jill Pipher
Joseph H. Silverman

An Introduction to Mathematical Cryptography

Springer

Jeffrey Hoffstein
Department of Mathematics
Brown University
151 Thayer St.
Providence, RI 02912
USA
jhoff@math.brown.edu

Jill Pipher
Department of Mathematics
Brown University
151 Thayer St.
Providence, RI 02912
USA
jpipher@math.brown.edu

Joseph H. Silverman
Department of Mathematics
Brown University
151 Thayer St.
Providence, RI 02912
USA
jhs@math.brown.edu

ISBN: 978-0-387-77993-5 e-ISBN: 978-0-387-77994-2
DOI: 10.1007/978-0-387-77994-2

Library of Congress Control Number: 2008923038

Mathematics Subject Classification (2000): 94A60, 11T71, 14G50, 68P25

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The creation of public key cryptography by Diffie and Hellman in 1976 and the subsequent invention of the RSA public key cryptosystem by Rivest, Shamir, and Adleman in 1978 are watershed events in the long history of secret communications. It is hard to overestimate the importance of public key cryptosystems and their associated digital signature schemes in the modern world of computers and the Internet. This book provides an introduction to the theory of public key cryptography and to the mathematical ideas underlying that theory.

Public key cryptography draws on many areas of mathematics, including number theory, abstract algebra, probability, and information theory. Each of these topics is introduced and developed in sufficient detail so that this book provides a self-contained course for the beginning student. The only prerequisite is a first course in linear algebra. On the other hand, students with stronger mathematical backgrounds can move directly to cryptographic applications and still have time for advanced topics such as elliptic curve pairings and lattice-reduction algorithms.

Among the many facets of modern cryptography, this book chooses to concentrate primarily on public key cryptosystems and digital signature schemes. This allows for an in-depth development of the necessary mathematics required for both the construction of these schemes and an analysis of their security. The reader who masters the material in this book will not only be well prepared for further study in cryptography, but will have acquired a real understanding of the underlying mathematical principles on which modern cryptography is based.

Topics covered in this book include Diffie–Hellman key exchange, discrete logarithm based cryptosystems, the RSA cryptosystem, primality testing, factorization algorithms, probability theory, information theory, collision algorithms, elliptic curves, elliptic curve cryptography, pairing-based cryptography, lattices, lattice-based cryptography, the NTRU cryptosystem, and digital signatures. A final chapter very briefly describes some of the many other aspects of modern cryptography (hash functions, pseudorandom number generators, zero-knowledge proofs, digital cash, AES,...) and serves to point the reader toward areas for further study.

Electronic Resources: The interested reader will find additional material and a list of errata on the Mathematical Cryptography home page:

`www.math.brown.edu/~jhs/MathCryptoHome.html`

This web page includes many of the numerical exercises in the book, allowing the reader to cut and paste them into other programs, rather than having to retype them.

No book is ever free from error or incapable of being improved. We would be delighted to receive comments, good or bad, and corrections from our readers. You can send mail to us at

`mathcrypto@math.brown.edu`

Acknowledgments: We, the authors, would like the thank the following individuals for test-driving this book and for the many corrections and helpful suggestions that they and their students provided: Liat Berdugo, Alexander Collins, Samuel Dickman, Michael Gartner, Nicholas Howgrave-Graham, Su-Ion Ih, Saeja Kim, Yuji Kosugi, Yesem Kurt, Michelle Manes, Victor Miller, David Singer, William Whyte. In addition, we would like to thank the many students at Brown University who took Math 158 and helped us improve the exposition of this book.

Contents

Introduction

A Principal Goal of (Public Key) Cryptography
is to allow two people to exchange confidential information, even if they have never met and can communicate only via a channel that is being monitored by an adversary.

The security of communications and commerce in a digital age relies on the modern incarnation of the ancient art of codes and ciphers. Underlying the birth of modern cryptography is a great deal of fascinating mathematics, some of which has been developed for cryptographic applications, but much of which is taken from the classical mathematical canon. The principal goal of this book is to introduce the reader to a variety of mathematical topics while simultaneously integrating the mathematics into a description of modern public key cryptography.

For thousands of years, all codes and ciphers relied on the assumption that the people attempting to communicate, call them Bob and Alice, shared a *secret key* that their adversary, call her Eve, did not possess. Bob would use the secret key to encrypt his message, Alice would use the same secret key to decrypt the message, and poor Eve, not knowing the secret key, would be unable to perform the decryption. A disadvantage of these *private key cryptosystems* is that Bob and Alice need to exchange the secret key before they can get started.

During the 1970s, the astounding idea of *public key cryptography* burst upon the scene.[1] In a public key cryptosystem, Alice has two keys, a public encryption key K^{Pub} and a private (secret) decryption key K^{Pri}. Alice publishes her public key K^{Pub}, and then Adam and Bob and Carl and everyone else can use K^{Pub} to encrypt messages and send them to Alice. The idea underlying public key cryptgraphy is that although everyone in the world knows K^{Pub} and can use it to encrypt messages, only Alice, who knows the private key K^{Pri}, is able to decrypt messages.

The advantages of a public key cryptosystem are manifold. For example, Bob can send Alice an encrypted message even if they have never previously been in direct contact. But although public key cryptography is a fascinating

[1] A brief history of cryptography is given is Sections 1.6, 2.1, 5.5, and 6.7.

theoretical concept, it is not at all clear how one might create a public key cryptosystem. It turns out that public key cryptosystems can be based on hard mathematical problems. More precisely, one looks for a mathematical problem that is hard to solve a priori, but that becomes easy to solve if one knows some extra piece of information.

Of course, private key cryptosystems have not disappeared. Indeed, they are more important than ever, since they tend to be significantly more efficient than public key cryptosystems. Thus in practice, if Bob wants to send Alice a long message, he first uses a public key cryptosystem to send Alice the key for a private key cryptosystem, and then he uses the private key cryptosystem to encrypt his message. The most efficient modern private key cryptosystems, such as DES and AES, rely for their security on repeated application of various mixing operations that are hard to unmix without the private key. Thus although the subject of private key cryptography is of both theoretical and practical importance, the connection with fundamental underlying mathematical ideas is much less pronounced than it is with public key cryptosystems. For that reason, this book concentrates almost exclusively on public key cryptography.

Modern mathematical cryptography draws on many areas of mathematics, including especially number theory, abstract algebra (groups, rings, fields), probability, statistics, and information theory, so the prerequisites for studying the subject can seem formidable. By way of contrast, the prerequisites for reading this book are minimal, because we take the time to introduce each required mathematical topic in sufficient depth as it is needed. Thus this book provides a self-contained treatment of mathematical cryptography for the reader with limited mathematical background. And for those readers who have taken a course in, say, number theory or abstract algebra or probability, we suggest briefly reviewing the relevant sections as they are reached and then moving on directly to the cryptographic applications.

This book is not meant to be a comprehensive source for all things cryptographic. In the first place, as already noted, we concentrate on public key cryptography. But even within this domain, we have chosen to pursue a small selection of topics to a reasonable mathematical depth, rather than providing a more superficial description of a wider range of subjects. We feel that any reader who has mastered the material in this book will not only be well prepared for further study in cryptography, but will have acquired a real understanding of the underlying mathematical principles on which modern cryptography is based.

However, this does not mean that the omitted topics are unimportant. It simply means that there is a limit to the amount of material that can be included in a book (or course) of reasonable length. As in any text, the choice of particular topics reflects the authors' tastes and interests. For the convenience of the reader, the final chapter contains a brief survey of areas for further study.

A Guide to Mathematical Topics: This book includes a significant amount of mathematical material on a variety of topics that are useful in cryptography. The following list is designed to help coordinate the topics that we cover with subjects that the class or reader may have already studied.

Congruences, primes, and finite fields	§§1.2, 1.3, 1.4, 1.5, 2.10.4
The Chinese remainder theorem	§2.8
Euler's formula	§3.1
Primality testing	§3.4
Quadratic reciprocity	§3.9
Factorization methods	§§3.5, 3.6, 3.7, 5.6
Discrete logarithms	§§2.2, 3.8, 4.4, 4.5, 5.3
Group theory	§2.5
Rings, polynomials, and quotient rings	§2.10, 6.9
Combinatorics and probability	§§4.1, 4.3
Information and complexity theory	§§4.6, 4.7
Elliptic curves	§§5.1, 5.2, 5.7, 5.8
Linear algebra	§6.3
Lattices	§§6.4, 6.5, 6.6, 6.12

Intended Audience and Prerequisites: This book provides a self-contained introduction to public key cryptography and to the underlying mathematics that is required for the subject. It is suitable as a text for advanced undergraduates and beginning graduate students. We provide enough background material so that the book can be used in courses for students with no previous exposure to abstract algebra or number theory. For classes in which the students have a stronger background, the basic mathematical material may be omitted, leaving time for some of the more advanced topics.

The formal prerequisites for this book are few, beyond a facility with high school algebra and, in Chapter 5, analytic geometry. Elementary calculus is used here and there in a minor way, but is not essential, and linear algebra is used in a small way in Chapter 3 and more extensively in Chapter 6. No previous knowledge is assumed for mathematical topics such as number theory, abstract algebra, and probability theory that play a fundamental role in modern cryptography. They are covered in detail as needed.

However, it must be emphasized that this is a mathematics book with its share of formal definitions and theorems and proofs. Thus it is expected that the reader has a certain level of mathematical sophistication. In particular, students who have previously taken a proof-based mathematics course will find the material easier than those without such background. On the other hand, the subject of cryptography is so appealing that this book makes a good text for an introduction-to-proofs course, with the understanding that the instructor will need to cover the material more slowly to allow the students time to become comfortable with proof-based mathematics.

Suggested Syllabus: This book contains considerably more material than can be comfortably covered by beginning students in a one semester course. However, for more advanced students who have already taken courses in number theory and abstract algebra, it should be possible to do most of the remaining material. We suggest covering the majority of the topics in Chapters 1, 2, and 3, possibly omitting some of the more technical topics, the optional material on the Vigènere cipher, and the section on ring theory, which is not used until much later in the book. The next four chapters on information theory (Chapter 4), elliptic curves (Chapter 5), lattices (Chapter 6), and digital signatures (Chapter 7) are mostly independent of one another, so the instructor has the choice of covering one or two of them in detail or all of them in less depth. We offer the following syllabus as an example of one of the many possibilities. We have indicated that some sections are optional. Covering the optional material leaves less time at the end for the later chapters.

Chapter 1 An Introduction to Cryptography.
Cover all sections.

Chapter 2 Discrete Logarithms and Diffie–Hellman.
Cover Sections 2.1–2.7. Optionally cover the more mathematically sophisticated Sections 2.8–2.9 on the Pohlig–Hellman algorithm. Omit Section 2.10 on first reading.

Chapter 3 Integer Factorization and RSA.
Cover Sections 3.1–3.5 and Sections 3.9–3.10. Optionally, cover the more mathematically sophisticated Sections 3.6–3.8, dealing with smooth numbers, sieves, and the index calculus.

Chapter 4 Probability Theory and Information Theory.
Cover Sections 4.1, 4.3, and 4.4. Optionally cover the more mathematically sophisticated sections on Pollard's ρ method (Section 4.5), information theory (Section 4.6), and complexity theory (Section 4.7). The material on the Vigenère cipher in Section 4.2 nicely illustrates the use of statistics theory in cryptanalysis, but is somewhat off the main path.

Chapter 5 Elliptic Curves.
Cover Sections 5.1–5.4. Cover other sections as time permits, but note that Sections 5.7–5.10 on pairings require finite fields of prime power order, which are described in Section 2.10.4.

Chapter 6 Lattices and Cryptography.
Cover Sections 6.1–6.8. (If time is short, it is possible to omit either or both of Sections 6.1 and 6.2.) Cover either Sections 6.12–6.13 or Sections 6.10–6.11, or both, as time permits. Note that Sections 6.10–6.11 on NTRU require the material on polynomial rings and quotient rings covereed in Section 2.10.

Chapter 7 Digital Signatures.
Cover Sections 7.1–7.2. Cover the remaining sections as time permits.

Chapter 8 Additional Topics in Cryptography.
The material in this chapter points the reader toward other important areas of cryptography. It provides a good list of topics and references for student term papers and presentations.

Further Notes for the Instructor: Depending on how much of the harder mathematical material in Chapters 2–4 is covered, there may not be time to delve into both Chapters 5 and 6, so the instructor may need to omit either elliptic curves or lattices in order to fit the other material into one semester.

We feel that it is helpful for students to gain an appreciation of the origins of their subject, so we have scattered a handful of sections throughout the book containing some brief comments on the history of cryptography. Instructors who want to spend more time on mathematics may omit these sections without affecting the mathematical narrative.

Chapter 1

An Introduction to Cryptography

1.1 Simple substitution ciphers

As Julius Caesar surveys the unfolding battle from his hilltop outpost, an exhausted and disheveled courier bursts into his presence and hands him a sheet of parchment containing gibberish:

`j s j r d k f q q n s l g f h p g w j f p y m w t z l m n r r n s j s y q z h n z x`

Within moments, Julius sends an order for a reserve unit of charioteers to speed around the left flank and exploit a momentary gap in the opponent's formation.

How did this string of seemingly random letters convey such important information? The trick is easy, once it is explained. Simply take each letter in the message and shift it five letters up the alphabet. Thus `j` in the *ciphertext* becomes `e` in the *plaintext*,[1] because `e` is followed in the alphabet by `f`,`g`,`h`,`i`,`j`. Applying this procedure to the entire ciphertext yields

```
j s j r d k f q q n s l g f h p g w j f p y m w t z l m n r r n s j s y q z h n z x
e n e m y f a l l i n g b a c k b r e a k t h r o u g h i m m i n e n t l u c i u s
```

The second line is the decrypted plaintext, and breaking it into words and supplying the appropriate punctuation, Julius reads the message

`Enemy falling back. Breakthrough imminent. Lucius.`

There remains one minor quirk that must be addressed. What happens when Julius finds a letter such as `d`? There is no letter appearing five letters before `d`

[1]The *plaintext* is the original message in readable form and the *ciphertext* is the encrypted message.

J. Hoffstein et al., *An Introduction to Mathematical Cryptography*,
DOI: 10.1007/978-0-387-77994-2_1,

in the alphabet. The answer is that he must wrap around to the end of the alphabet. Thus d is replaced by y, since y is followed by z,a,b,c,d.

This wrap-around effect may be conveniently visualized by placing the alphabet abcd...xyz around a circle, rather than in a line. If a second alphabet circle is then placed within the first circle and the inner circle is rotated five letters, as illustrated in Figure 1.1, the resulting arrangement can be used to easily encrypt and decrypt Caesar's messages. To decrypt a letter, simply find it on the inner wheel and read the corresponding plaintext letter from the outer wheel. To encrypt, reverse this process: find the plaintext letter on the outer wheel and read off the ciphertext letter from the inner wheel. And note that if you build a cipherwheel whose inner wheel spins, then you are no longer restricted to always shifting by exactly five letters. Cipher wheels of this sort have been used for centuries.[2]

Although the details of the preceding scene are entirely fictional, and in any case it is unlikely that a message to a Roman general would have been written in modern English(!), there is evidence that Caesar employed this early method of cryptography, which is sometimes called the *Caesar cipher* in his honor. It is also sometimes referred to as a *shift cipher*, since each letter in the alphabet is shifted up or down. *Cryptography*, the methodology of concealing the content of messages, comes from the Greek root words kryptos, meaning hidden,[3] and graphikos, meaning writing. The modern scientific study of cryptography is sometimes referred to as *cryptology*.

In the Caesar cipher, each letter is replaced by one specific substitute letter. However, if Bob encrypts a message for Alice[4] using a Caesar cipher and allows the encrypted message to fall into Eve's hands, it will take Eve very little time to decrypt it. All she needs to do is try each of the 26 possible shifts.

Bob can make his message harder to attack by using a more complicated replacement scheme. For example, he could replace every occurrence of a by z and every occurrence of z by a, every occurrence of b by y and every occurrence of y by b, and so on, exchanging each pair of letters c $\leftrightarrow$ x,..., m $\leftrightarrow$ n.

This is an example of a *simple substitution cipher*, that is, a cipher in which each letter is replaced by another letter (or some other type of symbol). The Caesar cipher is an example of a simple substitution cipher, but there are many simple substitution ciphers other than the Caesar cipher. In fact, a

[2]A cipher wheel with mixed up alphabets and with encryption performed using different offsets for different parts of the message is featured in a 15$^{\text{th}}$ century monograph by Leon Batista Alberti [58].

[3]The word cryptic, meaning hidden or occult, appears in 1638, while crypto- as a prefix for concealed or secret makes its appearance in 1760. The term cryptogram appears much later, first occurring in 1880.

[4]In cryptography, it is traditional for Bob and Alice to exchange confidential messages and for their adversary Eve, the eavesdropper, to intercept and attempt to read their messages. This makes the field of cryptography much more personal than other areas of mathematics and computer science, whose denizens are often X and Y!

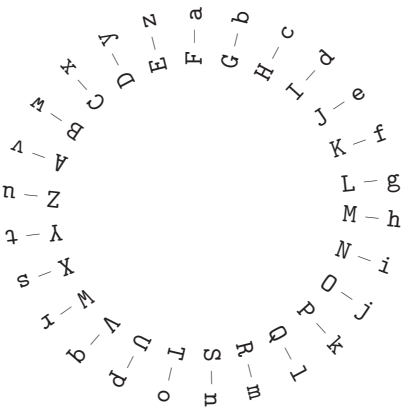

Figure 1.1: A cipher wheel with an offset of five letters

simple substitution cipher may be viewed as a rule or function

$$\{\texttt{a,b,c,d,e,...,x,y,z}\} \longrightarrow \{\texttt{A,B,C,D,E,...,X,Y,Z}\}$$

assigning each plaintext letter in the domain a different ciphertext letter in the range. (To make it easier to distinguish the plaintext from the ciphertext, we write the plaintext using lowercase letters and the ciphertext using uppercase letters.) Note that in order for decryption to work, the encryption function must have the property that no two plaintext letters go to the same ciphertext letter. A function with this property is said to be *one-to-one* or *injective*.

A convenient way to describe the encryption function is to create a table by writing the plaintext alphabet in the top row and putting each ciphertext letter below the corresponding plaintext letter.

Example 1.1. A simple substitution encryption table is given in Table 1.1. The ciphertext alphabet (the uppercase letters in the bottom row) is a randomly chosen permutation of the 26 letters in the alphabet. In order to encrypt the plaintext message

```
Four score and seven years ago,
```

we run the words together, look up each plaintext letter in the encryption table, and write the corresponding ciphertext letter below.

f	o	u	r	s	c	o	r	e	a	n	d	s	e	v	e	n	y	e	a	r	s	a	g	o
N	U	R	B	K	S	U	B	V	C	G	Q	K	V	E	V	G	Z	V	C	B	K	C	F	U

It is then customary to write the ciphertext in five-letter blocks:

```
NURBK SUBVC GQKVE VGZVC BKCFU
```

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
C	I	S	Q	V	N	F	O	W	A	X	M	T	G	U	H	P	B	K	L	R	E	Y	D	Z	J

Table 1.1: Simple substitution encryption table

j	r	a	x	v	g	n	p	b	z	s	t	l	f	h	q	d	u	c	m	o	e	i	k	w	y
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

Table 1.2: Simple substitution decryption table

Decryption is a similar process. Suppose that we receive the message

```
GVVQG VYKCM CQQBV KKWGF SCVKV B
```

and that we know that it was encrypted using Table 1.1. We can reverse the encryption process by finding each ciphertext letter in the second row of Table 1.1 and writing down the corresponding letter from the top row. However, since the letters in the second row of Table 1.1 are all mixed up, this is a somewhat inefficient process. It is better to make a decryption table in which the ciphertext letters in the lower row are listed in alphabetical order and the corresponding plaintext letters in the upper row are mixed up. We have done this in Table 1.2. Using this table, we easily decrypt the message.

G	V	V	Q	G	V	Y	K	C	M	C	Q	Q	B	V	K	K	W	G	F	S	C	V	K	V	B
n	e	e	d	n	e	w	s	a	l	a	d	d	r	e	s	s	i	n	g	c	a	e	s	e	r

Putting in the appropriate word breaks and some punctuation reveals an urgent request!

```
Need new salad dressing. -Caesar
```

1.1.1 Cryptanalysis of simple substitution ciphers

How many different simple substitution ciphers exist? We can count them by enumerating the possible ciphertext values for each plaintext letter. First we assign the plaintext letter `a` to one of the 26 possible ciphertext letters `A`–`Z`. So there are 26 possibilities for `a`. Next, since we are not allowed to assign `b` to the same letter as `a`, we may assign `b` to any one of the remaining 25 ciphertext letters. So there are $26 \cdot 25 = 650$ possible ways to assign `a` and `b`. We have now used up two of the ciphertext letters, so we may assign `c` to any one of the remaining 24 ciphertext letters. And so on.... Thus the total number of ways to assign the 26 plaintext letters to the 26 ciphertext letters, using each ciphertext letter only once, is

$$26 \cdot 25 \cdot 24 \cdots 4 \cdot 3 \cdot 2 \cdot 1 = 26! = 403291461126605635584000000.$$

There are thus more than 10^{26} different simple substitution ciphers. Each associated encryption table is known as a *key*.

Suppose that Eve intercepts one of Bob's messages and that she attempts to decrypt it by trying every possible simple substitution cipher. The process of decrypting a message without knowing the underlying key is called *cryptanalysis*. If Eve (or her computer) is able to check one million cipher alphabets per second, it would still take her more than 10^{13} years to try them all.[5] But the age of the universe is estimated to be on the order of 10^{10} years. Thus Eve has almost no chance of decrypting Bob's message, which means that Bob's message is secure and he has nothing to worry about![6] Or does he?

It is time for an important lesson in the practical side of the science of cryptography:

> Your opponent always uses her best strategy to defeat you, not the strategy that you want her to use. Thus the security of an encryption system depends on the best known method to break it. As new and improved methods are developed, the level of security can only get worse, never better.

Despite the large number of possible simple substitution ciphers, they are actually quite easy to break, and indeed many newspapers and magazines feature them as a companion to the daily crossword puzzle. The reason that Eve can easily cryptanalyze a simple substitution cipher is that the letters in the English language (or any other human language) are not random. To take an extreme example, the letter `q` in English is virtually always followed by the letter `u`. More useful is the fact that certain letters such as `e` and `t` appear far more frequently than other letters such as `f` and `c`. Table 1.3 lists the letters with their typical frequencies in English text. As you can see, the most frequent letter is `e`, followed by `t`, `a`, `o`, and `n`.

Thus if Eve counts the letters in Bob's encrypted message and makes a frequency table, it is likely that the most frequent letter will represent `e`, and that `t`, `a`, `o`, and `n` will appear among the next most frequent letters. In this way, Eve can try various possibilities and, after a certain amount of trial and error, decrypt Bob's message.

In the remainder of this section we illustrate how to cryptanalyze a simple substitution cipher by decrypting the message given in Table 1.4. Of course the end result of defeating a simple substitution cipher is not our main goal here. Our key point is to introduce the idea of statistical analysis, which will prove to

[5] Do you see how we got 10^{13} years? There are $60 \cdot 60 \cdot 24 \cdot 365$ seconds in a year, and 26! divided by $10^6 \cdot 60 \cdot 60 \cdot 24 \cdot 365$ is approximately $10^{13.107}$.

[6] The assertion that a large number of possible keys, in and of itself, makes a cryptosystem secure, has appeared many times in history and has equally often been shown to be fallacious.

By decreasing frequency			
E	13.11%	M	2.54%
T	10.47%	U	2.46%
A	8.15%	G	1.99%
O	8.00%	Y	1.98%
N	7.10%	P	1.98%
R	6.83%	W	1.54%
I	6.35%	B	1.44%
S	6.10%	V	0.92%
H	5.26%	K	0.42%
D	3.79%	X	0.17%
L	3.39%	J	0.13%
F	2.92%	Q	0.12%
C	2.76%	Z	0.08%

In alphabetical order			
A	8.15%	N	7.10%
B	1.44%	O	8.00%
C	2.76%	P	1.98%
D	3.79%	Q	0.12%
E	13.11%	R	6.83%
F	2.92%	S	6.10%
G	1.99%	T	10.47%
H	5.26%	U	2.46%
I	6.35%	V	0.92%
J	0.13%	W	1.54%
K	0.42%	X	0.17%
L	3.39%	Y	1.98%
M	2.54%	Z	0.08%

Table 1.3: Frequency of letters in English text

```
LOJUM YLJME PDYVJ QXTDV SVJNL DMTJZ WMJGG YSNDL UYLEO SKDVC
GEPJS MDIPD NEJSK DNJTJ LSKDL OSVDV DNGYN VSGLL OSCIO LGOYG
ESNEP CGYSN GUJMJ DGYNK DPPYX PJDGG SVDNT WMSWS GYLYS NGSKJ
CEPYQ GSGLD MLPYN IUSCP QOYGM JGCPL GDWWJ DMLSL OJCNY NYLYD
LJQLO DLCNL YPLOJ TPJDM NJQLO JWMSE JGGJG XTUOY EOOJO DQDMM
YBJQD LLOJV LOJTV YIOLU JPPES NGYQJ MOYVD GDNJE MSVDN EJM
```

Table 1.4: A simple substitution cipher to cryptanalyze

have many applications throughout cryptography. Although for completeness we provide full details, the reader may wish to skim this material.

There are 298 letters in the ciphertext. The first step is to make a frequency table listing how often each ciphertext letter appears.

	J	L	D	G	Y	S	O	N	M	P	E	V	Q	C	T	W	U	K	I	X	Z	B	A	F	R	H
Freq	32	28	27	24	23	22	19	18	17	15	12	12	8	8	7	6	6	5	4	3	1	1	0	0	0	0
%	11	9	9	8	8	7	6	6	6	5	4	4	3	3	2	2	2	2	1	1	0	0	0	0	0	0

Table 1.5: Frequency table for Table 1.4—Ciphertext length: 298

The ciphertext letter J appears most frequently, so we make the provisional guess that it corresponds to the plaintext letter e. The next most frequent ciphertext letters are L (28 times) and D (27 times), so we might guess from Table 1.3 that they represent t and a. However, the letter frequencies in a short message are unlikely to exactly match the percentages in Table 1.3. All that we can say is that among the ciphertext letters L, D, G, Y, and S are likely to appear several of the plaintext letters t, a, o, n, and r.

th	he	an	re	er	in	on	at	nd	st	es	en	of	te	ed
168	132	92	91	88	86	71	68	61	53	52	51	49	46	46

(a) Most common English bigrams (frequency per 1000 words)

LO	OJ	GY	DN	VD	YL	DL	DM	SN	KD	LY	NG	OY	JD	SK	EP	JG	SV	JM	JQ
9	7	6 each		5 each					4 each										

(b) Most common bigrams appearing in the ciphertext in Table 1.4

Table 1.6: Bigram frequencies

There are several ways to proceed. One method is to look at *bigrams*, which are pairs of consecutive letters. Table 1.6(a) lists the bigrams that most frequently appear in English, and Table 1.6(b) lists the ciphertext bigrams that appear most frequently in our message. The ciphertext bigrams `LO` and `OJ` appear frequently. We have already guessed that `J` = `e`, and based on its frequency we suspect that `L` is likely to represent one of the letters `t`, `a`, `o`, `n`, or `r`. Since the two most frequent English bigrams are `th` and `he`, we make the tentative identifications

$$\texttt{LO} = \texttt{th} \qquad \text{and} \qquad \texttt{OJ} = \texttt{he}.$$

We substitute the guesses `J` = `e`, `L` = `t`, and `O` = `h`, into the ciphertext, writing the putative plaintext letter below the corresponding ciphertext letter.

```
LOJUM YLJME PDYVJ QXTDV SVJNL DMTJZ WMJGG YSNDL UYLEO SKDVC
the-- -te-- ----e ----- --e-t ---e- --e-- ----t --t-h -----
GEPJS MDIPD NEJSK DNJTJ LSKDL OSVDV DNGYN VSGLL OSCIO LGOYG
---e- ----- --e-- --e-e t---t h---- ----- ---tt h---h t-h--
ESNEP CGYSN GUJMJ DGYNK DPPYX PJDGG SVDNT WMSWS GYLYS NGSKJ
----- ----- --e-e ----- ----- -e--- ----- ----- --t-- ----e
CEPYQ GSGLD MLPYN IUSCP QOYGM JGCPL GDWWJ DMLSL OJCNY NYLYD
----- ---t- -t--- ----- -h--- e---t ----e --t-t he--- --t--
LJQLO DLCNL YPLOJ TPJDM NJQLO JWMSE JGGJG XTUOY EOOJO DQDMM
te-th -t--t --the --e-- -e-th e---- e--e- ---h- -hheh -----
YBJQD LLOJV LOJTV YIOLU JPPES NGYQJ MOYVD GDNJE MSVDN EJM
--e-- tthe- the-- --ht- e---- ----e -h--- ---e- ----- -e-
```

At this point, we can look at the fragments of plaintext and attempt to guess some common English words. For example, in the second line we see the three blocks

```
VSGLL OSCIO LGOYG,
---tt h---h t-h--.
```

Looking at the fragment th---ht, we might guess that this is the word thought, which gives three more equivalences,

$$\mathtt{S} = \mathtt{o}, \qquad \mathtt{C} = \mathtt{u}, \qquad \mathtt{I} = \mathtt{g}.$$

This yields

```
LOJUM YLJME PDYVJ QXTDV SVJNL DMTJZ WMJGG YSNDL UYLEO SKDVC
the-- -te-- ----e ----- o-e-t ---e- --e-- -o--t --t-h o---u
GEPJS MDIPD NEJSK DNJTJ LSKDL OSVDV DNGYN VSGLL OSCIO LGOYG
---eo --g-- --eo- --e-e to--t ho--- ----- -o-tt hough t-h--
ESNEP CGYSN GUJMJ DGYNK DPPYX PJDGG SVDNT WMSWS GYLYS NGSKJ
-o--- u--o- --e-e ----- ----- -e--- o---- --o-o --t-o --o-e
CEPYQ GSGLD MLPYN IUSCP QOYGM JGCPL GDWWJ DMLSL OJCNY NYLYD
u---- -o-t- -t--- g-ou- -h--- e-u-t ----e --tot heu-- --t--
LJQLO DLCNL YPLOJ TPJDM NJQLO JWMSE JGGJG XTUOY EOOJO DQDMM
te-th -tu-t --the --e-- -e-th e--o- e--e- ---h- -hheh -----
YBJQD LLOJV LOJTV YIOLU JPPES NGYQJ MOYVD GDNJE MSVDN EJM
--e-- tthe- the-- -ght- e---o ----e -h--- ---e- -o--- -e-
```

Now look at the three letters ght in the last line. They must be preceded by a vowel, and the only vowels left are a and i, so we guess that Y = i. Then we find the letters itio in the third line, and we guess that they are followed by an n, which gives N = n. (There is no reason that a letter cannot represent itself, although this is often forbidden in the puzzle ciphers that appear in newspapers.) We now have

```
LOJUM YLJME PDYVJ QXTDV SVJNL DMTJZ WMJGG YSNDL UYLEO SKDVC
the-- ite-- --i-e ----- o-ent ---e- --e-- ion-t -it-h o---u
GEPJS MDIPD NEJSK DNJTJ LSKDL OSVDV DNGYN VSGLL OSCIO LGOYG
---eo --g-- n-eo- -ne-e to--t ho--- -n-in -o-tt hough t-hi-
ESNEP CGYSN GUJMJ DGYNK DPPYX PJDGG SVDNT WMSWS GYLYS NGSKJ
-on-- u-ion --e-e --in- ---i- -e--- o--n- --o-o -itio n-o-e
CEPYQ GSGLD MLPYN IUSCP QOYGM JGCPL GDWWJ DMLSL OJCNY NYLYD
u--i- -o-t- -t-in g-ou- -hi-- e-u-t ----e --tot heuni niti-
LJQLO DLCNL YPLOJ TPJDM NJQLO JWMSE JGGJG XTUOY EOOJO DQDMM
te-th -tunt i-the --e-- ne-th e--o- e--e- ---hi -hheh -----
YBJQD LLOJV LOJTV YIOLU JPPES NGYQJ MOYVD GDNJE MSVDN EJM
i-e-- tthe- the-- ight- e---o n-i-e -hi-- --ne- -o--n -e-
```

So far, we have reconstructed the following plaintext/ciphertext pairs:

	J	L	D	G	Y	S	O	N	M	P	E	V	Q	C	T	W	U	K	I	X	Z	B	A	F	R	H
	e	t	-	-	i	o	h	n	-	-	-	-	-	u	-	-	-	-	g	-	-	-	-	-	-	-
Freq	32	28	27	24	23	22	19	18	17	15	12	12	8	8	7	6	6	5	4	3	1	1	0	0	0	0

Recall that the most common letters in English (Table 1.3) are, in order of decreasing frequency,

e, t, a, o, n, r, i, s, h.

We have already assigned ciphertext values to e, t, o, n, i, h, so we guess that D and G represent two of the three letters a, r, s. In the third line we notice that GYLYSN gives -ition, so clearly G must be s. Similarly, on the fifth line we have LJQLO DLCNL equal to te-th -tunt, so D must be a, not r. Substituting these new pairs G = s and D = a gives

```
LOJUM YLJME PDYVJ QXTDV SVJNL DMTJZ WMJGG YSNDL UYLEO SKDVC
the-- ite-- -ai-e ---a- o-ent a--e- --ess ionat -it-h o-a-u
GEPJS MDIPD NEJSK DNJTJ LSKDL OSVDV DNGYN VSGLL OSCIO LGOYG
s--eo -ag-a n-eo- ane-e to-at ho-a- ansin -ostt hough tshis
ESNEP CGYSN GUJMJ DGYNK DPPYX PJDGG SVDNT WMSWS GYLYS NGSKJ
-on-- usion s-e-e asin- a--i- -eass o-an- --o-o sitio nso-e
CEPYQ GSGLD MLPYN IUSCP QOYGM JGCPL GDWWJ DMLSL OJCNY NYLYD
u--i- sosta -t-in g-ou- -his- esu-t sa--e a-tot heuni nitia
LJQLO DLCNL YPLOJ TPJDM NJQLO JWMSE JGGJG XTUOY EOOJO DQDMM
te-th atunt i-the --ea- ne-th e--o- esses ---hi -hheh a-a--
YBJQD LLOJV LOJTV YIOLU JPPES NGYQJ MOYVD GDNJE MSVDN EJM
i-e-a tthe- the-- ight- e---o nsi-e -hi-a sane- -o-an -e-
```

It is now easy to fill in additional pairs by inspection. For example, the missing letter in the fragment atunt i-the on the fifth line must be l, which gives P = l, and the missing letter in the fragment -osition on the third line must be p, which gives W = p. Substituting these in, we find the fragment e-p-ession on the first line, which gives Z = x and M = r, and the fragment -on-lusion on the third line, which gives E = c. Then consi-er on the last line gives Q = d and the initial words the-riterclai-e- must be the phrase "the writer claimed," yielding U = w and V = m. This gives

```
LOJUM YLJME PDYVJ QXTDV SVJNL DMTJZ WMJGG YSNDL UYLEO SKDVC
thewr iterc laime d--am oment ar-ex press ionat witch o-amu
GEPJS MDIPD NEJSK DNJTJ LSKDL OSVDV DNGYN VSGLL OSCIO LGOYG
scleo ragla nceo- ane-e to-at homam ansin mostt hough tshis
ESNEP CGYSN GUJMJ DGYNK DPPYX PJDGG SVDNT WMSWS GYLYS NGSKJ
concl usion swere asin- alli- leass oman- propo sitio nso-e
CEPYQ GSGLD MLPYN IUSCP QOYGM JGCPL GDWWJ DMLSL OJCNY NYLYD
uclid sosta rtlin gwoul dhisr esult sappe artot heuni nitia
LJQLO DLCNL YPLOJ TPJDM NJQLO JWMSE JGGJG XTUOY EOOJO DQDMM
tedth atunt ilthe -lear nedth eproc esses --whi chheh adarr
YBJQD LLOJV LOJTV YIOLU JPPES NGYQJ MOYVD GDNJE MSVDN EJM
i-eda tthem the-m ightw ellco nside rhima sanec roman cer
```

It is now a simple matter to fill in the few remaining letters and put in the appropriate word breaks, capitalization, and punctuation to recover the plaintext:

> The writer claimed by a momentary expression, a twitch of a muscle or a glance of an eye, to fathom a man's inmost thoughts. His

> conclusions were as infallible as so many propositions of Euclid. So startling would his results appear to the uninitiated that until they learned the processes by which he had arrived at them they might well consider him as a necromancer.[7]

1.2 Divisibility and greatest common divisors

Much of modern cryptography is built on the foundations of algebra and number theory. So before we explore the subject of cryptography, we need to develop some important tools. In the next four sections we begin this development by describing and proving fundamental results from algebra and number theory. If you have already studied number theory in another course, a brief review of this material will suffice. But if this material is new to you, then it is vital to study it closely and to work out the exercises provided at the end of the chapter.

At the most basic level, *Number Theory* is the study of the natural numbers

$$1, 2, 3, 4, 5, 6, \ldots,$$

or slightly more generally, the study of the integers

$$\ldots, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, \ldots.$$

The set of integers is denoted by the symbol $\mathbb{Z}$. Integers can be added, subtracted, and multiplied in the usual way, and they satisfy all the usual rules of arithmetic (commutative law, associative law, distributive law, etc.). The set of integers with their addition and multiplication rules are an example of a *ring*. See Section 2.10.1 for more about the theory of rings.

If a and b are integers, then we can add them $a + b$, subtract them $a - b$, and multiply them $a \cdot b$. In each case, we get an integer as the result. This property of staying inside of our original set after applying operations to a pair of elements is characteristic of a ring.

But if we want to stay within the integers, then we are not always able to divide one integer by another. For example, we cannot divide 3 by 2, since there is no integer that is equal to $\frac{3}{2}$. This leads to the fundamental concept of divisibility.

Definition. Let a and b be integers with $b \neq 0$. We say that b *divides* a, or that a *is divisible by* b, if there is an integer c such that

$$a = bc.$$

We write $b \mid a$ to indicate that b divides a. If b does not divide a, then we write $b \nmid a$.

[7] *A Study in Scarlet* (Chapter 2), Sir Arthur Conan Doyle.

Example 1.2. We have $847 \mid 485331$, since $485331 = 847 \cdot 573$. On the other hand, $355 \nmid 259943$, since when we try to divide 259943 by 355, we get a remainder of 83. More precisely, $259943 = 355 \cdot 732 + 83$, so 259943 is not an exact multiple of 355.

Remark 1.3. Notice that every integer is divisible by 1. The integers that are divisible by 2 are the *even integers*, and the integers that are not divisible by 2 are the *odd integers*.

There are a number of elementary divisibility properties, some of which we list in the following proposition.

Proposition 1.4. *Let $a, b, c \in \mathbb{Z}$ be integers.*
(a) *If $a \mid b$ and $b \mid c$, then $a \mid c$.*
(b) *If $a \mid b$ and $b \mid a$, then $a = \pm b$.*
(c) *If $a \mid b$ and $a \mid c$, then $a \mid (b + c)$ and $a \mid (b - c)$.*

Proof. We leave the proof as an exercise for the reader; see Exercise 1.6. □

Definition. A *common divisor* of two integers a and b is a positive integer d that divides both of them. The *greatest common divisor* of a and b is, as its name suggests, the largest positive integer d such that $d \mid a$ and $d \mid b$. The greatest common divisor of a and b is denoted $\gcd(a, b)$. If there is no possibility of confusion, it is also sometimes denoted by (a, b). (If a and b are both 0, then $\gcd(a, b)$ is not defined.)

It is a curious fact that a concept as simple as the greatest common divisor has many applications. We'll soon see that there is a fast and efficient method to compute the greatest common divisor of any two integers, a fact that has powerful and far-reaching consequences.

Example 1.5. The greatest common divisor of 12 and 18 is 6, since $6 \mid 12$ and $6 \mid 18$ and there is no larger number with this property. Similarly,

$$\gcd(748, 2024) = 44.$$

One way to check that this is correct is to make lists of all of the positive divisors of 748 and of 2024.

$$\begin{aligned} \text{Divisors of } 748 &= \{1, 2, 4, 11, 17, 22, 34, 44, 68, 187, 374, 748\}, \\ \text{Divisors of } 2024 &= \{1, 2, 4, 8, 11, 22, 23, 44, 46, 88, 92, 184, 253, \\ &\qquad 506, 1012, 2024\}. \end{aligned}$$

Examining the two lists, we see that the largest common entry is 44. Even from this small example, it is clear that this is not a very efficient method. If we ever need to compute greatest common divisors of large numbers, we will have to find a more efficient approach.

The key to an efficient algorithm for computing greatest common divisors is *division with remainder*, which is simply the method of "long division" that you learned in elementary school. Thus if a and b are positive integers and if you attempt to divide a by b, you will get a quotient q and a remainder r, where the remainder r is smaller than b. For example,

$$\begin{array}{r} 13\text{ R }9 \\ 17\ )\ 230 \\ \underline{17} \\ 60 \\ \underline{51} \\ 9 \end{array}$$

so 230 divided by 17 gives a quotient of 13 with a remainder of 9. What does this last statement really mean? It means that 230 can be written as

$$230 = 17 \cdot 13 + 9,$$

where the remainder 9 is strictly smaller than the divisor 17.

Definition. (Division Algorithm) Let a and b be positive integers. Then a *divided by* b *has quotient* q *and remainder* r means that

$$a = b \cdot q + r \qquad \text{with } 0 \le r < b.$$

The values of q and r are uniquely determined by a and b.

Suppose now that we want to find the greatest common divisor of a and b. We first divide a by b to get

$$a = b \cdot q + r \qquad \text{with } 0 \le r < b. \tag{1.1}$$

If d is any common divisor of a and b, then it is clear from equation (1.1) that d is also a divisor of r. (See Proposition 1.4(c).) Similarly, if e is a common divisor of b and r, then (1.1) shows that e is a divisor of a. In other words, the common divisors of a and b are the same as the common divisors of b and r; hence

$$\gcd(a, b) = \gcd(b, r).$$

We repeat the process, dividing b by r to get another quotient and remainder, say

$$b = r \cdot q' + r' \qquad \text{with } 0 \le r' < r.$$

Then the same reasoning shows that

$$\gcd(b, r) = \gcd(r, r').$$

Continuing this process, the remainders become smaller and smaller, until eventually we get a remainder of 0, at which point the final value $\gcd(s, 0) = s$ is equal to the gcd of a and b.

We illustrate with an example and then describe the general method, which goes by the name *Euclidean algorithm*.

Example 1.6. We compute $\gcd(2024, 748)$ using the Euclidean algorithm, which is nothing more than repeated division with remainder. Notice how the quotient and remainder on each line become the new a and b on the subsequent line:

$$\begin{aligned} 2024 &= 748 \cdot 2 + 528 \\ 748 &= 528 \cdot 1 + 220 \\ 528 &= 220 \cdot 2 + 88 \\ 220 &= 88 \cdot 2 + 44 \quad \leftarrow \boxed{\gcd = 44} \\ 88 &= 44 \cdot 2 + 0 \end{aligned}$$

Theorem 1.7 (The Euclidean Algorithm)**.** *Let a and b be positive integers with $a \geq b$. The following algorithm computes $\gcd(a, b)$ in a finite number of steps.*

(1) *Let $r_0 = a$ and $r_1 = b$.*

(2) *Set $i = 1$.*

(3) *Divide r_{i-1} by r_i to get a quotient q_i and remainder r_{i+1},*

$$r_{i-1} = r_i \cdot q_i + r_{i+1} \qquad \textit{with} \quad 0 \leq r_{i+1} < r_i.$$

(4) *If the remainder $r_{i+1} = 0$, then $r_i = \gcd(a, b)$ and the algorithm terminates.*

(5) *Otherwise, $r_{i+1} > 0$, so set $i = i + 1$ and go to Step* 3.

The division step (Step 3*) is executed at most*

$$2\log_2(b) + 1 \quad \textit{times.}$$

Proof. The Euclidean algorithm consists of a sequence of divisions with remainder as illustrated in Figure 1.2 (remember that we set $r_0 = a$ and $r_1 = b$).

$$\begin{array}{rcl} a = b \ \cdot q_1 + r_2 & & \text{with } 0 \leq r_2 < b, \\ b = r_2 \cdot q_2 + r_3 & & \text{with } 0 \leq r_3 < r_2, \\ r_2 = r_3 \cdot q_3 + r_4 & & \text{with } 0 \leq r_4 < r_3, \\ r_3 = r_4 \cdot q_4 + r_5 & & \text{with } 0 \leq r_5 < r_4, \\ \vdots & \vdots & \vdots \\ r_{t-2} = r_{t-1} \cdot q_{t-1} + r_t & & \text{with } 0 \leq r_t < r_{t-1}, \\ r_{t-1} = r_t \cdot q_t & & \end{array}$$

Then $r_t = \gcd(a, b)$.

Figure 1.2: The Euclidean algorithm step by step

The r_i values are strictly decreasing, and as soon as they reach zero the algorithm terminates, which proves that the algorithm does finish in a finite

number of steps. Further, at each iteration of Step 3 we have an equation of the form

$$r_{i-1} = r_i \cdot q_i + r_{i+1}.$$

This equation implies that any common divisor of r_{i-1} and r_i is also a divisor of r_{i+1}, and similarly it implies that any common divisor of r_i and r_{i+1} is also a divisor of r_{i-1}. Hence

$$\gcd(r_{i-1}, r_i) = \gcd(r_i, r_{i+1}) \qquad \text{for all } i = 1, 2, 3, \ldots. \tag{1.2}$$

However, as noted above, we eventually get to an r_i that is zero, say $r_{t+1} = 0$. Then $r_{t-1} = r_t \cdot q_t$, so

$$\gcd(r_{t-1}, r_t) = \gcd(r_t \cdot q_t, r_t) = r_t.$$

But equation (1.2) says that this is equal to $\gcd(r_0, r_1)$, i.e., to $\gcd(a, b)$, which completes the proof that the last nonzero remainder in the Euclidean algorithm is equal to the greatest common divisor of a and b.

It remains to estimate the efficiency of the algorithm. We noted above that since the r_i values are strictly decreasing, the algorithm terminates, and indeed since $r_1 = b$, it certainly terminates in at most b steps. However, this upper bound is far from the truth. We claim that after every two iterations of Step 3, the value of r_i is at least cut in half. In other words:

$$\textbf{Claim:} \quad r_{i+2} < \tfrac{1}{2} r_i \quad \text{for all } i = 0, 1, 2, \ldots.$$

We prove the claim by considering two cases.

Case I: $r_{i+1} \le \frac{1}{2} r_i$

We know that the r_i values are strictly decreasing, so

$$r_{i+2} < r_{i+1} \le \tfrac{1}{2} r_i.$$

Case II: $r_{i+1} > \frac{1}{2} r_i$

Consider what happens when we divide r_i by r_{i+1}. The value of r_{i+1} is so large that we get

$$r_i = r_{i+1} \cdot 1 + r_{i+2} \quad \text{with} \quad r_{i+2} = r_i - r_{i+1} < r_i - \tfrac{1}{2} r_i = \tfrac{1}{2} r_i.$$

We have now proven our claim that $r_{i+2} < \frac{1}{2} r_i$ for all i. Using this inequality repeatedly, we find that

$$r_{2k+1} < \frac{1}{2} r_{2k-1} < \frac{1}{4} r_{2k-3} < \frac{1}{8} r_{2k-5} < \frac{1}{16} r_{2k-7} < \cdots < \frac{1}{2^k} r_1 = \frac{1}{2^k} b.$$

Hence if $2^k \ge b$, then $r_{2k+1} < 1$, which forces r_{2k+1} to equal 0 and the algorithm to terminate. In terms of Figure 1.2, the value of r_{t+1} is 0, so we

have $t+1 \le 2k+1$, and thus $t \le 2k$. Further, there are exactly t divisions performed in Figure 1.2, so the Euclidean algorithm terminates in at most $2k$ iterations. Choose the smallest such k, so $2^k \ge b > 2^{k-1}$. Then

$$\#\text{ of iterations} \le 2k = 2(k-1) + 2 < 2\log_2(b) + 2,$$

which completes the proof of Theorem 1.7. □

Remark 1.8. We proved that the Euclidean algorithm applied to a and b with $a \ge b$ requires no more than $2\log_2(b)+1$ iterations to compute $\gcd(a,b)$. This estimate can be somewhat improved. It has been proven that the Euclidean algorithm takes no more than $1.45\log_2(b) + 1.68$ iterations, and that the average number of iterations for randomly chosen a and b is approximately $0.85\log_2(b) + 0.14$. (See [61].)

Remark 1.9. One way to compute quotients and remainders is by long division, as we did on page 12. You can speed up the process using a simple calculator. The first step is to divide a by b on your calculator, which will give a real number. Throw away the part after the decimal point to get the quotient q. Then the remainder r can be computed as

$$r = a - b \cdot q.$$

For example, let $a = 2387187$ and $b = 27573$. Then $a/b \approx 86.57697748$, so $q = 86$ and

$$r = a - b \cdot q = 2387187 - 27573 \cdot 86 = 15909.$$

If you need just the remainder, you can instead take the decimal part (also sometimes called the *fractional part*) of a/b and multiply it by b. Continuing with our example, the decimal part of $a/b \approx 86.57697748$ is 0.57697748, and multiplying by $b = 27573$ gives

$$27573 \cdot 0.57697748 = 15909.00005604.$$

Rounding this off gives $r = 15909$.

After performing the Euclidean algorithm on two numbers, we can work our way back up the process to obtain an extremely interesting formula. Before giving the general result, we illustrate with an example.

Example 1.10. Recall that in Example 1.6 we used the Euclidean algorithm to compute $\gcd(2024, 748)$ as follows:

$$\begin{aligned}
2024 &= 748 \cdot 2 + 528 \\
748 &= 528 \cdot 1 + 220 \\
528 &= 220 \cdot 2 + 88 \\
220 &= 88 \cdot 2 + 44 \quad \leftarrow \boxed{\gcd = 44} \\
88 &= 44 \cdot 2 + 0
\end{aligned}$$

We let $a = 2024$ and $b = 748$, so the first line says that

$$528 = a - 2b.$$

We substitute this into the second line to get

$$b = (a - 2b) \cdot 1 + 220, \qquad \text{so} \qquad 220 = -a + 3b.$$

We next substitute the expressions $528 = a - 2b$ and $220 = -a + 3b$ into the third line to get

$$a - 2b = (-a + 3b) \cdot 2 + 88, \qquad \text{so} \qquad 88 = 3a - 8b.$$

Finally, we substitute the expressions $220 = -a + 3b$ and $88 = 3a - 8b$ into the penultimate line to get

$$-a + 3b = (3a - 8b) \cdot 2 + 44, \qquad \text{so} \qquad 44 = -7a + 19b.$$

In other words,

$$-7 \cdot 2024 + 19 \cdot 748 = 44 = \gcd(2024, 748),$$

so we have found a way to write $\gcd(a, b)$ as a linear combination of a and b using integer coefficients.

In general, it is always possible to write $\gcd(a, b)$ as an integer linear combination of a and b, a simple sounding result with many important consequences.

Theorem 1.11 (Extended Euclidean Algorithm)**.** *Let a and b be positive integers. Then the equation*

$$au + bv = \gcd(a, b)$$

always has a solution in integers u and v. (See Exercise 1.12 for an efficient algorithm to find a solution.)

If (u_0, v_0) is any one solution, then every solution has the form

$$u = u_0 + \frac{b \cdot k}{\gcd(a, b)} \quad \text{and} \quad v = v_0 - \frac{a \cdot k}{\gcd(a, b)} \qquad \text{for some } k \in \mathbb{Z}.$$

Proof. Look back at Figure 1.2, which illustrates the Euclidean algorithm step by step. We can solve the first line for $r_2 = a - b \cdot q_1$ and substitute it into the second line to get

$$b = (a - b \cdot q_1) \cdot q_2 + r_3, \qquad \text{so} \qquad r_3 = -a \cdot q_2 + b \cdot (1 + q_1 q_2).$$

Next substitute the expressions for r_2 and r_3 into the third line to get

$$a - b \cdot q_1 = \big(-a \cdot q_2 + b \cdot (1 + q_1 q_2)\big) q_3 + r_4.$$