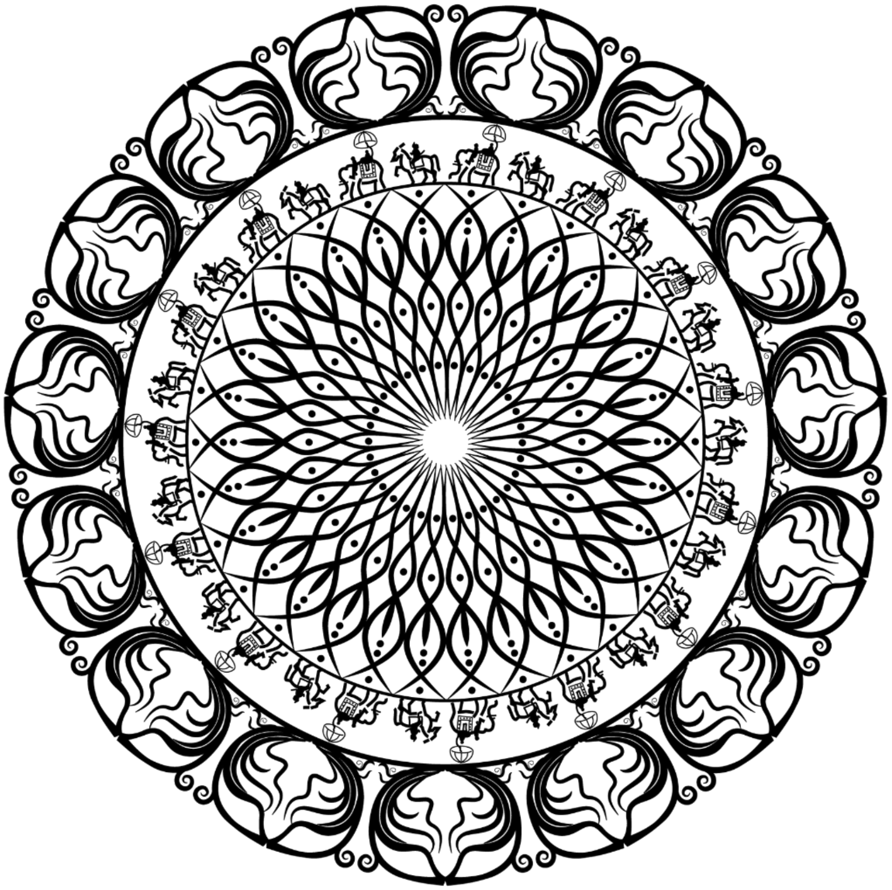
Mathematical Vignettes



by **Stephen Fratini**

Table of Contents

List of Figures

List of Tables

List of Equations

Preface

"The best teachers are those who show you where to look, but don't tell you what to see."

— Alexandra K. Trenfor

"The ultimate test of your knowledge is your ability to convey it to another."

— Richard Feynman

This book is a collection of short introductions (vignettes) about assorted topics of mathematics and related fields such as cryptography. My goal is to expose the reader to the wonderful world of mathematics and encourage further study.

Some of the topics are considered to be pure mathematics (e.g., number theory, Diophantine equations and continued fractions) while other topics fit into the realm of applied mathematics (e.g., stochastic processes, linear programming and cryptography). In some sense, the distinction is arbitrary since much of pure mathematics is required in support of applied mathematics, e.g., modern day cryptography is highly dependent on number theory.

I've tried to keep the necessary background for reading each section to a minimum, and have listed prerequisite knowledge for each section (in some cases, pointing to sections within the book).

There are many references for further study. I've used as many free sources as possible, e.g., Wikipedia, free online articles, YouTube videos and books that are available for electronic borrowing from the Internet Archive.

Where possible, I've provided references to online tools where the reader can experiment with the concepts being explained. I've tried to reference what I think are stable online sources, but (as we all know) there is no guarantee that an online link will continue to exist indefinitely.

I've included proofs of theorems in cases where the proofs are constructive (i.e., the proof entails an algorithm or method of solution) or where the proof adds further understanding of the topic. I've avoided complex existence proofs.

As this is more of a survey book than a textbook, I've not included many exercises or problems.

Keep in mind the intent here is to provide a brief introduction to various topics in mathematics. As such, most of the sections in this book are less than 30 pages in length.

"To learn one must be humble. But life is the great teacher."

— James Joyce, Ulysses

"Bodily exercise, when compulsory, does no harm to the body; but knowledge which is acquired under compulsion obtains no hold on the mind."

— Plato, The Republic

Acknowledgements

A special thanks goes to Dave Milham who reviewed and commented on the entire book. Thanks to Tony Clark for his review of the section on game theory, and several helpful discussions on the topic. Thanks to Carl Piva for suggestions concerning the potential audience for this book, and for reviewing the section entitled “From Natural Numbers to Quaternions.”

Stephen Fratini
Sole Proprietor of The Art of Managing Things
Eatontown, New Jersey (USA)
Email: sfratini@artofmanagingthings.com
LinkedIn: www.linkedin.com/in/stephenfratini

Other books by the author:

- *The Art of Managing Things (2nd edition)*, self-published on Amazon, https://www.amazon.com/Art-Managing-Things-Stephen-Fratini-ebook/dp/B07N4H4YWH/, January 2019.
- *Mathematical Thinking: Exercises for the Mind*, self-published on Amazon, https://www.amazon.com/Mathematical-Thinking-Exercises-Stephen-Fratini-ebook/dp/B08F75CDD6/, August 2020.
- *Financial Mathematics with Python*, self-published on Amazon, https://www.amazon.com/gp/product/B08VKQR141, February 2021.
- *Math in Art, and Art in Math*, self-published on Amazon, https://www.amazon.com/dp/B091D1F8MB, March 2021.
- *Algebra through Discovery and Experimentation*, self-published on Amazon, https://www.amazon.com/dp/B09B5L9WL5, July 2021.
- *The Struggle Against Chaos*, self-published on Amazon, https://www.amazon.com/dp/B09BLPQ86Q, July 2021.

1 Introduction

"It is, in fact, nothing short of a miracle that the modern methods of instruction have not yet entirely strangled the holy curiosity of inquiry; for this delicate little plant, aside from stimulation, stands mainly in need of freedom. Without this it goes to wrack and ruin without fail."

— Albert Einstein

1.1 Purpose

The purpose of the book is to expose the reader to a variety of interesting topics from mathematics and to encourage further learning.

1.2 Intended Audience

This book is intended for those with a desire to learn about assorted topics in mathematics. Many of the topics are not typically covered in an undergraduate mathematics or science major curriculum. So, even those with a technical background are likely to be exposed to new ideas and concepts.

1.3 Prerequisites

The prerequisites are listed at the beginning of each section. Some sections require very minimal background, e.g., just arithmetic, algebra and vectors, while other sections require a bit more background and a higher level of mathematical maturity (most notably, the sections on stochastic processes and linear programming).

1.4 Outline

The outline for the book is as follows:

- Section 1 is this introduction.
- Section 2 provides a brief introduction to the concept of mathematical induction (which is used in several proofs throughout the book).
- Section 3 covers various types of numbers such as integers, fractions, real numbers and complex (imaginary) numbers.
- Section 4 introduces difference equations. Difference equations are used to model various number sequences such as the Fibonacci sequence, i.e., $1,1,2,3,5,8,13,21,34,55, \dots$
- In Section 5, we discuss continued fractions (fractions with several layers, and usually having a repeating pattern).
- Section 6 provides a short introduction to nested radicals, e.g., nested sequences of square roots, typically with a repeating pattern. Both continued fractions and nested radicals can be modeled with difference equations.
- Diophantine equations (equations with integer solutions) are covered in Section 7.
- Random processes (also known as stochastic processes), whose behavior are best modeled using probability, are discussed in Section 8.
- Section 9 introduces game theory.

- Section 10 covers linear programming (solving systems of linear inequalities).
- Section 11 touches on a topic related to linear programming known as integer programming. Integer programming entails the optimization of systems of linear inequalities where the solution is restricted to the integers.
- Section 12 covers some topics in number theory (with a focus on aspects of number theory used in cryptography).
- Section 13 introduces cryptography. The focus here is on encryption and decryption algorithms, but not on code breaking techniques.
- There is also a short epilogue in Section 14.

2 Principle of Finite Induction

The principle of finite induction is very important in mathematics as it is used to prove many theorems, including several theorems in this book. By way of analogy, finite induction is like dominoes. If you know (1) the dominoes are equally spaced so that if any given domino falls, the next will fall and so on, and (2) the first domino has fallen, then you can conclude eventually every domino will fall. In finite induction, we have a statement with variable n (rather than a domino). If the statement can be shown true for $n = 1$, and it can be proved that "if the statement is true for $n = k$ then the statement is true for $n = k + 1$," then finite induction tells us the statement is true for all values of n.

Another analog comes from the book Concrete Mathematics [1]:

> Mathematical induction proves that we can climb as high as we like on a ladder, by proving that we can climb onto the bottom rung (the basis) and that from each rung we can climb up to the next one (the step).

Although its name may suggest otherwise, mathematical induction should not be considered as a form of inductive reasoning as defined earlier in this book. Mathematical induction is, in fact, an example of a deductive reasoning technique. The confusing terminology is unfortunate but the terms "inductive reasoning" and "mathematical induction" are firmly embedded in the literature and not likely to change.

The principle of finite induction is stated more formally in the following theorem:

Theorem 1. (First Principle of Finite Induction) Let S be a set of positive integers such that

- $1 \in S$
- *whenever* $k \in S$, *it must be that* $k + 1 \in S$

then S is necessarily the set of all positive integers.

Proof: By way of contradiction, assume that the set T (of all positive integers not in S) is nonempty. By the well-ordering principle, T must have a least element (call it x). We are given that $1 \in S$ and so it must be that $x > 1$ and thus, $0 < x - 1 < x$. Since x is the least element in T, $x - 1 \notin T$ which implies that $x - 1 \in S$. By hypothesis, S must contain $(x - 1) + 1 = x$ which contradicts the fact that $x \in T$. So, T must be empty and thus S is the set of all positive integers. ■

As an easy illustration of the first principle of finite induction, we prove that the sum of the first n odd numbers is n^2, i.e., $1 + 3 + 5 + \cdots + (2n - 1) = n^2$ (for all positive integer values of n).

> **Proof**: Clearly, the formula holds for $n = 1$. Assume the formula is true for $n = k$, i.e., $1 + 3 + 5 + \cdots + 2k - 1 = k^2$. Consider the case of $n = k + 1$, i.e., $[1 + 3 + 5 + \cdots + (2k - 1)] + (2k + 1) = k^2 + (2k + 1) = (k + 1)^2$ (which was to be proved). ■

There is an alternate version of the principle of finite induction that strengthens the second hypothesis:

Theorem 2. (Second Principle of Finite Induction) Let S be a set of positive integers such that

- $1 \in S$
- *whenever* $1,2,\dots,k \in S$*, it must be that* $k+1 \in S$

then S is necessarily the set of all positive integers.

Proof: By way of contradiction, assume that set T (of all positive integers not in S) is nonempty. By the well-ordering principle, T must have a least element (call it x). By hypothesis, x must be greater than 1. Further, since x is the least element in T, $1,2,\dots,(x-1)$ are not in T and are thus in S. But the second hypothesis implies that $(x-1)+1=x \in S$ which contradicts the fact that $x \in T$. So, T must be empty and thus S is the set of all positive integers. ■

Some statements do require the second principle of finite induction (as opposed to the first principle of induction). For example, consider the Lucas sequence: 1,3,4,7,11,18,29,47,76, …

The general pattern (after the first two terms) is $x_n = x_{n-1} + x_{n-2}$ (basically add the previous two numbers to get the next number in the sequence). We use the second principle of finite induction to prove that $x_n < \left(\frac{7}{4}\right)^n$. We have for $n=1$ that $x_1 = 1 < \left(\frac{7}{4}\right)^1$. Next, assume that the statement holds for $n = 1,2,\dots,k-1$. This gives us $x_{k-1} < \left(\frac{7}{4}\right)^{k-1}$ and $x_{k-2} < \left(\frac{7}{4}\right)^{k-2}$. It then follows that

$$x_k = x_{k-1} + x_{k-2} < \left(\frac{7}{4}\right)^{k-1} + \left(\frac{7}{4}\right)^{k-2} = \left(\frac{7}{4}\right)^{k-2}\left(\frac{7}{4}+1\right) = \left(\frac{7}{4}\right)^{k-2}\left(\frac{11}{4}\right) < \left(\frac{7}{4}\right)^{k-2}\left(\frac{7}{4}\right)^{2} = \left(\frac{7}{4}\right)^{k}$$

Thus, given that the statement is true for $n = 1,2,\dots,k-1$, we have proven the statement true for the case $n = k$, and the statement must therefore hold true for all values of n by the second principle of finite induction. ■

3 From Natural Numbers to Quaternions

Prerequisites: arithmetic, algebra, basic understanding of vector spaces

This section provides an elementary survey of numbers from simple counting numbers to the Quaternions. For a more advanced discussion of this same topic, see the book "From Natural Numbers to Quaternions" [1].

The **natural numbers** (or counting numbers) are the set of numbers $1, 2, 3, 4, 5, \dots$

The set of natural numbers is denoted by the symbol $\mathbb{N}$.

The number zero was initially only used as a placeholder, e.g., as in the number 1007. The earliest use of the number zero can be traced back to the year 683 AD in Cambodia [3].

The natural numbers and zero are known as the set of non-negative numbers.

The set of negative whole numbers $(\dots, -5, -4, -3, -2, -1)$, zero and the set of natural numbers comprise the set of **integers**, which can be represented as

$$\dots, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, \dots$$

The set of integers is denoted by the symbol $\mathbb{Z}$, standing originally for the German word Zahlen ("numbers").

The natural numbers can also be described as the set of positive integers.

All solutions to equations of the form $ax + b = 0$ where x is a variable, and a and b are integer constants make up the set of **rational numbers**. For example, the solution to $11x - 3 = 0$ is $\frac{3}{11}$ which is a member of the set of rational numbers. The integers are a subset of the rational numbers, e.g., 7 is the solution to the equation $x - 7 = 0$.

An equivalent definition of rational numbers is the set of all numbers of the form $\frac{a}{b}$ where a and b are integers.

The set of rational numbers is denoted by the symbol $\mathbb{Q}$. The set of rational numbers was thus denoted in 1895 by Giuseppe Peano after quoziente, meaning "quotient" in Italian.

All rational numbers can be written as either a decimal with a finite number of digits (e.g., $\frac{1}{8} = .125$) or as a repeating sequence of digits (e.g., $\frac{37}{189} = .\overline{195767}$ where the bar over the numbers means that the pattern repeats an infinite number of times. This can also be written as $.195767\dots$

When there is but a finite number of terms in the representation of a rational number, conversion back to fraction form is easy. For example, $.03125$ is equal to the fraction $\frac{3125}{100{,}000} = \frac{1}{32}$. Conversion from a repeating decimal to the associated fraction is more complex, see the explanation in the Wikipedia article entitled Fraction [5].

The set of all numbers that cannot be represented in the form $\frac{a}{b}$, where a and b are integers, are known as the **irrational numbers**. In decimal form, the representation of an irrational number continues indefinitely with no repeating pattern.

For example, $\sqrt{p}$ is irrational for any prime number p (i.e., a number whose only factors are 1 and itself). For a proof of this result, see Theorem 9-19 in Mathematical Thinking [4]. In this case, the irrational number $\sqrt{p}$ can be seen as the solution to polynomial equation with integer coefficients, i.e., $x^2 - p = 0$.

The set of all rational and irrational numbers are known as the **real numbers**. A Venn diagram depicting the set of real numbers and its key subsets is shown in Figure 1.

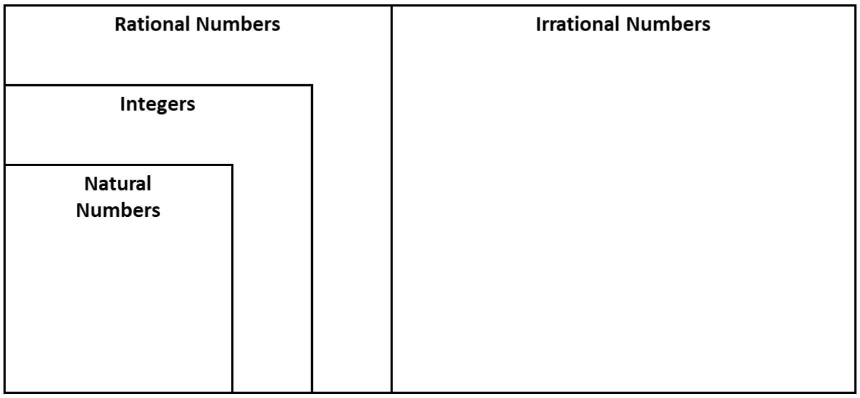


Figure 1. Venn Diagram for Real Numbers

Yet another classification depends on whether a number is a root to a polynomial with integer coefficients. Such numbers are known as **algebraic numbers.**

All rational numbers are algebraic numbers. Some irrational numbers are algebraic, e.g., $\sqrt{7}$ is irrational and it is also the root of $x^2 - 7 = 0$. Irrational numbers which are not algebraic are known as **transcendental numbers**. The numbers π and Euler's number e are examples of transcendental numbers. A list of known transcendental number classes is provided in the Wikipedia article on this topic [6].

It can be shown that the algebraic numbers are countable (i.e., can be put into a one-to-one correspondence with the natural numbers). Countable sets have measure zero. So, for example, on the interval $[0,1]$, the algebraic numbers have measure zero, and the complement (the transcendental numbers) have measure 1. Thus, the vast majority of numbers are transcendental but strangely, they are not easy to find. In general, it is difficult to prove that a number is transcendental. For most sums, products and powers of the number π and e (e.g., $\pi e, e + \pi, \pi - e, \frac{\pi}{e}, \pi^\pi, e^e, \pi^e, \pi^{\sqrt{2}}, e^{\pi^2}$), it is not known whether these numbers are rational, algebraic, irrational or transcendental.

Figure 2 shows an alternate view of the real numbers with a focus on the algebraic and transcendental numbers. The dark gray area represents the rational numbers (all of which are algebraic numbers) and the light gray area represents the irrational numbers.

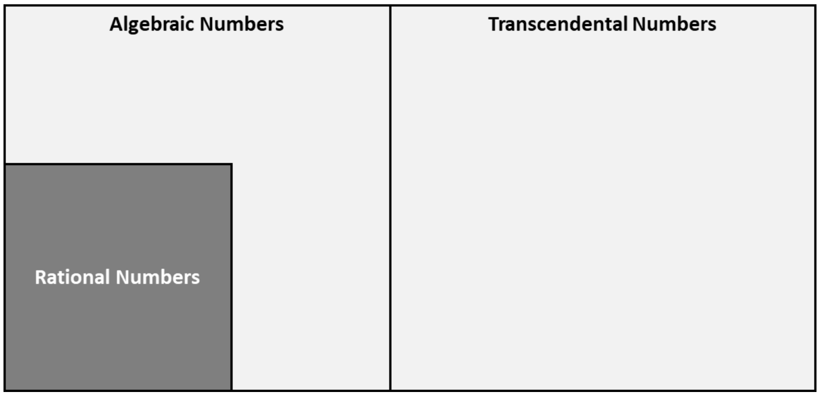


Figure 2. Alternate Venn Diagram for Real Numbers

Complex numbers are of the form $r\ +\ si$, where r and s are real numbers, and i represents the number which satisfies the equation $x^2 + 1 = 0$. Because no real number satisfies this equation, the entity i was invented to represent a solution ($-i$ is also a solution). The entity i was dubbed "an imaginary number" by the mathematician René Descartes (1596-1650). For the complex number $r + si$, r is called the real part and s is called the imaginary part. The set of complex numbers is denoted by either of the symbol C or $\mathbb{C}$.

Complex analysis, also known as the theory of functions of a complex variable, is the branch of mathematical analysis that studies functions of complex numbers. For an accessible introduction, see "Complex Analysis: A Visual and Interactive Introduction" [7].

In the realm of the complex numbers, all polynomial equations of a single variable and of degree n have n roots. Further, for every complex number $z = r + si$ there is an quadratic equation of the form $x^2 + bx + c$ that has z as one of its roots. To see this, we let

$$r = -\frac{b}{2}$$

$$s = \frac{i}{2}\sqrt{b^2 - 4c}$$

From the first equation, we have $b = -2r$. Plug this result into the second equation, square both sides of the equation and solve for c to get

$$c = r^2 + s^2$$

So, $x^2 - 2rx + (r^2 + s^2)$ has $z = r + si$ as a root.

The real numbers are the subset of the complex number whose imaginary part is equal to zero.

The distributive and commutative laws for addition and multiplication apply to complex numbers.

To add complex numbers, we add the real and imaginary parts separately, e.g.,

$$(3 - 7i) + (2 + 9i) = 5 + 2i$$

Subtraction is handled similarly, e.g.,

$$(8 - 2i) - (12 + 4i) = -4 - 6i$$

Multiplication for complex number is define as follows:

$$(a+bi)(c+di)=a(c+di)+bi(c+di)=ac+adi+bci-bd=(ac-bd)+i(ad+bc)$$

For example,

$$(3+4i)(1+2i)=(3-8)+10i=-5+10i$$

Given the above definitions for addition and multiplication, the commutative, associative and distributive laws hold true for the complex numbers.

The **conjugate of the complex number** $a+bi$ is defined to be $a-bi$. The product of a complex number and its conjugate has imaginary part equal to zero, i.e.,

$$(a+bi)(a-bi)=a^2+b^2$$

Division of complex numbers is achieved as follows:

$$\frac{(a+bi)}{(c+di)}=\frac{(a+bi)}{(c+di)}\cdot\frac{(c-di)}{(c-di)}=\frac{(ac+bd)+i(bc-ad)}{c^2+d^2}$$

For example,

$$\frac{1+7i}{2-3i}=\frac{1+7i}{2-3i}\cdot\frac{2+3i}{2+3i}=\frac{-19+17i}{13}$$

Complex numbers can be represented graphically by letting the horizontal axis represent the real part of complex numbers and the vertical axis represent the imaginary part of complex numbers.

In Figure 3, complex numbers $z_1=8+3i$ and $z_2=2+6i$, and their sum $w=10+9i$ are shown. Notice that complex number addition is done in the same way as vector addition. In fact, the set of complex numbers is a vector space with 1 and i being a basis (i.e., every member of $\mathbb{C}$ can be written as a linear combination of 1 and i).

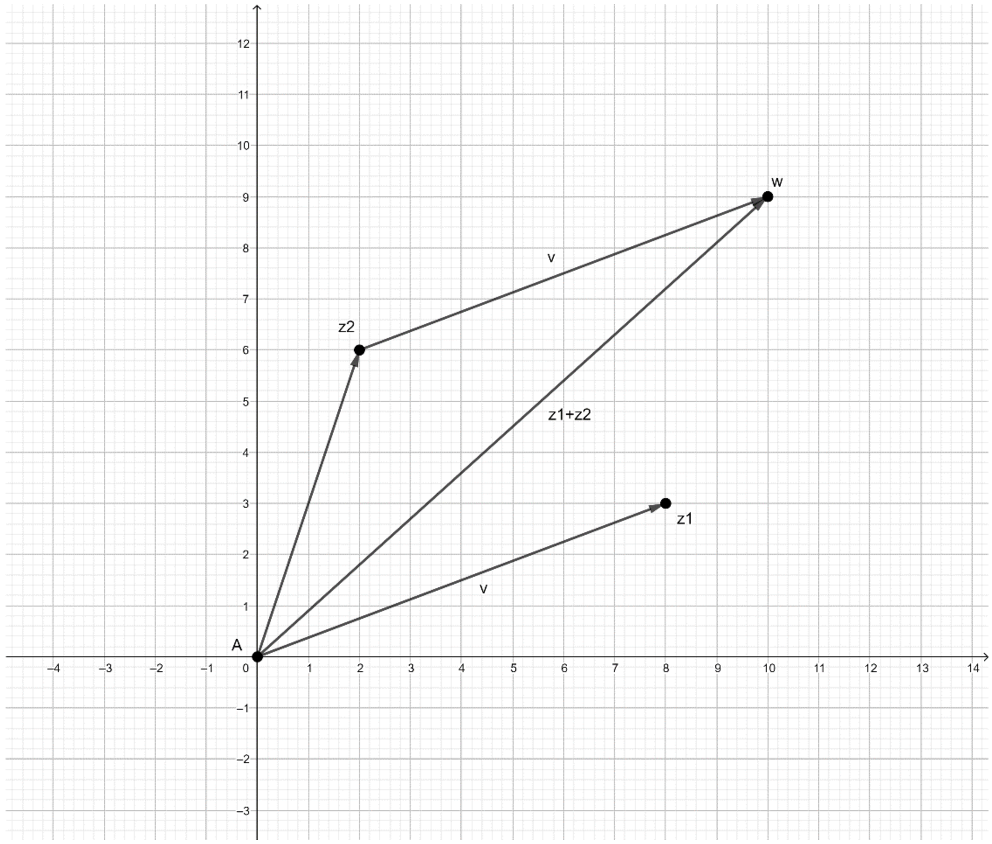


Figure 3. Graphic view of complex number addition

Complex numbers can also be represented in polar coordinates (i.e., by a radius from the origin and a given angle). As shown in Figure 4, the complex number $z = x + iy$ can be written in terms of terms of trigonometric ratios as $z = r\cos\theta + ir\sin\theta$ where $r = \sqrt{x^2 + y^2}$ is the distance of z from the origin, and θ is the angle between the real axis the line segment from the origin to z.

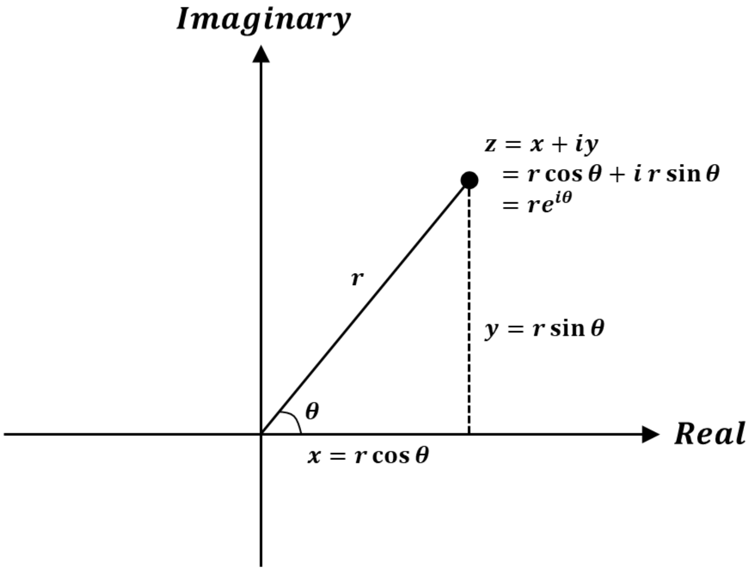

$\boldsymbol{Imaginary}$
$\boldsymbol{z = x + iy}$
$\boldsymbol{= r\cos\theta + i\,r\sin\theta}$
$\boldsymbol{= re^{i\theta}}$
$\boldsymbol{r}$
$\boldsymbol{y = r\sin\theta}$
$\boldsymbol{\theta}$
$\boldsymbol{Real}$
$\boldsymbol{x = r\cos\theta}$

Figure 4. Complex number in polar coordinates

We also have the very useful identity $e^{i\theta} = \cos\theta + i\sin\theta$ known as **Euler's formula** [9]. Euler's formula provides insight concerning the multiplication and division of complex numbers. Consider complex numbers $z = re^{i\alpha}$ and $w = se^{i\beta}$ as written in polar coordinates. We have that

$$zw = rse^{i(\alpha+\beta)}$$

$$\frac{z}{w} = \frac{r}{s}e^{i(\alpha-\beta)}$$

So, the product of z and w is the complex number with distance from the origin rs and at an angle $\alpha + \beta$. The quotient of z divided by w is the complex number with distance $\frac{r}{s}$ from the origin and at an angle $\alpha - \beta$ from the real axis.

. . .

Quaternions are an extension of the real numbers to 4-dimensions. The extension is achieved by adding three imaginary numbers i, j and k with all three equal to $\sqrt{-1}$. The general form of a quaternion is

$$a1 + bi + cj + dk$$

where a, b, c and d are real numbers. The numbers 1, i, j and k are unit vectors that form a basis for the quaternions. Quaternions are represented by the symbol $\mathbb{H}$ in honor of Irish mathematician William Rowan Hamilton who first described Quaternions in 1843. It was Hamilton who chose the name "quaternions."

The set of quaternions forms a 4-dimensional vector space over the real numbers, with $\{1, i, j, k\}$ as a basis and with the operations defined below.

Addition and subtraction of quaternions is straightforward, i.e., just add or subtract like terms. For example,

$$(2 + 3i - 4j + 2k) + (3 - 7i + 8j + k) = 5 - 4i + 4j + 3k$$

$$(7 + 6i + 5j + 4k) - (6 + 5i + 4j + 3k) = 1 + i + j + k$$

Definition of multiplication was a more challenging task for Hamilton. He eventually defined a scheme based on the multiplication characteristics of the basis vectors (as shown in Table 1).

Table 1. Quaternion multiplication table for unit vectors

$\times$	**1**	***i***	***j***	***k***
1	1	i	j	k
i	i	-1	k	$-j$
j	j	$-k$	-1	i
k	k	j	$-i$	-1

Using the distributive law and Table 1, the product of $a_1 + b_1 i + c_1 j + d_1 k$ and $a_2 + b_2 i + c_2 j + d_2 k$ can be shown to be

$$\begin{gathered} a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2 \\ +(a_1b_2 + b_1a_2 + c_1d_2 - d_1c_2)i \\ +(a_1c_2 - b_1d_2 + c_1a_2 + d_1b_2)j \\ +(a_1d_2 + b_1c_2 - c_1b_2 + d_1a_2)k \end{gathered}$$

Because of the asymmetry in Table 1 (e.g., $ji = -k$ but $ij = k$), multiplication is not commutative for Quaternions.

Scalar multiplication for the quaternion $q = a + bi + cj + dk$ is defined as

$$\lambda q = \lambda a + \lambda bi + \lambda cj + \lambda dk \text{ where } \lambda \text{ is a real number}$$

If we take the quaternion $q = a + bi + cj + dk$, then is conjugate is defined as $q^* = a - bi - cj - dk$. The norm or length of q is given by

$$\|q\| = \sqrt{qq^*} = \sqrt{a^2 + b^2 + c^2 + d^2}$$

Further, every non-zero quaternion has an inverse, which is defined as follows:

$$q^{-1} = \frac{1}{\|q\|} q^*$$

While it is a bit tedious, one can check that $qq^{-1} = q^{-1}q = 1$ using the definition of quaternion multiplication.

Division of Quaternions is ambiguous since $\frac{p}{q}$ can be interpreted as either pq^{-1} or $q^{-1}p$ but pq^{-1} does not necessarily equal $q^{-1}p$ since multiplication is not commutative for the Quaternions.

This non-commutativity has some surprising consequences, e.g., a polynomial equation over the quaternions can have more distinct solutions than the degree of the polynomial. For example, the equation $q^2 + 1 = 0$, has infinitely many quaternion solutions of the form

$$q = bi + cj + dk \text{ such that } b^2 + c^2 + d^2 = 1$$

Each solution can be seen as a point on the unit sphere in the three-dimensional space spanned by $\{i, j, k\}$.

It is impossible to construct a 3-dimensional real vector space that contains the complex numbers and which extends the multiplication defined for the complex numbers. To see this, choose another imaginary number j (in addition to the imaginary number i in $\mathbb{C}$) such that set $\{1, i, j\}$ forms a basis for the 3-dimensional real vector space given by

$$\mathbb{V} = \{v = a_1 1 + a_2 i + a_3 j \mid a_1, a_2, a_3 \in \mathbb{R}\}$$

We define $j^2 = -1$ and we want ij to be in $\mathbb{V}$ so that we have a closed system. Thus, it must be possible to write ij in the form of an element of $\mathbb{V}$, i.e., $ij = b_1 1 + b_2 i + b_3 j$ for $b_1, b_2, b_3 \in \mathbb{R}$. We can then write

$$\begin{aligned}(-1)j = (ii)j = i(ij) &= b_1 i - b_2 + b_3 ij \\ &= b_1 i - b_2 + b_3(b_1 1 + b_2 i + b_3 j) \\ &= (b_1 b_3 - b_2) + (b_1 + b_2 b_3)i + (b_3)^2 j\end{aligned}$$

Thus, $(-1)j = (b_1 b_3 - b_2) + (b_1 + b_2 b_3)i + (b_3)^2 j$.

Equating like terms in the above equation, we have that $(b_3)^2 = -1$ which contradict b_3 being an element of $\mathbb{R}$.

. . .

Further extensions of the real numbers are possible but some basic properties need to be sacrificed in the process.

Octonions, which have eight dimensions, are an extension of the quaternions. The commutative and associative properties do not hold for the octonions. The octonions do satisfy a weaker form of associativity known as alternative. The octonions were discovered in 1843 by John T. Graves, inspired by his friend Hamilton's work on quaternions. Recently, some physicists are considering octonions as a model for the symmetries associated with subatomic particles [8].

The **sedenions** form a 16-dimensional non-commutative and non-associative algebra over the real numbers. They are constructed as an extension to the octonions. Unlike the octonions, the alternative property does not hold.

4 Difference Equations

4.1 Overview

Prerequisites: arithmetic, algebra, sequences and infinite series, concept of a limit from calculus, basic trigonometry

A **recurrence relation** (or more commonly, **difference equation**) is an equation that recursively defines a sequence of numbers. Given one or more initial terms of the sequence, the recurrence relation allows one to determine the other numbers in the sequence.

As an example, consider the sequence defined by

$$x_{n+1} - x_n = 3, n = 1, 2, 3, \dots$$

$$x_0 = 4$$

From the recurrence relationship, we see that $x_1 - x_0 = 3$ and we know that $x_0 = 4$. Thus, $x_1 = 7$. Continuing in this manner, we have the following sequence:

$$4, 7, 10, 13, \dots$$

This is an example of an **arithmetic sequence**.

Many recurrence relations are represented as the difference of terms (one possible reason for the term “difference equation”). However, difference equations can also be viewed as the discrete analog of differential equations (yet another reason for the terminology). For more details on the relationship between difference equations and differential equations, see the section on Difference Calculus in the book “An Introduction to Difference Equations” [10]. The discussion here is confined to difference equations.

4.2 Classifications

Difference equations are classified based on several characteristics, including order, linearity, homogeneity and whether the equation is autonomous (there does not appear to be a noun that expresses this concept).

Let’s start with order. The order of a difference equation refers to the number of previous terms whose value one needs to know in order to compute the next term. If only the previous value of a sequence is needed to compute the next term, the difference equation is of first order. If *n* past values in the sequence are needed, then the difference equation is of order *n*.

For example, $x_{n+1} = 5x_n$ is a first order difference equation since only one previous number in the sequence is needed to compute the next term. If we are given $x_0 = 2$, then the sequence is

$$2, 10, 50, 250, 1250, \dots$$

This is an example of a **geometric sequence**.

The following difference equation is of order 3 since three previous terms are needed to define the next term. We also need three initial conditions to get started.

$$x_{n+1} = x_n + x_{n-1} + x_{n-2}$$
$$x_0 = 1$$
$$x_1 = 1$$
$$x_2 = 1$$

The above equation and initial conditions yield the following sequence:

$$1, 1, 1, 3, 5, 9, 17, 31, 57, 105, 193, 355, 653, 1201, \ldots$$

This is known as the Tribonacci sequence [11].

What about $x_{n+1} = 3x_{n-1}$? This seems a bit odd since we are skipping a term. This is considered to be a second order difference equation. Let's say the first two terms are 1 and 1, then the sequence would be

$$1, 1, 3, 3, 9, 9, 27, 27, \ldots$$

. . .

A difference equation is linear if the terms representing the sequence are raised to the first power. The examples that we have seen so far are all linear sequences. The following difference equation is not linear since the x_n term is squared:

$$x_{n+1} = 5(x_n)^2 + 1$$
$$x_0 = 0$$

The following is also classified as a linear difference equation even though the coefficients are functions of n. The only criterion is that the terms representing the sequence be linear.

$$x_{n+1} = nx_n + 3^n$$
$$x_0 = 1$$

In general, a linear difference equation of order k has the form shown in Equation 1.

Equation 1. General linear difference equation of order k

$$x_{n+k} + a_1(n)x_{n+k-1} + a_2(n)x_{n+k-2} + \cdots + a_k(n)x_n = b(n)$$

The terms $a_i(n)$ and $b(n)$ are real-valued functions. To determine the values of the terms in the sequence, k consecutive values must be known (does not need to start with $n = 0$).

If $b(n) = 0$, the linear difference equation is said to be **homogeneous**.

If all the $a_i(n)$ and $b(n)$ are constant, the linear difference equation is said to be **autonomous**.

The general form of a non-linear difference equation of order k is similar to the equation above for linear difference equations except that the terms $x_{n+k-1}, \ldots, x_{n+1}, x_n$ can be raised to any power. For example, the following is a non-linear difference equation of order 3:

$$x_{n+3} = \frac{3}{x_{n+2}} - \sqrt{n}(x_{n+1})^3 + n^2\sqrt{x_n} + 7^n$$

Non-linear difference equations can also be classified as homogeneous (or non-homogeneous), and autonomous (or non-autonomous). The above equation is non-homogeneous (because of the 7^n term) and non-autonomous (because of the coefficients $\sqrt{n}$ and n^2, and the 7^n term).

4.3 First Order Linear Difference Equations

The general format of a first order linear difference equation is

$$x_{n+1} = a(n) \cdot x_n + b(n)$$

This follows from the general equation of a linear difference equation of order k (see Equation 1).

Consider the computation of each term in the above sequence, starting from x_0

$$\begin{gathered} x_0 \\ x_1 = a(0)x_0 + b(0) \\ x_2 = a(1)x_1 + b(1) = a(1)\big(a(0)x_0 + b(0)\big) + b(1) = a(1)a(0)x_0 + a(1)b(0) + b(1) \\ x_3 = a(2)a(1)a(0)x_0 + a(2)a(1)b(0) + a(2)b(1) + b(2) \\ x_4 = a(3)a(2)a(1)a(0)x_0 + a(3)a(2)a(1)b(0) + a(3)a(2)b(1) + a(3)b(2) + b(3) \\ \dots \end{gathered}$$

The above pattern is summarized in Equation 2.

Equation 2: Formula for a first order linear difference equation

$$x_n = \left[\prod_{i=0}^{n-1} a(i)\right] x_0 + \sum_{j=0}^{n-2} \left[\prod_{i=j+1}^{n-1} a(i)\right] b(j) + b(n-1)$$

This equation is quite a monster and not so easy to use. However, we can derive several useful special cases from the general formula.

. . .

As a special case of the general first order linear difference equation, we consider equations of the form

$$x_{n+1} = ax_n$$

From the general formula applied to this case, we have that

$$x_n = a^n x_0$$

This is the general formula for a geometric sequence.

This formula can be applied to compound interest. Consider an investment of P dollars that is compounded k times per year at a rate of r% per year. That is to say that interest is computed k times per year, and added to the principal thereby also earning interest. In terms of a difference equation, we have

$$x_0 = P$$

$$x_{n+1} = x_n + \left(\frac{r}{k}\right) x_n = \left(1 + \frac{r}{k}\right) x_n$$

where x_n is the value of the investment after n compounding periods.

Using the above formula, we can compute a closed form for x_n:

$$x_1 = P\left(1 + \frac{r}{k}\right)$$

$$x_2 = P\left(1 + \frac{r}{k}\right)^2$$

$$\cdots$$

$$x_n = P\left(1 + \frac{r}{k}\right)^n$$

For example, if someone invests \$1000 at a nominal yearly rate of 5% which is compound monthly, then in 5 years, the investment (assuming no withdraws) is worth

$$x_{60} = 1000\left(1 + \frac{.05}{12}\right)^{60} \cong \$1283.36$$

. . .

The following is the general form of a first order linear non-homogeneous, autonomous difference equation. The only difference from the geometric sequence is the b term.

$$x_{n+1} = ax_n + b$$

From the general formula applied to this case, we have that

$$x_n = a^n x_0 + (a^{n-1} + a^{n-2} + \cdots + a + 1)b$$

Since $(1-a)(\, a^{n-1} + a^{n-2} + \cdots + a + 1) = 1 - a^n$, we have the following (for $a \neq 1$)

$$a^{n-1} + a^{n-2} + \cdots + a + 1 = \frac{1 - a^n}{1 - a}$$

Thus, for $a \neq 1$,

Equation 3: Formula for first order linear non-homogeneous, autonomous difference equation

$$x_n = a^n x_0 + \left(\frac{1 - a^n}{1 - a}\right)b$$

When $a = 1$, $x_n = x_0 + nb$ which is the arithmetic sequence.

As an application of this special case, consider a loan of Q dollars (or your favorite currency) that is to be paid back over N equal periods with the interest rate per period being i. The amount A is to be paid back by the borrower at the end of each period. Find a formula for A.

In the language of difference equations, we have

- $x_0 = Q$
- $x_N = 0$, the amount still owed on the loan is to be 0 after N periods
- $x_{n+1} = x_n + ix_n - A = (1 + i)x_n - A$. At the end of period $n + 1$, the borrower still owes x_n plus the amount of interest owed over the period (i.e., ix_n) minus the amount paid back (i.e., A).

Using Equation 3, we get

$$x_n = (1+i)^n Q - \left(\frac{1-(1+i)^n}{1-(1+i)}\right)A = (1+i)^n Q + \left(\frac{A}{i}\right)[1-(1+i)^n]$$

To determine the formula for A, we need to solve the following equation for A

$$x_N = (1+i)^N Q + \left(\frac{A}{i}\right)[1-(1+i)^N] = 0$$

$$(1+i)^N Q = \left(\frac{A}{i}\right)[(1+i)^N - 1]$$

$$A = \frac{iQ}{1-(1+i)^{-N}}$$

For example, consider a 30 year mortgage in the amount of \$250,000 at a 3% yearly rate, paid in monthly installments. What is the monthly payment?

We have that $N = 30 \cdot 12 = 360$, $Q = 250{,}000$ and $i = \frac{.03}{12} = .0025$.

So,

$$A = \frac{(.0025)(250{,}000)}{1-1.0025^{-360}} \cong \$1054.01$$

. . .

In general, an **equilibrium point** for a difference equation is a value $\bar{x}$ such at if $x_0 = \bar{x}$, then $x_n = \bar{x}$ for all values of n.

For first order linear non-homogeneous equations, when $a \neq 1$, there is one equilibrium point, i.e., $\bar{x} = \frac{b}{1-a}$. To see this, note that when $x_0 = \frac{b}{1-a}$, we have that

$$x_n = \frac{a^n b}{1-a} + \left(\frac{1-a^n}{1-a}\right)b = \frac{b}{1-a}$$

When $a = 1$, there is no equilibrium point. No matter what value is chosen for x_0, the sequence never returns to that value (unless b also equals 0 too, in which case the sequence is constant).

Further, it is possible for a point x to not be an equilibrium point but to reach (equal) an equilibrium point after finitely many iterations. In other words, a non-equilibrium point may transition to an equilibrium point in a finite time. Such points are referred to as **eventual equilibrium points**.

For example, define $x_{n+1} = f(x_n)$ where

$$f(x) = \begin{cases} 2x, & 0 \le x \le \frac{1}{2} \\ 2(1-x), & \frac{1}{2} < x \le 1 \end{cases}$$

There are two equilibrium points, i.e., 0 and $\frac{2}{3}$. However, there are an infinite number of eventual equilibrium points. For example, all points of the form $\frac{1}{2^k}$ are eventual equilibrium points. To see this, consider the sequence

$$x_0 = \frac{1}{2^k}$$

$$x_1 = f\left(\frac{1}{2^k}\right) = \frac{1}{2^{k-1}}$$

$$x_2 = f\left(\frac{1}{2^{k-1}}\right) = \frac{1}{2^{k-2}}$$

Eventually, we hit $x_j = \frac{1}{2}$ for iteration j which implies

$$x_{j+1} = f\left(\frac{1}{2}\right) = 1$$

$$x_{j+2} = f(1) = 0$$

$$x_{j+3} = x_{j+4} = \cdots = 0$$

. . .

The two special cases that we have seen so far are both classified as autonomous. This is a fancy way of saying that the next term in the sequence only depends on the previous term times a constant or with a constant added to it. If the next term in the sequence depends on the previous term and some function of n (i.e., the $a(n)$ or $b(n)$ in Equation 2), then the difference equation is said to be non-autonomous. The following are some examples of non-autonomous linear difference equations:

- $x_{n+1} = n^2 x_n$
- $x_{n+1} = 3x_n + (3n + \sqrt{n})$
- $x_{n+1} = \left(\frac{1}{n}\right) x_n - 5^n$

The first equation above is also homogeneous since $b(n) = 0$. The second and third equations are non-homogeneous.

If $b(n) = 0$, then we have the following formula for x_n from Equation 2:

$$x_n = \left[\prod_{i=0}^{n-1} a(i)\right] x_0$$

For example, $x_{n+1} = r^n x_n$ where r is any real number, then $a(n) = r^n$ and we have the following explicit formula for x_n

$$x_n = \left[\prod_{i=0}^{n-1} r^i\right] x_0 = (r \cdot r^2 \cdot r^3 \dots r^{n-1})x_0 = r^{(1+2+\cdots+(n-1))}x_0$$

However,

$$1 + 2 + \cdots + n - 1 = \frac{n(n-1)}{2}$$

and so, we have

$$x_n = x_0 \cdot r^{\frac{n(n-1)}{2}}$$

There is an interesting story associated with the sum of integers formula above. As a primary school student, the famous mathematician Carl Friedrich Gauss was given the assignment of adding all the numbers from 1 to 100. He surprised the teaching by almost immediately computing the answer. Rather than adding 1 to 2 to 3 and so on, he noticed the pattern $1 + 100 = 101, 2 + 99 = 101, 3 + 98 = 101, \dots$ and then used the following method to compute the answer:

$$1 + 2 + \cdots + 100 = \frac{101(100)}{2} = 5050$$

For more on this topic, see the article “Young Gauss and the sum of the first n positive integers” [12].

Let’s consider another example of a homogeneous, non-autonomous first order linear difference equation, i.e.,

$$x_{n+1} = \frac{x_n}{n+1}$$

Using Equation 2 and noting that $a(n) = \frac{1}{n}$ and $b(n) = 0$, we have

$$x_n = \left[\prod_{i=0}^{n-1} a(i)\right] x_0 = \left(1 \cdot \frac{1}{2} \cdot \frac{1}{3} \dots \cdot \frac{1}{n}\right) x_0 = \frac{x_0}{n!}$$

. . .

The following is an example of a non-homogeneous, non-autonomous first order linear difference equation:

$$x_{n+1} = x_n + \left(\frac{1}{5}\right)^n$$

Using Equation 2 and noting that $a(n) = 1$ and $b(n) = \left(\frac{1}{5}\right)^n$, we have

$$x_n = \left[\prod_{i=0}^{n-1} a(i)\right] x_0 + \sum_{j=0}^{n-2} \left[\prod_{i=j+1}^{n-1} a(i)\right] b(j) + b(n-1) = x_0 + \left[1 + \frac{1}{5} + \left(\frac{1}{5}\right)^2 + \cdots + \left(\frac{1}{5}\right)^{n-1}\right]$$

$$= x_0 + \frac{1-\left(\frac{1}{5}\right)^n}{1-\frac{1}{5}} = x_0 + \frac{5}{4}\left[1-\left(\frac{1}{5}\right)^n\right]$$

Thus, $\lim_{n\to\infty} x_n = x_0 + \frac{5}{4}$. For those not familiar with the limit concept from calculus, $\lim_{n\to\infty} x_n$ means that as n becomes very large the term x_n approaches $x_0 + \frac{5}{4}$. This is intuitively true since one can see that the term $\left(\frac{1}{5}\right)^n$ approaches 0 as n becomes large.

[Author's Remark: Surprisingly, at least to me, that Wolfram Alpha (https://www.wolframalpha.com/) can solve the above difference equation. Enter

x(n+1)=x(n)+(1/5)^n; x(0)=c

and the following solution is returned

$$x(n) = c + \frac{5}{4}(1 - 5^{-n})$$

The reader may want to check some of the other examples using this online resource.**]**

As a second example of a non-homogeneous, non-autonomous first order linear difference equation, consider the following

$$x_{n+1} = (n+1)x_n + 2^n(n+1)!$$

In terms of Equation 2, $a(n) = n + 1$ and $b(n) = 2^n(n+1)!$

and so, we have

$$x_n = \left[\prod_{i=0}^{n-1}(i+1)\right]x_0 + \sum_{j=0}^{n-2}\left[\prod_{i=j+1}^{n-1}(i+1)\right]2^j(j+1)! + 2^{n-1}n!$$

$$= n!\,x_0 + \sum_{j=0}^{n-2} n(n-1)\ldots(j+2)(j+1)!\,2^j + 2^{n-1}n! = n!\,x_0 + n!\sum_{j=0}^{n-2}2^j + n!\,2^{n-1}$$

$$= n!\,x_0 + \sum_{j=0}^{n-1} n!\,2^j = n!\,x_0 + \frac{n!\,(2^n-1)}{2-1} = n!\,(x_0 + 2^n - 1)$$

. . .

The lazy caterer's sequence [13], also known as the central polygonal numbers, describes the maximum number of pieces in which a circle can be cut with straight lines. The first few cases are shown in Figure 5. The sequence (number of pieces) is 1, 2, 4, 7, 11, 16, 22, …

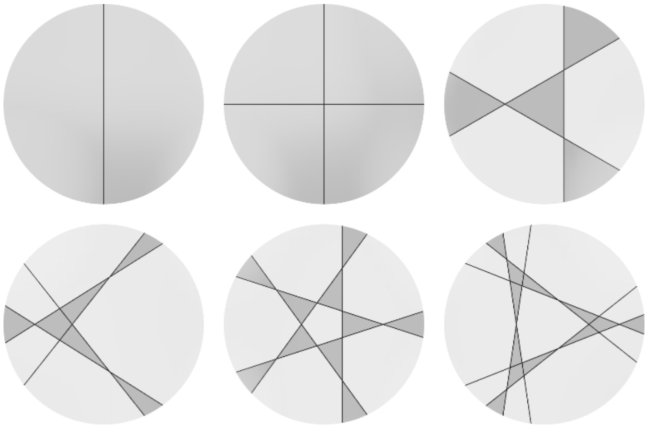


Figure 5. Dividing a pie with straight lines

To obtain the maximum number of pieces, the $n + 1st$ line should cross all the other previous lines inside the circle, but not cross any intersection of previous lines. Thus, the $n + 1st$ line is cut in n places, and into $n + 1$ line segments. Each segment divides one piece of circle (from step n) into 2 parts, adding exactly $n + 1$ to the number of pieces. In terms of a difference equation, we have

$$x_{n+1} = x_n + (n + 1)$$

$$x_0 = 1$$

where x_n is the number of the maximum number of pieces in which the circle can be divided with n lines. Using Equation 2, we have

$$x_n = 1 + (1 + 2 + \cdots + n) = 1 + \frac{n(n + 1)}{2} = \frac{n^2 + n + 2}{2}$$

This problem has the same solution as the problem of dividing the plane into the maximum number of regions by n lines [14].

4.4 Higher Order Linear Difference Equations

If each term in a linear difference equation depends on the previous k terms, then the order of the difference equation is k. If $k \geq 2$, the difference equation is said to be of higher order. The general format of a higher order difference equation is shown in Equation 1.

. . .

We start with the simple case of a second order linear difference equation that is homogeneous and autonomous, i.e., equations of the form

$$x_{n+2} + a_1 x_{n+1} + a_2 x_n = 0$$

In order to solve the above equation, we assume the solution is of the form $x_n = \lambda^n$ and plug into the difference equation to get

$$\lambda^{n+2} + a_1 \lambda^{n+1} + a_2 \lambda^n = 0$$

which simplifies to the quadratic equation (known as the **characteristic polynomial** in this context)

$$\lambda^2 + a_1 \lambda + a_2 = 0$$

The quadratic formula is then applied to find solutions. There are three cases, i.e., 2 distinct real solutions, repeated real solutions, or 2 complex solutions. In what follows, we show examples of each case.

Our first example is the difference equation $x_{n+2} - x_{n+1} - x_n = 0$. The characteristic equation is

$$\lambda^2 - \lambda - 1 = 0$$

From the quadratic formula, we have the following two solutions:

$$\lambda_1 = \frac{\left(1 + \sqrt{5}\right)}{2}, \lambda_2 = \frac{\left(1 - \sqrt{5}\right)}{2}$$

All solutions to the difference equation can be written in the form

$$x_n = c{\lambda_1}^n + d{\lambda_2}^n$$

To see this, compute x_{n+1} and x_{n+2} and then substitute back into the difference equations.

$$x_{n+1} = c{\lambda_1}^{n+1} + d{\lambda_2}^{n+1} = c\lambda_1{\lambda_1}^n + d\lambda_2{\lambda_2}^n$$

$$x_{n+2} = c{\lambda_1}^{n+2} + d{\lambda_2}^{n+2} = c{\lambda_1}^2{\lambda_1}^n + d{\lambda_2}^2{\lambda_2}^n$$

Substituting back into the difference equations, we get

$$\left(c{\lambda_1}^2{\lambda_1}^n + d{\lambda_2}^2{\lambda_2}^n\right) - \left(c\lambda_1{\lambda_1}^n + d\lambda_2{\lambda_2}^n\right) - \left(c{\lambda_1}^n + d{\lambda_2}^n\right)$$

$$= c{\lambda_1}^n\left({\lambda_1}^2 - \lambda_1 - 1\right) + d{\lambda_2}^n\left({\lambda_2}^2 - \lambda_2 - 1\right) = c{\lambda_1}^n \cdot 0 + d{\lambda_2}^n \cdot 0 = 0$$

For each set of two initial conditions, we can determine the values for c and d. For $x_0 = 1$ and $x_1 = 1$, we get the **Fibonacci sequence**, i.e., $1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$ This is gotten by directly substituting $x_0 = 1$ and $x_1 = 1$ into $x_{n+2} - x_{n+1} - x_n = 0$, solving for x_2 and then continuing the process to find additional terms in the sequence.

We also have the following two equations that can be solved for c and d:

$$x_0 = 1 = c + d$$

$$x_1 = 1 = c\frac{1 + \sqrt{5}}{2} + d\frac{1 - \sqrt{5}}{2}$$

$\varphi = \frac{1+\sqrt{5}}{2}$ is known as the **Golden Ratio**, and $-\psi = \frac{\sqrt{5}-1}{2}$ is known as the golden ratio conjugate or silver ratio. Further, $\psi = 1 - \varphi = -\frac{1}{\varphi}$.

Using these symbols, the above equations can be written as

$$c + d = 1$$

$$c\varphi + d\psi = 1$$

Solving this system of equations and making use of the golden ratio identities, we get that $c = \frac{1}{\sqrt{5}}$ and $d = -\frac{1}{\sqrt{5}}$. This gives us the general formula for x_n (known as Binet's formula):

$$x_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right] = \frac{1}{\sqrt{5}}(\varphi^n - \psi^n)$$

This may be hard to believe at first since on one hand, x_n is a sequence of natural numbers and on the other hand, the closed formula for x_n entails irrational numbers in a complex expression. This is easy to verify with a spreadsheet or calculator. Try computing x_{13} using the above formula and you will get 233.

If we apply the initial conditions $x_0 = 2$ and $x_1 = 1$ to the difference equation $x_{n+2} - x_{n+1} - x_n = 0$, we get the **Lucas numbers**, i.e., $2, 1, 3, 4, 7, 11, 18, 29, 47, 76, 123, \ldots$ The initial conditions yield the following equations:

$$c + d = 2$$

$$c\varphi + d(1-\varphi) = 1$$

The solution is $c = d = 1$, which gives us the following expression for x_n

$$x_n = \left(\frac{1+\sqrt{5}}{2}\right)^n + \left(\frac{1-\sqrt{5}}{2}\right)^n = \varphi^n + \psi^n$$

. . .

In some cases, the characteristic equation has repeated roots, e.g., the difference equation

$$x_{n+2} - 6x_{n+1} + 9x_n = 0$$

has characteristic equation

$$\lambda^2 - 6\lambda + 9 = (\lambda - 3)^2 = 0$$

So, $\lambda = 3$ is a repeated root. In this case, the format of the general solution is slight modified (i.e., multiple the second term by n):

$$x_n = c3^n + dn3^n = 3^n(dn + c)$$

We can confirm our answer using Wolfram Alpha by entering x(n+2) - 6x(n+1) + 9x(n) = 0.

The initial conditions $x_0 = -2$ and $x_1 = -3$ yield the following two equations

$$-2 = c3^0 + 0 \ \Rightarrow c = -2$$

$$-3 = 3(-2) + 3d \ \Rightarrow d = 1$$

Thus, the solution for the given initial conditions is

$$x_n = -2 \cdot 3^n + n3^n = 3^n(n-2)$$

The corresponding sequence is $-2, -3, 0, 27, 162, 729, 2916, 10935, 39366, 137781, \ldots$

There is one additional case, i.e., complex roots to the quadratic characteristic equation. Keep in mind that complex roots come in conjugate pairs of the for $\alpha \pm i\beta$. In such cases, the general solution is of the form

$$x_n = c(\alpha + i\beta)^n + d(\alpha - i\beta)^n$$

Converting to polar coordinates, we have

$$\alpha = r\cos\theta\,, \beta = r\sin\theta\,, r = \sqrt{\alpha^2 + \beta^2}, \theta = \tan^{-1}\left(\frac{\beta}{\alpha}\right)$$

and so,

$$x_n = c(r\cos\theta + ir\sin\theta)^n + d(r\cos\theta - ir\sin\theta)^n$$

Using Euler's formula, i.e., $e^{i\theta} = \cos\theta + i\sin\theta$, we have that

$$(r\cos\theta + ir\sin\theta)^n = r^n e^{in\theta} = r^n(\cos n\theta + i\sin n\theta)$$

$$(r\cos\theta - ir\sin\theta)^n = r^n e^{in\theta} = r^n(\cos n\theta - i\sin n\theta)$$

Applying the above to the formula for x_n yields

$$x_n = cr^n(\cos n\theta + i\sin n\theta) + dr^n(\cos n\theta - i\sin n\theta)$$

$$= r^n[(c+d)\cos n\theta + i(c-d)\sin n\theta]$$

For example, consider the difference equation

$$x_{n+2} - 2x_{n+1} + 2x_n = 0$$

$$x_0 = 2$$

$$x_1 = 5$$

The characteristic equation is

$$\lambda^2 - 2\lambda + 2 = 0$$

and the roots are $1 + i$ and $1 - i$. In polar coordinates, $r = \sqrt{2}$ and $\theta = \frac{\pi}{4}$.

Using the general equation for the case of complex roots, we have

$$x_n = (\sqrt{2})^n\left[(c+d)\cos\left(\frac{n\pi}{4}\right) + i(c-d)\sin\left(\frac{n\pi}{4}\right)\right]$$

Using the initial conditions, we get

$$x_0 = 2 = c + d$$

$$x_1 = 5 = \sqrt{2}\left[\frac{c+d}{\sqrt{2}} + \frac{i(c-d)}{\sqrt{2}}\right] = (c+d) + i(c-d) = 2 + i(c-d)$$

$$3 = i(c-d)$$

Plugging the above back into the formula for x_n, we have

$$x_n = (\sqrt{2})^n\left[2\cos\left(\frac{n\pi}{4}\right) + 3\sin\left(\frac{n\pi}{4}\right)\right]$$

Using either the formula above or the difference equation yields the same sequence, i.e.,

$$2, 5, 6, 2, -8, -20, -24, -8, 32, 80, 96, 32, -128, -320, -384, \ldots$$

. . .

For orders higher than 2, a similar procedure can be used to solve homogeneous, autonomous linear difference equations with constant coefficients. The characteristic equation will be of higher degree and in some cases, have a combination of distinct real roots, repeated roots and complex roots. Further details can be found in Chapter 5 of the Schaum's Outlines on this topic [15].

4.5 Non-homogeneous Linear Difference Equations

Recall from Equation 1 that a general linear difference equation is of the form

$$x_{n+k} + a_1(n)x_{n+k-1} + a_2(n)x_{n+k-2} + \cdots + a_k(n)x_n = b(n)$$

If $b(n) \neq 0$, the difference equation is said to be non-homogeneous.

The general solution of the above equation when $b(n) = 0$ is known as the **complementary solution** of the associated non-homogeneous equation (i.e., when $b(n) \neq 0$) and is denoted by y_n. A solution of the non-homogeneous equation is known as a **particular solution** and is denoted by z_n. The general solution to Equation 1 is of the form

$$x_n = y_n + z_n$$

For a proof of this result, see Theorem 2.30 in the book by Elaydi [16].

The method of solution for non-homogeneous linear difference equations is to first find a general solution to the associated homogeneous linear difference equation (using methods discussed previously in this book) and then find a particular solution to the non-homogeneous linear difference equation. One approach is to make an intelligent guess concerning the particular solution and then substitute the guess into the difference equation. This approach is known as the **method of undetermined coefficients**. If $b(n)$ is in certain forms, there are known particular solutions, as shown in Table 2.

Table 2. Possible particular solutions to non-homogeneous linear difference equations

$\boldsymbol{b(n)}$	**Particular Solution**
c, where c is a constant	An
c^n, where c is a constant	Ac^n
$\sin(cn)$ or $\cos(cn)$	$A\cos(cn) + B\sin(cn)$
polynomial $p(n)$ of degree m	$A_0n^m + A_1n^{m-1} + \cdots + A_m$
$c^n\, p(n)$ where $p(n)$ is of degree m	$c^n(A_0n^m + A_1n^{m-1} + \cdots + A_m)$
$c^n \sin(cn)$ or $c^n\cos(cn)$	$c^n(A\cos(cn) + B\sin(cn))$

. . .

For example, consider the difference equation

$$px_{n+1} - x_n + qx_{n-1} = -1$$

where $p + q = 1, p \neq q$, $0 \leq p \leq 1$ and $0 \leq q \leq 1$. These conditions may appear a bit obscure but are related to a problem that we will see in Section 8.2.2.

We first find the complementary solution, i.e., the general solution to $px_{n+1} - x_n + qx_{n-1} = 0$. Keeping in mind that $p + q = 1$, we can factor the associated characteristic equation as follows

$$p\lambda^2 - \lambda + q = (p\lambda - q)(\lambda - 1) = 0$$

which has solutions $\lambda_1 = 1$ and $\lambda_2 = \frac{q}{p}$. If we let $r = \frac{q}{p}$, the general solution to the homogeneous equation is of the form

$$y_n = a\lambda_1^n + b\lambda_1^n = a + br^n$$

Next, we find the particular solution to the non-homogeneous equation. The suggestion from Table 2 is to try $z_n = An$. Substituting into the non-homogeneous equation, we get

$$pA(n + 1) - An + qA(n - 1) = -1$$

$$pAn + pA - An + qAn - qA = -1$$

$$An(p + q) - An + A(p - q) = -1$$

However, we are given that $p + q = 1$, and so the above equation reduces to

$$A(p - q) = -1$$

$$A = \frac{1}{q - p}$$

So, the particular solution is $z_n = An = \frac{n}{q-p}$ and the general solution to the non-homogeneous equation is

$$x_n = \ a + br^n + \frac{n}{q - p}$$

. . .

As a second example, let us determine the general solution to

$$x_{n+2} + 4x_{n+1} + 3x_n = 3n$$

We first find the complementary solution. The characteristic equation for the homogeneous equation is

$$\lambda^2 + 4\lambda + 3 = 0$$

which can be factored as

$$(\lambda + 3)(\lambda + 1) = 0$$

The roots are $\lambda_1 = -1$ and $\lambda_1 = -3$.

Thus, the complementary solution is

$$y_n = a(-1)^n + b(-3)^n$$

Given the form of $b(n)$, the suggestion from Table 2 is to $z_n = An + B$ as a possible particular solution for the non-homogeneous difference equation. Substituting z_n into $x_{n+2} + 4x_{n+1} + 3x_n = 3n$, we get

$$A(n+2) + B + 4A(n+1) + 4B + 3An + 3B = 3n$$

Collecting like terms, the above equation reduces to

$$8An + (6A + 8B) = 3n$$

This gives use a system of two equations with two unknowns, i.e.,

$$8A = 3$$

$$6A + 8B = 0$$

Solving the system of equations, we get $A = \frac{3}{8}$ and $B = -\frac{9}{32}$ and thus, $z_n = \frac{3}{8}n - \frac{9}{32}$.

Putting the above results together, we determine that the general solution to the non-homogeneous difference equation is

$$x_n = \ a(-1)^n + b(-3)^n + \frac{3}{8}n - \frac{9}{32}$$

4.6 Non-linear Difference Equations

When the relationship between the x_i terms in a difference equation is other than a linear combination, we have what is called a non-linear difference equation. Some examples of non-linear difference equations

- $x_{n+1} + \boldsymbol{x_n}^{\boldsymbol{2}} = 0$
- $x_{n+2} + \frac{\boldsymbol{1}}{\boldsymbol{x_{n+1}}} - 3x_n = n^2$
- $nx_{n+2} + \sqrt{\boldsymbol{x_{n+1}}} + 2x_1 = 0$
- $\boldsymbol{x_n x_{n+1}} - x_{n-1} = 3$

In the above examples, the terms that makes the equation non-linear are in bold. The concepts of order, homogeneity and autonomous-ness also apply to non-linear difference equations. The first example above is of first order, homogeneous and autonomous. The second example is of order 2, non-homogeneous and non-autonomous. The third example is of order 2, homogeneous and non-autonomous. The fourth example is 2nd order, non-homogeneous and autonomous.

Consider the difference equation

$$x_{n+1} = \frac{x_n}{5x_n + 2}$$

$$x_0 = 1$$

The corresponding sequence is

$$\frac{1}{1}, \frac{1}{7}, \frac{1}{19}, \frac{1}{43}, \frac{1}{91}, \frac{1}{187}, \dots$$

The denominators in the above sequence can be represented by a difference equation. The recurrence relationship is given by

$$y_n = 2y_{n-1} + 5$$

$$y_0 = 1$$

Using Equation 3, we get the following closed form for y_n

$$y_n = 2^n + 5\left(\frac{2^n - 1}{2 - 1}\right) = 2^n + 5(2^n - 1) = 6 \cdot 2^n - 5 = 3 \cdot 2^{n+1} - 5$$

More generally, we can find a closed form given any constant value for x_0, i.e.,

$$x_{n+1} = \frac{x_n}{5x_n + 2}$$

$$x_0 = c$$

Applying some algebra, we get the sequence

$$c, \frac{c}{5c + 2}, \frac{c}{15c + 4}, \frac{c}{35c + 8}, \frac{c}{75c + 16}, \dots$$

The general pattern is

$$x_n = \frac{c}{5c(2^n - 1) + 2^n}$$

For all real values of c, the sequence converges to 0 as n approaches infinity, i.e., $\lim_{n \to \infty} x_n = 0$.

. . .

For first order, non-autonomous difference equations (can be linear or non-linear) of the form $x_{n+1} = f(x_n)$, we state the following definitions:

- An equilibrium point $\bar{x}$ is **stable** if for any $\varepsilon > 0$ there exists $\delta > 0$ such that $|x_0 - \bar{x}| < \delta$ implies $|f^n(x_0) - \bar{x}| < \varepsilon$ for all $n > 0$. If $\bar{x}$ is not stable, then it is referred to as an unstable equilibrium point. Note that $f^n(x_0)$ is the function f applied to x_0 a total of n times. For example, $f^3(x_0) = f\left(f\big(f(x_0)\big)\right)$.

- An equilibrium point $\bar{x}$ is an **attractor** if there exists $\gamma > 0$ such that $|x_0 - \bar{x}| < \gamma$ implies $\lim\limits_{n\to\infty} x_n = \bar{x}$. If $\gamma = \infty$, then $\bar{x}$ is a global attractor. Basically, no matter where you start, the sequence converges to $\bar{x}$. If no such γ exists, then $\bar{x}$ is a **repelling equilibrium point**.
- An equilibrium point $\bar{x}$ is an **asymptotically stable equilibrium point** if it is stable and attracting. If $\gamma = \infty$, then $\bar{x}$ is said to be globally asymptotically stable.

The following difference equation exhibits some of the terms defined above:

$$x_{n+1} = \begin{cases} \dfrac{3}{4}, & \dfrac{1}{2} < x_n \le 1 \\ \dfrac{1}{2}, & x_n = \dfrac{1}{2} \\ \dfrac{1}{4}, & 0 \le x_n < \dfrac{1}{2} \end{cases}$$

The equilibrium points are $\frac{1}{4}, \frac{1}{2}$ and $\frac{3}{4}$. The points $\frac{1}{4}$ and $\frac{3}{4}$ are asymptotically stable equilibrium points, and $\frac{1}{2}$ is a repelling equilibrium point. Even a slight variation from $\frac{1}{2}$ gets mapped to either $\frac{1}{4}$ or $\frac{3}{4}$. There are no global attractors in this example.

The following difference equation has an unstable equilibrium:

$$x_{n+1} = \begin{cases} 0, & x_n = 0 \\ (-1)^n \left(\dfrac{x_n}{2} + \dfrac{1}{2}\right), & x_n \ne 0 \end{cases}$$

There is but one equilibrium point and that is at 0. Regardless of the value for x_0, the sequence eventually oscillates between 0.2 and $-.6$. The first few terms of the sequence (with $x_0 = 2$) are shown in Table 3. For $\varepsilon = .1$, there is no value of δ such that $|x_n - 0| < \varepsilon$ for all $n > 0$. By definition, the equilibrium point at 0 is unstable.

Table 3. Unstable equilibrium point at 0

n	x_n
0	2.0000
1	-1.5000
2	-0.2500
3	-0.3750
4	0.3125
5	-0.6563
6	0.1719
7	-0.5859
8	0.2070
9	-0.6035
10	0.1982
11	-0.5991

n	x_n
12	0.2004
13	-0.6002
14	0.1999
15	-0.5999
16	0.2000
17	-0.6000
18	0.2000
19	-0.6000
20	0.2000
21	-0.6000
22	0.2000
23	-0.6000

The points in Table 3 are shown graphically in Figure 6 (n is along the horizontal axis, and x_n is along the vertical axis). The oscillating pattern emerges after a few step, regardless of the value chosen for x_0. Even for values of x_0 chosen arbitrarily close to 0, the sequence will eventually start to oscillate between $-.6$ and $.2$. So, the equilibrium point at 0 is repelling.

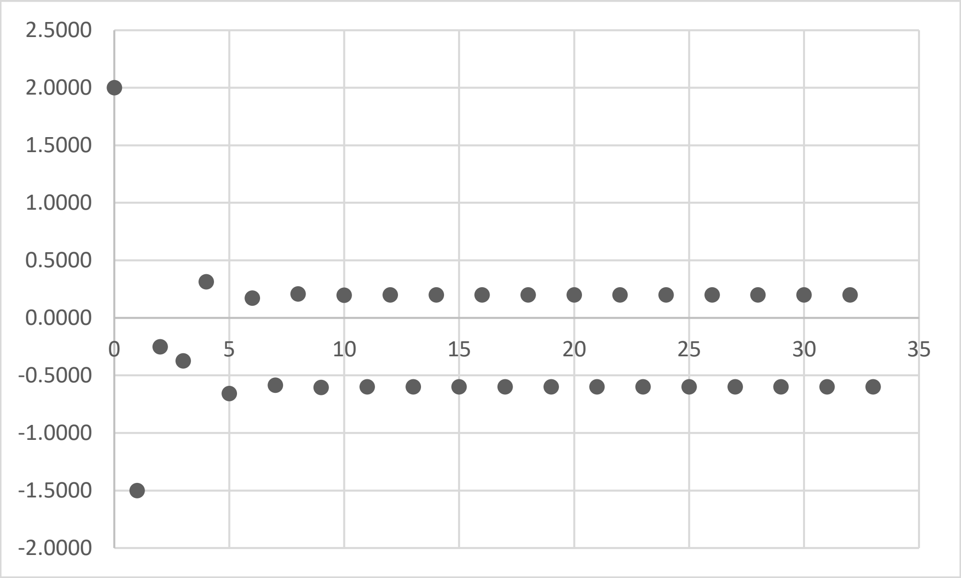


Figure 6. Unstable equilibrium point at 0

In general, a **period-k point** is a point $\bar{x}$ of the sequence $\{x_n\}$ for which a subsequence of period k converges to $\bar{x}$. In other words, $\bar{x}$ is a period-k point of the sequence $\{x_n\}$ if there exists a subsequence of the form $x_\alpha, x_{\alpha+k}, x_{\alpha+2k}, x_{\alpha+3k}, \ldots$ such that $\lim_{i\to\infty} x_{\alpha+ik} = \bar{x}$.

In the above example, $-.6$ and 0.2 are period-2 points.

. . .

The following theorem is useful for classifying equilibrium points.

Theorem 3. Let $\bar{x}$ be an equilibrium point of the difference equation $x_{n+1} = f(x_n)$ where f is continuous and differentiable at $\bar{x}$. In this case, we have

- If $|f'(\bar{x})| < 1$, then $\bar{x}$ is asymptotically stable.
- If $|f'(\bar{x})| > 1$, then $\bar{x}$ is unstable.
- If $|f'(\bar{x})| = 1$, the theorem provides no information concerning the stability of the equilibrium point.

. . .

The **logistic map** is a non-linear difference equation of order 2 given by the following formula

$$x_{n+1} = rx_n(1 - x_n)$$

with $x_0 \in (0,1)$ and $r \in (0,4]$.

In general, this difference equation cannot be solved in closed form.

The logistic map is used to model population dynamics where x_n is the ratio of an existing population to the maximum possible population and r is the growth rate (average number of offspring for each individual in the population).

Solving the equation $x = rx(1 - x)$, we find that the logistic map has two equilibrium points, i.e., $x = 0$ and $x = \frac{r-1}{r}$. The first equilibrium exists for all values of r and the second equilibrium point only exist for $r > 1$. For $r < 1$, we get $x < 0$ which is out of range by definition.

Applying Theorem 3 to $f(x) = rx(1 - x)$ and noting that $f'(x) = r - 2rx$, we see that

- $f'\left(\frac{r-1}{r}\right) = r - 2r\left(\frac{r-1}{r}\right) = 2 - r$. The equilibrium point at $\frac{r-1}{r}$ is asymptotically stable when $|2 - r| < 1$ which is equivalent to $1 < r < 3$.
- $f'(0) = r$, and so, $x = 0$ is asymptotically stable when $|r| < 1$ which is equivalent to $0 < r < 1$, noting that we are given $r \in (0,4]$.

At $r = 3.05$, there are two period-2 points at approximately 0.5902 and 0.7377, see Figure 7.

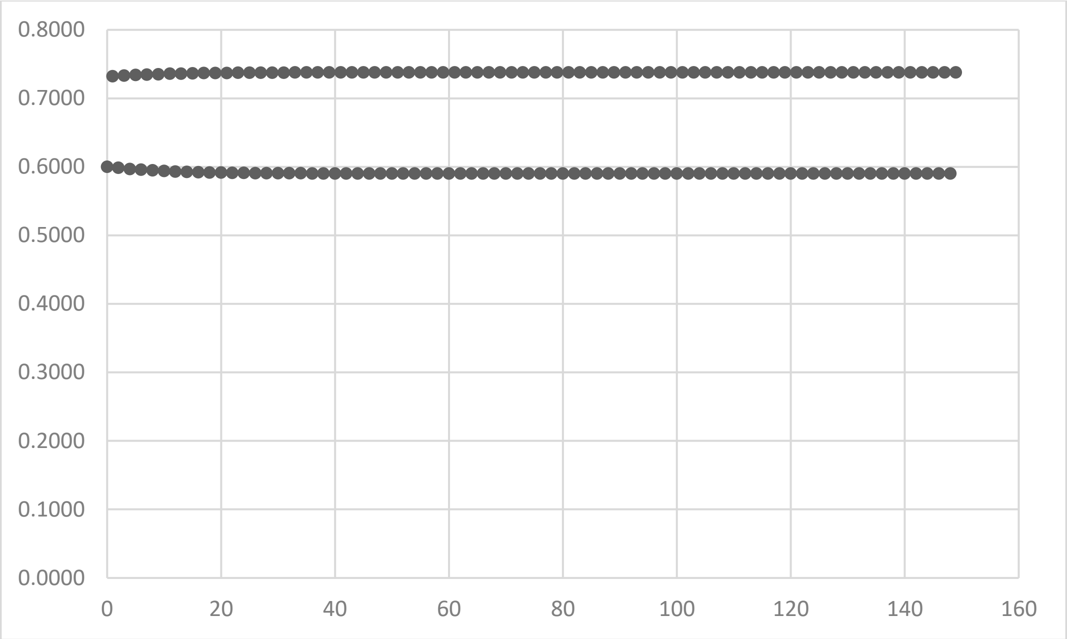


Figure 7. Logistic Map for $r = 3.05$

At $r = 3.5$, there are four period-4 points at approximately 0.3828, 0.8269, 0.5009 and 0.8750, see Figure 8.

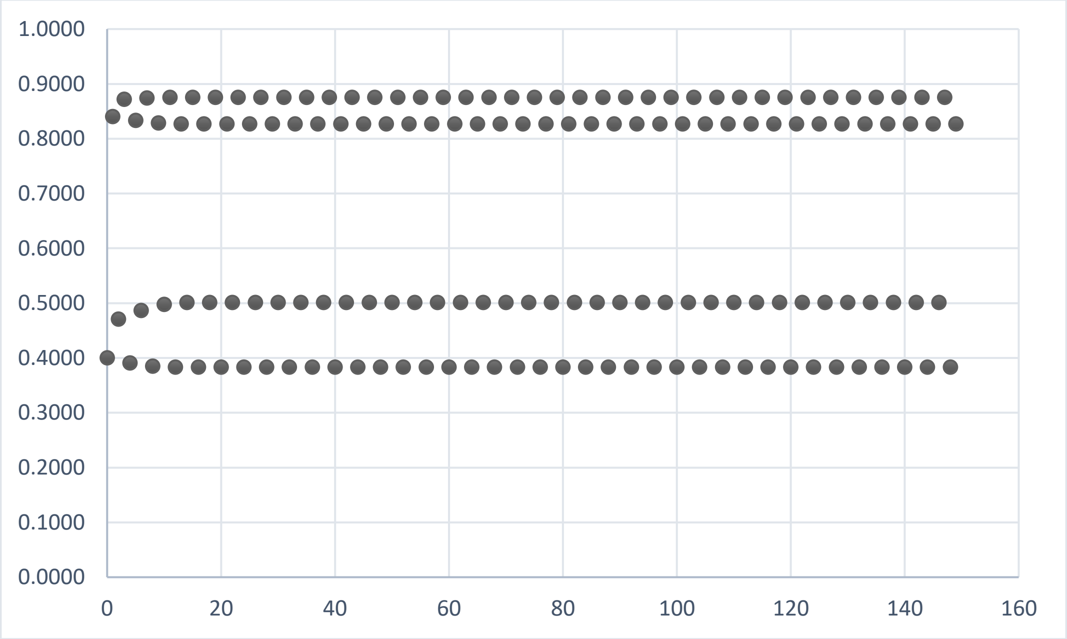


Figure 8. Logistic map for $r = 3.5$

As *r* increases beyond 3.54409, from almost all initial conditions, the sequence will oscillate among 8 values, then 16 values and so on. At $r \cong 3.56995$ chaotic behavior begins. For almost all initial conditions, there are no longer oscillations of finite period. Slight variations of x_0 lead to vastly different sequences over time. This chaotic behavior is further explained in the Wikipedia article entitled “Logistic map” [17].

5 Continued Fractions

5.1 Overview

Prerequisites: algebra, proof by mathematical induction, concept of a greatest common divisor (see Section 12.4)

A **continued fraction** is an expression of the form

$$a_1 + \cfrac{b_1}{a_2 + \cfrac{b_2}{a_3 + \cfrac{b_3}{a_4 + \cfrac{b_4}{1 + \cfrac{b_5}{a_5}}}}}$$
$$\cdots$$

When all the b_i terms are equal to 1, the expression is known as a **simple continued fraction**. In this book, only simple continued fractions are considered. When not all of the b_i terms are equal to 1, the expression is referred to as a **generalized continued fraction**.

In a finite continued fraction (simple or not), the iteration stops in a finite number of steps by using an integer in place of another continued fraction. An infinite continued fraction (simple or not) extends indefinitely. In both cases, all numbers in the sequence, **other than the first**, must be positive. The integers a_i are called the coefficients or terms of the continued fraction.

The representation of $\frac{53}{19}$ as a continued fraction is shown below.

$$\frac{53}{19} = 2 + \frac{15}{19} = 2 + \cfrac{1}{\left(\frac{19}{15}\right)} = 2 + \cfrac{1}{1 + \cfrac{4}{15}} = 2 + \cfrac{1}{1 + \cfrac{1}{\left(\frac{15}{4}\right)}}$$

$$= 2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{3}{4}}} = 2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{\left(\frac{4}{3}\right)}}} = \mathbf{2} + \cfrac{1}{\mathbf{1} + \cfrac{1}{\mathbf{3} + \cfrac{1}{\mathbf{1} + \cfrac{1}{\mathbf{3}}}}}$$

The above can also be represented in a more concise notation, i.e., $[2,1,3,1,3]$. This notation is sufficient to reconstruct the final iteration of the continued fraction.

Note that

$$\frac{1}{3} = \cfrac{1}{\left(\frac{3}{1}\right)} = \cfrac{1}{3 + \cfrac{0}{1}} = \frac{1}{3}$$

which is why we stopped at $\frac{1}{3}$.

The above representation is not unique as there is a variation, i.e., we can write the last term as

$$\frac{1}{3} = \cfrac{1}{2 + \cfrac{1}{1}}$$

With this variation (the only one possible), the resulting continued fraction representation is $[2,1,3,1,2,1]$. This variation is useful in some proofs when either an even or odd number of terms is required.

If the faction is less than one, a very similar procedure is used. For example, take $\frac{19}{53}$. The first term in the continued fraction is 0. The second term and beyond have already been computed in the previous problem. The final result is

$$\frac{19}{53} = 0 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{3}}}}}$$

In shorthand notation, $\frac{19}{53} = [0,2,1,3,1,3]$.

. . .

In the case of negative fractions, we need to determine the negative quotient that leaves the smallest positive remainder. For example, consider $-\frac{59}{13}$ which can be written as $-5 + \frac{6}{13}$. From here, we follow the procedure for positive continued fractions, i.e.,

$$-5 + \frac{6}{13} = -5 + \frac{1}{\left(\frac{13}{6}\right)} = -5 + \cfrac{1}{2 + \cfrac{1}{6}}$$

In notation, $-\frac{59}{13} = [-5,2,6]$.

As a second example, we determine the continued fraction for the reciprocal of the fraction in the previous problem. First, we write $-\frac{13}{59} = -1 + \frac{46}{59}$ and then proceed to determine the continued fraction as follows:

$$-\frac{13}{59} = -1 + \frac{46}{59} = -1 + \frac{1}{\left(\frac{59}{46}\right)} = -1 + \cfrac{1}{1 + \cfrac{13}{46}} = -1 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{7}{13}}}$$

$$= -1 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{6}{7}}}} = -1 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{6}}}}}$$

In notation, $-\frac{13}{59} = [-1,1,3,1,1,6]$.

As practice, try expressing the following as continued fractions:

- $\frac{11}{7}$ Answer: [1,1,1,3]
- $\frac{7}{11}$ Answer: [0,1,1,1,3]
- $\frac{23}{6}$ Answer: [3,1,5]
- $\frac{29}{18}$ Answer: [1,1,1,1,1,3]
- $-\frac{34}{13}$ Answer: [-3,2,1,1,2]

The following website was just to compute the above answers: https://www.alpertron.com.ar/CONTFRAC.HTM.

5.2 Concepts

By working backward, it is apparent that every finite continued fraction (simple or not) is a rational number. It is also true that every rational number can be represented by a finite simple continued fraction. Before proving this fact, we state an important result from basic number theory.

Theorem 4. (Euclid's division lemma) Given two integers a and b, such that $b \neq 0$, there exist unique integers c and r such that

$$a = bc + r$$

and

$$0 \le r < |b|.$$

In the above, a is called the dividend, b is called the divisor, c is called the quotient and r is called the remainder.

For a proof of this result, see the Wikipedia article entitled "Euclidean division" [18].

. . .

Theorem 5. Every rational number can be represented as a finite simple continued fraction. The solution is unique, except for the last term in the continued fraction expansion.

Proof: Taken any rational number $\frac{p}{q}$ such that $q > 0$ (p can be positive, negative or zero). From Euclid's division lemma, there exists a unique integer a_1 such that

$$p = qa_1 + r_1, \qquad 0 \le r_1 < q$$

which can be written as

$$\frac{p}{q} = a_1 + \frac{r_1}{q}, \qquad 0 \le r_1 < q$$

If $r_1 = 0$, then we are done and $\frac{p}{q} = [a_1]$. Otherwise, we continue the process of representing $\frac{p}{q}$ as a continued fraction. Noting that $\frac{p}{q} = a_1 + \frac{1}{\left(\frac{q}{r_1}\right)}$, we then apply Euclid's division lemma to $\frac{q}{r_1}$ to get

$$\frac{q}{r_1} = a_2 + \frac{r_2}{r_1}, \qquad 0 \leq r_2 < r_1$$

where a_2 is the unique largest positive integer such $0 \leq r_2 < r_1$. If $r_2 = 0$, then we are done and $\frac{p}{q} = [a_1, a_2]$. Otherwise, we continue the process.

Eventually, the process must stop since at each step $0 \leq r_{i+1} < r_i < \cdots < r_1$ and there can only be a finite number of integers between r_1 and 0. If the process stops at step n, then

$$\frac{p}{q} = [a_1, a_2, \dots, a_n]$$

At each step, the selection of a_i was unique. However, it is possible to vary the representation of the last term.

If $a_n > 1$, then the last term in the expansion can be written as

$$\frac{1}{a_n} = \frac{1}{(a_n - 1) + \frac{1}{1}}$$

and in this case, we have the alternate continued fraction representation

$$\frac{p}{q} = [a_1, a_2, \dots, a_{n-1}, a_n - 1,1]$$

If $a_n = 1$, then

$$\frac{1}{a_{n-1} + \frac{1}{a_n}} = \frac{1}{a_{n-1} + 1}$$

and in this case, we have the alternate continued fraction representation

$$\frac{p}{q} = [a_1, a_2, \dots, a_{n-1} + 1]$$

. . .

As we shall see, the truncated continued fraction representations for a number provide progressively better approximations for the given number. These truncated representations are known as **convergents**.

Given a continued fraction $[a_1, a_2, a_3, \dots]$, the convergents are defined as follows:

$$c_1 = a_1;\ c_2 = a_1 + \frac{1}{a_2};\ c_3 = a_1 + \frac{1}{a_2 + \frac{1}{a_3}}, \dots$$

The above definition applies to both finite and infinite continued fractions.

For example, the convergents $-\frac{13}{59}$ are

$$c_1 = -1$$

$$c_2 = 0$$

$$c_3 = -\frac{1}{4} = -.25$$

$$c_4 = -\frac{1}{5} = -.2$$

$$c_5 = -\frac{2}{9} = -.22\dots$$

$$c_6 = -\frac{13}{59} \cong -.220338$$

Each convergent is a rational number and so, we can write c_i as $\frac{p_i}{q_i}$ where p_i and q_i are integers.

The following theorem is helpful in computing convergents.

Theorem 6. The convergents for the continued fraction $[a_1, a_2, a_3, \dots]$, represented as $c_i = \frac{p_i}{q_i}$ satisfy the following equations for $i = 1,2,3,4,5,\dots$

$$p_i = a_i p_{i-1} + p_{i-2}$$

$$q_i = a_i q_{i-1} + q_{i-2}$$

with initial values

$$p_{-1} = 0, p_0 = 1$$

$$q_{-1} = 1, q_0 = 0$$

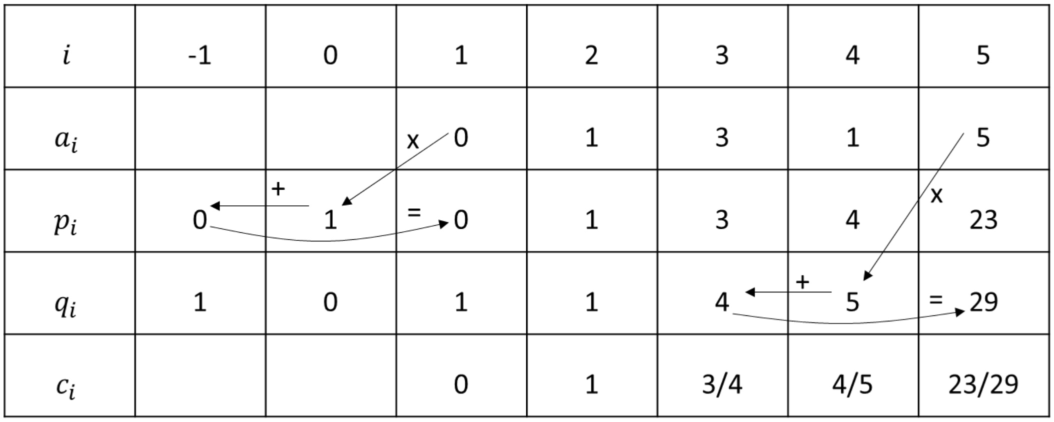
The proof is by induction. Details can be found in Chapter 1, Section 2 of the book “Continued Fractions” [19].

Remarks concerning Theorem 6:

- All the q_i terms are positive, since $q_1 = a_1 q_0 + q_{-1} = a_1(0) + 1 = 1$, $q_2 = a_2 q_1 + q_0 = a_2(1) + 0 = a_2$ and subsequent terms are given by $q_i = a_i q_{i-1} + q_{i-2}$ where $a_i > 0$ for $i \geq 2$. Recall that we defined the continued fraction representation such that only $a_1$could possibly be negative.

- From the formulas stated in the theorem, we can verify that $c_1 = \frac{p_1}{q_1} = \frac{a_1(1)+0}{1} = a_1;\ c_2 = \frac{p_2}{q_2} = \frac{a_2p_1+1}{a_2q_1+0} = \frac{a_1a_2+1}{a_2} = a_1 + \frac{1}{a_2}$ and so on.

The formulas from Theorem 6 can be put in table form for easier calculation. Figure 9 shows the calculation of the convergents for $\frac{23}{29}$. The first step is to calculate the continued fraction representation and place that in the table. For the problem at hand, $\frac{23}{29} = [0, 1, 3, 1, 5]$. Starting with the initial values (which are the same in all cases), we start to calculate the values of the p_i and q_i terms. For example, $p_1 = a_1p_0 + p_{-1}$ (this is shown pictorially in the table). The same visual pattern repeats for other p_i terms. An example of the calculation for $q_5 = a_5q_4 + q_3$ is shown to the right of the figure. This same visual pattern can be used to compute the other q_i terms, with the computations going from left to right (starting with q_1). Computation of the c_i terms is simple, once we have computed the p_i and q_i terms. Notice that the c_i terms get closer to $\frac{23}{29}$ at each step. This is true in all cases.



i	-1	0	1	2	3	4	5
a_i			0	1	3	1	5
p_i	0	1	0	1	3	4	23
q_i	1	0	1	1	4	5	29
c_i			0	1	3/4	4/5	23/29

Figure 9. Calculation of convergents for 23/29

To further highlight the patterns exhibited by convergents, we expand $\frac{101}{129}$ into a continued fraction and then compute the convergents. This particular fraction has a rather long continued fraction representation, i.e., $\frac{101}{129} = [0,1,3,1,1,1,1,5]$. The convergents are shown in Table 4.

Table 4. Calculation for convergents for 101/129

i	-1	0	1	2	3	4	5	6	7	8
a_i			0	1	3	1	1	1	1	5
p_i	0	1	0	1	3	4	7	11	18	101
q_i	1	0	1	1	4	5	9	14	23	129
c_i			0	1	$\frac{3}{4}$	$\frac{4}{5}$	$\frac{7}{9}$	$\frac{11}{14}$	$\frac{18}{23}$	$\frac{101}{129}$

Each convergent gets closer to the actual fraction, i.e., $0, 1, .75, .8, .7778, .7857, .7826, .7829$ (numbers shown to 4 decimal place accuracy). Further, the convergents with odd indices, i.e., $0, .75, .7778, .7826$, approach $\frac{101}{129}$ from below and the convergents with even indices, i.e., $1, .8, .7857$, approach $\frac{101}{129}$ from above. As we shall prove, this is always the case but first, we need some preliminary results.

Theorem 7. If the convergents for a continued fraction are $c_i = \frac{p_i}{q_i}$, then

$$p_i q_{i-1} - p_{i-1} q_i = (-1)^i, \qquad i \geq 0$$

Proof: The proof is by induction.

From the initial conditions defined in Theorem 6, we have

$$p_0 q_{-1} - p_{-1} q_0 = 1 \cdot 1 - 0 \cdot 0 = 1$$

and with some additional calculations

$$p_1 q_0 - p_0 q_1 = a_1 \cdot 0 - 1 \cdot 1 = -1$$

So, the theorem holds true for the cases $i = 0$ and $i = 1$.

Next, assume the theorem is true for the case $i = k$ and show that this implies the theorem must be true for the case $i = k + 1$.

From Theorem 6, with $i = k + 1$, we have

$$p_{k+1} = a_{k+1} p_k + p_{k-1}$$

$$q_{k+1} = a_{k+1} q_k + q_{k-1}$$

Using the above two equations,

$$p_{k+1} q_k - p_k q_{k+1} = (a_{k+1} p_k + p_{k-1}) q_k - p_k (a_{k+1} q_k + q_{k-1})$$

$$= a_{k+1}p_kq_k + p_{k-1}q_k - p_ka_{k+1}q_k - p_kq_{k-1}$$

$$= (-1)(\, p_{k-1}q_k - p_kq_{k-1})$$

Given the assumption that the theorem holds for $i = k$, i.e.,

$$p_kq_{k-1} - p_{k-1}q_i = (-1)^k$$

and substituting into the above expression yields

$$p_{k+1}q_k - p_kq_{k+1} = (-1)^{k+1}$$

Thus, we have shown that the theorem holds for $i = k + 1$ and that completes the induction proof. ■

The next theorem is essentially a corollary to the previous theorem.

Theorem 8. Each convergent $c_i = \frac{p_i}{q_i}$ of a continued fraction is in reduced form, i.e., $gcd(p_i, q_i) = 1$.

Proof:

From Theorem 7, we know that $p_iq_{i-1} - p_{i-1}q_i = (-1)^i$. Thus, if any number exactly divides p_i and q_i, it must divide $(-1)^i$, i.e., the number must be -1 or 1. In other words, the only divisors of p_i and q_i are -1 or 1, i.e., $gcd(\mathrm{p_i}, \mathrm{q_i}) = 1$ (recall that the GCD is defined to be a positive number). ■

When dividing the equation from Theorem 7 by q_nq_{n-1}, we get

$$\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^n}{q_nq_{n-1}}$$

By definition, $c_i = \frac{p_i}{q_i}$ and so

Equation 4. Difference of two adjacent convergents

$$c_n - c_{n-1} = \frac{(-1)^n}{q_nq_{n-1}}, \qquad n \geq 2$$

Next, let's determine a formula for $c_n - c_{n-2}$.

By definition,

$$c_n - c_{n-2} = \frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{p_nq_{n-2} - p_{n-2}q_n}{q_nq_{n-2}}$$

Next, substitute the equations for p_n and q_n (from Theorem 6) into the numerator of the above equation to get

$$p_nq_{n-2} - p_{n-2}q_n = (a_np_{n-1} + p_{n-2})q_{n-2} - p_{n-2}(a_nq_{n-1} + q_{n-2})$$

$$= a_n(p_{n-1}q_{n-2} - p_{n-2}q_{n-1})$$

From Theorem 7, $p_{n-1}q_{n-2} - p_{n-2}q_{n-1} = (-1)^{n-1}$ and so, we have the result

Equation 5. Difference of convergents that are two index values apart

$$c_n - c_{n-2} = \frac{a_n(-1)^{n-1}}{q_n q_{n-2}}, \qquad n \geq 3$$

Equation 4 and Equation 5 can be used to prove that, in general,

Theorem 9. The following inequality holds true for the convergents of a simple continued fraction (finite or infinite)

$$c_1 < c_3 < c_5 < \cdots < c_{2n+1} < \cdots < c_{2n} < \cdots < c_6 < c_4 < c_2$$

Proof:

Substituting $n = 2$ and then $n = 3$ into Equation 4, and noting that $q_i > 0$ for $i \geq 1$, we have that

$$c_2 - c_1 = \frac{1}{q_1 q_2} > 0, \qquad c_3 - c_2 = -\frac{1}{q_2 q_3} < 0$$

Setting $n = 3$ in Equation 5, and noting that $a_i \geq 0$ for all $i \geq 2$, we get

$$c_3 - c_1 = \frac{a_3(-1)^2}{q_1 q_3} > 0$$

The above three inequalities imply $c_1 < c_3 < c_2$.

Next, substitute $n = 4$ into Equation 4, and $n = 4$ into Equation 5 to get

$$c_4 - c_3 = \frac{1}{q_3 q_4} > 0$$

$$c_4 - c_2 = \frac{-a_4}{q_2 q_4} < 0$$

Thus, we have $c_1 < c_3 < c_4 < c_2$.

Continuing in this manner, we have the desired string of inequalities. ■

We have now shown that the odd indexed convergents strictly increase and the even indexed convergents strictly decrease, but we have not shown that they converge (from below and above) to the continued fraction under consideration. Clearly, in the case of a finite simple continued fraction, the sequence of convergents does converge since the last convergent (say c_n) equals the continued fraction $[a_1, a_2, \ldots, a_n]$. If the simple continued fraction is infinite, the sequence of convergents also converges. We state this result in the following theorem.

Theorem 10. The sequence of convergents for the infinite simple continued fraction $[a_1, a_2, a_3, \dots]$ *converges to the number being represented by the continued fraction.*

For a proof, see Theorem 3.6 and the associated development in the book by Olds [20].

5.3 Continued Fractions for Irrational Numbers

Thus far, all of the examples in this section have involved rational numbers, which as we saw, always have a finite representation as a simple continued fraction. However, keep in mind that the theorems stated in Section 5.2 apply to both finite and infinite simple continued fractions.

The expansion of an irrational number follows the same process as that for a rational number, except that the expansion is always infinite. That the expansion for an irrational number is infinite should be clear since if the expansion stopped at a finite number of steps, one would have a rational number – a contradiction.

The process goes as follows for irrational number x.

At the first step, we choose largest integer less than x (call it a_1) such that

$$x = a_1 + \frac{1}{x_1}, \qquad 0 < \frac{1}{x_1} < 1$$

Since $x_1 = \frac{1}{x - a_1} > 1$, x_1 is also irrational.

Next, we choose the largest integer less than x_1 (call it a_2) such that

$$x_1 = a_2 + \frac{1}{x_2}, \qquad 0 < \frac{1}{x_2} < 1, \qquad a_1 \geq 1$$

Again, x_2 is irrational.

The process continues indefinitely, and we get the following representation

$$x = [a_1, a_2, a_3, \dots], a_i \geq 1 \text{ for } i = 2, 3, 4, \dots$$

At each step, the x_i term is irrational.

As an example of the above process, consider $\sqrt{7} \cong 2.645751311$.

Step 1. The largest integer less than $\sqrt{7}$ is 2, and so we write

$$\sqrt{7} = 2 + \frac{1}{x_1}$$

$$\Rightarrow \; x_1 = \frac{1}{\sqrt{7} - 2} = \frac{1}{\sqrt{7} - 2} \cdot \frac{\sqrt{7} + 2}{\sqrt{7} + 2} = \frac{1}{3}(\sqrt{7} + 2) \cong 1.54858377$$

Step 2. The largest integer less than x_1 is 1, and so we write

$$\frac{1}{3}\left(\sqrt{7} + 2\right) = 1 + \frac{1}{x_2}$$

$$\Longrightarrow\ x_2 = \frac{3}{\sqrt{7}-1} = \frac{3}{\sqrt{7}-1}\cdot\frac{\sqrt{7}+1}{\sqrt{7}+1} = \frac{\sqrt{7}+1}{2} \cong 1.822875656$$

Step 3. The largest integer less than x_2 is 1, and so we write

$$\frac{\sqrt{7}+1}{2} = 1 + \frac{1}{x_3}$$

$$\Longrightarrow\ x_3 = \frac{2}{\sqrt{7}-1} = \frac{2}{\sqrt{7}-1}\cdot\frac{\sqrt{7}+1}{\sqrt{7}+1} = \frac{\sqrt{7}+1}{3} \cong 1.215250437$$

Step 4. The largest integer less than x_3 is 1, and so we write

$$\frac{\sqrt{7}+1}{3} = 1 + \frac{1}{x_4}$$

$$\Longrightarrow\ x_4 = \frac{3}{\sqrt{7}-2} = \frac{3}{\sqrt{7}-2}\cdot\frac{\sqrt{7}+2}{\sqrt{7}+2} = \sqrt{7}+2 \cong 4.645751311$$

Step 4. The largest integer less than x_4 is 4, and so we write

$$\sqrt{7}+2 = 4 + \frac{1}{x_5}$$

$$\Longrightarrow\ x_5 = \frac{1}{\sqrt{7}-2} = \frac{1}{\sqrt{7}-2}\cdot\frac{\sqrt{7}+2}{\sqrt{7}+2} = \frac{1}{3}(\sqrt{7}+2)$$

This brings us back to the same result as in Step 1, and so the process repeats the same pattern over and over. The final result is

$$\sqrt{7} = [2, \overline{1,1,1,4}]$$

where $\overline{1,1,1,4}$ means that the sequence 1,1,1,4 repeats indefinitely.

Here are some additional continued fraction expansions for irrational numbers (computed using the online application at https://www.alpertron.com.ar/CONTFRAC.HTM):

- $\sqrt{23} = [4, \overline{1, 3, 1, 8}]$
- $\sqrt{33} = [\ 5, \overline{1, 2, 1, 10}]$
- $\sqrt{1291} = [\ 35, \overline{1, 13, 2, 1, 1, 2, 3, 1, 1, 1, 1, 4, 1, 11, 6, 2, 4, 3, 23, 1, 1, 1, 4, 7, 1, 3, 2, 1,}$ $\overline{6, 2, 35, 2, 6, 1, 2, 3, 1, 7, 4, 1, 1, 1, 23, 3, 4, 2, 6, 11, 1, 4, 1, 1, 1, 1, 3, 2, 1, 1, 2, 13, 1, 70}]$

[**Author's Remark**: There are several online applications that compute continued fractions. For example, try the command "continued fraction for sqrt(1291)" at wolframalpha.com.**]**

The repeating pattern is no accident. In 1770, Lagrange proved the following theorem for **quadratic irrational numbers**, i.e., numbers of the form $\frac{a+\sqrt{b}}{c}$ where a, b and c are integers and b is not a square.

Theorem 11. All quadratic irrational numbers have a continued fraction expansion which is periodic (after possibly some initial non-periodic terms).

The proof (which relies on several preliminary results) can be found in Section 4.6 of the book by Olds [20].

It is possible to work in reverse, i.e., given a continued fraction, find the associated number. For example, take the continued fraction

$$x = 2 + \cfrac{1}{6 + \cfrac{1}{3 + \cfrac{1}{6 + \cfrac{1}{3 + \cfrac{1}{6 + \cdots}}}}}$$

Let

$$y = 6 + \cfrac{1}{3 + \cfrac{1}{6 + \cfrac{1}{3 + \cfrac{1}{6 + \cdots}}}}$$

Thus, we have

$$x = 2 + \frac{1}{y}, \qquad y = 6 + \cfrac{1}{3 + \cfrac{1}{y}}$$

Working with the latter equation, gives us

$$y = 6 + \frac{y}{3y + 1}$$

$$3y^2 + y = 6(3y + 1) + y$$

$$y^2 - 6y - 2 = 0$$

Using the quadratic formula and choosing the positive alternative, we have that

$$y = \frac{6 + \sqrt{44}}{2} = \frac{6 + 2\sqrt{11}}{2} = 3 + \sqrt{11}$$

So, we have that

$$x = 2 + \frac{1}{3+\sqrt{11}} = 2 + \frac{1}{3+\sqrt{11}} \cdot \frac{3-\sqrt{11}}{3-\sqrt{11}} = 2 + \frac{3-\sqrt{11}}{-2} = 2 - \frac{3}{2} + \frac{\sqrt{11}}{2} = \frac{1+\sqrt{11}}{2}$$

In fact, the continued fraction representation for $\frac{1+\sqrt{11}}{2}$ is $[2,\overline{6,3}]$.

. . .

Continued fraction can be computed for transcendental numbers such as π and e. The first few terms for the simple continued fraction for π are

$$[3, 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, 84, 2, 1, 1, 15, 3, 13, 1, \dots]$$

There is no known pattern to this sequence and the only known way to obtain the terms is to compute them one-by-one using a known decimal approximation for π. According to http://oeis.org/A001203, 30,113,021,586 terms have been computed for the continued fraction representation of π by Syed Fahad on 27 April 2021. This is the current record.

However, there are several **generalized** continued fraction representations for π which do exhibit patterns, e.g.,

$$\pi = \cfrac{4}{1 + \cfrac{1^2}{3 + \cfrac{2^2}{5 + \cfrac{3^2}{7 + \cfrac{4^2}{9 + \cdots}}}}}$$

For additional examples, see https://en.wikipedia.org/wiki/Continued_fraction#Generalized_continued_fraction.

The simple continued fraction for e does exhibit a pattern, see the first few terms below

$$[2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, 1, 1, 12, 1, 1, 14, 1, 1, 16, 1, 1, 18, 1, 1, 20, 1, 1, 22, 1, 1, 24, 1, 1, 26, 1, 1, 28,$$
$$1, 1, 30, 1, 1, 32, 1, 1, 34, 1, 1, 36, 1, 1, 38, 1, 1, 40, 1, 1, 42, 1, 1, 44, 1, 1, 46, 1, 1, 48, 1, 1, 50, 1, 1, 52, 1, 1, 54, \dots]$$

6 Nested Radicals

Prerequisites: algebra, proof by mathematical induction, concept of a limit from calculus, suggest reading Sections 4 and 5 before this section

Similar in concept to continued fraction is the concept of nested radicals (basically radicals enclosing other radicals in either a finite or infinite sequence).

For example, $\sqrt{5+2\sqrt{6}}$ is a finite nested radical. Since $\left(\sqrt{2}+\sqrt{3}\right)^2 = 5+2\sqrt{6}$, we can simplify $\sqrt{5+2\sqrt{6}}$ to the expression $\sqrt{2}+\sqrt{3}$ where "simplification" means fewer layers of nesting.

The expression

$$\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}}$$

is an example of an infinite nested radical. As we shall see, this expression is equal to the Golden Ratio, i.e., $\frac{1+\sqrt{5}}{2}$.

The Indian mathematician Srinivasa Ramanujan is famous for his work in number theory, including many formulas related to nested radicals. Ramanujan posed the follow problem to readers of the *Journal of the Indian Mathematical Society*:

> Find the value of $\sqrt{1+2\sqrt{1+3\sqrt{1+\cdots}}}$

After no solution was proposed after six months, Ramanujan provided a solution to the more general problem:

$$x+n+a = \sqrt{ax+(n+a)^2+x\sqrt{a(x+n)+(n+a)^2+(x+n)\sqrt{\ldots}}}$$

Taking $x=2, n=1$ and $a=0$ in the above equation yields

$$3 = \sqrt{1+2\sqrt{1+3\sqrt{1+\cdots}}}$$

The Wikipedia article on nested radicals [21] states several other formulas discovered by Ramanujan, e.g.,

$$\sqrt[3]{\sqrt[5]{\frac{32}{5}}-\sqrt[5]{\frac{27}{5}}} = \sqrt[5]{\frac{1}{25}}+\sqrt[5]{\frac{3}{25}}-\sqrt[5]{\frac{9}{25}}$$

The process of simplifying a nested radical to an equivalent expression with no nesting is referred to as **denesting**.

In what follows, we consider a small subset of the set of nested radicals.

. . .

As was the case for continued fractions, we need to precisely define what is meant by an infinite nested radical and then show that a given infinite radical converges.

As a first case, consider infinite nested radicals of the form

$$\sqrt{a+\sqrt{a+\sqrt{a+\sqrt{a+\cdots}}}}$$

such that a is a real number greater than zero. We define the above expression as the limit of the sequence

$$\sqrt{a}, \sqrt{a+\sqrt{a}}, \sqrt{a+\sqrt{a+\sqrt{a}}}, \sqrt{a+\sqrt{a+\sqrt{a+\sqrt{a}}}}, \dots$$

Let us define $c_1(a)$ as $\sqrt{a}$ and for each $n \geq 1$, define $c_{n+1}(a) = \sqrt{a + c_n(a)}$, giving us something similar in concept to the convergents that we defined for continued fractions.

Theorem 12. The sequence $\{c_n(a)\}$ is strictly increasing and bounded, and thus $c(a) = \lim_{n\to\infty} c_n(a)$ exists.

Proof:

We use induction to prove that the sequence is strictly increasing. Since $a > 0$, $c_1(a) = \sqrt{a} < \sqrt{a+\sqrt{a}} = c_2(a)$. If $c_{n-1}(a) < c_n(a)$, then $c_n(a) = \sqrt{a + c_{n-1}(a)} < \sqrt{a + c_n(a)} = c_{n+1}(a)$. So, the induction hypothesis is true, and we have shown that $\{c_n(a)\}$ is strictly increasing.

To show that $\{c_n(a)\}$ is bounded, we consider two cases, i.e., $a \geq 2$ and $0 < a < 2$. In each case, the proof is by induction.

For $a \geq 2$, we have $0 < c_1(a) = \sqrt{a} < a$.

By way of induction, assume $0 < c_{n-1}(a) \leq a$, and prove the same for $c_n(a)$. Given the induction hypothesis, we have

$$0 < c_n(a) = \sqrt{a + c_{n-1}(a)} \leq \sqrt{2a} \leq \sqrt{a^2} = a$$

This proves the induction hypothesis. Thus, for the case $a \geq 2$, $\{c_n(a)\}$ is bounded above by a and below by 0.

For $0 < a < 2$, we have $c_1(a) = \sqrt{a} < \sqrt{2} < 2$.

By way of induction, assume $0 < c_{n-1}(a) \leq 2$, and prove the same for $c_n(a)$. Given the induction hypothesis, we have

$$0 < c_n(a) = \sqrt{a + c_{n-1}(a)} \leq \sqrt{a+2} \leq \sqrt{4} = 2.$$

This proves the induction hypothesis. Thus, for the case $0 < a < 2$, $\{c_n(a)\}$ is bounded above by 2 and below by 0.

Since $\{c_n(a)\}$ is strictly increasing and bounded, $\lim_{n\to\infty} c_n(a)$ exists. ■

Let $\lim_{n\to\infty} c_n(a) = c(a)$ and take the limit as n approaches infinity on both sides of the equation $c_{n+1}(a) = \sqrt{a + c_n(a)}$ to get $c = \sqrt{a+c}$ which can be written as $c^2 - c - a = 0$. Using the quadratic formula, we get that $c(a) = \frac{1\pm\sqrt{1+4a}}{2}$. However, we know that $c(a) > 0$ and $a > 0$, and so the unique solution is $\frac{1+\sqrt{1+4a}}{2}$. The function defined by $c(a) = \frac{1+\sqrt{1+4a}}{2}$ is 1-1, i.e., if $c(a_1) = c(a_2)$ then it must be that $a_1 = a_2$.

This verifies our earlier assertion that $\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}}$ is equal to the Golden Ratio.

We can use the above analysis to determine an equivalent infinite nested radical for any number greater than 1. (The condition of being greater than one comes from the fact that $\frac{1+\sqrt{1+4a}}{2} > 1$.) For example, take $\frac{7}{3}$. To determine an infinite nested radical that converges to $\frac{7}{3}$, we want $c(a) = \lim_{n\to\infty} c_n(a) = \frac{7}{3}$ which means that $a = c(a)^2 - c(a) = \frac{49}{9} - \frac{7}{3} = \frac{28}{9}$. Thus,

$$\frac{7}{3} = \sqrt{\frac{28}{9} + \sqrt{\frac{28}{9} + \sqrt{\frac{28}{9} + \sqrt{\frac{28}{9} + \cdots}}}}$$

. . .

Next, we consider nested radicals of the form

$$\sqrt{\frac{a}{b^2} + \sqrt{\frac{a}{b^2} + \sqrt{\frac{a}{b^2} + \sqrt{\frac{a}{b^2} + \cdots}}}}$$

where a and b are positive real numbers. The associated convergents are

$$\sqrt{\frac{a}{b^2}}, \sqrt{\frac{a}{b^2} + \sqrt{\frac{a}{b^2}}}, \sqrt{\frac{a}{b^2} + \sqrt{\frac{a}{b^2} + \sqrt{\frac{a}{b^2}}}}, \sqrt{\frac{a}{b^2} + \sqrt{\frac{a}{b^2} + \sqrt{\frac{a}{b^2} + \sqrt{\frac{a}{b^2}}}}}, \dots$$

From our previous analysis, we know that the limit of the convergents is

$$\frac{1 + \sqrt{1 + \frac{4a}{b^2}}}{2}$$

Multiplying b times the above nested radical, we get

$$\sqrt{a+b\sqrt{a+b\sqrt{a+b\sqrt{a+\cdots}}}}$$

The associated convergents are

$$\sqrt{a}, \sqrt{a+b\sqrt{a}}, \sqrt{a+b\sqrt{a+b\sqrt{a}}}, \sqrt{a+b\sqrt{a+b\sqrt{a+b\sqrt{a}}}}, \ldots$$

and the limit of the convergents is

$$b\cdot\frac{1+\sqrt{1+\frac{4a}{b^2}}}{2}=\frac{b+\sqrt{b^2+4a}}{2}$$

Some examples

$$\sqrt{1+2\sqrt{1+2\sqrt{1+2\sqrt{1+\cdots}}}}=\frac{2+\sqrt{4+4}}{2}=1+\sqrt{2}$$

$$\sqrt{44+7\sqrt{44+7\sqrt{44+7\sqrt{44+\cdots}}}}=\frac{7+\sqrt{49+176}}{2}=\frac{7+15}{2}=11$$

$$\sqrt{7+6\sqrt{7+6\sqrt{7+6\sqrt{7+\cdots}}}}=\frac{6+\sqrt{36+28}}{2}=\frac{6+8}{2}=7$$

$$\sqrt{14+5\sqrt{14+5\sqrt{14+5\sqrt{14+\cdots}}}}=\frac{5+\sqrt{25+56}}{2}=\frac{5+9}{2}=7$$

In general, for a given integer n, there are multiple values of a and b such that

$$n=\sqrt{a+b\sqrt{a+b\sqrt{a+b\sqrt{a+\cdots}}}}$$

To determine the integer values for a and b that satisfy the above equation, we solve the following equation for a in terms of b and n:

$$n = \frac{b + \sqrt{b^2 + 4a}}{2}$$

$$2n - b = \sqrt{b^2 + 4a}$$

$$4n^2 - 4nb + b^2 = b^2 + 4a$$

$$a = n^2 - nb$$

$$a = n(n - b)$$

Keeping in mind that $a > 0$ and $b > 0$, the only possible integer values for b are $1, 2, \ldots, (n-1)$. In the previous examples, we saw that $b = 6, a = 7$ and $b = 5, a = 14$ yield a solution of $n = 7$. The full set of integer solutions that yield $n = 7$ are

$\boldsymbol{b}$	1	2	3	4	5	6
$\boldsymbol{a}$	42	35	28	21	14	7

. . .

Infinite nested radicals with alternating negative numbers are valid and converge under certain conditions. In particular, the nested radical

$$\sqrt{a - b\sqrt{a - b\sqrt{a - b\sqrt{a - \cdots}}}}, a > 0, b > 0$$

is valid (i.e., does not involve imaginary numbers) under the condition $a \geq b^2$. The associated sequence is

$$\sqrt{a}, \sqrt{a - b\sqrt{a}}, \sqrt{a - b\sqrt{a - b\sqrt{a}}}, \ldots$$

Let $x_1(a, b) = \sqrt{a}$ (the first term in the above sequence) and then define the subsequent terms by $x_{n+1}(a, b) = \sqrt{a - bx_n(a, b)}$. We have the following theorem concerning convergence of $x_n(a, b)$.

Theorem 13. For positive numbers a, b and x, $x = \lim_{n\to\infty} x_n(a, b)$ if and only if

- $0 < b < \phi x$ and
- $a = x(x + b)$.

Details of the proof can be found in the paper by Zimmerman and Ho [22]. Note that ϕ in the above theorem is the Golden Ratio, i.e., $\frac{1+\sqrt{5}}{2}$.

For example, if we choose $x = 7$ and $b = 3$, then $a = 7(7 + 3) = 70$. Note that $0 < b < \phi x = \left(\frac{1+\sqrt{5}}{2}\right)(7) \cong 11.326$. So, we can apply Theorem 9 to get

$$7 = \sqrt{70 - 3\sqrt{70 - 3\sqrt{70 - 3\sqrt{70 - \cdots}}}}$$

The associated sequence converges very quickly. The first several terms (to 6 decimal place accuracy) are as follows

> 8.366600, 6.700761, 7.063832, 6.986308,7.002933, 6.999371, 7.000135, 6.999971, 7.000006, 6.999999, 7.000000, 7.000000, 7.000000

. . .

[Author's Remark: To the best of my knowledge, there are no elementary textbooks dedicated to nested radicals. For further information on this topic, the best resource appears to be the Wikipedia article on this topic [21]. The Wolfram MathWorld article entitled "Nested Radical" [23] also has some introductory material and a good reference list.**]**

7 Diophantine Equations

7.1 Overview

Prerequisites: algebra, continued fractions (as described in Section 5), concept of a greatest common divisor (see Section 12.4), some basic linear algebra is needed for Section 7.3

A **Diophantine equation** is a polynomial equation with two or more unknowns, such that the only solutions of interest are integers. Diophantine problems typically have fewer equations than unknowns and involve finding integers that simultaneously solve all equations. The word *Diophantine* refers to the Hellenistic mathematician of the 3rd century, Diophantus of Alexandria, who made a study of such equations.

. . .

The equation $y = 3x + 2$ is an example of a linear Diophantine equation. It has an infinite number of integer solutions. In this example, each integer value of x gives rise to a unique integer value solution for y. In general, the solution of a linear Diophantine equation takes a bit more effort (see Section 7.2).

. . .

As has been known for centuries that are an infinite number of integer solutions to the equation $x^2 + y^2 = z^2$. This is an example of a non-linear Diophantine equation. The solutions are known as Pythagorean triples. A right triangle whose sides form a Pythagorean triple is called a Pythagorean triangle. For example, there are 16 primitive Pythagorean triples of numbers up to 100:

(3, 4, 5)	(5, 12, 13)	(8, 15, 17)	(7, 24, 25)
(20, 21, 29)	(12, 35, 37)	(9, 40, 41)	(28, 45, 53)
(11, 60, 61)	(16, 63, 65)	(33, 56, 65)	(48, 55, 73)
(13, 84, 85)	(36, 77, 85)	(39, 80, 89)	(65, 72, 97)

. . .

Non-trivial integer solutions to equations of the form $x^n + y^n = z^n$ for $n \geq 3$ are not possible. [An example of a trivial solution is $x = 0$ and $y = z = 1$.] The proof of this conjecture remained an open issue for centuries. The problem, known as Fermat's Last Theorem, was initially posed by Pierre de Fermat in 1637 but not proved true until 1995 by Andrew Wiles. The Wikipedia article "Wiles's proof of Fermat's Last Theorem" [24] highlights the difficulty in proving this theorem:

> Wiles's proof of Fermat's Last Theorem is a proof by British mathematician Andrew Wiles of a special case of the modularity theorem for elliptic curves. Together with Ribet's theorem, it provides a proof for Fermat's Last Theorem. Both Fermat's Last Theorem and the modularity theorem were almost universally considered inaccessible to proof by contemporaneous mathematicians, meaning that they were believed to be impossible to prove using current knowledge.
>
> Together, the two papers which contain the proof are 129 pages long, and consumed over seven years of Wiles's research time. John Coates described the proof as one of the highest achievements of number theory, and John Conway called it "the proof of the [20th] century."

. . .

The smallest solution to the Diophantine equation $x^3 + y^3 = w^3 + z^3$ is $12^3 + 1^3 = 9^3 + 10^3 = 1729$. The number 1729 is known as the second taxicab number. More generally, the n^{th} **taxicab number**, denoted $Ta(n)$, is defined to be the smallest integer that can be expressed as the sum of two positive integer cubes in n ways. For example,

$$Ta(1) = 2 = 1^3 + 1^3$$

$$Ta(2) = 1729 = 12^3 + 1^3 = 9^3 + 10^3$$

$$Ta(3) = 87539319 = 167^3 + 436^3 = 228^3 + 423^3 = 255^3 + 414^3$$

Only three additional taxicab numbers are known, i.e., $Ta(4), Ta(5)$ and $Ta(6)$, but it has been proven that an infinite number exists, see the Wikipedia article on this topic [25].

Upper bounds are known for taxicab numbers from 7 to 12, see the journal article by Boyer [26]. For example, the upper bound for $Ta(12)$ is

$$73914858746493893996583617733225161086864012865017882136931801625152$$

[Author's Remark: Such an enormous number for humans to comprehend but almost nothing compared to infinity!]

The term "taxicab number" comes from a conversation between mathematicians G. H. Hardy and Srinivasa Ramanujan. As told by Hardy:

> I remember once going to see him [Ramanujan] when he was lying ill at Putney. I had ridden in taxi-cab No. 1729, and remarked that the number seemed to be rather a dull one, and that I hoped it was not an unfavourable omen. "No," he replied, "it is a very interesting number; it is the smallest number expressible as the sum of two [positive] cubes in two different ways."

[Author's Remark: For the reader interested in learning more about Ramanujan and his friendship with Professor G.H. Hardy, there is an excellent movie on this topic, i.e., "The Man Who Knew Infinity".**]**

. . .

In 1769, Leonhard Euler conjectured that the Diophantine equation $x^4 + y^4 + z^4 = w^4$ (or more generally, the equation $x_1^n + x_2^n + \cdots + x_{n-1}^n = x_n^n$) has no non-trivial integer solutions. It was not until 1966 that the conjecture was proved incorrect via a computer search by Lander and Parkin [27]. The counterexample that they found is

$$27^5 + 84^5 + 110^5 + 113^5 = 144^5$$

In 1988, Noam Elkies provided a recursive scheme for constructing infinitely many solutions to $x^4 + y^4 + z^4 = w^4$. In the paper by Elkies [28], the following counterexample was given

$$2{,}682{,}440^4 + 15{,}365{,}639^4 + 18{,}796{,}760^4 = 20{,}615{,}673^4$$

Elkies notes "This solution was beyond the range of earlier exhaustive searches. We could only find it by restricting the variables to lie on an appropriate curve." In other words, the counterexample was found via Elkies' recursive scheme and manual computation (not using a computer).

Roger Frye later translated the scheme by Elkies into a computer program and determined the minimal counterexample to Euler's conjecture [29]:

$$95{,}800\ ^4 + \ 217{,}519\ ^4 + \ 414{,}560\ ^4 = \ 422{,}481^4$$

This is the only solution to $x^4 + y^4 + z^4 = w^4$ in which all the variables are less than $1{,}000{,}000$.

. . .

Pell's equation, sometimes referred to as the Pell–Fermat equation, is any Diophantine equation of the form $x^2 - ny^2 = 1$ where n is a positive non-square integer. Lagrange (in 1766-69) proved that Pell's equation has a non-trivial integer solution for any non-square natural number n.

Fundamental solutions to Pell's equation for $n = 1, 2, \dots, 10$ are shown in Table 5.

Table 5. Fundamental solutions to Pell's equation

n	*x*	*y*
1	–	–
2	3	2
3	2	1
4	–	–
5	9	4
6	5	2
7	8	3
8	3	1
9	–	–
10	19	6

The 7th century Indian mathematician Brahmagupta is credited with discovering an identity that allows for additional solutions to be generated from known solutions. Consider two solutions for a given n

$$a^2 - nb^2 = 1$$

$$c^2 - nd^2 = 1$$

Multiply the two equations together to get

$$(a^2 - nb^2)(c^2 - nd^2) = 1$$

$$a^2c^2 + nb^2d^2 - nb^2c^2 - na^2d^2 = 1$$

$$a^2c^2 + 2na^2b^2c^2d^2 + nb^2d^2 - (nb^2c^2 - 2na^2b^2c^2d^2 + na^2d^2) = 1$$

$$(ac + nbd)^2 - n(bc + ad)^2 = 1$$

The last equation in the above sequence of equations is another solution for the case n. This even works if one uses the same starting solution twice. For $n = 8$, we have the solution

$$3^2 - 8 \cdot 1^2 = 1$$

We can use the above equation twice in **Brahmagupta's identity** to get another solution. Substituting $n = 8$, $a = c = 3$ and $b = d = 1$ into Brahmagupta's identity, we get

$$(3 \cdot 3 + 8 \cdot 1 \cdot 1)^2 - 8(1 \cdot 3 + 3 \cdot 1)^2 = 1$$

$$17^2 - 8 \cdot 6^2 = 1$$

In general, given a solution (x_1, y_1), we can find a second solution using Brahmagupta's identity, i.e.,

$$x_2 = x_1^2 + ny_1^2$$

$$y_2 = 2(x_1 y_1)^2$$

Next we can determine a third solution using (x_1, y_1) and (x_2, y_2). The process can be iterated indefinitely, using the recursive formula

Equation 6. Solutions to Pell's equation for a given n

$$x_k = x_{k-1}x_1 + ny_{k-1}y_1$$
$$y_k = x_{k-1}y_1 + y_{k-1}x_1$$

If one calculates the convergents $c_i = \frac{p_i}{q_i}$ for the continued fraction $\sqrt{n}$, it can be proven that there will be a pair (p_j, q_j) solving Pell's equation. The pair with the smallest value for p_j is called the fundamental solution of Pell's equation for a given n. In particular, if the continued fraction representation of $\sqrt{n}$ is of period m, then the fundamental solution is

$$(x_1, y_1) = \begin{cases} (p_m, q_m), & \text{if } m \text{ is even} \\ (p_{2m}, q_{2m}), & \text{if } m \text{ is odd} \end{cases}$$

It can also be proven that all solutions of Pell's equation for a given n can be generated by using the fundamental solution in the iteration described in Equation 6. For the said proofs and associated derivations, see Section 3.2 of "An Introduction to Diophantine Equations" [30].

As an example, consider the Pell's equation $x^2 - 13y^2 = 1$.

The continued fraction representation of $\sqrt{13}$ is $[3, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{6}]$ and so, the period is 5. Since 13 is odd, we know from the previously stated formula that convergent $c_{10} = \frac{p_{10}}{q_{10}}$ corresponds to the fundamental solution for $x^2 - 13y^2 = 1$. The convergents for $\sqrt{13}$ are

$$\frac{3}{1}, \frac{4}{1}, \frac{7}{2}, \frac{11}{3}, \frac{18}{5}, \frac{119}{33}, \frac{137}{38}, \frac{256}{71}, \frac{393}{109}, \frac{649}{180}, \dots$$

Noting that the number for convergents starts at 0, we have that $c_{10} = \frac{p_{10}}{q_{10}} = \frac{649}{180}$. Thus, the fundamental solution for our problem is $(x_1, y_1) = (649{,}180)$. As a check, we see that

$$649^2 - 13 \cdot 180^2 = 421{,}201 - 421{,}200 = 1$$

All the other solutions can be generated by the recursive formulas below:

$$x_k = 649x_{k-1} + 2340y_{k-1}$$

$$y_k = 180x_{k-1}y_1 + 649y_{k-1}$$

Let's try one more example. This time for a number with an even period. The continued fraction representation of $\sqrt{19}$ is $[4, \mathbf{2}, \mathbf{1}, \mathbf{3}, \mathbf{1}, \mathbf{2}, \mathbf{8}]$ and so, the period is 6. The convergents for $\sqrt{19}$ are

$$\frac{4}{1}, \frac{9}{2}, \frac{13}{3}, \frac{48}{11}, \frac{61}{14}, \frac{170}{39}, \frac{1421}{326}$$

Since the period is even, the previous formula tells us that $(p_6, q_6) = (170{,}39)$ is the fundamental solution for $x^2 - 19y^2 = 1$. As a check,

$$170^2 - 19 \cdot 39^2 = 28{,}900 - 28{,}899 = 1$$

. . .

For some values of n, the fundamental solution is surprisingly large. For $n = 61$, the fundamental solution to Pell's equation is $(1766319049, 226153980)$. For $n = 109$, the fundamental solution to Pell's equation is $(158070671986249, 15140424455100)$.

. . .

The Erdős–Straus conjecture (i.e., not proven yet) claims that for every positive integer $n \geq 2$, there exists at least one integer solution to the equation

$$4xyz = n(yz + xz + xy)$$

The conjecture is usually written in the following equivalent form

$$\frac{4}{n} = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}$$

Some examples

$$\frac{4}{5} = \frac{1}{2} + \frac{1}{4} + \frac{1}{20}$$

$$\frac{4}{17} = \frac{1}{5} + \frac{1}{34} + \frac{1}{170}$$

$$\frac{4}{269} = \frac{1}{68} + \frac{1}{9146} + \frac{1}{18292}$$

Although a solution is not known for all values of n, solutions are known for infinitely many values of n. For example, if $n \equiv 2 \pmod 3$, i.e., n is of the form $2 + 3k$ with k being a positive integer, then

$$\frac{4}{n} = \frac{1}{n} + \frac{1}{\left(\frac{n+1}{3}\right)} + \frac{1}{\left(\frac{n(n+1)}{3}\right)}$$

For example, take $n = 2 + 3(3) = 11$, then the above equation gives us

$$\frac{4}{11} = \frac{1}{11} + \frac{1}{4} + \frac{1}{44}$$

For some values of n, there are multiple solutions. For example, $n = 23$ has the following solutions:

$$(6, 139, 19182), (6, 140, 9600), (6, 141, 6486), (6, 142, 4899), (6, 144, 3312), (6, 147, 2254), (6, 150, 1725),$$

$$(6, 156, 1196), (6, 161, 966), (6, 174, 667), (6, 184, 552), (6, 207, 414), (6, 230, 345), (6, 276, 276),$$

$$(7, 42, 138), (8, 23, 184), (8, 24, 138), (9, 16, 3312), (9, 18, 138), (10, 15, 138), (12, 12, 138)$$

Where, for example, $(6, 139, 19182)$ is shorthand for

$$\frac{4}{23} = \frac{1}{6} + \frac{1}{139} + \frac{1}{19182}$$

The number of distinct solutions for a given n follows an irregular pattern. The first few terms in the sequence (i.e., number of solutions for a given n) are

$$1, 1, 2, 5, 5, 6, 4, 9, 7, 15, 4, 14, 33, 22, 4, 21, 9, 30, 25, 22, 19, 45, 10, 17, 25$$

Also, see https://oeis.org/A073101, in the Online Encyclopedia of Integer Sequences®.

7.2 Linear Diophantine Equations

A linear Diophantine equation has the form $ax + by = c$, where a, b and c are integers.

Consider the following ancient number theory puzzles, all of which can be formulated as linear Diophantine equation problems:

- (Attributed to Alcuin of York, 775 AD) One hundred bushels of grain are distributed to 100 people in three groups (A, B and C) such that each person in Group A receives 3 bushels, each person in Group B receives 2 bushels and each person in Group C receives ½ bushel. How many people are in each group? If we let x, y and z be the number of people in Groups A, B and C, respectively, then we have $x + y + z = 100$ and $3x + 2y + \frac{1}{2}\ z = 100$.
- (Attributed to Mahaviracarya, 850 AD) There are 63 baskets of bananas (each with the same number of bananas) and 7 additional bananas. The bananas are distributed equally among 23 people. How many bananas are there in each basket? Let x equal the number of bananas in each basket and y equal the number of bananas distributed to each person, then $63x + 7 = 23y$.
- (Attributed to Yen Kung, 1372 AD) Given a collection of coins (of unknown number). If 77 equal stacks of the coins are made, 27 coins are left over. However, if 78 equal stacks are made, there are no coins left over. How many coins are there? If we let x equal the number

of coins in the first type of stack and y equal the number of coins in the second type of stack, then the total number of coins equals $77x + 27 = 78y$.

With some rearrangement and simplification, all of the above problems can be reduced to linear Diophantine equations. We will solve all three of these problems but first we need to develop some concepts.

. . .

The following theorem gives necessary and sufficient conditions for when solutions exist for a given linear Diophantine equation.

The Greatest Common Divisor (GCD) of two integers is the largest positive integer that exactly (no remainder) divides the two integers. For example, $\gcd(3,18) = 3$, $\gcd(7,13) = 1$ and $\gcd(-2,-4) = 2$. We will further explore this concept in Section 12.4. However, we need a brief introduction here since the concept of a GCD is used in the following theorems.

Theorem 14. The linear Diophantine equation $ax + by = c$ has an integer solution if and only if $gcd(a,b)\,|c$.

Proof: Let $g = gcd(a,b)$. This implies that $g|a$ and $g|b$, or equivalently, there exists integers r and s such that $a = gr$ and $b = gs$.

If a solution does exist to the equation, say $ax_0 + by_0 = c$, then we have

$$c = ax_0 + by_0 = grx_0 + gsy_0 = g(rx_0 + sy_0)$$

which implies $g|c$.

Going in the other direction, assume that $g = gcd(a,b)\,|c$ or equivalently, $c = gt$ for some integer t. By Bézout's identity [31], there exists integers x_0 and y_0 such that $g = \ ax_0 + by_0$. Multiply by t to get

$$c = gt = (ax_0 + by_0)t = a(tx_0) + b(ty_0)$$

which implies that $ax + by = c$ has $x = tx_0$ and $y = ty_0$ as a particular solution ■

In the case $gcd(a,b)\,|1$, Theorem 14 implies there exists integers x and y such that $ax + by = 1$ if and only if $gcd(a,b) = 1$.

Once one solution is found to a linear Diophantine equation, it is possible to generate an infinite number of other solutions.

Theorem 15. If x_0, y_0 is a solution of the Diophantine equation $ax + by = c$ where $g = gcd(a,b)$ then all other solutions are given by

$$x = x_0 + \left(\frac{b}{g}\right)t, \quad y = y_0 - \left(\frac{a}{g}\right)t, \quad \textit{for any integer t.}$$

Proof: Assume there exists another solution to the equation (say x_1, y_1). Then we have

$$ax_0 + by_0 = c = ax_1 + by_1$$
$$\text{which implies } a(x_1 - x_0) = b(y_0 - y_1) \qquad (1)$$

By a theorem from number theory, $\gcd\left(\frac{a}{g}, \frac{b}{g}\right) = 1$. Let $r = \frac{a}{g} \Rightarrow a = gr$ and let $s = \frac{b}{g} \Rightarrow b = gs$.

Substituting into Equation 1 and cancelling g on both sides, we get

$$r(x_1 - x_0) = s(y_0 - y_1) \qquad (2)$$

So, $r|s(y_0 - y_1)$ and $\gcd(r,s) = 1$. By Euclid's lemma [32], we have that $r|(y_0 - y_1)$ or equivalently, there exist an integer t such that $y_0 - y_1 = rt$. Substituting into Equation 2, we get $x_1 - x_0 = st$. Thus, we have

$$x_1 = x_0 + st = x_0 + \left(\frac{b}{g}\right)t$$

$$y_1 = y_0 - rt = y_0 - \left(\frac{a}{g}\right)t$$

So, if x_1, y_1 is a solution other than the given solution x_0, y_0 then it must be of the form given in the above equations.

We can verify that x_1, y_1 is a solution to the given Diophantine equation, regardless of the value of the integer t, by the following line of reasoning

$$ax_1 + by_1 = a\left[x_0 + \left(\frac{b}{g}\right)t\right] + b\left[y_0 - \left(\frac{a}{g}\right)t\right] = (ax_0 + by_0) + \left(\frac{ab}{g} - \frac{ab}{g}\right)t = c.$$

Thus, each integer value of t gives a solution to the equation. ■

Since we can write $ax - by = c$ as $ax + (-b)y = c$ and $-b$ is still an integer, Theorem 15 also holds for equations of the form $ax - by = c$. In this case, all integer solutions to an equation of the form $ax - by = c$ are given by

$$x = x_0 - \left(\frac{b}{g}\right)t, \quad y = y_0 - \left(\frac{a}{g}\right)t, \text{ for any integer t}$$

. . .

We are still left with the problem of finding an initial solution for a given linear Diophantine equation. There are several methods, e.g., trial and error, and an approach based on continued fractions and convergents.

First, we will apply the trial and error method to the baskets and bananas problem, i.e., $63x - 23y = -7$. We solve for y in terms of x, i.e., $y = \frac{63x+7}{23}$, and then try different values of x until we find a corresponding integer value for y.

x	0	1	2	3	4	5	6	7
y	$\frac{7}{23}$	$\frac{70}{23}$	$\frac{133}{23}$	$\frac{196}{23}$	$\frac{259}{23}$	$\boldsymbol{\frac{322}{23} = 14}$	$\frac{385}{23}$	$\frac{448}{23}$

So, one solution is $x = 5, y = 14$. From Theorem 15 with $a = 63$ and $b = -23$, we have that all the solutions to the equation are given by

$$x = 5 - 23t$$

$$y = 14 - 63t$$

The value $t = 0$ yields the minimal solution with positive values for x and y. In this case, there are 5 bananas in each of the 63 baskets plus 7 more bananas, with each person getting 14 bananas.

However, as can be seen from the formula above, there are an infinite number of positive solutions. For example, we could have 28 bananas in each of 63 baskets plus 7 bananas, with each person getting 77 bananas.

. . .

There is a systematic approach for finding a solution to a linear Diophantine equation based on continued fractions. We start with the Diophantine equation $ax - by = 1$ and consider the continued fraction for the rational number $\frac{a}{b}$. The continued fraction convergents for $\frac{a}{b}$ are $c_i = \frac{p_i}{q_i}$.

Recall from Theorem 7

$$p_i q_{i-1} - p_{i-1} q_i = (-1)^i$$

Substituting $i = n$ in the above equation and noting that $c_n = \frac{p_n}{q_n} = \frac{a}{b}$, we get

$$a q_{n-1} - b p_{n-1} = (-1)^n$$

If n is even, then a solution to $ax - by = 1$ is given by the penultimate convergent, i.e., (q_{n-1}, p_{n-1}).

For example, take the equation $23x - 78y = 1$. The associated continued fraction expansion is $\frac{23}{78} = [0,3,2,1,1,4]$, $n = 6$ is even, and the penultimate convergent is $\frac{5}{17}$. [The convergents can be computed using Wolfram Alpha, at www.wolframalpha.com. For the problem at hand, just type in "convergents for 23/78". Another online resource for continued fractions is the Continued Fraction Calculator at www.alpertron.com.ar/CONTFRAC.HTM.]

This gives us the solution $(17,5)$. As a check, we have that

$$23(17) - 78(5) = 391 - 390 = 1$$

By Theorem 15 (with $a = 23$ and $b = -78$), we can represent the set of all integer solutions to $23x - 78y = 1$ as

$$x = 17 - 78t$$

$$y = 5 - 23t$$

If n is odd, then we have

$$a q_{n-1} - b p_{n-1} = -1$$

which can be written as

$$a(-q_{n-1}) - b(-p_{n-1}) = 1$$

and so, $(-q_{n-1}, -p_{n-1})$ is a solution in this case.

For example, take the equation $23x - 77y = 1$. The associated continued fraction expansion is $\frac{23}{77} = [0,3,2,1,7]$, $n = 5$ is odd, and the penultimate convergent is $\frac{3}{10}$. This yields the solution $(-10, -3)$. As a check, we have that

$$23(-10) - 77(-3) = -230 + 231 = 1$$

By Theorem 15 (with $a = 23$ and $b = -77$), we can represent the set of all integer solutions to $23x - 78y = 1$ as

$$x = -10 - 77t$$

$$y = -3 - 23t$$

For $t = -1$, we get a solution with x and y being positive, i.e., $(67,20)$.

. . .

If the equation is of the form $ax + by = 1$, we have two more cases to consider.

When n is even, we rewrite $aq_{n-1} - bp_{n-1} = 1$ as

$$aq_{n-1} + b(-p_{n-1}) = 1$$

and thus $(q_{n-1}, -p_{n-1})$ is a solution to $ax + by = 1$.

For example, $(17, -5)$ is a solution to $23x + 78y = 1$.

When n is odd, we rewrite $aq_{n-1} - bp_{n-1} = -1$ as

$$a(-q_{n-1}) + bp_{n-1} = 1$$

and thus $(-q_{n-1}, p_{n-1})$ is a solution to $ax + by = 1$ in this case.

For example, $(-10,3)$ is a solution to $23x + 77y = 1$.

. . .

The previous four cases can be used to solve Diophantine equations that have an integer other than 1 on the right-hand side of the equation. For example, let's take another look at the baskets and bananas problem, i.e., $63x - 23y = -7$. We first solve the $63x - 23y = 1$. The continued fraction expansion of $\frac{63}{23}$ is $[2,1,2,1,5]$ and penultimate convergent is $\frac{11}{4}$. Since there are an odd number of terms in the expansion, we have that

$$63(4) - 23(11) = -1$$

Multiplying by 7 on both sides of the above equation yields

$$63(4 \cdot 7) - 23(11 \cdot 7) = -7$$

Thus, $(28,77)$ is a solution to $63x - 23y = -7$. By Theorem 15, all solutions are of the form

$$x = 28 - 23t$$

$$y = 77 - 63t$$

For $t = 1$, we get the minimal positive solution of $(5,14)$ which agrees with our previous analysis.

. . .

Next, consider the bushels of grain problem attributed to Alcuin of York. We have that $x + y + z = 100$ which we can rewrite as $z = 100 - x - y$ and then substitute into $3x + 2y + \frac{1}{2}z = 100$ to get $3x + 2y + \frac{1}{2}(100 - x - y) = 100$ which simplifies to $5x + 3y = 100$. This equation is simple enough to guess at a solution, i.e., $x = 5, y = 25$, which implies $z = 70$. There are no other solutions that meet the additional requirement that $x + y + z = 100$.

We can also solve this using the continued fraction method. The continued fraction corresponding to $\frac{5}{3}$ is $[1,1,2]$ and the penultimate convergent is $\frac{2}{1}$. Since the number of convergents is odd, we have

$$5(1) - 3(2) = -1$$

which can be recast as

$$5(1) + 3(-2) = -1$$

Now, multiple by -100 to get

$$5(-100) + 3(200) = 100$$

Thus, $(-100, 200)$ is a solution to $5x + 3y = 100$. By Theorem 15, all solutions are given by

$$x = -100 + 3t$$

$$y = 200 - 5t$$

For $t = 35$, we get the solution $(5,25)$ which, as noted, is the only solution that satisfies the conditions of the problem.

7.3 Systems of Linear Diophantine Equations

In this section, we explore a method for solving systems of linear Diophantine equations. We first describe an alternate approach for solving a single linear Diophantine equation and then extend that approach to a system of linear Diophantine equations.

The Euclidean algorithm [34] is a method for determining the Greatest Common Division (GCD) of two integers. The following theorem is an extension of the Euclidean algorithm.

Theorem 16. (Extended Euclidean Algorithm) Given any two integers a and b, it is possible to unimodular row reduce

$$\left\langle \begin{matrix} a \\ b \end{matrix} \middle| \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \right\rangle \text{ to } \left\langle \begin{matrix} d \\ 0 \end{matrix} \middle| \begin{matrix} s & t \\ u & v \end{matrix} \right\rangle$$

such that $\gcd(a, b) = d$. Further, the general solution to the Diophantine equation $ax + by = d$ is given by

$$x = s + ku$$

$$y = t + kv$$

with k being any integer.

Proof: See the preprint article entitled “Linear Diophantine Equations” [33].

Unimodular row operations are a subset of matrix row operations, and include only the following operations:

- Adding an integer multiple of one row of a matrix to another row
- Interchanging two rows of a matrix
- Multiplication of a row of a matrix by -1.

[**Author’s Remark**: Sometimes I use ⟨ ⟩ for matrices and other times I use [ ]. The issue is with the editing tool that I am using. There is no intended difference in terms of meaning.]

We can use Theorem 16 to find a solution to $63x - 23y = -7$. We start with

$$\left\langle \begin{array}{r|rr} 63 & 1 & 0 \\ -23 & 0 & 1 \end{array} \right\rangle$$

Adding twice the second row to the first, we get

$$\left\langle \begin{array}{r|rr} 17 & 1 & 2 \\ -23 & 0 & 1 \end{array} \right\rangle$$

Adding twice the first row to the second, we get

$$\left\langle \begin{array}{r|rr} 17 & 1 & 2 \\ 11 & 2 & 5 \end{array} \right\rangle$$

Subtracting the second row from the first, we get

$$\left\langle \begin{array}{r|rr} 6 & -1 & -3 \\ 11 & 2 & 5 \end{array} \right\rangle$$

Subtracting the first row from the second row, we get

$$\left\langle \begin{array}{r|rr} 6 & -1 & -3 \\ 5 & 3 & 8 \end{array} \right\rangle$$

Subtracting the second row from the first row, we get

$$\left\langle \begin{array}{r|rr} 1 & -4 & -11 \\ 5 & 3 & 8 \end{array} \right\rangle$$

Finally, we subtract five times the first row from the second to get

$$\left\langle \begin{array}{r|rr} 1 & -4 & -11 \\ 0 & 23 & 63 \end{array} \right\rangle$$

So, $\gcd(63, -23) = 1$ and the solution to the Diophantine equation $63x - 23y = 1$ is given by

$$x = -4 + 23t$$

$$y = -11 + 63t$$

For $t = 0$, we get the solution $(-4, -11)$ which is verified below

$$63(-4) - 23(-11) = -252 + 253 = 1$$

If we multiple the above equation by -7, we get

$$63(28) - 23(77) = -7$$

and so, a solution to the Diophantine equation $63x - 23y = -7$ is $(28,77)$. By Theorem 15, we get the general solution

$$x = 28 - 23t$$

$$y = 77 + 63t$$

When $t = -1$, we get the minimum positive solution $(5,14)$ which agrees with our previous solution to the problem.

. . .

The following theorem is a bit out of the main flow here but nevertheless a useful result.

Theorem 17. The GCD of $a_1, a_2, \ldots, a_n$ can be determined by the unimodular row reduction of the matrix

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \text{ to } \begin{bmatrix} d \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where $d = \gcd\,(a_1, a_2, \ldots, a_n)$.

Proof: See the preprint article entitled “Linear Diophantine Equations” [33].

As an example, we determine the GCD of 2,7 and 21 using Theorem 17.

$$\begin{bmatrix} 2 \\ 7 \\ 21 \end{bmatrix} \overset{R_3 - 3R_2}{\Rightarrow} \begin{bmatrix} 2 \\ 7 \\ 0 \end{bmatrix} \overset{R_2 - 3R_1}{\Rightarrow} \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \overset{R_1 - 2R_2}{\Rightarrow} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \overset{R_1 \leftrightarrow R_2}{\Rightarrow} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

So, $\gcd(2,7,21) = 1$.

. . .

Consider the system of linear Diophantine equations:

$$x + y + z = 100$$

$$5x + 3y = 100$$

This can be written in matrix form as $AX = B$, where

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 5 & 3 & 0 \end{bmatrix}, \quad X = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \text{ and } B = \begin{bmatrix} 100 \\ 100 \end{bmatrix}$$

The following theorem provides a way to solve systems of linear Diophantine equations of the form $AX = B$.

Theorem 18. For a given system of linear Diophantine equations $AX = B$, unimodular reduce $[A^t|I]$ to $[R|T]$, where R is in relaxed row-echelon form. The system $AX = B$ has integer solutions if and only if the system $R^tK = B$ has integer solutions for K, and in this case, all the solutions of $AX = B$ are of the form $X = T^tK$.

Proof: See the preprint article entitled “Linear Diophantine Equations” [33].

Some remarks concerning Theorem 18:

- The superscript t indicates the transpose of the given matrix, i.e., row #1 becomes column #1, row #2 becomes column #2, and so on.
- When a matrix is in **relaxed row-echelon form**, we have the following
  - All the rows with all zeros are at the bottom of the matrix.
  - The leading entry in each non-zero row is to the right of all leading entries in the rows above. (In strict row-echelon form, the leading entries in non-zero rows must be 1.)
- The matrix I is the identity matrix. It is a square matrix with all ones along the main diagonal and zeros elsewhere.

Returning to the example that we stated just before Theorem 18, the first step to the solution is to unimodular reduce $[A^t|I]$ to $[R|T]$.

$$\left\langle\begin{array}{cc|ccc}1&5&1&0&0\\1&3&0&1&0\\1&0&0&0&1\end{array}\right\rangle\rightarrow\left\langle\begin{array}{cc|ccc}1&5&1&0&0\\0&-2&-1&1&0\\0&-5&-1&0&1\end{array}\right\rangle\rightarrow\left\langle\begin{array}{cc|ccc}1&5&1&0&0\\0&-2&-1&1&0\\0&1&2&-3&1\end{array}\right\rangle\rightarrow\left\langle\begin{array}{cc|ccc}1&5&1&0&0\\0&1&2&-3&1\\0&0&3&-5&2\end{array}\right\rangle$$

In going from the first matrix to the second, we subtracted row #1 from row #2 and #3. In going from the second to the third matrix, we add minus three times the second row to the third. In computing the last matrix, we took two steps, i.e., switched rows #2 and #3, and then added twice row #2 to row #3.

The above computation gives us R and T:

$$R=\begin{bmatrix}1&5\\0&1\\0&0\end{bmatrix},T=\begin{bmatrix}1&0&0\\2&-3&1\\3&-5&2\end{bmatrix}$$

Next, we compute $R^tK=B$:

$$\begin{bmatrix}1&0&0\\5&1&0\end{bmatrix}\begin{bmatrix}k_1\\k_2\\k_3\end{bmatrix}=\begin{bmatrix}100\\100\end{bmatrix}$$

Thus, $k_1=100$ and $k_2=-400$. We are free to choose any integer value for k_3. So, we have

$$K=\begin{bmatrix}100\\-400\\k_3\end{bmatrix}$$

and

$$X=T^tK=\begin{bmatrix}1&2&3\\0&-3&-5\\0&1&2\end{bmatrix}\begin{bmatrix}100\\-400\\k_3\end{bmatrix}=\begin{bmatrix}-700+3k_3\\1200-5k_3\\-400+2k_3\end{bmatrix}$$

If we take $k_3=235$, then we get the minimum positive solution for X:

$$x=5$$

$$y=25$$

$$z=70$$

This corresponds to the solution of the bushels of grain problem from Section 7.2.

The approach described in Theorem 18 can be used even when there is only one equation. For example, consider $6x-14y+21z=11$. This is a plane in three-space. Our task is to find points in three-space that have integer components that fall on the given plane.

We first compute R and T.

$$\left\langle\begin{array}{c|ccc}6&1&0&0\\-14&0&1&0\\21&0&0&1\end{array}\right\rangle\rightarrow\left\langle\begin{array}{c|ccc}6&1&0&0\\-14&0&1&0\\7&0&1&1\end{array}\right\rangle\rightarrow\left\langle\begin{array}{c|ccc}6&1&0&0\\-14&0&1&0\\1&-1&1&1\end{array}\right\rangle\rightarrow\left\langle\begin{array}{c|ccc}1&-1&1&1\\0&-14&15&14\\0&7&-6&-6\end{array}\right\rangle$$

Thus,

$$R = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad T = \begin{bmatrix} -1 & 1 & 1 \\ -14 & 15 & 14 \\ 7 & -6 & -6 \end{bmatrix}$$

and

$$R^t K = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix} = B = [11]$$

So, $k_1 = 11$ and we are free to choose any integer values for k_2 and k_3.

From Theorem 18, we have that

$$X = T^t K = \begin{bmatrix} -1 & -14 & 7 \\ 1 & 15 & -6 \\ 1 & 14 & -6 \end{bmatrix} \begin{bmatrix} 11 \\ k_2 \\ k_3 \end{bmatrix} = \begin{bmatrix} -11 - 14k_2 + 7k_3 \\ 11 + 15k_2 - 6k_3 \\ 11 + 14k_2 - 6k_3 \end{bmatrix}$$

We are free to select any integer values for k_2 and k_3, and this will result in an integer solution for X.

For example, if $k_2 = -1$ and $k_3 = 1$, then we have the solution

$$X = \begin{bmatrix} 10 \\ -10 \\ -9 \end{bmatrix}$$

However, there are no values for k_2 and k_3 that yield all positive values for X.

8 Stochastic Processes

8.1 Overview and Background

Prerequisites: probability theory (including random variables, expected value and variance), calculus (limits, derivatives and integrals), difference equations (see Section 4)

Notation: In this section, we use P to represent probability, e.g., $P(A = 1) = .5$ read as "the probability that some variable A equals the value 1 is .5."

For some processes, the main interest is on some numerical result based on the outcomes (or states) of the process. For example, consider a simple game of rolling two dice, where a player is awarded $2 on a $1 bet if the sum is odd, and $0 otherwise. In this example, there is a variable (call it X) which can take values $2 or $0. The value of X depends on chance, with associated probabilities for various alternatives. In general, a **random variable** is a numerical function defined on a **sample space** S (i.e., the set of all possible outcomes of some process) such that its specific value at any particular instance depends on chance and can be assigned a probability. A collection of elements in S is called an **event**. For the dice game, the event of getting an even sum (call the event A) includes the following elements from the sample space

$$(1,1), (1,3), (1,5), (2,2), (2,4), (2,6), (3,1), (3,3), (3,5),$$
$$(4,2), (4,4), (4,6), (5,1), (5,3), (5,5), (6,2), (6,4), (6,6)$$

In terms of probabilities, we have $P(A) = P(X = \$0) = \frac{1}{2}$. Let B be the event of an odd roll of the two dice. $P(B) = P(X = \$2) = \frac{1}{2}$ since A and B are mutually exclusive and the two events cover the entire sample space, i.e., $AUB = S$.

In the language of mathematics, a random variable is a function X that maps from a sample space S to a set of numbers Y. Using function notation, we write $X: S \longrightarrow Y$.

The expected value of a random variable X is represented as $E(X)$ and the variance is represented as $V(X)$. It is assumed that the reader is familiar with these concepts.

. . .

A process is classified as being stochastic (i.e., **stochastic process)** if when observed at various times, the outcome (or state) can be modeled as a random variable. The term "random process" is another term with the same meaning as "stochastic process."

For a stochastic process, at each observed time, there is a probability distribution for the various outcomes. In general, the probability distribution at a given time depends on past outcomes. A stochastic process is typically represented as $X(t)$ or X_t, with the meaning that at each time, there is a different random variable representing (modeling) the stochastic process. The times can be discrete or continuous, leading to discrete-time and continuous-time stochastic processes, respectively. The values of $X(t)$ (i.e., state space) can be discrete (either a finite or countable set such as the integers) or continuous (an uncountable set of values such as the real number), known as discrete-valued and continuous-valued stochastic processes, respectively. Further, one can classify stochastic processes based on whether the future of a process depends on its past (or not). A stochastic process is said to have the **Markov property** if the conditional probability distribution of future states of the process (conditional on both past and present values) depends only upon the present state, i.e., given the present, the future does not depend on the past.

Table 6 provides examples for the various types of Markovian processes.

Table 6. Examples of Markovian Processes

	Discrete State Space	**Continuous State Space**
Discrete-time	Random walk along a straight line where the state is the number of steps from the starting point. At each transition, the “walker” goes in the forward direction with probability p and in the backward direction with probability $1-p$. All steps are of the same unit length. **State Space**: {…, -3, -2, -1, 0, 1, 2, 3, …} representing the location on the real number line **Time**: measured by the number of steps	Random walk where the steps are of varying length between 0 and some constant A. This type of process is uncommon in the literature. **State Space**: real numbers in the interval $(-\infty, \infty)$ **Time**: measured by the number of steps
Continuous-time	A queue with a single server, where we are concerned with the number of customers in the queue (discrete state), and the duration between changes in the number of customers in the queue which depends on a continuous arrival rate (input) and a continuous serving rate (output). It is assumed that at most a finite number of customers can arrive in any instant. **State Space**: {0, 1, 2, …} representing the number of customers in the queue **Time**: real numbers in the interval $(0, \infty)$	A queue with a single server but instead of customers, we are interested in some continuous quantity that comes or goes at various points in time. For example, consider a business (server) that buys, holds, and sells gold. **State Space**: real numbers in the interval [0,M] where M is the maximum amount of gold the server can hold **Time**: real numbers in the interval $(0, \infty)$

Table 7 provides examples for the various types of non-Markovian processes, i.e., stochastic processes whose transitions depend on the past state (or states) of the process.

Table 7. Examples of non-Markovian Processes

	Discrete State Space	**Continuous State Space**
Discrete-time	Random walk along a straight line where the state is the number of steps from the starting point. At each transition X_n, the "walker" determines its next state related to where it was two transitions ago, i.e., from state X_{n-2}. At transition n, it moves one step to the right of X_{n-2} with probability p, and one step to the left of X_{n-2} with probability $1 - p$. All steps are of the same unit length. **State Space**: {…, -3, -2, -1, 0, 1, 2, 3, …} representing the location on the real number line **Time**: measured by the number of steps	Same as the discrete state space and discrete-time example, but with the steps being of varying length between 0 and some constant A. **State Space**: real numbers in the interval $(-\infty, \infty)$ **Time**: measured by the number of steps
Continuous-time	Same as the discrete state space and discrete-time example, but with the timing of the steps based on a probability distribution, e.g., the exponential or uniform distribution. **State Space**: {…, -3, -2, -1, 0, 1, 2, 3, …} **Time**: real numbers in the interval $(0, \infty)$	Same as the continuous state space and discrete-time example, but with the timing of the steps based on a probability distribution. **State Space**: real numbers in the interval $(-\infty, \infty)$ **Time**: real numbers in the interval $(0, \infty)$

In this book, the focus is on discrete-valued stochastic processes that have the Markov property.

8.2 Discrete-time Markov Chains

8.2.1 Definitions and Concepts

A discrete-time, discrete state space Markovian process is known as a **Markov Chain**.

More formally, let $\{X_n\}$ for $n = 0, 1, 2, \dots$ be a sequence of random variables with a finite or countably infinite state space $I = \{i_0, i_1, i_2, \dots\}$. The random variable X_n represents the state of a stochastic process system at time n. A stochastic process is a Markov Chain if for each $n = 0, 1, 2, \dots$

$$P(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \dots, X_n = i_n\,\} = P(X_{n+1} = i_{n+1} \mid X_n = i_n\,\}$$

for all possible values of $i_0, i_1, i_2 \dots \in I$. In other words, given the past history of a process, the transition to the next step only depends on the present state.

If the transition probability $P(X_{n+1} = j \mid X_n = i\,\} = p_{ij}$ does not depend on n (as emphasized in the shorthand notation p_{ij}), then the Markov chain is said to be **time-homogeneous**.

Let $S = \{i_0, i_1, i_2, \dots\}$ be the set of states for a given Markov chain. This is known as the **state space** for the Markov chain. If a Markov chain is in a given state (say i), then the sum of the transition probabilities from i must add to 1, i.e.,

$$\sum_{j \in S} p_{ij} = 1$$

Two states in a Markov chain are said to **communicate** if they are reachable from one another (bidirectional) by a sequence of transitions with positive probability. This is an equivalence relation [35] which yields a set of communicating classes. A Markov chain is said to be **irreducible** if there is one communicating class, i.e., the entire state space.

The next two subsections concern specific examples of Markov chains. We will cover additional concepts concerning Markov chains in Section 8.2.4.

8.2.2 Gambler's Ruin Scenarios

The **gambler's ruin problem** concerns a probabilistic game between two players (call them A and B). The game consists of a series of plays where player A either wins one unit from player B with probability p or loses one unit to B with probability $q = 1 - p$. Each play of the game is independent of all previous plays. Assume that, at the start of the game, A has k units of money, and B has $a - k$ units. The total amount of money between the two players remains at the constant amount a. Let X_n be a random variable representing A's amount of money after n plays. We are given that $X_o = k$.

- If $X_n = 0$ for some value of n, then player A has lost all (i.e., ruin) and player B has a units.
- If $X_n = a$ for some value of n, then player A has won and player B has 0 units (i.e., ruin).

The sequence of random variables $\{X_n\}$ with state space $S = \{0, 1, \ldots, a\}$ and transition probabilities as described above is a Markov chain. The transition probabilities among the states are depicted in Figure 10. If $X_n = 0$, the game is over (B has won all the money) and there is zero probability of leaving state 0. If $X_n = a$, the game is over (A has won all the money) and there is zero probability of leaving state a. States 0 and a are known as **absorbing states** since once entered, there is zero probability of leaving.

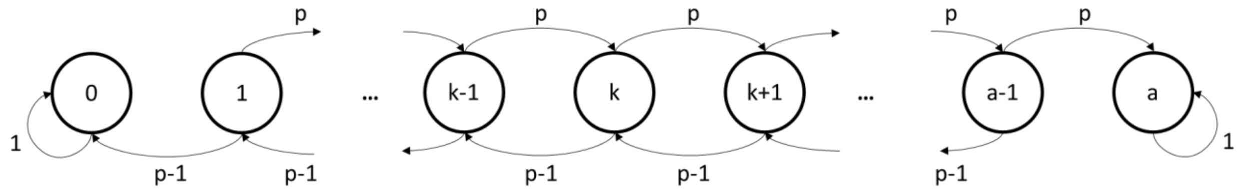


Figure 10. State transition diagram for gambler's ruin

Let L_k be the event that player A eventually losses all of his or her money (starting with k units). We have that

$$P(L_k) = \sum_{n=k}^{\infty} P(X_n = 0)$$

In other words, the probability of event L_k is the sum of the probabilities of going broke at play number k, or play number $k + 1$, or play number $k + 2$ and so on. It is now possible for player A to go broke in less than k plays. Thus, the summation starts at $n = k$.

Consider the first play of the game. Let W be the event that A wins the first play and $\neg W$ be the event A loses. Using the law of total probably, we can write

$$P(L_k) = P(L_k|W)P(W) + P(L_k|\neg W)P(\neg W)$$

The event L_k, given that A has won the first play, is equivalent to L_{k+1}. The event L_k, given that A has lost the first play, is equivalent to L_{k-1}. So, we can write the above equation as

$$P(L_k) = P(L_{k+1})P(W) + P(L_{k-1})P(\neg W)$$

For easy of notation, let $P(L_k) = x_k$. Noting that $P(W) = p$ and $P(\neg W) = q$, we can write the ruin probability equation as

$$x_k = x_{k+1}p + x_{k-1}q$$

This can be rearranged into the form of a second-order homogeneous linear difference equation, i.e.,

Equation 7. Difference equation for gambler's ruin problem

$$px_{k+1} - x_k + qx_{k-1} = 0$$

Since ruin is guaranteed if A starts with 0 units of money, $x_0 = 1$. Since ruin is impossible if A starts with a units of money, $x_a = 0$. We now have sufficient information to solve the difference equation using the techniques from Section 4.4.

We start by assuming a solution of the form $x_k = \lambda^k$. Substituting into Equation 7, we get

$$p\lambda^{k+1} - \lambda^k + q\lambda^{k-1} = 0$$

$$\lambda^{k-1}(p\lambda^2 - \lambda + q) = 0$$

$$\lambda^{k-1}(p\lambda - q)(\lambda - 1) = 0$$

$$(p\lambda - q)(\lambda - 1) = 0$$

For the case $p \neq q$, there are two non-zero solutions to the above characteristic equation, i.e., $\lambda_1 = 1$ and $\lambda_2 = q/p$. Since the difference equation is linear, the general solution is a linear combination of the two solutions, i.e.,

$$x_k = A\lambda_1^k + B\lambda_2^k = A + B\left(\frac{q}{p}\right)^k$$

Substituting the initial conditions into the above equation yields the following (with $r = q/p$)

$$A + B = 1$$

$$A + Br^a = 0$$

Solving the above systems of equations, we get

$$A = -\frac{r^a}{1 - r^a}; \quad B = \frac{1}{1 - r^a}$$

Thus, the probability of ruin (starting with k units of money) is

Equation 8. Probability of ruin when $r \neq 1$

$$x_k = \frac{r^k - r^a}{1 - r^a}$$

Table 8 provides an example of the above equation when the total amount of money between the two gamblers is 40 ($a = 40$). The top row lists values of k and the left column lists values of p.

Table 8. Gambler's ruin example – probabilities of ruin

	5	10	15	20	25	30	35
0.1	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.999983
0.2	1.000000	1.000000	1.000000	1.000000	1.000000	0.999999	0.999023
0.3	1.000000	1.000000	1.000000	1.000000	0.999997	0.999791	0.985542
0.4	0.999999	0.999995	0.999960	0.999699	0.997716	0.982659	0.868313
0.6	0.131687	0.017341	0.002284	0.000301	0.000040	0.000005	0.000001
0.7	0.014458	0.000209	0.000003	0.000000	0.000000	0.000000	0.000000
0.8	0.000977	0.000001	0.000000	0.000000	0.000000	0.000000	0.000000
0.9	0.000017	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

For $0 < r < 1$ and as a approaches infinity, we have

$$\lim_{a\to\infty} x_k = \lim_{a\to\infty} \frac{r^k - r^a}{1 - r^a} = \frac{r^k - 0}{1 - 0} = r^k$$

This is the case where player A is going against an opponent with an infinite amount money but where the odds favor A. For example, if $p = .6$ and $q = .4$, then $r = \frac{2}{3}$. Assume A starts with 100 units of money, then A's probability of ruin is $\left(\frac{2}{3}\right)^{100} \cong 0$. Even if A starts with only 5 units, A's probably of ruin is still small, i.e., $\left(\frac{2}{3}\right)^{5} \cong .13169$.

[Author's Remark: This seems counterintuitive to me. I would expect player A to have a higher probability of losing if he or she starts with only 5 units while playing against an opponent with infinite cash, even given that A's probability of winning each game is 0.6.**]**

For $r > 1$ and as a approaches infinite, we have

$$\lim_{a\to\infty} x_k = \lim_{a\to\infty} \frac{r^k - r^a}{1 - r^a} = \lim_{a\to\infty} \frac{\left(\frac{r^k}{r^a} - 1\right)}{\frac{1}{r^a} - 1} = \frac{0 - 1}{0 - 1} = 1$$

which is as expected since we assumed $p < q$ and that player B has infinite cash.

. . .

For the case that $p = q = \frac{1}{2}$, the characteristic equation becomes

$$(\lambda - 1)^2 = 0$$

One solution is $\lambda = 1$. To determine a second independent solution, try k times the repeated root, i.e., $x_k = k\lambda^k = k(1)^k = k$. Substituting into Equation 7, we get

$$px_{k+1} - x_k + qx_{k-1} = \frac{1}{2}(k+1) - k + \frac{1}{2}(k-1) = 0$$

This verifies that $x_k = k$ is a second solution to the difference equation. So, the general solution in this case is

$$x_k = A + Bk$$

Substituting the initial conditions into the above equation yields

$$A = 1$$

$$A + \mathrm{a}B = 0$$

So, $A = 1$, $B = -\frac{1}{a}$ and

$$x_k = A + Bk = 1 - \frac{k}{a}$$

which can be written as

Equation 9. Probability of ruin when $r = 1$

$$x_k = \frac{a-k}{a}$$

For example, if the pot is 40 units of money and A starts with 20 units, then the probability of A losing all of his or her money is

$$x_{20} = \frac{40-20}{40} = \frac{1}{2}$$

which makes sense intuitively.

As a approaches infinite, player A is sure to lose all of his or her money since

$$\lim_{a\to\infty} x_k = \lim_{a\to\infty} \frac{a-k}{a} = \lim_{a\to\infty} \frac{1-\frac{k}{a}}{1} = \frac{1-0}{1} = 1$$

. . .

We can use the above analysis to determine the probability of ruin **for player B**.

In the case that $r \neq 1$, the probability of ruin for player B can be gotten from Equation 8 as modified for Player B, i.e., using $\frac{1}{r}$ rather than r. (Recall that Equation 7 was derived from the perspective of Player A. If we did the derivation from the perspective of Player B, the equation would simply have p and q reversed, i.e., $qx_{k+1} - x_k + px_{k-1} = 0$, which leads to using $\frac{1}{r}$ rather than r in Equation 8 when recast for Player B.)

$$x_{a-k} = \frac{\left(\frac{1}{r}\right)^{a-k} - \left(\frac{1}{r}\right)^{a}}{1 - \left(\frac{1}{r}\right)^{a}} = \frac{r^{k-a} - r^{-a}}{1 - r^{-a}} = \frac{r^k - 1}{r^a - 1}$$

In the case $r = 1$, the probability of ruin for player B can be gotten from Equation 9

$$x_{a-k} = \frac{a - (a - k)}{a} = \frac{k}{a}$$

In both cases, $x_k + x_{a-k} = 1$. So, with probably 1, the game must terminate with one of the players losing all their money.

. . .

The probability of eventual overall win or ruin only tells part of the story. A gambler may also want to know the expected numbers of plays before he or she wins all or goes to ruin. To solve this problem, we introduce two random variables, i.e., N (the number of plays until the end of the game) and K (the initial units of money held by player A). The total amount of money between the two players is a. We seek the expected number of plays until the game ends given that player A starts with $K = k$ units of money, which can be written as the following conditional expectation:

Equation 10. Expected number of plays until ruin, starting with k units of money

$$E(N|K) = \sum_{n=0}^{\infty} n\, p(n|k) = y_k$$

where $p(n|k)$ is the conditional probability that the game ends in n plays given that player A starts with k units of money.

Since $x_k + x_{a-k} = 1$, it is certain that the game will end with one of the players winning all the money. Thus, it must be that

$$\textstyle\sum_{n=0}^{\infty} p(n|k) = 1, \text{ each value of } k$$

and so, $p(n|k)$ is a probability density function.

By the law of total probability, we have

$$p(n|k) = p(n-1|k+1) \cdot p + p(n-1|k-1) \cdot q$$

In words, the event (state) of n steps being left in the game with player A having k units of money, is equivalent to (and has the same probability as)

- there being $n - 1$ steps left in the game with player A having $k + 1$ units of money and when player A wins the current game, or
- there being $n - 1$ steps left in the game with player A having $k - 1$ units of money and when player A loses the current game.

Substituting the above equation into Equation 10 yields

$$y_k = \sum_{n=1}^{\infty} n\,[p(n-1|k+1) \cdot p + p(n-1|k-1) \cdot q]$$

Let $s = n - 1$ in the above equation to get

$$y_k = p\sum_{s=0}^{\infty}(s+1)\,p(s|k+1) + q\sum_{s=0}^{\infty}(s+1)\,p(s|k-1)$$

$$= \;p\sum_{s=1}^{\infty} s\,p(s|k+1) + \;q\sum_{s=1}^{\infty} s\,p(s|k-1) + p\sum_{s=0}^{\infty} p(s|k+1) + q\sum_{s=0}^{\infty} p(s|k-1)$$

$$= \;p\sum_{s=1}^{\infty} s\,p(s|k+1) + \;q\sum_{s=1}^{\infty} s\,p(s|k-1) + p\cdot 1 + q\cdot 1$$

$$= \;p\sum_{s=1}^{\infty} s\,p(s|k+1) + \;q\sum_{s=1}^{\infty} s\,p(s|k-1) + 1$$

The above can be written as the following difference equation

Equation 11. Difference equation for expect number of plays in gambler's ruin

$$p\cdot y_{k+1} - y_k + q\cdot y_{k-1} = -1$$

The solution to the above difference equation (for $p \neq q$) was derived in Section 4.5, i.e.,

$$y_k = \;a + br^k + \tfrac{k}{q-p}, \quad \text{where } r = \tfrac{q}{p}$$

If $k = 0$ or $k = a$, the game terminates before any plays. This observation gives us two boundary conditions, i.e., $y_0 = y_a = 0$. Substituting into the general equation for y_k, we get

$$a + b = 0$$

$$a + br^a + \frac{a}{q-p} = 0$$

Thus,

$$a = -b = -\frac{a}{(q-p)(1-r^a)}$$

and noting that $q - p = (1-p) - p = 1 - 2p$, we have

$$y_k = -\frac{a\left(1-r^k\right)}{(q-p)(1-r^a)} + \frac{k}{q-p} = \frac{1}{1-2p}\left[k - \frac{a\left(1-r^k\right)}{(1-r^a)}\right]$$

For $a = 40$, Table 9 shows the expected number of plays before one of the players goes to ruin. Values of p are listed in the left column and values of k are listed in the top row. Take note of the symmetry in the table. The formula used to compute the row corresponding to $p = q = .5$ in Table 9 (shown in gray) is developed in the discussion just below the table. As expected, the longest expected time to ruin for one of the players occurs at $p = q = .5$ and with each player starting with the same amount of money (in this case, 20 units). The other rows in the table were computed using the formula directly above (case $p \neq q$).

Table 9. Gambler's ruin example – expected number of plays

	5	10	15	20	25	30	35
0.1	6.25	12.50	18.75	25.00	31.25	37.50	43.75
0.2	8.33	16.67	25.00	33.33	41.67	50.00	58.27
0.3	12.50	25.00	37.50	50.00	62.50	74.98	86.05
0.4	25.00	50.00	74.99	99.94	124.54	146.53	148.66
0.5	175	300	375	400	375	300	175
0.6	148.66	146.53	124.54	99.94	74.99	50.00	25.00
0.7	86.05	74.98	62.50	50.00	37.50	25.00	12.50
0.8	58.27	50.00	41.67	33.33	25.00	16.67	8.33
0.9	43.75	37.50	31.25	25.00	18.75	12.50	6.25

For the case that $p = q = \frac{1}{2}$ (implying $r = 1$), Equation 11 becomes

$$\frac{1}{2}y_{k+1} - y_k + \frac{1}{2}y_{k-1} = -1$$

which can be written as

$$y_{k+1} - 2y_k + y_{k-1} = -2$$

The characteristic equation for the homogeneous difference equation $y_{k+1} - 2y_k + y_{k-1} = 0$ is

$$\lambda^2 - 2\lambda + 1 = 0$$

The characteristic equation has a double root, i.e., $\lambda = 1$. In this case, the complementary solution is

$$A(1)^k + Bk(1)^k = A + Bk$$

For a solution to the non-homogeneous difference equation (the particular solution), try Ck^2. This works since when substituted into $x_{k+1} - 2x_k + x_{k-1}$, we get

$$C(k+1)^2 - 2Ck^2 + C(k-1)^2 = 2C$$

Let $C = -1$ to satisfy the equation. Thus, particular solution is $-k^2$.

The general solution to the non-homogeneous difference equation has the form

$$y_k = A + Bk - k^2$$

The boundary conditions, i.e., $y_0 = y_a = 0$, yield the following two equations

$$A = 0$$

$$A + aB - a^2 = 0$$

So, $A = 0$ and $B = a$. Thus, the general solution for the expected number of plays (when $p = q = \frac{1}{2}$) is

$$y_k = ak - k^2 = k(a - k)$$

8.2.3 Random Walks

In Section 8.2.1 (Table 6 and Table 7), we discussed several examples of random walks. In this section, we will focus on random walks with a discrete state space and in discrete time. The gambler's ruin process in the previous section can be seen as a random walk with a finite state space and with absorbing states at either extreme of the state space. The general discrete state space / discrete time random walk allows for the following cases:

- Infinite walk in both directions, with state space $\{\dots, -2, -1, 0, 1, 2, \dots\}$
- Infinite walk in one direction and bounded on one side, with state space of the form $\{a, a + 1, a + 2, \dots\}$ or $\{\dots, a - 2, a - 1, a\}$. State a can be either **absorbing** (i.e., zero probability of leaving once entered) or **reflecting** (probability greater than zero of leaving once entered).
- Finite walk with state space of the form $\{b, b + 1, b + 2, \dots, a\}$. States a and b can be any combination of absorbing or reflecting. When state $b = 0$, and states a and b are both absorbing, we have the equivalent of the gambler's ruin process.

. . .

In the case of an unbounded (unrestricted) random walk in one dimension, let random variable X_n represent the position of the random walk at step n. Let W_n be a modified Bernoulli random variable (i.e., having values -1 or 1 as opposed to the usual Bernoulli random variable with possible values 0 or 1) representing the *n*th step. We then have the following equation

$$X_n = X_0 + \sum_{i=1}^{n} W_i = X_0 + \sum_{i=1}^{n-1} W_i + W_n$$

which implies that

$$X_n = X_{n-1} + W_n$$

So, the position of the walk at step n only depends on position at step $n - 1$, meaning that this process is a Markov chain.

Without loss of generality, we assume $X_0 = 0$ in the following.

Let the probability of going one step to the right be

$$P(X_n = i \,|X_{n-1} = i - 1) = p$$

and the probability of going one step to the left be

$$P(X_n = i \,|X_{n-1} = i + 1) = q = 1 - p$$

Since the expected value function is linear, we have

$$E(X_n) = E\left(\sum_{i=1}^{n} W_i\right) = \sum_{i=1}^{n} E(W_i) = \sum_{i=1}^{n} 1 \cdot p + (-1) \cdot q = n(p-q)$$

Since the W_i random variables are independent and identically distributed, we have by Bienaymé's identity [36]

$$V(X_n) = V\left(\sum_{i=1}^{n} W_i\right) = \sum_{i=1}^{n} V(W_i) = \sum_{i=1}^{n} E\left({W_i}^2\right) - [E(W_i)]^2 = \sum_{i=1}^{n} 1 - (p-q)^2$$
$$= \sum_{i=1}^{n} (1-p^2) - q^2 + 2pq = \sum_{i=1}^{n} (1-p)(1+p) - q^2 + 2pq$$
$$= \sum_{i=1}^{n} q(1+p) - q^2 + 2pq = \sum_{i=1}^{n} 4pq = 4npq$$

Note that $q(1+p) - q^2 = q(1+p-q) = q\big((1-q)+p\big) = q(p+p) = 2qp$.

For example, if the bias is to the right (say $p = .7$), then after $n = 100$ steps, the expected location is $100(.7 - .3) = 40$ and the standard deviation is $\sqrt{4(100)(.7)(.3)} \cong 9.17$.

. . .

Next, we determine the probability distribution of an unbounded (unrestricted) random walk. X_n is defined as above and we again assume that $X_0 = 0$. Let the probability of taking a step to the right be p and to the left be $q = 1 - p$. Let R_n be the random variable representing the number of steps to the right after a total of n steps. Define L_n similarly for steps to the left after a total of n steps. So, if at step $n = 10$ there have been 7 steps to the right and 3 steps to the left, the $R_{10} = 7$ and $L_{10} = 3$, and $X_{10} = 7 - 3 = 4$. In general,

$$X_n = R_n - L_n$$

Let the random variable N represent the total number of steps, i.e., $N = R_n + L_n$. Thus,

$$N = R_n + L_n = (X_n + L_n) + L_n = X_n + 2L_n$$

which implies

$$L_n = \frac{1}{2}(N - X_n)$$

and

$$R_n = \frac{1}{2}(N + X_n)$$

We seek $P(X_n = x \mid X_0 = 0)$ which we will denote as p_n. Using the above two equations, we see that being at position x at time n is equivalent to $r = \frac{1}{2}(n+x)$ steps to the right and $l = \frac{1}{2}(n-x)$ steps to the left.

The number of ways in which r steps can be chosen from a total of n steps is

$$\binom{n}{r} = \binom{n}{\frac{1}{2}(n+x)}$$

X_n follows the binomial distribution, i.e.,

$$p_n = \binom{n}{\frac{1}{2}(n+x)} p^{\frac{1}{2}(n+x)} q^{\frac{1}{2}(n-x)}$$

For example, find the probability that a random walk, starting in position 0, is in position 0 at step number 100, given that $p = q = .5$. From the equation above, we have

$$p_{100} = P(X_{100} = 0) = \binom{100}{50}(.5)^{50}(.5)^{50} = \binom{100}{50}(.5)^{100} \cong 0.0796.$$

Keep in mind, this is not necessarily the first return to 0. The above probability takes into account the possibility of several returns to 0 before step 100.

. . .

Calculation of first return probabilities for random walks is a bit involved. The return to a given state (without loss of generality, we will assume that state to be 0) can only happen after an even number of steps. In the following we assume the starting state is 0.

Let A_{2m} be the event that the random walk is in state 0 at step $2m$, and B_{2i} be the event that the random walk has return to state 0 for the first time at step $2i$. Using the law of total probability, we can relate the return to state 0 to the first return to state 0 as follows:

$$P(A_{2m}) = \sum_{i=1}^{m} P(A_{2m} | \, B_{2i}) \, P(B_{2i})$$

In words, the probability of the event being in state 0 at step $2m$ is the sum of the probabilities of

- returning to state 0 at step 2 for the first time, and again being in state 0 at step $2m$
- returning to state 0 at step 4 for the first time, and again being in state 0 at step $2m$
- …
- returning to state 0 for the first time at step $2m$.

We've already calculated $P(A_{2m})$; it is p_{2m}. Also, $P(A_{2m}| \, B_{2i})$ is the same as p_{2m-2i}.

If we let $f_{2i} = P(B_{2i})$, we can express the previous equation as

$$p_{2m} = \sum_{i=1}^{m} p_{2m-2i} \, f_{2i}$$

We seek an expression for f_{2i}. The generating function for f_{2i} is stated below. The derivation is fairly complex. For the details, see Section 3.4 of "Stochastic Processes: An Introduction" [43].

$$G(s) = 1 - (1 - s^2)^{\frac{1}{2}}$$

[Recall that the probability generating function for a discrete random variable X is defined as

$$G(s) = \sum_{n=0}^{\infty} P(X = n)s^n$$

Once the generating function is determined, the various probabilities for X are coefficients of the powers of s.]

For the problem at hand, the Taylor series for $G(s)$ is

$$\frac{s^2}{2} + \frac{s^4}{2^3} + \frac{s^6}{2^4} + \frac{5s^8}{2^7} + \frac{7s^{10}}{2^8} + \frac{21s^{12}}{2^{10}} + \frac{33s^{14}}{2^{11}} + \frac{429s^{16}}{2^{15}} + \cdots$$

The above series was computed using Wolfram Alpha, with the following command

```
taylor series for 1-sqrt(1-x^2)
```

The coefficients of the above series correspond to the first return probabilities. For example, the probability of the first return to 0 at step #12 is the coefficient of the s^{12}, i.e., $\frac{21}{2^{10}} \cong .0205$.

The numerators of the coefficients in the above series are related to the Catalan numbers, see https://oeis.org/A098597.

. . .

Random walk processes are possible in higher dimensions. Figure 11 shows an example of a 2-dimensional walk. The transition probabilities are only shown for transitions to and from the state $(0,0)$. The diagram extends indefinitely, with the lattice points (i.e., points with integer-valued coordinates) representing the states of the random walk. The transition probabilities are the same for all states, i.e., u (up), d (down), r (right) and l (left). Further, we have that

$$u + d + r + l = 1$$

If $u = d = r = l = \frac{1}{4}$, then the random walk is said to be symmetric.

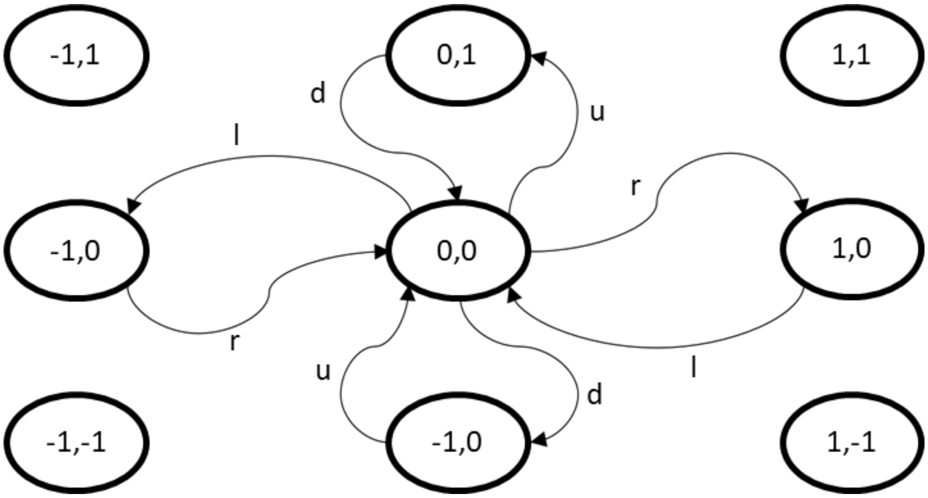


Figure 11. Example of a 2-dimensional random walk

For symmetric random walks in 1 and 2 dimensions, the probability of eventually returning to the starting point is 1. For dimensions 3 and higher, eventual return to the starting point is not certain, i.e., probability less than 1. Eventual return probabilities are given in Table 10. See the article “Pólya's Random Walk Constants” [37] for further details.

Table 10. Return probabilities for higher-dimension random walks

Dimension	Probability of eventual return to starting point
3	0.340537
4	0.193206
5	0.135178
6	0.104715
7	0.0858449
8	0.072912

. . .

There are many applications of random walk theory in the sciences. In fact, the term “random walk” was originally proposed by mathematician and biostatistician Karl Pearson in a letter (1905) to the journal *Nature* concerning a model for mosquito infestation in a forest. Pearson wanted to understand the distribution of the mosquitos after a given number of movements (steps). Pearson’s letter (open request for solution) was answered by Lord Rayleigh who had already solved a more general form of this problem (1880) in the context of sound waves in heterogeneous materials. Both problems relate to diffusion, which is but one area of application for random walks. The journal article “Random walks and diffusion on networks” [39] provides an extensive survey on the theory and applications of random walks to networks with a variety of structures.

The note “A bibliography on applications of random walks in theoretical chemistry and physics” [38] provides a lengthy list (20 pages worth) of references concerning the application of random walk theory to theoretical chemistry and physics.

The journal article ““Random Walks and Their Applications: Widely Used as Mathematical Models” [40] covers applications of random walk theory to polymer configurations, solid physics, inhomogeneous media, and biology.

The monograph “Random Walks and Electric Networks” [41] focuses on the interplay between physics and mathematics in terms of an example, i.e., electric networks.

8.2.4 General Theory

After having seen several specific examples of Markov chains (i.e., gambler’s ruin and random walk), we return to some general concepts and theorems. The focus here is on Markov chains with countably infinite or a finite number of states.

8.2.4.1 Transition Probabilities

Consider a Markov chain $\{X_n\}$ with state space $S = \{i_1, i_2, i_3, \dots\}$ and transition probabilities given by $P(X_{k+1} = j \mid X_k = i\,\} = p_{ij}$. The transition probabilities can be represented as the matrix T below:

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

T is known as the transition matrix for the Markov chain. Each row represents the probabilities of transitioning from a given state to any of the other states. For example, row 2 has the probabilities of transitioning from state i_2 to each of the other states. As defined, each row is a discrete probability distribution. A matrix with all positive entries and whose rows sum to 1 is said to be **row-stochastic**.

For example, take the random walk with states $\{0,1,2,\ldots\}$, and with probability $\frac{3}{5}$ of going right (for states $1,2,3,\ldots$), probability $\frac{2}{5}$ of going left (for states $1,2,3,\ldots$) and probability 1 of going to the right if in state 0. The transition matrix is shown below:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ \frac{2}{5} & 0 & \frac{3}{5} & 0 & 0 & 0 & \cdots \\ 0 & \frac{2}{5} & 0 & \frac{3}{5} & 0 & 0 & \cdots \\ 0 & 0 & \frac{2}{5} & 0 & \frac{3}{5} & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Also of interest are the probabilities of going from one state to another in a n steps, i.e.,

$$p_{ij}^{(n)} = P(X_{k+n} = j \mid X_k = i\,\}$$

It can be proven that $p_{ij}^{(n)}$ is the (i,j) element of T^n (see the section concerning the Chapman-Kolmogorov Equations in [42]). To be clear, $p_{ij}^{(n)}$ is the probability of starting in state i and being in state j at step n. This includes the possibility of earlier visits to state j. This is different from the probability of the first visit to state j starting from state i (which will be discussed later).

For the random walk in the example above, the two-step transition probabilities are given by the matrix below (which is the square of the one-step transition matrix). For example, the probability of a transition from state 2 to state 2 in two steps is the entry in position $(2,2)$, i.e., $\frac{16}{25}$. (Note that our Markov chain in this example started number at 0, and so we number the rows and columns starting with 0.)

$$\begin{bmatrix} \frac{2}{5} & 0 & \frac{3}{5} & 0 & 0 & 0 & \dots \\ 0 & \frac{16}{25} & 0 & \frac{9}{25} & 0 & 0 & \dots \\ 0 & 0 & \frac{16}{25} & 0 & \frac{9}{25} & 0 & \dots \\ 0 & 0 & 0 & \frac{16}{25} & 0 & \frac{9}{25} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

. . .

The probabilities in matrix T are all conditioned on a given state. It is also possible to compute the probability of being in a given state after n steps based on an initial probability distribution. Let $p_j^{(0)}$ be the probability that a given Markov chain is initially in state j. We can write the set of initial probabilities in the form of a row vector, i.e.,

$$p^{(0)} = [p_1^{(0)}, p_2^{(0)}, p_3^{(0)}, \dots]$$

The probability of being in state j at step n given initial probability distribution $p^{(0)}$ is represented by the notation $p_j^{(n)}$. Using the law of total probability, we have

$$p_j^{(n)} = \sum_{i=1}^{\infty} p_i^{(0)}\, p_{ij}^{(n)}$$

The above formula is $p^{(0)}$ times column j in T^n. This is true for all values of j. Thus, we have

$$p^{(n)} = p^{(0)} T^n$$

where we define $p^{(n)}$ as follows

$$p^{(n)} = [p_1^{(n)}, p_2^{(n)}, p_3^{(n)}, \dots]$$

$p_j^{(n)}$ is known as absolute or unconditional probability of being in state j at step n, given the initial distribution $p^{(0)}$.

Returning to our bounded random walk example, assume the initial probability distribution is given by the row matrix

$$p^{(0)} = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0,0,0, \dots\right]$$

Applying the formula for absolute probability, we get

$$p^{(n)} = p^{(0)} T^2 = \left[\frac{10}{75}, \frac{16}{75}, \frac{31}{75}, \frac{9}{75}, \frac{9}{75}, 0,0,0, \dots\right]$$

As a check, we see that $\frac{10}{75}+\frac{16}{75}+\frac{31}{75}+\frac{9}{75}+\frac{9}{75}=\frac{75}{75}=1$. For example, the absolute probability of being in state 2 at step 2 (given the initial probability distribution noted above) is $\frac{31}{75}$. (For this example, keep in mind that first entry in $p^{(n)}$ represents state 0, the second entry represents state 1, and so on.)

. . .

Markov chains can be used to solve fairly complex probability problems. For example, consider a container that initially contains two balls. At each step,

- a ball is randomly taken from the container and
- replaced with a ball of the same color (with probability .6) or replaced with a ball of the opposite color (with probability .4)

The number of balls in the container remains constant at 2.

a) What is the probability of having 2 white balls in the container after the 4th step is complete if we started with 2 white balls?

b) Same question as above if we started with 1 white ball and 1 black ball.

c) What is the probability of having 1 white and 1 black ball in the container after the 4th step is complete if we started with 2 black balls?

d) If initially both balls are white, what is the probability that the 5th ball selected for replacement is white?

To model this problem as a Markov chain, let X_n be the number of white balls in the container at step n. In this case, the Markov chain has three states, i.e., $\{0,1,2\}$. The associated transition matrix is as follows:

$$T=\begin{bmatrix} .6 & .4 & 0 \\ .2 & .6 & .2 \\ 0 & .4 & .6 \end{bmatrix}$$

For example, a transition from state 1 to 1 occurs if a white ball is selected and replaced by a white ball with probability $(.5)(.6)=.3$, or a black ball is selected and replaced by a black ball with probability $(.5)(.6)=.3$. Thus, the probability of going from state 1 to state 1 is $.3+.3=.6$.

Questions a,b and c can be answer by computing the 4th power of the transition matrix

$$T^4=\begin{bmatrix} .3152 & .4992 & .1856 \\ .2496 & .5008 & .2496 \\ .1856 & .4992 & .3152 \end{bmatrix}$$

For question a, we seek the probability of being in state 2 after 4 steps if the process started in state 2. The answer is entry (2,2) in T^4, i.e., .3152. (Note that we label the rows and columns of the transition matrix from 0 to 2 in this case – just to align with the state numbering.)

For question b, we are looking for the probability of the transition $1 \to 2$, which is entry (1,2) in T^4, i.e., .2496.

For question c, we seek the probability of the transition $0 \to 1$, which is entry $(0,1)$ in T^4, i.e., .4992.

To answer question d, we condition on the number of white balls in the container after the fourth selection.

$$P(5th\ selection\ is\ white) = \sum_{i=0}^{2} P(5th\ selection\ is\ white \mid X_4 = i)P(X_4 = i \mid X_0 = 2)$$

$$= 0 \cdot p_{2,0}^{(4)} + (.5)p_{2,1}^{(4)} + 1 \cdot p_{2,2}^{(4)} \cong 0 + (.5)(.4992) + (1)(.3152) = .5648$$

As a check, we can determine the probability that the 5th selection is a black ball, given that we started with two white balls.

$$P(5th\ selection\ is\ white) = \sum_{i=0}^{2} P(5th\ selection\ is\ black \mid X_4 = i)P(X_4 = i \mid X_0 = 2)$$

$$= 1 \cdot p_{2,0}^{(4)} + (.5)p_{2,1}^{(4)} + 0 \cdot p_{2,2}^{(4)} \cong .1856 + (.5)(.4992) + 0 = .4352$$

The two probabilities for the above two cases add to 1, as expected.

. . .

At each step of a process, we randomly place a ball into one of 8 urns. All the balls are identical. Placement in an urn is done one ball at a time. All the urns are empty when the process starts. What is the probability that exactly 3 urns will be occupied after the first 9 balls have been distributed?

Let X_n be a Markov chain with states $\{0,1,2,\ldots,8\}$ where the state represents the number of occupied urns. The transition probabilities are given by

$$p_{i,i} = \frac{i}{8}$$

$$p_{i,i+1} = 1 - \frac{i}{8}$$

It is not possible to move up more than one state since only one ball is placed in an urn at each step. The current state cannot decrease since we never remove any balls.

After 1 step, one of the urns must have a ball. So, we can reduce the problem as follows:

$$p_{0,3}^{(9)} = p_{0,1} \cdot p_{1,3}^{(8)} = 1 \cdot p_{1,3}^{(8)} = p_{1,3}^{(8)}$$

This probability can be computed by determining the transition matrix for the Markov chain and then raising the matrix to the 8th power. However, we can make a simplification here since we are only interested in whether 3 urns are occupied or not. In particular, we can collapse states 4,5,6,7 and 8 into a new state (call it 4*). This leaves us with the following $4\ x\ 4$ transition matrix:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1/8 & 7/8 & 0 & 0 \\ 0 & 0 & 2/8 & 6/8 & 0 \\ 0 & 0 & 0 & 3/8 & 5/8 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Using the matrix multiplication online application at symbolab.com, the above matrix raised to the 8th power (numbers in the matrix are approximated to 5 decimal places) is as follows:

$$\begin{bmatrix} 0 & 0 & .00042 & .01934 & .98022 \\ 0 & 0 & .00010 & \mathbf{.00757} & .99232 \\ 0 & 0 & .00001 & .00225 & .99772 \\ 0 & 0 & 0 & .00039 & .99960 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The answer to our problem is approximately $.00757$.

8.2.4.2 Communication Classes and Irreducibility

Recall that two states (i and j) of a Markov chain are said to communicate if they are reachable from one another by a sequence of transitions with positive probability. This is equivalent to the condition that $p_{i,j}^{(n)} > 0$ for some value of $n \geq 1$. A Markov chain may be divided into collections of states that communicate among themselves. More formally, we have the following theorem.

Theorem 19 A Markov chain with a finite or countably infinite number of states can be uniquely decomposed into collection of disjoint sub-chains $C_1, C_2, \ldots$ whose union is the entire Markov chain such that all states within a sub-chain communicate with each other but with no other states.

Each sub-chain is known as a **communication class** of the Markov chain.

A Markov chain with only one communication class is referred to as an **irreducible Markov chain**.

For the gambler's ruin Markov chain, we have three communications classes, i.e., $\{0\}, \{1,2, \ldots, a - 1\}$ and $\{a\}$.

For the unbounded, 1-dimensional random walk, there is but one communication class and so the Markov chain is irreducible.

The Markov chain in Figure 12 has three communication classes, i.e., $\{0,1,2\}, \{3\}$ and $\{4,5,6\}$.

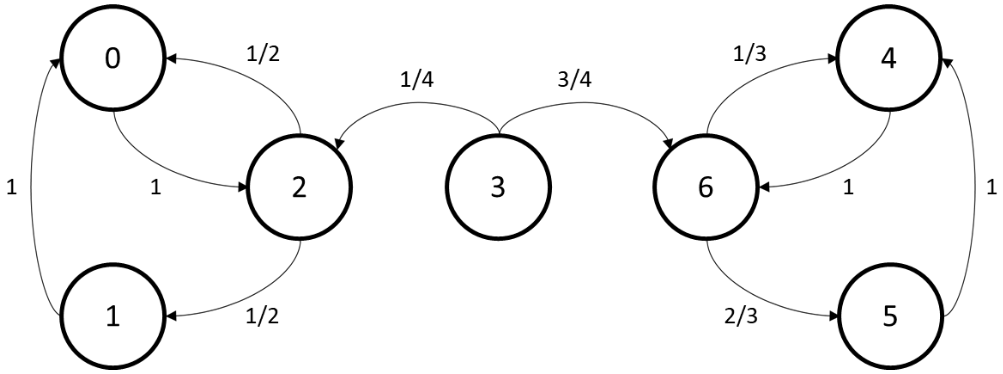


Figure 12. Markov chain with 3 communication classes

8.2.4.3 Recurrence and Transience

In addition to the probability of a Markov chain returning to a given state, we are also interested in the number of steps it takes to return. To that end, we define the return time to state i as follows:

$$\tau_{i,i} = \min \{n \geq 1 \text{ such that } X_n = i \text{ given that } X_0 = i\}$$

If no such n exists (i.e., the Markov chain never returns to state i), then $\tau_{i,i} = \infty$.

Further, we define $f_{i,i} = P(\tau_{i,i} < \infty)$ as the probability of ever returning to state i given that the chain started in state i.

A state i is defined to be **recurrent** if $f_{i,i} = 1$; and **transient** if $f_{i,i} < 1$. [Some sources use the term “persistent” to mean the same thing as “recurrent.”]

By the Markov property, each time a Markov chain visits a state i it will return with the same probability $f_{i,i}$. So, if $f_{i,i} = 1$, state i will be visited an infinite number of times (thus, the term “recurrent”). If $f_{i,i} < 1$, state i will only be visited a finite number of times.

The total number of visits to state i, starting from state i, is given by the equation

$$N_{i,i} = \sum_{n=0}^{\infty} I\{X_n = i | X_0 = i\}$$

[The indicator function $I\{A\}$ equals 1 if A occurs and 0 if A does not occur. So, $P(I\{A\} = 1) = P(A)$ and $P(I\{A\} = 0) = P(\neg A)$. Further,

$$E(I\{A\}) = 1 \cdot P(A) + 0 \cdot P(\neg A) = P(A)$$

So, there is a one-to-one correspondence between the expected value of the indicator function of an event and the probability of that event.]

The random variable $N_{i,i}$ has a geometric distribution when $f_{i,i} < 1$, i.e.,

$$P\left(N_{i,i} = n\right) = f_{i,i}^{n-1}\left(1 - f_{i,i}\right), \qquad n \geq 1$$

The above equation can be interpreted as the probability of starting in state i, returning $n - 1$ times, and never returning again for a total of n visits (including the initial visit).

The expected number of visits to state i is $E\left(N_{i,i}\right) = \frac{1}{1-f_{i,i}}$ (known result for geometric distributions). This gives us an equivalent definition for recurrent and transient states:

- State i is recurrent ($f_{i,i} = 1$) if and only if $E\left(N_{i,i}\right) = \infty$.
- State i is transient ($f_{i,i} < 1$) if and only if $E\left(N_{i,i}\right) = \frac{1}{1-f_{i,i}} < \infty$.

Using the indicator function, we see that

$$E\left(N_{i,i}\right) = \sum_{n=1}^{\infty} E(I\{X_n = i|X_0 = i\}) = \sum_{n=1}^{\infty} P(\{X_n = i|X_0 = i\}) = \sum_{n=1}^{\infty} p_{i,i}^{(n)}$$

Thus, state i is recurrent if and only if

$$\sum_{n=1}^{\infty} p_{i,i}^{(n)} = \infty$$

and otherwise, state i is transient.

Theorem 20. If state i is recurrent, and state i communicates with state j, then state j is recurrent. Further, if state i is transient, and state i communicates with state j, then state j is transient

Proof:

State i is recurrent:

Since i communicates with j, there exists positive integers k and m such that $p_{i,j}^{(k)} > 0$ and $p_{j,i}^{(m)} > 0$. Thus, for any positive integer n, we have

$$p_{j,j}^{(m+n+k)} \geq p_{j,i}^{(m)} p_{i,i}^{(n)} p_{i,j}^{(k)}$$

The above inequality is true since the probability on the left allows for any route between j back to j in $m + n + k$ steps, and the probability is the probability of returning to j also in $m + n + k$ steps but with the added restrictions of visiting state i at step m, returning to i in n steps, and then traveling to j in k steps.

Summing the above inequality over all values of n, we have

$$\sum_{n=1}^{\infty} p_{j,j}^{(m+n+k)} \geq p_{j,i}^{(m)} p_{i,j}^{(k)} \sum_{n=1}^{\infty} p_{i,i}^{(n)} = \infty$$

Thus, state j is also recurrent.

State i is transient:

This follows easily from the above result. For if j were recurrent, then we know that i would also be recurrent (contradiction to our assumption. ■

Theorem 20 implies the following result:

Theorem 21. If i and j are in the same communication class, then both are either recurrent or both are transient. Further, for an irreducible Markov chain, either all states are recurrent or all states are transient.

For an irreducible Markov chain, if all states are recurrent, then the Markov chain is said to be recurrent; otherwise, the Markov chain is said to be transient.

Theorem 22. An irreducible Markov chain with a finite state space is recurrent.

Proof:

If a Markov chain has a finite state space, then not all the states can be transient; otherwise, after a finite number of steps the chain would leave every state never to return and thus have nowhere to go. ■

8.2.4.4 Positive and Null Recurrence

The expected return time to a recurrent state i, i.e., $E(\tau_{i,i})$, may be infinite. In such cases, the state is said to be **null recurrent**. If the expected return time for a state is a positive number less than infinity, then the state is said to be **positive recurrent** (also known as non-null recurrent).

The states of the symmetric random walk (i.e., $p = q = .5$) are null recurrent. For a derivation of this fact, see Section 3.4 of the book “Stochastic Processes: An Introduction” [43].

Theorem 23. If two states i and j are recurrent and communicate (i.e., in the same communication class), and state i is positive recurrent, then state j is also positive recurrent. Further, the states in a recurrent communication class are either all positive recurrent or all together null recurrent.

For a proof of this theorem, see Proposition 4.5 in the book “Introduction to Probability Models” [42].

Combining Theorem 21 and Theorem 23, we have the following summary result:

Theorem 24. The states in a communication class are either all positive recurrent, all null recurrent or all transient.

Further, in an irreducible Markov chain, the states are either all positive recurrent, all null recurrent or all transient.

8.2.4.5 Periodicity

The states of a Markov chain can be periodic in terms of return visits. For example, the states in a 1-dimensional random walk are all of period 2 since the return time (in steps) is always an even number.

To define periodicity, we use the function

$$d(x) = \gcd\{n \mid p_{x,x}^{(n)} > 0\}$$

For a given state i, if $d(x) > 1$, the state is said to be **periodic**.

If for a state x in a Markov chain $p_{x,x}^{(1)} > 0$ (i.e., there is a positive probability of a return to state x in one step), then state x is **aperiodic**, since $d(x) = 1$.

Figure 13 depicts a Markov chain all of whose states are of period three. For example, consider state 1. Starting in state 1, one can return in 3,6,9, … steps. The GCD of these numbers is 3 and thus state 1 is of period 3

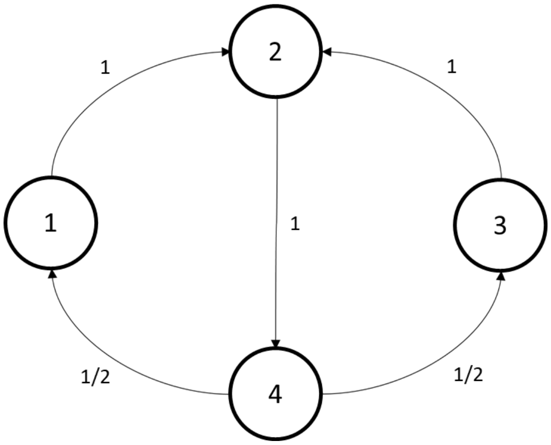


Figure 13. Markov chain with states of period 3

Periodicity, like recurrence and transience, is a class property that is shared by all states in a communication class. In particular, we have the following theorem:

Theorem 25. If states x and y are in the same communication class, then $d(x) = d(y)$.

Proof:

Since x and y communicate, we know there exists positive integers k and m such that

$$p_{x,y}^{(k)} > 0, \qquad p_{y,x}^{(m)} > 0$$

We have that

$$p_{x,x}^{(k+m)} \geq p_{x,y}^{(k)} p_{y,x}^{(m)} > 0$$

since the paths from x to x of length $k + m$ include (are a superset of) the set of paths from x to x going first from x to y in k steps and then from y to x in m steps. Since x has period $d(x)$, $k + m$ must be a multiple of $d(x)$.

Next, let l be any positive integer such that $p_{y,y}^{(l)} > 0$. Using a path argument similar to the above, we have that

$$p_{x,x}^{(k+l+m)} \geq p_{x,y}^{(k)}\, p_{y,y}^{(l)}\, p_{y,x}^{(m)} > 0$$

Since x has period $d(x)$, $k + l + m$ must be a multiple of $d(x)$.

Thus, $(k + l + m) - (k_m) = l$ must be a multiple of $d(x)$. Since l was chosen arbitrarily, it must be that $d(x)$ is a divisor of all l such that $p_{y,y}^{(l)} > 0$. So, $d(x) \leq d(y)$ by the definition of GCD.

If we reverse the roles of x and y in the above arguments, then we also have $d(y) \leq d(x)$.

Putting the two results together gives us $d(x) = d(y)$. ■

The Markov chain in Figure 14 is irreducible, i.e., every state can be reached by every other state. So, by Theorem 25, we only need to determine the periodicity of one state to determine the periodicity of all the states.

Let's analyze state 4. We have the following paths for returning to state 4:

- $4 \rightarrow 3 \rightarrow 2 \rightarrow 1$ (3 steps)
- $4 \rightarrow 3 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 3 \rightarrow 2 \rightarrow 4$ (8 steps)

There are other return loops but they are all combinations of the above loops. The GCD of 3 and 8 is 1, and so, $d(4) = 1$. Thus, state 4 is aperiodic and by Theorem 25, all the states of the Markov chain are aperiodic.

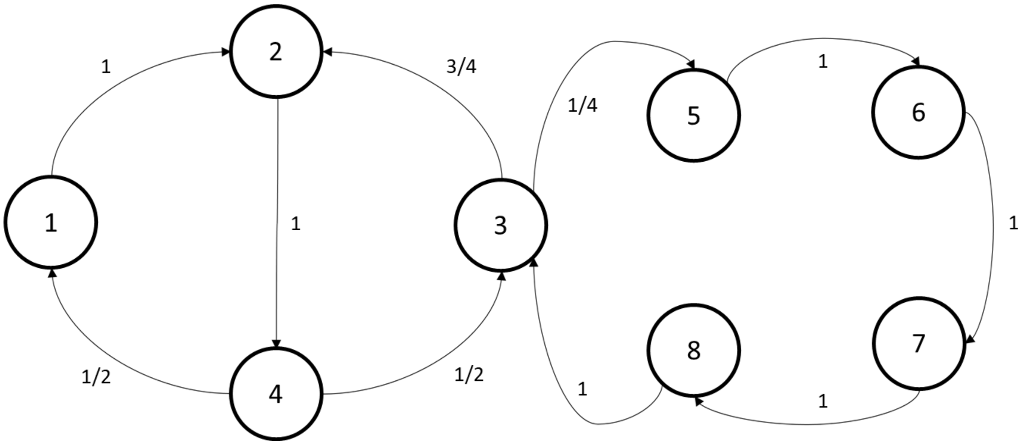


Figure 14. Aperiodic Markov chain

The irreducible Markov chain in Figure 15 is a variation of the Markov chain in Figure 14, with one additional state. With this modification, the Markov chain is now periodic. If we consider state 1, the following are some example loops:

- $1 \rightarrow 2 \rightarrow 4 \rightarrow 1$ (3 steps)
- $1 \rightarrow 2 \rightarrow 4 \rightarrow 3 \rightarrow 2 \rightarrow 4 \rightarrow 1$ (6 steps)
- $1 \rightarrow 2 \rightarrow 4 \rightarrow 3 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 9 \rightarrow 3 \rightarrow 2 \rightarrow 4 \rightarrow 1$ (12 steps)

All other loops back to 1 are also multiples of 3 and so, $d(1) = 0$. By Theorem 25, all the states in the Markov chain are of period 3.

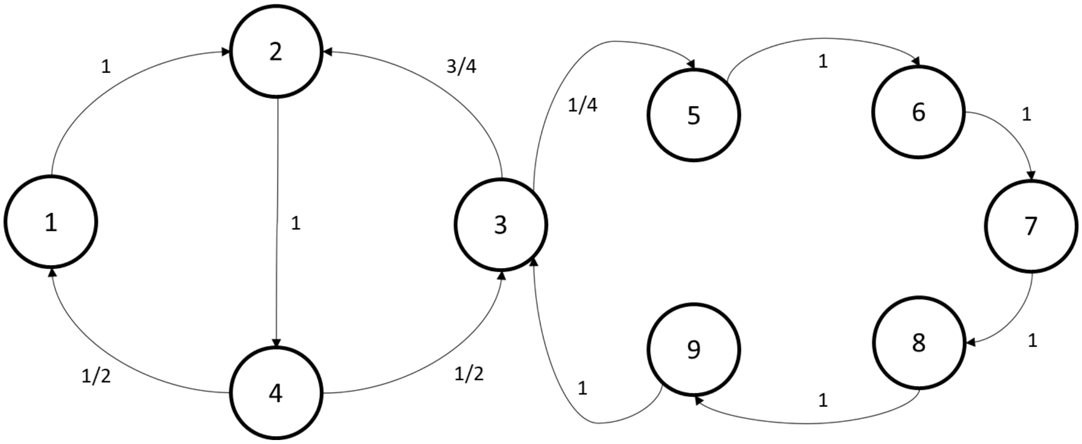


Figure 15. Markov chain with period 3

8.2.4.6 Stationary (Limiting) distributions

In some cases, a Markov chain will exhibit limiting behavior in the sense that the long-term probabilities of being in each state are stationary. In what follows, we let π_j be the long-term proportion of time that a Markov chain spends in state j. More formally, we define π_j (assuming a starting state of i) as

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} I\{X_k = j \mid X_0 = i\}$$

which can be shown to be equivalent to

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} P(\{X_k = j \mid X_0 = i) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} p_{i,j}^{(k)}$$

We state the following two theorems (proofs can be found in Section 4.4 of [42]).

Theorem 26. If a Markov chain is irreducible and positive recurrent, then for any initial state (i.e., ***independent*** *of the initial state), the long-term proportion of the time that the Markov chain is in state j is given by*

$$\pi_j = \frac{1}{E(\tau_{j,j})}$$

Theorem 27. The long-term proportions for the states in an irreducible, positive recurrent Markov chain, with state space S, are the solutions to the equations

$$\pi_j = \sum_{i\in S}\pi_i p_{i,j}$$

$$\sum_{i\in S}\pi_i = 1$$

If we define the row matrix $\pi = (\pi_1, \pi_2, \pi_3, \dots)$ and column matrix $\overline{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$, then the above equations can be written in matrix form as

$$\pi = \pi T$$

$$\pi\overline{1} = 1$$

where T is the transition matrix for the Markov chain.

For example, consider the Markov chain in Figure 13. The transition matrix T is

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

The equations $\pi = \pi T$ and $\pi\overline{1} = 1$ can be expanded to

$$\pi_1 = \frac{1}{2}\pi_4$$

$$\pi_2 = \pi_1 + \pi_3$$

$$\pi_3 = \frac{1}{2}\pi_4$$

$$\pi_4 = \pi_2$$

$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

Solving the above equations, we get $\pi_1 = \pi_3 = \frac{1}{6}$ and $\pi_2 = \pi_4 = \frac{1}{3}$, and so $\pi = (\frac{1}{6}, \frac{1}{3}, \frac{1}{6}, \frac{1}{3})$.

One can check the solution by computing πT and verifying that it equals π.

8.3 Continuous-time Markov Chains

8.3.1 Definitions and Concepts

In this section, we consider Markov chains that have a discrete state space, and continuous time intervals between state transitions. In particular, we assume the interval between state transitions is governed by a continuous random variable $X > 0$ with the associated memoryless property:

$$P(X > a + b \mid X > b) = P(X > a)$$

It should be emphasized that in the case of discrete state space Markov chains with discrete times between transitions, time is measured in steps (transitions). The real-time duration between transitions is not considered in this case. In this section, we are looking at the case where the time it takes to transition from one state to another is a continuous random process.

The continuous version of the memoryless property (i.e., the above expression) implies that random variable X must follow the exponential distribution, which we prove in the following theorem.

Theorem 28. If random variable X with range $[0, \infty)$ satisfies the continuous memoryless property, then X is exponentially distributed.

Proof:

From the memoryless property and the definition of conditional probability, we have

$$P(X > a) = P(X > a + b \mid X > b) = \frac{P(X > a + b, X > b)}{P(X > b)} = \frac{P(X > a + b)}{P(X > b)}$$

which implies

$$P(X > a + b) = P(X > a) \cdot P(X > b)$$

(Note that the comma between $X > a + b$ and $X > b$ mean "and". The same notation will be used in the following text.)

Letting $G(a) = P(X > a)$, we can write the above equation as $G(a + b) = G(a) \cdot G(b)$.

$G(a)$ is monotonically decreasing, as is the case for all survival functions [44]. Thus, $F(a)$, the cumulative probability distribution for X, is monotonically increasing. We will show that $F(a)$ is exponential.

We will prove that the above equation $G(a + b) = G(a) \cdot G(b)$ implies $G(x) = [G(1)]^x$ for any positive real number x. This result can be written as

$$G(x) = e^{-\lambda x}$$

where $\lambda = -\ln[G(1)]$, i.e., X is exponential distribution with cumulative probability distribution $F(x) = 1 - e^{-\lambda x}$.

We first prove $G(x) = [G(1)]^x$ when x is a rational number.

Assuming x is a rational number, we can write $x = \frac{m}{n}$ for positive integers m and n. By repeated application of the result $G(a + b) = G(a) \cdot G(b)$, we have

$$G\left(\frac{m}{n}\right) = G\left(\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}\right) = \left[G\left(\frac{1}{n}\right)\right]^m$$

Raising the above equation to the power n yields

$$\left[G\left(\frac{m}{n}\right)\right]^n = \left[G\left(\frac{1}{n}\right)\right]^{nm} = \left(\left[G\left(\frac{1}{n}\right)\right]^n\right)^m = [G(1)]^m$$

Raising the above to the $\frac{1}{n}$ power gives us

$$G\left(\frac{m}{n}\right) = [G(1)]^{\frac{m}{n}}$$

Next, we prove the above result for x being any real number. To that end, we choose sequences of rational numbers $\{q_n\}$ and $\{r_n\}$ such that $q_n < x < r_n$ and both q_n and r_n approach x as $n \to \infty$. From the result that we proved for the rational case, we have

$$[G(1)]^{q_n} = G(q_n) \geq G(x) \geq G(r_n) = [G(1)]^{r_n}$$

Letting $n \to \infty$ yields our desired result, i.e., $[G(1)]^x = G(x)$. ■

The converse of the above theorem is also true. If random variable X is exponentially distributed with rate λ, then the memoryless property holds, i.e.,

$$P(X > a + b \mid X > b) = \frac{P(X > a + b, X > b)}{P(X > b)}$$

$$= \frac{P(X > a + b)}{P(X > b)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda b}} = e^{-\lambda a} = P(X > a)$$

Thus, the only continuous random variable defined on $[0, \infty)$ that satisfies the memoryless property is the exponential distribution.

In summary, a **continuous-time Markov chain** is defined by a set of states and associated state transition probabilities, along with exponential **holding times** for each state (i.e., the amount of time the process remains in a state before transiting to another state). Further, the holding times are independent of each other.

For example, take the continuous-time Markov chain with states $\{0,1,2\}$ and associated state transition probabilities given by the following matrix:

$$\begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/6 & 0 & 5/6 \\ 1/3 & 2/3 & 0 \end{bmatrix}$$

We also need to be given the holding times for each state.

- The amount of time before leaving state 0 is governed by an exponential distribution with rate 6.
- The holding times for states 1 and 2 have exponential distributions with rates 18 and 12, respectively.

Notice the zeros on the main diagonal. This is always the case for a continuous-time Markov chain.

8.3.2 Counting Processes

A stochastic process $\{N(t), t \geq 0\}$ is classified as a **counting process** if $N(t)$ is the total number of *events* (sometimes "arrivals" is used) that have occurred by time t . It follows that $N(0) = 0$, and $N(t) \geq N(s)$ if $t > s$ (i.e., $N(t)$ is non-decreasing). For example, the number of people arriving at a vending machine in a hotel lobby is a counting process. The number of vehicles crossing an intersection is another example.

Without further conditions, a counting process is not a continuous-time Markov process. However, it can be proven that if a random process has the two conditions described below then it also has the Markov property (see Item #6 in "Processes with Stationary, Independent Increments" [45].)

A counting process is said to have **independent increments** if the number of events occurring in disjoint time intervals is independent. More formally, a counting process (or for that matter any stochastic process) has independent increments if and only if for every positive integer m and any choice of times $t_0 < t_1 < t_2 < \cdots < t_{m-1} < t_m$ the random variables

$$N(t_1) - N(t_0), N(t_2) - N(t_1), \ldots, N(t_m) - N(t_{m-1})$$

are independent.

The vending machine example is likely to have independent increments. The traffic example may not since if an intersection is very busy during a time interval, drivers may use an alternate route during several subsequent intervals.

A counting process is said to have **stationary increments** if the distribution of the number of events occurring in any time interval depends only on the length of the time interval. In other words, the process has stationary increments if the number of events in the interval $(s, s\ + t)$ has the same distribution for all values of s.

If the vending machine (or perhaps a money dispensing machine) is in an area used throughout the day and night (e.g., a mini-market open $24\ x\ 7$) with a constant rate of customer arrivals (i.e., events), the associated counting process (number of people using the machine) will have stationary increments.

It is possible to have independent increments which are not stationary. For example, the counting processes associated with the intersection crossing example, may have independent increments (one time period not affecting another) but the traffic flow could vary throughout the day (thus, implying different probability distributions for vehicle arrivals at the intersection).

It is also possible to have stationary increments that are dependent. For example, let the distribution for the process be $N(t) = tZ$ where Z is the normal distribution. The distribution for each increment of length s has the same normal distribution:

$$N(s + t) - N(s) = (s + t)Z - tZ = sZ$$

However, we have (for example)

$$P[N(1) - N(0) > 0, N(3) - N(2) < 0] = P[Z > 0, Z < 0] = 0$$

but

$$P[N(1) - N(0) > 0]\ P[N(3) - N(2) < 0] = P[Z > 0]\ P[Z < 0] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

So, the increments $N(1) - N(0)$ and $N(3) - N(2)$ are dependent. Also, see the associated discussion on Mathematics StackExchange [47].

Stochastic processes with the added conditions of independent and stationary increments are known as Lévy processes. From the Wikipedia article on Lévy processes [46]:

> A Lévy process, named after the French mathematician Paul Lévy, is a stochastic process with independent, stationary increments: it represents the motion of a point whose successive displacements are random, in which displacements in pairwise disjoint time intervals are independent, and displacements in different time intervals of the same length have identical probability distributions. A Lévy process may thus be viewed as the continuous-time analog of a random walk.

The most well-known examples of Lévy processes are the Wiener process, often called the Brownian motion process, and the Poisson process.

8.3.3 Poisson Processes

In what follows, we use an axiomatic approach to define Poisson processes. In particular, a counting process $\{N(t), t \geq 0\}$ is a **Poisson process** with rate $\lambda > 0$ if the following axioms hold:

i. $N(0) = 0$

ii. $N(t)$ has independent increments

iii. $P\big[\big(N(t+h) - N(t)\big) = 1\big] = \lambda h + o(h)$

iv. $P\big[\big(N(t+h) - N(t)\big) \geq 2\big] = o(h)$

Regarding axioms iii and iv, the notation $o(h)$ is known as "Little o" and is defined relative to a given function. The function $f(x)$ is said to be of order $o(h)$ if

$$\lim_{h \to 0} \frac{f(h)}{h} = 0$$

In words, this means that $f(h)$ is dominated by h as h approaches 0.

For example, $f(x) = x^2$ is $o(h)$ since

$$\lim_{h \to 0} \frac{h^2}{h} = \lim_{h \to 0} h = 0$$

The linear combination of two functions that are $o(h)$ is also $o(h)$, i.e.,

$$\lim_{h \to 0} \frac{af(h) + bg(h)}{h} = a \lim_{h \to 0} \frac{f(h)}{h} + b \lim_{h \to 0} \frac{g(h)}{h} = 0 + 0 = 0$$

Axiom iii is saying that the probability of having one event in an interval of length h is approximately the arrival rate λ times the length of the interval h. Axiom iv is saying that the probability of two arrives in an interval is approximately 0. In both cases, we assume h is small, i.e., approaching 0.

Let $N_s(t)$ be the process when a given Poisson process $N(t)$ is observed from time $s > 0$. In terms of a formula, we have

$$N_s(t) = N(s+t) - N(s)$$

Theorem 29. $N_s(t)$ is a Poisson process with rate λ.

Proof:

i. $N_s(0) = N(s+0) - N(s) = 0$

ii. Since non-overlapping intervals are independent for all values of $t \geq 0$, they are also independent if we start from time $s \geq t \geq 0$.

iii. $P[N_s(t+h) - N_s(t) = 1] = P\big[N\big((s+t)+h\big) - N(s+t) = 1\big] = \lambda h + o(h)$

iv. similar analysis to iii. ■

Let T_n be a random variable that represents the interarrival time between event $n-1$ and event n.

Using the axioms that define a Poisson process, we can prove that the interarrival times are exponentially distributed. We first prove the result for T_1.

Theorem 30. The random variable T_1 is exponentially distributed.

Proof: We want to show that $P(T_1 > t) = P(N(t) = 0) = e^{-\lambda t}$.

We have that

$$P(N(t+h) = 0) = P[N(t) = 0, N(t+h) - N(t) = 0]$$

By Axiom ii, the above expression is equal to

$$P(N(t) = 0) \cdot P(N(t+h) - N(t) = 0)$$

which is equivalent to

$$P(N(t) = 0) \cdot [1 - P(N(t+h) - N(t) \geq 0)]$$

By Axioms iii and iv, and the fact the linear combinations of $o(h)$ functions are also $o(h)$, the above expression is equal to

$$P(N(t) = 0) \cdot \big(1 - \lambda h + o(h)\big)$$

Thus,

$$P(N(t+h) = 0) = P(N(t) = 0) \cdot \big(1 - \lambda h + o(h)\big)$$

which implies

$$P(N(t+h) = 0) - P(N(t) = 0) = -\lambda h P(N(t) = 0) + o(h)$$

Letting $f(t) = P(N(t) = 0)$, dividing both sides of the above by h and taking the limit as h approaches zero, we get

$$f'(t) = -\lambda\, f(t)$$

which can be written as

$$\frac{f'(t)}{f(t)} = -\lambda$$

Taking the indefinite integral on both sides of the above, we get

$$\ln\big(f(t)\big) = -\lambda t + A$$

which implies

$$f(t) = Be^{-\lambda t}$$

However, $f(0) = P(N(0) = 0) = 1$ and so, $B = 1$.

Since T_1 is greater than t if and only if $N(t) = 0$, we conclude that $P(T_1 > t) = P(N(t) = 0) = e^{-\lambda t}$. ■

Using Theorem 30, we can show that all the interarrival times $\{T_n\}$ are exponentially distributed.

Theorem 31. The interarrival times $\{T_n\}$ are independent and exponentially distributed random variables with rate λ.

Proof: We already know from Theorem 31 that the result holds for T_1.

Next, we should the result holds for T_2.

We have that

$$P(T_2 > t \mid T_1 = s) = P(\text{no events in } (s, s+t] \mid T_1 = s)$$

Note that the two intervals $(0, s]$ and $(s, s+t]$ are disjoint, and by Axiom ii, we know that disjoint increments are independent. Thus, we have

$$P(\text{no events in } (s, s+t] \mid T_1 = s)$$
$$= P(\text{no events in } (s, s+t])$$

which is equivalent to

$$P(N(s+t) - N(t) = 0)$$
$$= P(N_s(t) = 0)$$
$$= e^{-\lambda t}$$

The last equality in the above follows since (by Theorem 29) $N_s(t)$ is Poisson process with rate λ, and by Theorem 30, the first interarrival time of $N_s(t)$ is exponentially distributed. (Note that $N_s(t) = 0$ is equivalent to the first arrival (after time s) taking longer than time t.)

So, T_2 is exponentially distributed with rate λ. Further, since $P(T_k > t \mid T_{k-1} = s)$ does not dependent on s, T_2 is independent of T_1. We can repeat the argument for $T_3, T_4 \ldots$ ■

So, we have established that a Poisson process is a continuous-time Markov process. The state space is $\{0,1,2, \ldots\}$. The transition probabilities are very simple, i.e.,

- the probability of going from state i to $i+1$ is 1
- all other transitions have probability 0.

Each state has the same holding time distribution, i.e., exponential distribution with rate λ.

. . .

It can be shown (see Theorem 5.1 in Ross [42]) that the probability distribution for a Poisson process $N(t)$ with rate λ is given by the following formula:

$$P(N(t) = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \qquad n \geq 0$$

Not surprisingly, this is known as the **Poisson distribution**. The expected value and variance of $N(t)$ are both equal to λt.

Since (by Theorem 29) $N_s(t) = N(s + t) - N(s)$ is a Poisson process with rate λ, the above result implies that the number of arrivals (events) in any interval of length t (for a given Poisson process $N(t)$) follows the Poisson distribution with rate λ. Thus, Poisson processes have stationary increments.

. . .

It is possible to **merge** two independent Poisson processes, with the result being another Poisson process. For example, consider two streams of Poisson process arrivals at a bank (e.g., business customers and residential customers). Let $N_1(t)$ be a Poisson process with rate λ_1, and $N_2(t)$ be a Poisson process with rate λ_2. It can be shown (see "Combining and Splitting Poisson Processes" [48]) that the merged process $N(t) = N_1(t) + N_2(t)$ is a Poisson process with rate $\lambda_1 + \lambda_2$. Further, independent of time t and of the preceding number of arrivals, the probability of the next arrival being of Type 1 (i.e., from the $N_1(t)$ Poisson process) is given by

$$\frac{\lambda_1}{\lambda_1 + \lambda_2}$$

and the probability of the next arrival being of Type 2 is

$$1 - \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

On the other hand, one can also **split** a Poisson process with rate λ into two independent Poisson processes with respective rates $p\lambda$ and $(1 - p)\lambda$. The assumption here is that the arrivals are of two types (Type 1 and 2) with the probability of a Type 1 arrival being p and the probability of a Type 2 arrival being $1 - p$.

. . .

The restriction of a constant rate λ can be lifted and replaced by a rate that varies with time. The time-varying rate is known as the intensity function and is represented as $\lambda(t)$. This modification gives rise to what is called a **nonhomogeneous Poisson process**. A key characteristic of nonhomogeneous Poisson processes is that the condition of stationary increments is not required. This allows for the arrival rates to vary over time.

Nonhomogeneous Poisson processes are a good fit for many counting processes, e.g., vehicles arriving at an intersection, arrivals at a bank or other businesses.

More formally, a counting process $\{N(t), t \geq 0\}$ is a nonhomogeneous Poisson process with rate $\boldsymbol{\lambda(t)} > 0$ if the following axioms hold:

i. $N(0) = 0$

ii. $N(t)$ has independent increments

iii. $P[N(t+h) - N(t) = 1] = \boldsymbol{\lambda(t)} \cdot h + o(h)$

iv. $P[N(t+h) - N(t) \geq 2] = o(h)$

The mean value function of a nonhomogeneous Poisson process is defined to be

$$m(t) = \int_0^t \lambda(x)dx$$

For a nonhomogeneous Poisson process $\{N(t), t \geq 0\}$ with rate $\lambda(t) > 0$, the following can be proven (see Theorem 5.3 in Ross [42]):

$$P(N(t) = n) = \frac{[m(t)]^n e^{-m(t)}}{n!}$$

. . .

The basic Poisson process only allows for arrivals one-at-a-time. If this restriction is lifted, we have what is called a **compound Poisson process**. A compound Poisson process is defined as

$$X(t) = \sum_{i=1}^{N(t)} Y_i$$

where $N(t)$ is a basic Poisson process, $\{Y_i\}$ is a collection identically distributed (usually discrete) random variables, and $\{Y_i\}$ is independent from $N(t)$.

For example, if we are interested in the number of transactions that each customer requests at a bank visit, this can be modeled as a compound Poisson process. In this case, the customers arrive according to the Poisson process $N(t)$ and the number of transitions for customer number i is modeled with the random variable Y_i.

When $Y_i = 1$ for all i, then $X(t) = N(t)$ reduces to a basic Poisson process.

Using Wald's equation [49], we have

$$E\big(X(t)\big) = E\big(Y_1 + Y_2 + \cdots + Y_{N(t)}\big) = E\big(N(t)\big) \cdot E(Y_1) = \lambda t \cdot E(Y_1)$$

[There is nothing special about Y_1 in the above formula. We could have used any element from the set $\{Y_i\}$ since they are identically distributed and all have the same expected value.]

The variance of $X(t)$ is $\lambda t \cdot E(Y_1^2)$.

8.3.4 Birth and Death Processes

Birth and death processes are another example of continuous-time Markov chains. In such a process, the state transitions are of two types, i.e.,

- births – increase the population by one
- deaths – decrease the state by one.

The terms "birth" and "death" are used to represent any process whose associated state can either increase or decrease by one at each transition, and where the associated transition time depends on the current state. One may view this as a continuous analog of a random walk with state space $\{0,1,2,\ldots\}$.

More precisely, whenever the system is in state n, new arrivals (births) occur according to an exponential distribution with birth rate λ_n and deaths occurs according to an exponential distribution with death rate μ_n. The two distributions are assumed to be independent of each other.

It is possible to have a pure birth process when $\mu_n = 0$ for all values of n, or to have a pure death process when $\lambda_n = 0$ for all values of n.

If $\mu_n = 0$ for all values of n and $\lambda_n = \lambda$, then we have a Poisson process.

If $\mu_n = 0$ for all values of n and $\lambda_n = n\lambda$, we have what is called a Yule process [50].

. . .

The following theorem is needed to further our analysis of birth and death processes.

Theorem 32. If $X_1, X_2, \ldots, X_n$ are independent and exponentially distributed random variables with respective rates $\lambda_1, \lambda_2, \ldots, \lambda_n$ then min{$X_1, X_2, \ldots, X_n$} is exponentially distributed with rate $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$ and the probability the next arrival comes from event stream k is

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_n}$$

Proof: See the Wikipedia article on the exponential distribution [51].

The probability of a transition from state 0 to 1 is 1 since there cannot be a death if the population size (i.e., current state) is 0, i.e., $p_{0,1} = 1$.

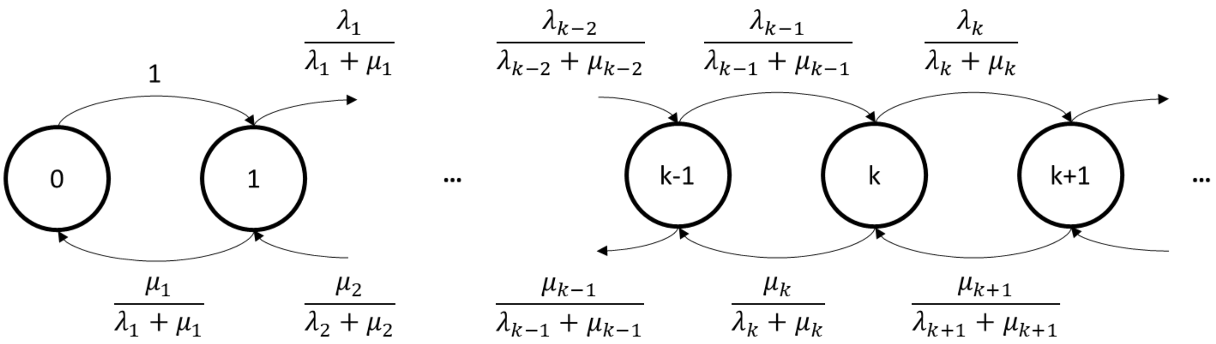
By Theorem 32 (for the case $n = 2$), the probability of a transition from state i to $i + 1$ (i.e., having a birth before the next death) is

$$p_{i,i+1} = \frac{\lambda_i}{\lambda_i + \mu_i}$$

Similarly, the probability of a transition from state i to $i - 1$ (i.e., having a death before the next birth) is

$$p_{i,i-1} = \frac{\mu_i}{\lambda_i + \mu_i}$$

Figure 16 shows the state transitions and associated probabilities for a birth and death process.



1
$\frac{\lambda_1}{\lambda_1+\mu_1}$
$\frac{\lambda_{k-2}}{\lambda_{k-2}+\mu_{k-2}}$
$\frac{\lambda_{k-1}}{\lambda_{k-1}+\mu_{k-1}}$
$\frac{\lambda_k}{\lambda_k+\mu_k}$
0
1
...
k-1
k
k+1
...
$\frac{\mu_1}{\lambda_1+\mu_1}$
$\frac{\mu_2}{\lambda_2+\mu_2}$
$\frac{\mu_{k-1}}{\lambda_{k-1}+\mu_{k-1}}$
$\frac{\mu_k}{\lambda_k+\mu_k}$
$\frac{\mu_{k+1}}{\lambda_{k+1}+\mu_{k+1}}$

Figure 16. State transition diagram for birth and death process

The above state diagram is not particularly interesting for a pure birth or a pure death process, since in the former, the probabilities are all 1 going to the right and in the latter, the probabilities are all 1 going to the left.

. . .

In Theorems 1 and 2 of the paper entitled “The Classification of Birth and Death Processes” [52], the author proves that a birth and death process is

- recurrent if and only if $\sum_{i=1}^{\infty}\prod_{n=1}^{i}\frac{\mu_n}{\lambda_n} = \infty$
- positive recurrent if and only if $\sum_{i=1}^{\infty}\prod_{n=1}^{i}\frac{\mu_n}{\lambda_n} = \infty$ and $\sum_{i=1}^{\infty}\prod_{n=1}^{i}\frac{\lambda_{n-1}}{\mu_n} < \infty$
- null recurrent if and only if $\sum_{i=1}^{\infty}\prod_{n=1}^{i}\frac{\mu_n}{\lambda_n} = \infty$ and $\sum_{i=1}^{\infty}\prod_{n=1}^{i}\frac{\lambda_{n-1}}{\mu_n} = \infty$.

For example, consider the birth and death process with rates

$$\lambda_0 = 1$$
$$\lambda_n = \alpha n, \qquad n \geq 1$$
$$\mu_0 = 0$$
$$\mu_n = \alpha n, \qquad n \geq 1$$

We have that

$$\sum_{i=1}^{\infty}\prod_{n=1}^{i}\frac{\mu_n}{\lambda_n} = \sum_{i=1}^{\infty}\prod_{n=1}^{i} 1 = \sum_{i=1}^{\infty} 1 = \infty$$

and

$$\sum_{i=1}^{\infty}\prod_{n=1}^{i}\frac{\lambda_{n-1}}{\mu_n} = \sum_{i=1}^{\infty}\frac{1}{\alpha}\cdot\frac{1}{2}\cdot\frac{2}{3}\dots\cdot\frac{i-1}{i} = \frac{1}{\alpha}\sum_{i=1}^{\infty}\frac{1}{i} = \infty$$

Thus, the process is null recurrent. This makes sense intuitively since the birth and death rates are the same.

As a second example, consider a slight modification of the previous example. With the modification, we get a process that is positive recurrent.

$$\lambda_0 = 1$$
$$\lambda_n = \alpha n, \qquad n \geq 1$$
$$\mu_0 = 0$$
$$\mu_n = \alpha(\boldsymbol{n} + \mathbf{1}), \qquad n \geq 1$$

Applying the classifications tests, we get

$$\sum_{i=1}^{\infty} \prod_{n=1}^{i} \frac{\mu_n}{\lambda_\mathrm{n}} = \sum_{i=1}^{\infty} \prod_{n=1}^{i} \frac{2\alpha}{\alpha} \cdot \frac{3\alpha}{2\alpha} \cdot \ldots \cdot \frac{(i+1)\alpha}{i\alpha} = \sum_{i=1}^{\infty} (i+1) = \infty$$

and

$$\sum_{i=1}^{\infty} \prod_{n=1}^{i} \frac{\lambda_{n-1}}{\mu_n} = \sum_{i=1}^{\infty} \frac{1}{2\alpha} \cdot \frac{\alpha}{3\alpha} \cdot \frac{2\alpha}{4\alpha} \ldots \cdot \frac{(i-1)\alpha}{(i+1)\alpha} = \frac{1}{\alpha} \sum_{i=1}^{\infty} \frac{1}{i(i+1)} = \frac{1}{\alpha} \sum_{i=1}^{\infty} \left[\frac{1}{i} - \frac{1}{i+1} \right] = \frac{1}{\alpha} \cdot 1 < \infty$$

So, the process is positive recurrent. This makes sense intuitively since the death rate is greater than the birth rate.

. . .

In the case when a birth and death process is positive recurrent, the stationary probabilities [53] are given by

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \prod_{n=1}^{i} \frac{\lambda_{n-1}}{\mu_n}}$$

$$\pi_k = \pi_0 \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i}, \;\; k = 1,2,3, \ldots$$

where π_k is the long-term proportion of time the process is in state k.

For the last example above, we have

$$\pi_0 = \frac{1}{1 + \frac{1}{\alpha}} = \frac{\alpha}{\alpha + 1}$$

$$\pi_k = \frac{\alpha}{\alpha + 1} \cdot \frac{1}{\alpha} \cdot \frac{1}{k(k+1)} = \frac{1}{\alpha + 1} \cdot \frac{1}{k(k+1)}$$

For $\alpha = 10$, $\pi_0 = 10/11$ and $\pi_k = \frac{1}{11} \cdot \frac{1}{k(k+1)}, k = 1,2,3, \ldots$

So, as one might expect, the largest proportion of time is spent in state 0, with diminishing amounts of time in state k as k increases.

8.3.5 Queues

The Investopedia definition of **queueing theory** [54]:

> Queuing theory is a branch of mathematics that studies how lines form, how they function, and why they malfunction. Queuing theory examines every component of waiting in line, including the arrival process, service process, number of servers, number of system places, and the number of customers—which might be people, data packets, cars, or anything else.
>
> Real-life applications of queuing theory cover a wide range of businesses. Its findings may be used to provide faster customer service, increase traffic flow, improve order shipments from a warehouse, or design data networks and call centers.

In this section, we cover just two examples from the expansive field of queueing theory. For further study see:

- The Wikipedia article "Queueing theory" [55] provides an overview of the topic.
- Kleinrock's two volume set entitled "Queueing System" is the definitive introduction to the topic [56] [57].

. . .

Consider a situation where arrivals (usually called "customers" in queueing theory jargon) follow a Poisson process with rate λ. The customers wish to have some task performed by a server (e.g., a bank teller). For this example, there is but one server who completes each task according to a Poisson process with rate μ. If the server is busy with an existing customer, an arrival will wait in line. If the server is not busy, the customer will go directly to the server. The state of the process is the number of customers waiting in line plus the customer being served.

In terms of the birth and death process variables, we have

$$\lambda_n = \lambda, \qquad n \geq 1$$

$$\mu_0 = 0$$

$$\mu_n = \mu, \qquad n \geq 1$$

As described, this particular birth and death process is known as an **M/M/1 queueing system** [58]. The first M refers to the arrival process which satisfies the **M**arkov property, and the second M refers to the process of the server which also satisfies the **M**arkov property. The "1" is an indication that we are dealing with a single server.

. . .

If the above example is modified to have c servers, then we get what is called an **M/M/c queueing system** [59]. The arrival rate is λ. If a server is available, a new arrival selects one of the open servers. If all servers are busy, a new arrival comes into the queue. There is but one queue. Each server follows a Poisson process with rate μ. Collectively, the collection of servers is also a Poisson process with rate $n\mu$ when there are fewer customers than servers, i.e., when the number of customers in the system n is less than c. When all the servers are busy, the collective rate is $c\mu$.

In terms of the birth and death process variables, we have

$$\lambda_n = \lambda, \qquad n \geq 1$$

$$\mu_0 = 0$$

$$\mu_n = n\mu, \qquad 1 \le n < c$$
$$\mu_n = c\mu, \qquad n \ge c$$
$$\cdots$$

In general, something known as **Kendall's Notation** [60] is used to describe queues. The format for Kendall's notation is $A/S/c/K/N/D$ where

- A is the distribution of the interarrival times, e.g., M for exponential (Markovian)
- S is the distribution of the service times, e.g., M for exponential (Markovian)
- c is the number of servers
- K is the buffer size (arrivals are turned away when the buffer is full)
- N is the size of the population from which customers are drawn to the queue
- D is the queue discipline, e.g., First In First Out (FIFO), Service In Random Order (SIRO)

For the M/M/1 and M/M/c types of queues, $K = N = \infty$ and $D = FIFO$. So, using Kendall's notation, we would write M/M/1/∞/∞/FIFO and M/M/s/∞/∞/FIFO, respectively.

8.3.6 General Theory

Now that we have studied several classes of continuous-time Markov chains, we return to the general theory pertaining to all such processes.

Let's start by returning to the example at the end of Section 8.3.1, i.e., continuous-time Markov chain with transition matrix

$$\begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/6 & 0 & 5/6 \\ 1/3 & 2/3 & 0 \end{bmatrix}$$

and associated exponential holding times with rates 6,18 and 12, respectively. The state space is $\{0,1,2\}$.

In general, a continuous-time Markov chain can be characterized by its state space, transition probabilities, and the rates of the exponentially distributed holding times for each state.

An alternate and more concise representation of a continuous-time Markov chain can be achieved with something called a **Q-matrix**. If I is a countable (possibly finite) set, we define the Q-matrix on I as the matrix $Q = (q_{ij}: i, j \in I\}$ such that

- $0 \le -q_{ii} < \infty$ for all i
- $q_{ij} > 0$ for all $i \ne j$
- $\sum_{j \in I} q_{ij} = 0$ for all i

The last condition is a convention with no specific interpretation, but is helpful with various computations.

For our example, the Q-matrix for the above example is

$$\begin{bmatrix} -6 & 3 & 3 \\ 3 & -18 & 15 \\ 4 & 8 & -12 \end{bmatrix}$$

The off-diagonal entries can be viewed as the rate of going from one state to another. For example, the entry 8 is the rate of going from state 2 to state 1.

One can work backwards from the Q-matrix to get the transition probabilities and the holding time rates for each state. For example, consider the third row in the above Q-matrix. To get the transition probabilities, we divide each off-diagonal entry by the sum of the off-diagonal entries for the row. So, the transition probability from state 2 to 0 is $\frac{4}{4+8} = \frac{4}{12} = \frac{1}{3}$ and the probability of going from state 2 to 1 is $\frac{8}{4+8} = \frac{8}{12} = \frac{2}{3}$. Once we've computed the transition probabilities for a row, we can get the rate associated with the holding time for the state corresponding to the row by dividing any non-diagonal entry in the row by its associated transition probability. For example, the holding time rate for state 2 is $\frac{4}{\left(\frac{1}{3}\right)} = 12$ (using the first element in the row). Using the second element in the row gives the same answer, i.e., $\frac{8}{\left(\frac{2}{3}\right)} = 12$.

In general, a continuous-time Markov chain can be characterized by

- a countable (possibly finite) state space S
- a square Q-matrix with number of rows (columns) equal to the size of S
- an initial state given by $X_0 = n$, or a probability distribution for the initial state.

If the probability of going from state i to j given by p_{ij} and the rate of the exponentially distributed holding time for state i is given by λ_i, then the following relationships with the Q-matrix can be stated:

$$\lambda_i = \sum_{j \neq i} q_{ij} = -q_{ii}, \qquad i \in I$$

$$p_{ij} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}} = -\frac{q_{ij}}{q_{ii}}, \qquad i \neq j$$

Knowing the Q-matrix allows one to compute the transition probabilities and holding time rates, and vice versa. The two representations are equivalent. In what follows, we will state several results that rely on the Q-matrix representation.

. . .

Before continuing with our study of the Q-matrix and continuous-time Markov chains, we need to introduce two preliminary results.

The exponent of a matrix A is defined in terms of the power series for e^x. For square matrix A (could even be an infinitely countable number of row and columns), the following definition is stated:

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!}$$

It is remarkable (at least to the author) that the above series converges for any matrix $A = (a_{ij} : i, j \in I)$ where I is a countable (possible infinite) set.

For example, take the matrix

$$A = \begin{bmatrix} 1.2 & 2 & 3 \\ 4 & 5.3 & 6 \\ 7 & 8 & -2.5 \end{bmatrix}$$

Using the online calculator at https://www.symbolab.com/solver/matrix-exponential-calculator, we get the following approximation for e^A

$$\begin{bmatrix} 19938.40869 & 25465.10894 & 15076.23394 \\ 48003.45466 & 61312.85801 & 36298.67065 \\ 37129.05468 & 47422.63980 & 28075.49086 \end{bmatrix}$$

Further, if matrices $A = (a_{ij} : i, j \in I)$ and $B = (b_{ij} : i, j \in I)$ commute (i.e., $AB = BA$), then

$$e^{A+B} = e^A e^B$$

which implies

$$e^{nA} = (e^A)^n, \qquad n = 0,1,2, \ldots$$

These results are proven in Section 2.10 of the book by Norris [61].

. . .

An n x n real-valued matrix A is said to be **diagonalizable** (or non-defective) if there exists an matrix P such that $A = P^{-1}DP$ where D is a diagonal matrix, i.e., a matrix with all zero entries except on the main diagonal.

The condition for when a square real-valued matrix is diagonalizable depends on the concepts of eigenvalue, eigenvector and eigenspace from linear algebra.

A real number λ is an **eigenvalue** of a square matrix A is there exists a column vector $\boldsymbol{v}$ such that

$$A\boldsymbol{v} = \lambda\boldsymbol{v}$$

A vector corresponding to an eigenvalue is called (believe it or not) an **eigenvector**.

The set of all eigenvectors corresponding to an eigenvalue, together with the zero vector, is called the **eigenspace** corresponding to the eigenvalue. Equivalently, an eigenspace is the space generated by the eigenvectors corresponding to the same eigenvalue, i.e., the space of all vectors that can be written as a linear combination of those eigenvectors.

A basic result from linear algebra is that an n x n real matrix A is diagonalizable if and only if the sum of the dimensions of its eigenspaces is equal to n which is true if and only if there exists a basis of $\mathbb{R}^n$ consisting of eigenvectors of A. If such a basis exists, then the matrix B, whose columns consist of the eigenvectors of A, satisfies the equation $A = B^{-1}DB$. The diagonal elements of D are the eigenvalues of A.

If a matrix A can be diagonalized, it is easy to compute higher powers of A since

$$A^n = (B^{-1}DB)(B^{-1}DB)\dots(B^{-1}DB) = B^{-1}D^nB$$

Computing D^n is just a matter of raising the diagonal entries to the power n and noting all the other entries are zero.

For example, consider the Q-matrix from our earlier example, i.e.,

$$Q = \begin{bmatrix} -6 & 3 & 3 \\ 3 & -18 & 15 \\ 4 & 8 & -12 \end{bmatrix}$$

Using the diagonalize capability at wolframalpha.com, we get

$$D = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -18-\sqrt{69} & 0 \\ 0 & 0 & -18+\sqrt{69} \end{bmatrix}$$

$$B^{-1} = \begin{bmatrix} 1 & \frac{1}{44}(5\sqrt{69}-36) & \frac{1}{44}(-36-5\sqrt{69}) \\ 1 & \frac{1}{44}(-15-8\sqrt{69}) & \frac{1}{44}(8\sqrt{69}-15) \\ 1 & 1 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} \frac{32}{85} & \frac{4}{17} & \frac{33}{85} \\ \frac{2(59\sqrt{69}-552)}{5865} & -\frac{2(69+16\sqrt{69})}{1173} & \frac{2(299+7\sqrt{69})}{1955} \\ -\frac{2(59\sqrt{69}+552)}{5865} & \frac{32}{17\sqrt{69}}-\frac{2}{17} & \frac{2(299-7\sqrt{69})}{1955} \end{bmatrix}$$

[**Author's Remark**: Agreed that this is a bit ugly. However, Wolfram Alpha can also provide approximate solutions in terms of decimal representations. The point is that we can raise A to fairly high powers with minimal computational demands. For example, raising A to power 100 (without the diagonalization technique) will take a lot of computational effort. In fact, Wolfram Alpha gives a computation limit error if you try request the value of A^{100}. However, using the diagonalization of A, we only need raise the diagonal elements of D to the 100th power, and then compute $P^{-1}D^{100}P$ using scientific notation given the large values along the diagonal in D^{100}.]

. . .

Back to our discussion of Q-matrices and associated topics.

The following theorem holds for any square (and finite) matrix (does not need to meet the conditions of a Q-matrix).

Theorem 33. If Q be a square matrix of finite dimension and let $P(t) = e^{tQ},\ t \geq 0$, then the following statements hold true:

i. $P(s+t) = P(s)P(t)$ for all real numbers $s, t \geq 0$ (this is known as the **semigroup property**).

ii. $P(t)$ is the unique solution to the forward equation $\frac{d}{dt}P(t) = P(t)Q,\ P(0) = I$, where I is the identity matrix.

iii. $P(t)$ is the unique solution to the backward equation $\frac{d}{dt}P(t) = QP(t),\ P(0) = I$

iv. for $n = 0{,}1{,}2, \ldots$, $P^{(k)}(0) = Q^k$ where $P^{(k)}(0)$ is the kth derivative of $P(t)$ evaluated at 0.

Proof: From our previous discussion, we know that e^{tQ} is well-defined, i.e., converges.

Also, since sQ and tQ commute, we have from our previous discussion that $e^{(s+t)Q} = e^{sQ+tQ} = e^{sQ}e^{tQ}$.

The remaining details of the proof can be found in the book by Norris [61] (see Theorem 2.1.1). ■

In the case of a Q-matrix, we have the following theorem.

Theorem 34. A square matrix Q of finite dimension is a Q-matrix if and only if $P(t) = e^{tQ}$ is a stochastic matrix (i.e., all entries are non-negative and each row adds to 1) for all $t \geq 0$.

For the details of the proof, see Theorem 2.1.2 in the book by Norris [61].

The key result here is the following:

Theorem 35. The transition probabilities for a continuous-time Markov process $\{X(t): t \geq 0\}$ with Q-matrix Q are given by

$$P(X(t) = j \mid X(0) = i) = p_{i,j}(t)$$

where $p_{i,j}(t)$ is entry (i,j) of e^{tQ}.

Proof: See Section 2.8 in the book by Norris [61].

The matrix $P(t)$ with entries $p_{i,j}(t)$ is called the **embedded Markov chain** (sometime jump matrix or jump process) associated with the continuous-time Markov process with Q-matrix Q. The row matrix π is said to be a stationary (or invariant) distribution of the embedded Markov chain if $\pi P(t) = \pi$ for every value of $t \geq 0$.

Theorem 36. The row vector π is a stationary distribution of the embedded Markov chain of a continuous-time Markov process with Q-matrix Q if and only if $\pi Q = 0$.

By definition, a continuous-time Markov chain is said to be irreducible if for any two states i and j it is possible to travel from i to j in a finite number of steps, and vice versa.

Theorem 37. If a continuous-time Markov chain $X(t)$ is irreducible and has a stationary distribution π, then

$$\lim_{t\to\infty} p_{i,j}(t) = \pi_j$$

The proofs for the previous two theorems are omitted. The interested reader can find the details in the Section 4.3 of the book by Durrett [62].

As an example, consider the continuous-time Markov chained characterized by the Q-matrix that we discussed earlier, i.e.,

$$Q = \begin{bmatrix} -6 & 3 & 3 \\ 3 & -18 & 15 \\ 4 & 8 & -12 \end{bmatrix}$$

We have that $Q = B^{-1}DB$ and $Q^n = B^{-1}D^nB$, with the previously computed results for B^{-1}, D and B. We seek the matrix $P(t)$. From Theorem 35,

$$P(t) = e^{tQ} = \sum_{n=0}^{\infty} \frac{(tQ)^n}{n!} = B^{-1}\left[\sum_{n=0}^{\infty} \frac{(tD)^n}{n!}\right]B$$

$$= B^{-1}\left[\sum_{n=0}^{\infty} \frac{1}{n!}\begin{pmatrix} 0^n & 0 & 0 \\ 0 & \alpha t^n & 0 \\ 0 & 0 & \beta t^n \end{pmatrix}\right]B$$

$$= B^{-1}\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{\alpha t} & 0 \\ 0 & 0 & e^{\beta t} \end{pmatrix}B$$

where $\alpha = -18 - \sqrt{69}$ and $\beta = -18 - \sqrt{69}$.

For example, doing the appropriate matrix multiplications, we get

$$p_{0,0}(t) = x + r \cdot y \cdot e^{\alpha t} + s \cdot z \cdot e^{\beta t}$$

where we have made the following substitutions for the first row of B^{-1} and the first column of B:

$$r = \frac{1}{44}\left(5\sqrt{69} - 36\right)$$

$$s = \frac{1}{44}\left(-36 - 5\sqrt{69}\right)$$

$$x = \frac{32}{85}$$

$$y = \frac{2\left(59\sqrt{69} - 552\right)}{5865}$$

$$z = -\frac{2\left(59\sqrt{69} + 552\right)}{5865}$$

As $t \to \infty$, $p_{0,0}(t) = \frac{32}{82}$. This can be interpreted as the long-term proportion of time that the process is in state 0.

Further, solving the simultaneous equations $\pi Q = 0$ and $\pi_0 + \pi_1 + \pi_2 = 1$, we get that

$$\pi = \left(\frac{32}{85}, \frac{20}{85}, \frac{33}{85}\right)$$

As expected (and as a check), we see that $\pi_0 = \lim_{t\to\infty} p_{0,0}(t) = \frac{32}{82}$.

From Theorem 37, we know that the stationary distribution does not depend on the initial state. So, we also know that $\lim_{t\to\infty} p_{1,0}(t) = \lim_{t\to\infty} p_{2,0}(t) = \frac{32}{82}$. We can make similar statements for the stationary distributions related to states 1 and 2.

9 Game Theory

9.1 Definitions and Examples

Prerequisites: algebra, basic understanding of graph theory terminology, mathematical induction

Game theory entails the applications of mathematics to the study of strategic interactions between two or more rational agents in situations with clearly defined rules and outcomes. A **rational agent** is a decision making entity (e.g., a person, company or software) that aims to achieve optimal results based on given premises and information.

Some additional definitions of "game theory":

- From The Free Dictionary by Farlex: *A mathematical method of decision-making in which a competitive situation is analyzed to determine the optimal course of action for an interested party, often used in political, economic, and military planning. Also called "theory of games."*
- From Britannica online: *branch of applied mathematics that provides tools for analyzing situations in which parties, called players, make decisions that are interdependent. This interdependence causes each player to consider the other player's possible decisions, or strategies, in formulating strategy. A solution to a game describes the optimal decisions of the players, who may have similar, opposed, or mixed interests, and the outcomes that may result from these decisions.*
- From Merriam-Webster online: *the analysis of a situation involving conflicting interests (as in business or military strategy) in terms of gains and losses among opposing players.*
- From Investopedia: *Game theory is a theoretical framework for conceiving social situations among competing players. In some respects, game theory is the science of strategy, or at least the optimal decision-making of independent and competing actors in a strategic setting.*
- From the Stanford Encyclopedia of Philosophy: *Game theory is the study of the ways in which interacting choices of economic agents produce outcomes with respect to the preferences (or utilities) of those agents, where the outcomes in question might have been intended by none of the agents.*

. . .

In terms of references, the book *Game Theory 101: The Complete Textbook* [63] and the associated videos [64] provide a good introduction to the topic using a wealth of examples. For a more theoretical introduction but still basic in terms of the mathematics involved, see the online videos and associated lectures from the MIT course Economic Applications of Game Theory [65].

. . .

The **prisoner's dilemma** is a common example used in many game theory textbooks and articles. The situation goes as follows:

Two individuals (call them Mr. A and Ms. B) are accused of a serious crime, arrested and placed in prison while awaiting trial. Before initial questioning, the prisoners have no means of communicating with the other. The prosecutors lack sufficient evidence to convict the accused of a major charge, but they have enough to convict both on a lesser charge. Simultaneously, the prosecutors offer each prisoner a deal to basically betray the other and in return, get a lesser sentence. The details of the possible outcomes, which are made known to A and B, are as follows:

- If A and B betray each other, they each get sentenced to two years in prison (for the major charge but at a reduced sentence since they cooperated with the prosecution team).
- If one betrays and the other remains silent, the betrayer will be set free and the other prisoner will serve three years in prison (for the major charge).
- If both remain silent, each will serve one year in prison (based on the lesser charge).

This game can be represented in matrix form (known as a **payoff matrix**), as shown in Table 11. Sentence time (the payoff) is represented with a negative number so that more desirable payoffs have higher associated values. The convention is to put Mr. A's payoff in the first element of each ordered pair, and place Ms. B's payoff in the second element. S1 and B1 are the possible strategies for Mr. A, and S2 and B2 are the possible strategies for Ms. B.

Table 11. Prisoner's dilemma – Version 1

	S2: B remains silent	B2: B betrays A
S1: A remains silent	$(-1,-1)$	$(-3,0)$
B1: A betrays B	$(0,-3)$	$(-2,-2)$

Table 11 is an example of what is called the **normal form of a game**. For two-player games, the normal form of game is typically represented as a matrix. For games with more than two players, the matrix representation is not used. In such cases, a game is represented by a function that maps each possible combination of strategies with an associated payoff for each player.

An equivalent approach is to add 4 to each outcome. This has the effect of rating outcomes on a scale of 1 to 4, with 4 being the most desirable outcome, see Table 12.

Table 12. Prisoner's dilemma – Version 2

	S2: B remains silent	B2: B betrays A
S1: A remains silent	$(3,3)$	$(1,4)$
B1: A betrays B	$(4,1)$	$(2,2)$

. . .

In **simultaneous games** all players move at the same time, or equivalent the later players are unaware of the earlier players' actions (making them effectively simultaneous). In **sequential games** later players have some (or all) knowledge about the actions of players who have acted on their strategies earlier. The prisoner's dilemma is a simultaneous game. Go, chess, checkers and tic-tac-toe are sequential games.

As an example of a sequential game, consider the situation where Aristotle wants to sell his laptop computer to Bernice. Assume that he is considering three possible asking prices, i.e., $1000, $2000 or $3000. Bernice can either accept or reject the offer from Aristotle. Aristotle makes his offer first and then Bernice decides. Bernice values Aristotle's computer at $2500. So, for example, if Aristotle offers the computer at $3000, Bernice would perceive a net loss of $500. The game is depicted in

Figure 17. The payoffs are listed at each leaf of the tree, with the payoff for Aristotle coming first. This type of representation is known as the **extensive form of a game**.

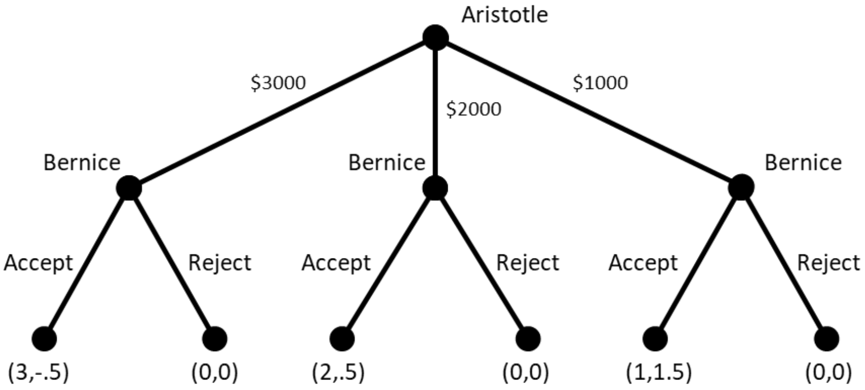


Figure 17. Laptop sale

In the laptop sale game, it is critical to note this approach assumes monetary returns as the only motivation for the players, but this does not need to be the case. It could be that Aristotle and Bernice are good friends. Aristotle wants to make some money on the exchange but does want to overcharge Bernice. Bernice wants a fair price but doesn't want to shortchange Aristotle. In this case, we can use the relative preferences of the two players for the payoffs rather than money. The three options for the sale price are the same but now, Aristotle favors $2000 as the best option, $1000 as the second best option and $3000 as the last option. The preferences for Aristotle are listed in Figure 18, with the favored option having a payoff of 3, the second option with a payoff of 2 and the third option with a payoff of 1. Also wanting a fair deal (for both parties in the deal), Bernice has the same preferences as Aristotle, i.e., $2000 as the best option, $1000 as the second best option and $3000 as the last option. (The dollar amounts are left in the figure just to remind the reader of what each branch represents from a monetary perspective, but it is the preferences (1,2 and 3) that are used in the decision making.)

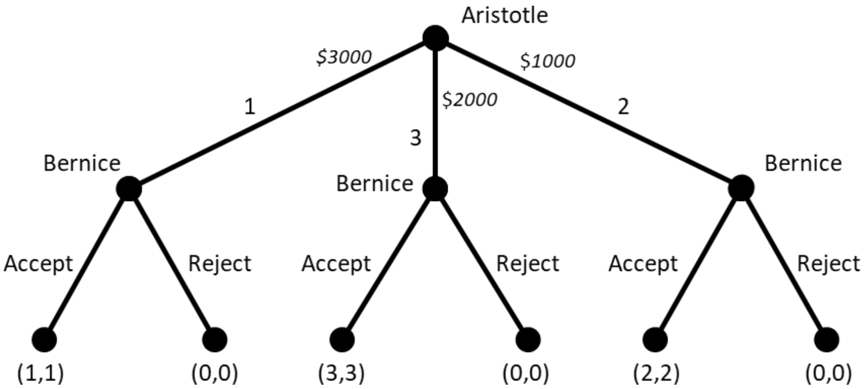


Figure 18. Laptop sale – alternate payoffs for Aristotle

. . .

In a **symmetric game**, the payoffs for playing a particular strategy depend only on the other strategies employed, and not on who is playing them. In other words, if the identities of the players can be changed without changing the payoffs to the strategies, then a game is symmetric. The prisoner's dilemma is a symmetric game. If for some reason, the deal offered to Mr. A was 6 months (.5 years) for betrayal of Ms. B (while Ms. B remained silent), then the game would be asymmetric and have the payoff matrix shown in Table 13.

Table 13. Prisoner's dilemma – Asymmetric modification

	S2: B remains silent	B2: B betrays A
S1: A remains silent	$(-1,-1)$	$(-3,0)$
B1: A betrays B	$(.5,-3)$	$(-2,-2)$

. . .

Pitching pennies is another game that is commonly used as an example in game theory. In this game, there are two players (A and B) who simultaneously pitch pennies. If both tosses result in heads or both tosses result in tails, A wins one point and B loses one point. If the coins do not match (one heads and the other tails), then A loses one point and B wins one point. The payoff matrix is shown in Table 14. The possible "strategies" for player A are s_{11} and s_{12}. The possible "strategies" for player B are s_{21} and s_{22}. The term "strategies" is in quotes to emphasize that the outcomes for each player are random events and not the result of active decisions on the part of the players.

Table 14. Pitching pennies – Zero-Sum Version

	s_{21}: Heads for B	s_{22}: Tails for B
s_{11}: Heads for A	$(1,-1)$	$(-1,1)$
s_{12}: Tails for A	$(-1,1)$	$(1,-1)$

If the sum of the payoffs for each combination of strategies is equal to zero, the game is said to be a **zero-sum game**. Pitching pennies is an example of a zero-sum game. Prisoner's dilemma is an example of a non-zero-sum game.

If the sum of the payoffs for each combination of strategies is equal to the same number, the game is said to be a **constant-sum game**. If we modify the payoff matrix for the pitching pennies game by adding 2 to each payoff (as shown in Table 15), we get a game with a constant payoff of 4 of each combination of strategies. Clearly, the set of zero-sum games is a subset of the set of constant-sum games. Further, it is also possible to modify a constant-sum game to get an equivalent zero-sum game, and vice versa.

Table 15. Pitching pennies – Zero-Sum Version

	s_{21}: Heads for B	s_{22}: Tails for B
s_{11}: Heads for A	$(3,1)$	$(1,3)$
s_{12}: Tails for A	$(1,3)$	$(3,1)$

. . .

Some games are intended to be played multiple times. For example, the children's game of rock-scissors-paper (also known by other orderings of the three items) is designed to be a repetitive game. In this game, there are two players (A and B). At each round of the game, each player selects rock, scissor or paper. The payoffs are as follows:

- Rock beats (dulls) scissors. The player who selected rock gets 1 point and the player who chose scissors gets -1 point.
- Scissors beat (cuts) paper. The player who selected scissors gets 1 point and the player who chose paper gets -1 point.
- Paper beats (wraps) rock. The player who selected paper gets 1 point and the player who chose rock gets -1 point.
- If both players make the same selection, there is a draw (with each player getting 0 points).

This is a simultaneous, zero-sum, symmetric game. The payoff matrix is shown in Table 16. In each of the payoff pairs, player A's payoff comes first.

Table 16. Payoff matrix for rock-scissors-paper game

	s_{21}: Rock for B	s_{22}: Scissors for B	s_{23}: Paper for B
s_{11}: Rock for A	$(0,0)$	$(1,-1)$	$(-1,1)$
s_{12}: Scissors for A	$(-1,1)$	$(0,0)$	$(1,-1)$
s_{13}: Paper for A	$(1,-1)$	$(-1,1)$	$(0,0)$

For a repeated game such as rock-scissors-paper, it would be unwise to continually play the same strategy since your opponent would catch-on and beat you every time. Deterministically playing a given strategy is known as a **pure strategy**. When a player selects among several strategies (based on a probability distribution), this is known as a **mixed strategy**. For the game at hand, each player may want to randomly select each strategy one third of the time. This could be done by rolling a die, where the player selects rock if 1 or 2 is rolled, scissors if 3 or 4 is rolled, and paper if 5 or 6 is rolled.

By design, the players in the pitching pennies game are using a mixed strategy, with the flip of the penny effectively randomizing the selection of a "strategy".

9.2 Normal Form of a Game

Games can be represented more formally using something known as the **normal form**. The normal form of a game includes the following information:

- A finite set of players $N = \{1,2, \dots n\}$
- A finite set of pure strategies for each player. The possible pure strategies for player i are represented as $S_i = \{s_{i1}, s_{i2}, \dots, s_{in_i}\}$. The first subscript represents the player and the second subscript is an index for the strategy, e.g., s_{14} is the 4th strategy for player 1. Also, n_i is the number of pure strategies for player i. The players do not necessarily have the same number of strategies.

- Once the set of possible pure strategies are known for each player, one can define the set of possible strategy combinations (known as **strategy profiles**) which is $S = S_1 \times S_2 \times \ldots \times S_n$, i.e., the cross-product of the S_i. In general, the cross product of n finite sets is a collection of n-tuples where the first term can be any element from the first set, the second term can be any element from the second set, and so on. For example, in the rock-scissor-paper example, (s_{11}, s_{23}) is a strategy profile, i.e., a formal way of indicating player A selects rock and player B selects paper.
    - While the double subscripting notation is more expressive and accurate, it is difficult to represent a general element of S using double subscripting. For example, $(s_{ia}, s_{ib}, \ldots, s_{iz})$ gives the impression that there are 26 players. We could go with the following $(s_{ia_1}, s_{ia_2}, \ldots, s_{ia_n})$ where s_{ia_i} is an unspecified (arbitrary) strategy for player i. An alternate approach (which we will use in some situations) is $s = (s_1, s_2, \ldots, s_n)$ where s_i is an unspecified (arbitrary) strategy for player i. The downside of this approach occurs when several strategy profiles need to be represented. In such cases, some sort of marking is typically used, e.g., strategy profiles $s = (s_1, s_2, \ldots, s_n)$ and $s^* = (s_1^*, s_2^*, \ldots, s_n^*)$.
- For each player i, there is a utility (or payoff) function u_i that maps from each possible strategy profile to an associated payoff for player i. For example, in the rock-scissor-paper example, $u_1(s_{11}, s_{23}) = -1$ and $u_2(s_{11}, s_{23}) = 1$.

. . .

A strategy s_{ix} for player i is said to **strictly dominate** a strategy s_{iy} (also for player i) if no matter what the other players do, playing s_{ix} is strictly better than playing s_{iy} for player i. The term “better than” means a strictly greater payoff. In terms of notation, strategy s_{ix} strictly dominates s_{iy} if and only if

$$u_i(s_{ix}, s_{\neg i}) > u_i(s_{iy}, s_{\neg i}) \text{ for every } s_{\neg i} \in S_{\neg i} = S_1 \times S_2 \times \ldots \times S_{i-1} \times S_{i+1} \times \ldots \times S_n$$

$S_{\neg i}$ is a notation that represents all strategy profiles for a game minus the strategy profiles for player i and $s_{\neg i} \in S_{\neg i}$.

For example, in the various versions of the prisoner’s dilemma, B1 strictly dominates S1 for Mr. A, and B2 strictly dominates S2 for Ms. B.

If we change the condition to $u_i(s_{ix}, s_{\neg i}) \geq u_i(s_{iy}, s_{\neg i})$, then s_{ix} is said to **weakly dominate** s_{iy}.

For example, in Table 17, s_{11} weakly dominates s_{12} for Player 1.

Table 17. Example of weak dominance

		Player 2		
		s_{21}	s_{22}	s_{23}
Player 1	s_{11}	$(0,0)$	$(1,3)$	$(4,2)$
	s_{12}	$(0,1)$	$(0,4)$	$(1,2)$
	s_{13}	$(1,-3)$	$(4,1)$	$(1,7)$

. . .

A strategy s_{ix} for player i is a **best response** to $s_{\neg i}$ if and only if

$$u_i(s_{ix}, s_{\neg i}) \geq u_i\left(s_{iy}, s_{\neg i}\right) \text{ for every other } s_{iy}.$$

In words, a strategy s_{ix} for player i is a best response to a given combination of strategies from the other players $s_{\neg i}$ if and only if the payoff for s_{ix} is greater than or equal to any other strategy that player i can play to a given combination of strategies from the other players. The "greater than" or "equal to" sign in the condition implies there can be several best responses to a given combination of strategies from the other players in a game. It should be emphasized that the above inequality **only needs to hold for one** $s_{\neg i} \in S_{\neg i}$.

For example, in Table 17, Player 1's best response to a play of s_{22} by Player 2 is s_{13}.

. . .

A strategy s_{ix} is a **strictly dominant strategy** for player i if and only if s_{ix} strictly dominates all the other strategies of player i.

For the game in Table 18, s_{23} provides a better payoff for Player 2 than any other strategy available to Player 2, regardless of the strategy played by Player 1. Thus, s_{23} is a strictly dominant strategy for Player 2.

Table 18. Examples of a dominant strategies

		Player 2		
		s_{21}	s_{22}	s_{23}
Player 1	s_{11}	$(0,0)$	$(1,3)$	$(4,5)$
	s_{12}	$(0,1)$	$(0,4)$	$(1,5)$
	s_{13}	$(1,-3)$	$(4,1)$	$(4,7)$

A strategy s_{ix} is a **weakly dominant strategy** for player i if and only if s_{ix} weakly dominates all the other strategies of player i. Clearly, a strictly dominant strategy is also a weakly dominant strategy.

For the game in Table 18, s_{13} is a weakly dominant strategy for Player 1, noting that Player 1 has best responses of s_{11} or s_{13} when Player 2 goes with s_{23}.

9.3 Solutions for Simultaneous Games

9.3.1 Dominant Strategy Equilibrium

A strategy profile $s = (s_1, s_2, \dots, s_n)$ is a **dominant strategy equilibrium**, if and only if for each player i, s_i is a weakly dominant strategy.

For the game in Table 18, (s_{13}, s_{23}) is a dominant strategy equilibrium since s_{13} is a weakly dominant strategy for Player 1 and s_{23} is a weakly (also strictly) dominant strategy for Player 2.

For the rock-scissor-paper game in Table 16, a dominant strategy equilibrium does not exist.

In the prisoner's dilemma game (Table 19), B1 is a weakly (also strictly) dominant strategy for Mr. A, and B2 is a weakly (also strictly) dominant strategy for Ms. B, and so, (B1, B2) is a dominant strategy equilibrium. If Mr. A is given knowledge that Ms. B will betray him, he has no incentive to vary from also betraying (i.e., will get a lower payoff of 1 if he decides to remain silent), and vice versa. It is true that (3,3) is a better payoff for both players but this requires trust and involves some risk, e.g., if Mr. A plays S1 but Ms. B plays B2, then Mr. A gets the lowest payoff possible in the game. (From the earlier discussion of this game in Section 9.1, recall that we used S to denote the Silent strategy and B to denote the Betrayal strategy.)

Table 19. Prisoner's dilemma – dominant strategy equilibrium

	S2: B remains silent	B2: B betrays A
S1: A remains silent	(3,3)	(1,4)
B1: A betrays B	(4,1)	(2,2)

In general, rational cautious players will play the dominant strategy equilibrium when it exists.

9.3.2 Strictly Dominated Pure Strategies

For example, in the prisoner's dilemma, Mr. A always gets a higher payout by playing strategy B1 regardless of whether Ms. B plays strategy S2 or B2 (see Table 20, with the payoffs for Mr. A in bold). So, B1 is a dominant pure strategy for Mr. A.

Table 20. Dominant pure strategy for Mr. A in prisoner's dilemma

	S2: B remains silent	B2: B betrays A
S1: A remains silent	(**3**,3)	(**1**,4)
B1: A betrays B	(**4**,1)	(**2**,2)

Using similar logic, B2 is a dominant pure strategy for Ms. B.

Since the two parties cannot coordinate and they have no reason to trust each other, they both betray each other. This results in payoff (2,2). This is a **no regret** solution for both players in the sense if one player is told the decision of the other, that player has no reason to change their decision. For example, if Mr. A is told that Ms. B has decided to be silent (strategy S2), he cannot improve his outcome by changing to S1.

The dilemma is that both prisoners could do better by remaining silent but without coordination and trust, they both make the “no regret” decision. Notice that (3,3) can lead to regrets. For example, if Ms. B remains silent and Mr. A betrays, then her outcome drops from 3 to 1.

Exercise for the reader: find the “no regret” solution(s) for the following game:

		Player 2	
		Left	Right
Player 1	Up	$(3,3)$	$(-25,8)$
	Down	$(7,-100)$	$(0,0)$

9.3.3 Iterated Elimination of Strictly Dominated Strategies (IESDS)

In the prisoner’s dilemma game, each player decided to betray the other without needing to consider the payouts of the other player. In other games, it takes several steps to arrive at a rational (no regret) solution. For example, consider the payoff matrix in Table 21. Player 1 has strategies s_{11}, s_{12} and s_{13}. Player 2 has strategies s_{21}, s_{22} and s_{23}. In each cell of the matrix, Player 1’s payoff is the first element and Player 2’s payoff is the second element.

Table 21. Iterated elimination – Step 1

		Player 2		
		s_{21}	s_{22}	s_{23}
Player 1	s_{11}	(8,**5**)	(7,**1**)	(2,7)
	s_{12}	(7,**4**)	(3,**3**)	(1,2)
	s_{13}	(-100,**1**)	(0,**0**)	(3,5)

As can be seen from the highlighted (bold) numbers in Table 21, strategy s_{21} dominates s_{22} for Player 2, and so, Player 2 has no reason to ever play s_{22}. Since the possible strategies and payoff matrix are known to both players, Player 1 can do the same analysis as Player 2 and realize that Player 2 will not play s_{22}. Thus, Player 1 only needs to consider the payoff matrix shown in Table 22.

Table 22. Iterated elimination – Step 2

	s_{21}	s_{23}
s_{11}	(**8**,5)	(**2**,7)
s_{12}	(**7**,4)	(**1**,2)
s_{13}	(-100,1)	(3,5)

Player 1 analyzes the payoff matrix in Table 22 and determines that strategy s_{11} has a better payoff than strategy s_{12} regardless of how Player 2 plays. Player 2 applies the same logic and determines that Player 1 will never play strategy s_{12}. Thus, Player 2 only needs to consider the payoff matrix shown in Table 23.

Table 23. Iterated elimination – Step 3

	s_{21}	s_{23}
s_{11}	(8,**5**)	(2,**7**)
s_{13}	(-100,**1**)	(3,**5**)

Continuing the analysis, Player 2 sees that strategy s_{23} strictly dominates s_{21}. Player 1 makes the same deduction, and concludes that Player 2 will only play strategy s_{23}. The game reduces to the situation shown in Table 24. Player 1 picks the dominant strategy s_{13} based on the deduction that Player 2 will select strategy s_{23}.

Table 24. Iterated elimination – Step 4

	s_{23}
s_{11}	(**2**,7)
s_{13}	(**3**,5)

The payoff (3,5) represents a “no regret” outcome for both players. If Player 1 is told that Player 2 has selected strategy s_{23}, no better outcome can be attained by changing from the selection of strategy s_{13}. Similarly, if Player 2 is told that Player 1 has selected strategy s_{13}, no better outcome can be attained by changing from the selection of strategy s_{23}.

This is a rational (no regret) approach on the part of both players, but it is not the only approach. For example, Player 1 may be terrified by the possibility of getting a -100 payoff if strategy s_{13} is played and Player 2 decides to act irrationally (with respect to the above analysis) and selects strategy s_{21}. In this scenario, Player 1 may decide to go with strategy s_{11} or s_{12}.

Exercise for the reader: Use iterated elimination of strictly dominated pure strategies to the “no regret” strategy profile for the following game:

		Player 2				
		F	G	H	I	J
Player 1	A	$(4,-1)$	$(3,0)$	$(-3,1)$	$(-1,4)$	$(-2,0)$
	B	$(-1,1)$	$(2,2)$	$(2,3)$	$(-1,0)$	$(2,7)$
	C	$(2,-2)$	$(-1,-1)$	$(-3,4)$	$(4,-1)$	$(0,2)$
	D	$(1,6)$	$(-3,0)$	$(-2,4)$	$(0,1)$	$(-3,4)$
	E	$(0,0)$	$(1,4)$	$(-3,1)$	$(-2,3)$	$(-1,-1)$

Solution: The order of elimination is D, F, E, G, A, I, C and H, with the “no regret” payoff being (2,7).

. . .

All the games that we have seen so far have the same number of strategies for each player. This condition is not necessary, e.g., see Table 25. The order of elimination is A, Y and C, with the “no regret” payoff being (1,4).

Table 25. Game where players have different number of strategies

		Player 2	
		X	Y
Player 1	A	$(0,1)$	$(-4,2)$
	B	$(1,4)$	$(3,3)$
	C	$(-2,2)$	$(3,-1)$

. . .

IESDS does not necessarily lead to a strategy profile. Consider the game in Table 26. Strategy B strictly dominates C for Player 1. Once strategy C is eliminated, strategy Y is seen to strictly dominate Z.

Table 26. IESDS leading to several rational strategies

		Player 2		
		X	Y	Z
Player 1	A	$(5,3)$	$(1,4)$	$(7,2)$
	B	$(4,3)$	$(5,2)$	$(4,0)$
	C	$(2,7)$	$(0,4)$	$(3,3)$

After eliminating strategies C and then Z, we have the game shown in Table 27. At this point, there are no strictly dominant pure strategies left to remove. Further analysis is possible if mixed strategies are considered (see Section 0).

Table 27. Game with several rational strategies

		Player 2	
		X	Y
Player 1	A	$(5,3)$	$(1,4)$
	B	$(4,3)$	$(5,2)$

. . .

Different orders of eliminating dominant strategies in 2-person games can lead to different reduced games. However, the order of elimination does not matter for strictly dominated strategies, and for elimination of weakly dominated strategies in zero-sum games, see the paper by Gilboa, Kalai and Zemel [66].

The game in Table 28 can be reduced, with varying results, using iterated elimination with a combination of weakly and strictly dominated strategies.

Table 28. Combination of strictly and weakly dominated strategies

		Player 2		
		X	Y	Z
Player 1	A	(1,0)	(2,0)	(3,0)
	B	(1,2)	(2,1)	(2,1)
	C	(1,1)	(0,3)	(4,2)

In one reduction path, we first eliminate strategy B since it is weakly dominated by strategy A.

	X	Y	Z
A	(1,0)	(2,0)	(3,0)
C	(1,1)	(0,3)	(4,2)

Next, we eliminate strategies X and Z since they are weakly dominated by strategy Y. We then see that strategy A strictly dominates C, which leaves us with the strategy profile (A, Y).

	Y
A	$(\mathbf{2}, \mathbf{0})$
C	(0,3)

On the other hand, we could start by eliminating strategy Z since it is weakly dominated by Y.

	X	Y
A	(1,0)	(2,0)
B	(1,2)	(2,1)
C	(1,1)	(0,3)

Next, we eliminate strategy C since it is weakly dominated by both A and B.

	X	Y
A	(1,0)	(2,0)
B	(1,2)	(2,1)

At this point, strategy Y can be eliminated since it is weakly dominated by X.

	X
A	$(1,0)$
B	$(1,2)$

In this reduced form of the game, Player 1 has no preference between strategies A or B. Either way, this gives a solution different from the previous elimination process which resulted in (A,Y).

9.3.4 Nash Equilibrium

A solution via iterated elimination of pure dominated strategies is not always possible. For example, consider the payoff matrix shown in Table 29.

- Players 1 and 2 (both AI applications) can decide to cooperate on a complex task A, with respective payoffs 7 and 5.
- Players 1 and 2 can decide to independently work on different instances of a less complex task B, with respective payoffs 3 and 2.
- If one player works on task A and the other on B, the player working on task A will get a score of -10 since task A requires both applications. However, the player working on task B does have positive payoff, i.e., 4 for Player 1 or 3 for Player 2.

Table 29. Trust Dilemma – Version 1

		Player 2	
		A	B
Player 1	A	$(7,5)$	$(-10,3)$
	B	$(4,-10)$	$(3,2)$

Player 1 cannot determine whether it is always better to play A or B, independent of what Player 2 decides. Similarly, Player 2 cannot determine whether it is always better to play A or B, independent of what Player 1 decides. So, there are no dominated strategies in this game. However, there are still two "no regret" solutions, i.e., (7,5) and (3,2). For example, if Player 1 knows that Player 2 will play strategy A, Player 1 will have "no regret" in also playing since strategy B gives Player 1 a lesser outcome in this case. The strategy profile (A,A) is said to be **payoff dominant** since it provides higher payoffs (for all players) than all other no regret solutions. The strategy profile (B,B) is said to be **risk dominant** since the loss from deviating from (B,B) is greater than deviating from (A,A). To see this, consider the following:

- If Player 1 deviates from (B,B) by playing A, it loses 13 points. If Player 2 deviates from (B,B) by playing A, it loses 12 points.
- On the other hand, Player 1 deviates from (A,A) by playing B, it loses 3 points. If Player 2 deviates from (A,A) by playing B, it loses 2 points.

For this game (even if the two players do not trust each other), strategy A makes most sense since the payoff is higher and the risk is less for both players.

In a variation of the trust dilemma (Table 30), there are two no regret solutions, i.e., strategy profile (A, A) and (B, B). However, in this game, strategy profile (B, B) is both payoff and risk dominate. So, if the players don't trust each other, they may go for the lower risk but lower payoff strategy profile (A, A).

Table 30. Trust Dilemma – Version 2

		Player 2	
		A	B
Player 1	A	$(7,5)$	$(-10,3)$
	B	$(4,-10)$	$(8,6)$

. . .

In general, if each player in a game has chosen a strategy and no player can increase their own expected payoff by changing their strategy while the other players keep theirs unchanged, then the current set of strategy choices constitutes what is called a **Nash equilibrium** (named after the mathematician John Nash). There are no regrets when a Nash equilibrium is played. A successful iterated elimination of strictly dominated pure strategies results in a Nash equilibrium, but as we have seen in the above example, it is possible to have Nash equilibriums even when iterated elimination of strictly dominated pure strategies does not apply.

More precisely, a Nash equilibrium is defined as follows:

A strategy profile $s^* = (s_1^*, s_2^*, \dots, s_n^*)$ is a **Nash equilibrium** if and only if s_i^* is a best response to $s_{\neg i}^* = (s_1^*, s_2^*, \dots, \mathrm{s}_{i-1}^*, \mathrm{s}_{i+1}^*, \dots, s_n^*)$ for each i.

This can also be stated using the expected payoff function:

For each player i,

$$u_i(s^*) = u_i(s_i^*, s_{\neg i}^*) \geq u_i(s_i, s_{\neg i}^*) \text{ for every } s_i \in S_i$$

The concepts of dominant strategy equilibrium and Nash equilibrium are related in the following theorem.

Theorem 38. If $s^* = (s_1^*, s_2^*, \dots, s_n^*)$ is a dominant strategy equilibrium, then s^* is a Nash equilibrium.

Proof: We are given that $s^* = (s_1^*, s_2^*, \dots, s_n^*)$ is a dominant strategy equilibrium. Take **any** player i. By the definition of dominant strategy equilibrium, s_i^* is a weakly dominant strategy for player i. So, for **any** given s_i, we have

$$u_i(\mathrm{s}_i^*, s_{\neg i}) \geq u_i(s_i, s_{\neg i}) \text{ for every } s_{\neg i} \in S_{\neg i}$$

In particular,

$$u_i(s_i^*, s_{\neg i}^*) \geq u_i(s_i, s_{\neg i}^*)$$

Since the i and s_i were chosen arbitrarily in the above argument, $s^* = (\mathrm{s}_1^*, \mathrm{s}_2^*, \dots, \mathrm{s}_\mathrm{n}^*)$ is a Nash equilibrium. ■

The converse of the above theorem is not true. In the game displayed in Table 31, (A,X) and (B,Y) are both Nash equilibrium (e.g., A is a best response to Player 2 employing strategy X, and X is a best response to Player 1 employing strategy A) but neither are a dominant strategy equilibrium (i.e., neither player has a weakly dominant strategy). Further, there can be at most one dominant strategy equilibrium, but as our examples demonstrates, there can be several Nash equilibria for a game.

Table 31. Nash equilibrium does not imply dominant strategy equilibrium

		Player 2	
		X	Y
Player 1	A	(7,1)	(0,0)
	B	(0,0)	(1,7)

. . .

In terms of an algorithm for finding the Nash equilibriums, do the following:

1. For Player 1, determine the best response for each combination of the other players strategies. Put a mark against each best strategy for Player 1.
2. Do the same for Player 2, 3, …, and n.
3. Each combination of strategies where each constituent strategy is a best strategy for all the players is a Nash equilibrium.

To illustrate the above procedure, consider a 2-player game where each player simultaneously chooses an integer from 0 to 3. Both players get a payoff equal to the smaller of the two selected integers. Further, if a player chooses a larger number than the other, then that player must give two points to the other player. The payoff matrix is shown Table 32.

Table 32. Integer choosing game

		Player 2			
		Choose 0	Choose 1	Choose 2	Choose 3
Player 1	Choose 0	(0,0)	(2,-2)	(2,-2)	(2,-2)
	Choose 1	(-2,2)	(1,1)	(3,-1)	(3,-1)
	Choose 2	(-2,2)	(-1,3)	(2,2)	(4,0)
	Choose 3	(-2,2)	(-1,3)	(0,4)	(3,3)

We first determine Player 1's best response for each possible choice by Player 2.

- If Player 2 chooses 0, the best response from Player 1 is to choose 0.
- If Player 2 chooses 1, the best response from Player 1 is to choose 0.
- If Player 2 chooses 2, the best response from Player 1 is to choose 1.
- If Player 2 chooses 3, the best response from Player 1 is to choose 2.

The best responses for Player 1 are marked with an asterisk in Table 33.

Table 33. Integer choosing game – Best responses for Player 1

	Choose 0	Choose 1	Choose 2	Choose 3
Choose 0	(0*,0)	(2*,-2)	(2,-2)	(2,-2)
Choose 1	(-2,2)	(1,1)	(3*,-1)	(3,-1)
Choose 2	(-2,2)	(-1,3)	(2,2)	(4*,0)
Choose 3	(-2,2)	(-1,3)	(0,4)	(3,3)

Next, we determine the best responses for Player 2 corresponding to each selection by Player 1. The combined results of best choices for both players are shown in Table 34. The only strategy combination where both players have a best response occurs when both players choose 0. So, the only Nash equilibrium for this game is $(0,0)$ and this is the only "no regret" solution. Other solutions (strategy combinations) are higher for both players, e.g., $(3,3)$, but these solutions are not stable. For example, if Player 1 is told that Player 2 will choose 3, then Player 1 will go with 2 for a higher payoff.

Table 34. Integer choosing game – Best responses for both players

	Choose 0	Choose 1	Choose 2	Choose 3
Choose 0	**(0*,0*)**	(2*,-2)	(2,-2)	(2,-2)
Choose 1	(-2,2*)	(1,1)	(3*,-1)	(3,-1)
Choose 2	(-2,2)	(-1,3*)	(2,2)	(4*,0)
Choose 3	(-2,2)	(-1,3)	(0,4*)	(3,3)

Exercise for the reader: find the Nash equilibria for the following game:

		Player 2		
		X	Y	Z
Player 1	A	$(1,2)$	$(13,19)$	$(-2,5)$
	B	$(23,47)$	$(-3,2)$	$(4,3)$
	C	$(-3,7)$	$(2,4)$	$(31,37)$

For 2-player games, there is a simpler algorithm for finding pure Nash equilibria. For each strategy combination, check to see if Player 1's outcome is greater than its other outcomes in the same row and check to see if Player 2's outcome is greater than its other outcomes in the same column. If so, then the strategy combination is a Nash equilibrium. Applying this procedure to the above exercise, we see that $(23,27)$, $(13,19)$ and $(31,37)$ are Nash equilibria.

. . .

Not all games have a Nash equilibrium consisting of pure strategies. For example, neither the pitching pennies game nor the rock-scissors-paper games have pure strategies Nash equilibriums (see the analysis in Table 35 and Table 36).

Table 35. Pitching pennies game – no Nash equilibrium in pure strategies

	s_{21}: Heads for B	s_{22}: Tails for B
s_{11}: Heads for A	(3*,1)	(1,3*)
s_{12}: Tails for A	(1,3*)	(3*,1)

Table 36. Rock-scissors-paper game – no Nash equilibrium in pure strategies

	s_{21}: Rock for B	s_{22}: Scissors for B	s_{23}: Paper for B
s_{11}: Rock for A	(0,0)	(1*,-1)	(-1,1*)
s_{12}: Scissors for A	(-1,1*)	(0,0)	(1*,-1)
s_{13}: Paper for A	(1*,-1)	(-1,1*)	(0,0)

All is not lost here. As we will see, it is possible to define mixed strategies that combine several pure strategies according to a probability distribution. For the two examples above, one can define mixed strategies that are Nash equilibria.

9.3.5 Mixed Strategies

For some games, it is advantageous for some (if not all) of the players to use a mixed strategy which involves playing each pure strategy according to a probability distribution. The underlying assumption are

- The game is played several times.
- There is no dominant strategy equilibrium and no pure strategy Nash equilibria.
- It is beneficial to keep your opponents guessing, i.e., when the opponents can benefit from knowing the next move.

. . .

In the pitching pennies game (as described in Table 35), the various strategies are played based on the probability each penny lands as a heads or tails. Let us assume the probability is $.5$ for heads and $.5$ for tails (for each penny and independent of who is pitching the penny). With these assumptions, Player 1 has mixed strategy $\sigma_1 = (.5, .5)$, i.e., play Heads 50% of the time and play Tails 50% of the time, and Player 2 has the same mixed strategy, i.e., $\sigma_2 = (.5, .5)$.

If Player 1 uses strategy σ_1 and Player 2 always pitches Heads, then

- the expected payoff for Player 1 is $.5(3) + .5(1) = 1.5 + .5 = 2$
- the expected payoff for Player 2 is $.5(1) + .5(3) = .5 + 1.5 = 2$

If Player 1 uses strategy σ_1 and Player 2 always pitches Tails, then

- the expected payoff for Player 1 is $.5(1) + .5(3) = .5 + 1.5 = 2$
- the expected payoff for Player 2 is $.5(3) + .5(1) = 1.5 + .5 = 2$

So, Player is indifferent as to whether Player 2 chooses Heads or Tails. [We assume Player 2 can make such a choice, just to make a point about Player 1 being indifferent about the outcome from Player 2 when Player 1 uses strategy σ_1.]

Since the game is symmetric, we can draw the same conclusions about Player 2 when he, she or it uses σ_2.

When both players use the mixed strategies, the expected payoff for Player 1 is

$$3 \cdot P(H,H) + 1 \cdot P(H,T) + 3 \cdot P(T,H) + 1 \cdot P(T,T) = 3(.25) + .25 + 3(.25) + .25 = 2$$

Again, since the game is symmetric, we get the same result for Player 2.

The resulting payoff matrix, with the mixed strategies included, is as shown in Table 37.

Table 37. Pitching pennies game with mixed strategies

		Player 2		
		s_{21}: Heads for B	s_{22}: Tails for B	σ_2
Player 1	s_{11}: Heads for A	(3*,1)	(1,3*)	(2*,2)
	s_{12}: Tails for A	(1,3*)	(3*,1)	(2*,2)
	σ_1	(2,2*)	(2,2*)	(2*,2*)

If we do a Nash equilibrium analysis on the updated table, we see that (σ_1, σ_2) is a Nash equilibrium.

. . .

In the rock-scissor-paper game, if Player 1 uses mixed strategy $\sigma_1 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and Player 2 uses mixed strategy $\sigma_2 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, then we have the payoffs shown in Table 38 (with the best responses for each player marked with an asterisk). As can been seen from the table, (σ_1, σ_2) is a Nash equilibrium.

Table 38. Rock-scissors-paper game with mixed strategies

	s_{21}: Rock for B	s_{22}: Scissors for B	s_{23}: Paper for B	σ_2
s_{11}: Rock for A	(0,0)	(1*,-1)	(-1,1*)	(0*,0)
s_{12}: Scissors for A	(-1,1*)	(0,0)	(1*,-1)	(0*,0)
s_{13}: Paper for A	(1*,-1)	(-1,1*)	(0,0)	(0*,0)
σ_1	(0,0*)	(0,0*)	(0,0*)	(0*,0*)

. . .

The existence of Nash equilibriums for the previous two examples is not an accident and is expected due to the following theorem.

Theorem 39 (Nash Equilibrium Existence) Every game with a finite number of players (each of which has a finite number of pure strategies) has at least one Nash equilibrium (which is possibly a mixed strategy).

This theorem stated and proved by John Nash in his PhD thesis (1950), see the Wikipedia article Nash equilibrium [67].

For the pitching pennies and rock-scissors-paper games, we were able to guess the mixed strategy equilibrium given the high-level of symmetry in the payoff matrices. Unfortunately, the determination of a mixed strategy equilibrium is typically not an easy task.

Consider the game shown in Table 39. First, we note that there are no dominated strategies. So, iterated elimination of pure dominated strategies does not apply. Next, we check for Nash equilibriums. Player 1's best response to X is A, and best response to Y is B. Player 2's best response to A is X, and best response to B is Y. The best responses are marked with an asterisk in the table. As can be seen, there is no strategy profile in which both players have a best response, and thus, there is no pure strategy equilibrium.

Table 39. Mixed strategy equilibrium example

		Player 2	
		X	Y
Player 1	A	(3*,1)	(2,4*)
	B	(0,3*)	(3*,1)

The next step is to find a mixed strategy profile that is a Nash equilibrium. This is equivalent to finding

- mixed strategy $\sigma_1 = (p, 1-p)$ for Player 1 such that Player 2 has no preferences in playing X or Y when Player 1 uses σ_1
- mixed strategy $\sigma_2 = (q, 1-q)$ for Player 2 such that Player 1 has no preferences in playing A or B when Player 2 uses σ_2

If Player 1 employs σ_1, then Player 2 has payoff $p + 3(1-p)$ when playing X, and payoff $4p + (1 - p)$ when playing Y. For Player 2 to be indifferent between playing X and Y, we need to have

$$p + 3(1-p) = 4p + (1-p)$$

$$p + 3 - 3p = 4p + 1 - p$$

$$5p = 2$$

$$p = \frac{2}{5} = .4 \text{ and } 1 - p = .6$$

If Player 2 employs σ_2, then Player 1 has payoff $3q + 2(1-q)$ when playing A, and payoff $0 \cdot q + 3(1-q)$ when playing B. For Player 1 to be indifferent between playing A or B, we need to have

$$3q + 2(1-q) = 0 \cdot q + 3(1-q)$$

$$3q + 2 - 2q = 3 - 3q$$

$$4q = 1$$

$$q = \frac{1}{4} = .25 \text{ and } 1 - q = .75$$

So, $\sigma_1 = (.4, .6)$ and $\sigma_2 = (.25, .75)$.

If we add the two mixed strategies to the game, compute the payoff for the addition strategy profiles (pairs of strategies in this case) and identify the best response for each player based on the other player's strategy, we see that (σ_1, σ_2) is a Nash equilibrium (as shown in Table 40).

Table 40. Mixed strategy game – Nash equilibrium solution

		Player 2		
		X	Y	σ_2
Player 1	A	(3*,1)	(2,4*)	(2.25*, 3.25)
	B	(0,3*)	(3*,1)	(2.25*, 1.5)
	σ_1	(1.2, 2.2*)	(2.6, 2.2*)	(2.25*, 2.2*)

. . .

A strategy that survives the iterated elimination of strictly dominated strategies is said to be **rationalizable**, noting that mixed strategies can be used to eliminate other strategies.

Theorem 40. If $s = (s_1, s_2, \dots, s_n)$ is a Nash equilibrium, then s_i is rationalizable for each i.

Proof: We need to show that none of the strategies s_i is eliminated at any stage of the iterated elimination of strictly dominated strategies. Since each s_i is available at the beginning of the procedure, it is sufficient to show that if the strategies $s_1, s_2, \dots, s_n$ are all available at round n, then they will remain available (i.e., not eliminated) at round $n + 1$ (basically an induction proof). Since s is a Nash equilibrium, for each i, s_i is a best response to $s_{\neg i}$ (which, by the induction hypothesis, are still available at round n of the elimination process). Therefore, s_i is not strictly dominated at round n, and remains available at round $n + 1$. ■

The converse of the above theorem is not true. Consider the game in Table 37. None of the pure strategies is eliminated since each provides a best response in some situations. For example, consider (s_{11}, s_{21}) which is not a Nash equilibrium but s_{11} and s_{21} are rationalizable since neither is eliminated in the IESDS process, noting that s_{11} is not eliminated since it is a best response to Player 2 playing s_{21}, and s_{21} is not eliminated since it is a best response to Player 1 playing s_{12}.

9.3.6 Minimax and Maximin

In this section, we discuss what is known as the Fundamental Theorem of Game Theory (also known as the Minimax theorem). The minimax theorem states that every finite, zero-sum, two-person game has an optimal mixed strategy. It was proved by John von Neumann in 1928. Minimax is another way of viewing solutions to games.

For every two-person, zero-sum game with finitely many strategies, there exists a value v and a mixed strategy for each player, such that

- Given Player 2's strategy, the best possible payoff for Player 1 is v, and
- Given Player 1's strategy, the best possible payoff for Player 2 is $-v$.

Further, for two-player zero-sum games, the minimax-maximin optimal strategy is the same as the Nash equilibrium.

. . .

In general (not necessarily 2-person, zero-sum games), there are two aspects to minimax-maximin approach, i.e.,

- Determination of maximin strategy profiles. In this approach, each player determines the minimum possible payoff for each of its possible strategies, and then identifies the maximum out of the minimum payoffs (there could be ties). The **maximin value** is the highest value that the player can be sure to get without knowing the actions of the other players; equivalently, it is the lowest value the other players can force the player to receive when they know the player's action.

- Determination of minimax strategy profiles. In this approach, each player determines its maximum possible payoff for each combination of its opponents' strategies, and then identifies the minimum out of the maximum payoffs (there could be ties). The **minimax value** of a player is the smallest value that the other players can force the player to receive, without knowing the player's actions; equivalently, it is the largest value the player can be sure to get when they know the actions of the other players.

For each player, the maximin payoff value is less than or equal to the minimax payoff value.

This approach can be used for non-zero-sum as well as zero-sum games.

. . .

The maximin concept is illustrated in Table 41. A non-zero-sum game was selected as an example to illustrate the point that the maximin payoff for a player can be strictly less than the minimax payoff.

- For each of Player 1's strategies, determine its minimum possible payoff. The minimum payoffs are listed in the column to the right.
- Do the same for Player 2, with the minimum payoffs listed in the bottom row.
- Player 1 can maximize its minimum possible payoff by choosing strategy A. In other words, Player 1 can do no worse than a payoff of 2 if it plays strategy A.
- Player 2 can maximize its minimum possible payoff by choosing strategy X.
- For this game, the maximin optimal strategy is (A, X). If Player 1 uses strategy A, it cannot do any worse than a payoff of 2, and if Player 2 uses strategy X, it cannot do any worse than -3.

Table 41. Maximin analysis – non-zero sum game

		Player 2		
		X	Y	Minimums for Player 1
Player 1	A	(3,1)	(2,-33)	2
	B	(5,0)	(-11,3)	-11
	C	(-99,-3)	(4,4)	-99
	Minimums for Player 2	-3	-33	

The minimax concept is illustrated in Table 42.

- Given each possible strategy for Player 2, determine the maximum possible payoff for Player 1 (shown in the bottom row).
- Given each possible strategy for Player 1, determine the maximum possible payoff for Player 2 (shown in the right column).

- Determine the minimum of the maximum payoffs for each player (underlined numbers in the table).
- For this game, the minimax optimal strategy is (A, Y). If Player 2 employs strategy Y, Player 1 cannot do any better than a payoff of 4, and if Player 1 employs strategy A, Player 2 cannot do any better than a payoff of 1.

Table 42. Minimax analysis – non-zero sum game

		Player 2		
		X	Y	Maximums for Player 2
Player 1	A	(3,1)	(2,-33)	1
	B	(5,0)	(-11,3)	3
	C	(-99,-3)	(4,4)	4
	Maximums for Player 1	5	4	

The Nash equilibrium is (C, Y) which doesn't match either the maximin or minimax strategy profiles.

In general, the maximin payoff is always less than or equal to the minimax payoff for each player [68]. In our example,

- The maximin payoff for Player 1 is 2, and the minimax payoff for Player 1 is 4.
- The maximin payoff for Player 2 is -3, and the minimax payoff for Player 2 is 1.

It can be proven that for finite (number of strategies), zero-sum, two-person games the minimax and maximin strategy profiles coincide, and are a Nash equilibrium, see Theorems 4.44 and 4.45 in the book "Game Theory" [69]. Further, the payoff of Player 1 is the negative of the payoff for Player 2 at the Nash equilibrium. So, in two-player, zero-sum games the concept of a Nash equilibrium (which is based on stability) and the concept of a minimax/maximin coincide.

The two-player, zero-sum game shown in Table 43 will be used to illustrate the above points. An analysis of best responses to pure strategies (see the asterisks in the table) shows that there is no pure strategy Nash equilibrium, i.e., a strategy profile where both players have a best response. By Nash's equilibrium theorem, we know at least one mixed strategy Nash equilibrium exists.

Table 43. Nash equilibrium analysis of a two-player, zero-sum game

		Player 2	
		X	Y
Player 1	A	(-2,2*)	(1*,-1)
	B	(3*,-3)	(0,0*)

Let $\sigma_1 = (p, 1 - p)$ and $\sigma_2 = (q, 1 - q)$ be mixed strategies for Players 1 and 2, respectively.

The requirement for σ_1 is that when used by Player 1, Player 2 has no preference between strategies X or Y (i.e., they yield the same payoff). This results in the equation

$$2p - 3(1 - p) = -p + 0 \cdot (1 - p)$$
$$p = \frac{1}{2}$$

So, $\sigma_1 = (\frac{1}{2}, \frac{1}{2})$.

The requirement for σ_2 is that when used by Player 2, Player 1 has no preference between strategies A or B. Thus, q must satisfy

$$-2q + (1 - q) = 3q + 0 \cdot (1 - q)$$
$$q = \frac{1}{6}$$

So, $\sigma_2 = (\frac{1}{6}, \frac{5}{6})$.

The following table shows the game with the mixed strategies added. The strategy profile (σ_1, σ_2) is a Nash equilibrium. For example, consider the case where Player 2 utilizes strategy X and Player 1 utilizes strategy σ_1. In the case, Player 2 plays X 100% of the time, and Player 1 plays A 50% of the time and B 50% of the time. So, the payoff for Player 1 is $\frac{1}{2}(-2) + \frac{1}{2}(3) = \frac{1}{2}$, and the payoff for Player 2 is $\frac{1}{2}(2) + \frac{1}{2}(-3) = -\frac{1}{2}$.

		Player 2		
		X	Y	σ_2
Player 1	A	$(-2,2)$	$(1,-1)$	$\left(\frac{1}{2}, -\frac{1}{2}\right)$
	B	$(3,-3)$	$(0,0)$	$\left(\frac{1}{2}, -\frac{1}{2}\right)$
	σ_1	$\left(\frac{1}{2}, -\frac{1}{2}\right)$	$\left(\frac{1}{2}, -\frac{1}{2}\right)$	$\left(\frac{1}{2}, -\frac{1}{2}\right)$

The following table shows the maximin analysis. We first find the minimums for each players' strategy and then take the maximum. Strategy profile (σ_1, σ_2) is the maximin solution, and the payoff for Player 1 is the negative of the playoff for Player 2.

		Player 2			
		X	Y	σ_2	Minimums for Player 1
Player 1	A	$(-2,2)$	$(1,-1)$	$\left(\frac{1}{2},-\frac{1}{2}\right)$	-2
	B	$(3,-3)$	$(0,0)$	$\left(\frac{1}{2},-\frac{1}{2}\right)$	0
	σ_1	$\left(\frac{1}{2},-\frac{1}{2}\right)$	$\left(\frac{1}{2},-\frac{1}{2}\right)$	$\left(\frac{1}{2},-\frac{1}{2}\right)$	$\mathbf{\frac{1}{2}}$
	Minimums for Player 2	-3	-1	$-\mathbf{\frac{1}{2}}$	

The following table shows the minimax analysis. We first find the maximums for each player based on each strategy of the other player, and then take the minimums. Strategy profile (σ_1, σ_2) is the minimax solution, and the payoff for Player 1 is the negative of the playoff for Player 2.

		Player 2			
		X	Y	σ_2	Maximums for Player 2
Player 1	A	$(-2,2)$	$(1,-1)$	$\left(\frac{1}{2},-\frac{1}{2}\right)$	2
	B	$(3,-3)$	$(0,0)$	$\left(\frac{1}{2},-\frac{1}{2}\right)$	0
	σ_1	$\left(\frac{1}{2},-\frac{1}{2}\right)$	$\left(\frac{1}{2},-\frac{1}{2}\right)$	$\left(\frac{1}{2},-\frac{1}{2}\right)$	$-\mathbf{\frac{1}{2}}$
	Maximums for Player 1	3	1	$\mathbf{\frac{1}{2}}$	

. . .

It is possible to have an infinite number of mixed strategy Nash equilibriums for a game. Consider the game depicted in Table 44. The best responses for each player (in response to each of the other player's strategies) are marked with asterisks. There are two pure strategy Nash equilibria, i.e., (B, X) and (A, Y).

Table 44. Game with an infinite number of mixed strategy equilibria

		Player 2	
		X	Y
Player 1	A	(2,3*)	(5*,3*)
	B	(3*,3*)	(1,1)

If Player 1 selects pure strategy A, any mixed strategy for Player 2, call it $\sigma_2 = (p, 1-p)$, yields a payoff of 3 of Player 2.

Under what conditions (i.e., under what mixed strategies of Player 2) would Player 1 do best by going with pure strategy A?

Player 1's payoff for employing A when Player 2 employs mixed strategy σ_2 is

$$2p + 5(1 - p)$$
$$= 5 - 3p$$

Player 1's payoff for employing B when Player 2 employs mixed strategy σ_2 is

$$3p + (1 - p)$$
$$= 2p + 1$$

Thus, Player 1's best response is A when

$$5 - 3p \geq 2p + 1$$
$$p \leq \frac{4}{5}$$

So, Player 2 has an infinite number of mixed strategies $\sigma_2 = (p, 1 - p)$ with $p \leq \frac{4}{5}$ such that Player 1's best response is pure strategy A. In other words, there are an infinite number of Nash equilibria (A, σ_2), with $0 \leq p \leq \frac{4}{5}$.

9.4 Solutions for Sequential Games

All the solution methods described in the previous section pertain to simultaneous games. In this section, we consider solution techniques for sequential games.

9.4.1 Backward Induction Example

The main method for finding optimal solutions to sequential games is something called **backward induction**. To see how backward induction works, we will use the computer sale example that we previously mentioned in Section 9.1 (reproduced in Figure 19 but with additional detail).

- We start from the bottom of the extensive form tree.
  - On the left, Bernice will choose Reject since that gives her a payoff 0 rather than -.5.
  - In the middle, Bernice will choose Accept, getting a payoff of .5 versus 0.
  - On the right, Bernice will choose Accept again, since 1.5 is a better payoff than 0.
- Aristotle also knows the details of the tree and assumes Bernice will select the highest payoff in each of the subtrees noted above. Out of the three subtrees (and assuming Bernice rejects Option 3), the best payoff for Aristotle is to choose Option 2 (i.e., an asking price of $2000 of his used computer). This choice gives Aristotle a net return of $2000. If he chose Option 3, his expected net return would only be $0. If he chose Option 1, his net return would be $1000.
- Each of Bernice's choices in the subtrees are shown with heavy arrows, and Aristotle's choice is also shown with a heavy arrow. The only path from the root of the tree to the one of the leaves (terminal points) is the strategic profile (Option 2, Accept), i.e., Aristotle offering the computer at $2000 and Bernice agreeing to buy the computer.

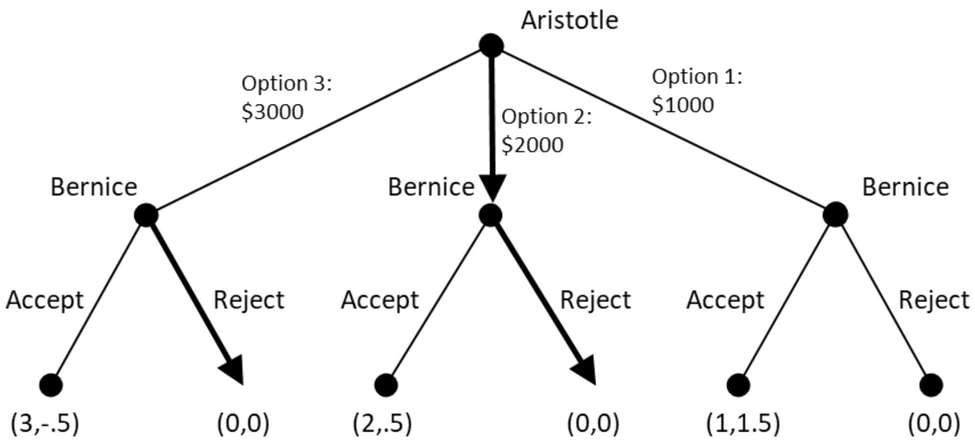


Figure 19. Backward induction for the compute sale game

9.4.2 Definitions and Concepts

The **extensive form of a game** (an example is shown in Figure 19) is a specification of a game that represents key aspects such as the sequencing of players' possible moves, their possible choices at every decision point, and the payoffs for all possible game outcomes. The mathematical structure known as a tree is used to represent a game in extensive form. The overall tree is known as the **game tree**. Each node (also referred to as vertex or point) of the game tree is marked with the name of the player who is to decide among several alternatives. The payoffs for each player are listed at the leaves of the game tree. Each subtree of the game tree is known as a **subgame**. For the game shown in Figure 19, there are 4 subgames, i.e., the entire game, and each subgame related to the three decision nodes for Beatrice. The subgames that are not the entire game are known as **proper subgames**. A game tree is said to be of **finite horizon** if the length from the root node to every leaf is finite. However, it is still possible for a finite horizon game tree to have an infinite number of numbers, e.g., when a player has an infinite number of possible options at a given decision point.

A game with **perfect information** is such that at any stage of the game, every player knows exactly what has taken place, earlier in the game, by all players, and also knows the complete structure and details of the game tree; otherwise, a game is said to be an **imperfect information game**. In this book, we only consider perfect information games. The book by Osborne [70] has detailed coverage of sequential games, including games with perfect and imperfect information.

A **subgame perfect equilibrium** is a refinement of the Nash equilibrium concept that pertains to sequential games. In particular, a strategy profile is a subgame perfect equilibrium if it represents a Nash equilibrium of every subgame of the original game. It has been proven that every finite extensive game with perfect information has at least one subgame perfect equilibrium. Further, the set of subgame perfect equilibria for a finite horizon extensive game with perfect information is equal to the set of strategy profiles determined by backward induction [70].

9.4.3 More Examples

In the game shown in Figure 20, Player 1 chooses a strategy first (either A or B), and then depending on Player 1's choice, Player 2 chooses a strategy. The payoffs are shown in the leaves of the tree, with the payoff for Player 1 coming first.

Using backward induction, we see that Player 2 will choose strategy Y in the subgame on the bottom left, and will choose strategy V in the subgame on the right. Player 1, assuming Player 2 is rational, will come to the same conclusions as Player 1 concerning the two subgames. Given Player 1's anticipated choices, Player 2 can choose either strategy A or B, and get a payoff of 2. Thus, we have two subgame perfect equilibria, i.e., strategy profiles (A, Y) and (B, V).

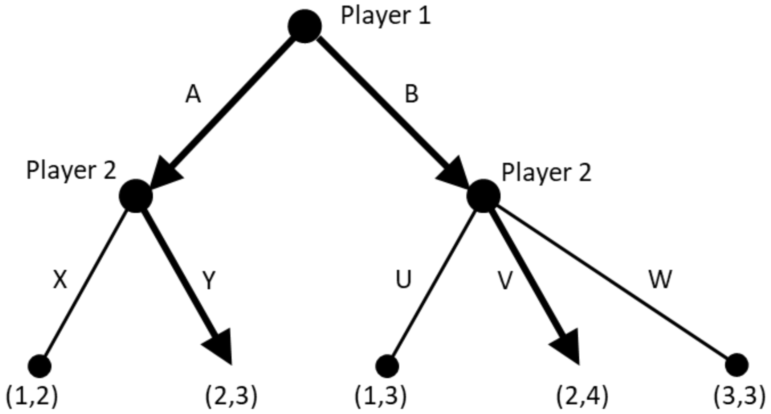


Figure 20. Multiple subgame perfect equilibria

. . .

Our next example is a variation of something called the **ultimatum game** (first proposed by Harsanyi [71]). The basic idea is that someone (an executor) is to divide a sum of money with another person (the decider). The executor can either make a fair 50-50 split, or can make an unequal split of the fortune in his or her favor (e.g., 70-30). Once the executor offers a distribution split, the decider can accept the offer as-is or reject. If rejected, the combined payout to the two parties is less than the full amount, e.g., 20% to each. Further, assume that the executor and decider do not know each other, their identities will not be revealed to each other, and in general, they have no ability to negotiate during the game or afterward. The situation is depicted in Figure 21.

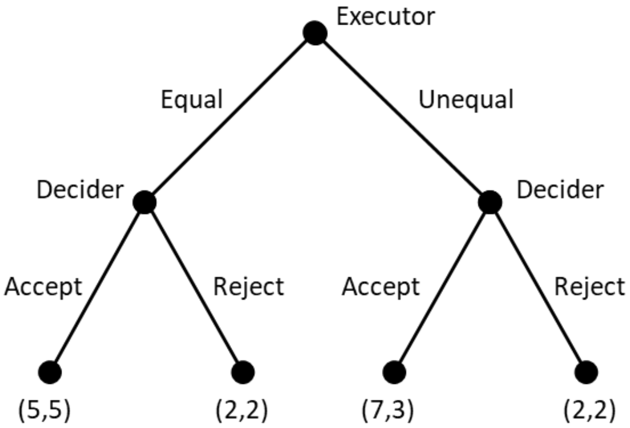


Figure 21. Ultimatum game

A backward induction analysis of the pure strategies of the game is as follows:

- If the executor takes the Equal strategy, then the strategy of Accept provides the best payoff for the decider.
- If the executor takes the Unequal route, then the decider again gets a higher payoff by accepting.
- So, without any fear of the decider rejecting an unequal division of the money, the executor will choose Unequal and the decider will choose Accept.

However, there is a variation of the ultimatum game where the decider may accept an unequal (unfair) offer with probability p and reject with probability $1-p$, as depicted in Figure 22.

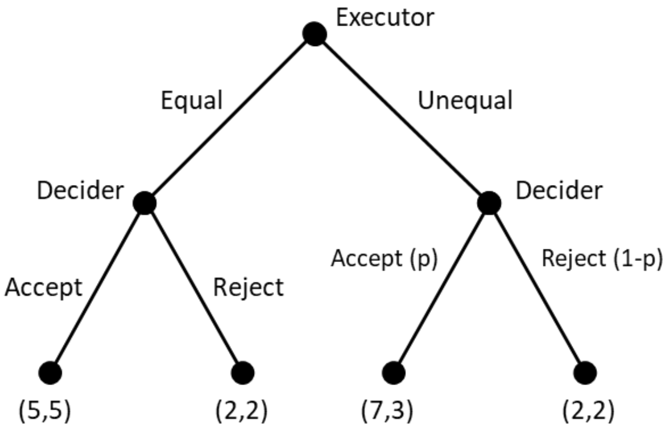


Figure 22. Ultimatum game with possibility of unequal offer being rejected

Whether Unequal is a viable strategy for the executor depends on the probability that the decider will Reject in that case. This is basically a mixed strategy. For what value of p does it make sense (more favorable) for the executor to select Unequal (with the hope of getting the bigger payoff)?

The payoff for the executor when he or she chooses Unequal is

$$7p + 2(1-p) = 5p + 2$$

The executor can be sure of getting a payoff of 5 in the Equal branch of the game. So, the executor at least wants an expected value greater than 5 to justify selecting the Unequal strategy. Thus, we want to solve

$$5p + 2 > 5$$

$$p > \frac{3}{5} = .6$$

This means there are an infinite number of subgame perfect equilibria with strategies of the form (Unequal, Accept (p>.6)).

The cutoff probability is not fixed with respect to the division of money in the Unequal case. In particular, the cutoff probability decreases as the unbalanced reward for the executor increases. For example, if the payoff for strategy $(Unequal, Accept(p))$ was (9,1) and the other Reject choice remained unchanged, then the executor needs $p > \frac{3}{7} \cong .43$ to have a higher expected payoff with the Unequal strategy.

Keep in mind that the payoffs used in the above analysis are closely associated with the financial breakdown of the money to be split. Other payoff schemes are possible. For example, the executor may be a very cautious person and wants to do an even split more than he or she wants to risk going for a higher percentage of the money. In this case, the payoffs for the executor indicate preferences (going from 1 to 4, with 4 being the most preferred outcome). The payoffs for the decider remain the same (based on the percentage of the monetary distribution). The modified game is as shown in Figure 23. In this case, the optimal strategy is now (Equal, Accept).

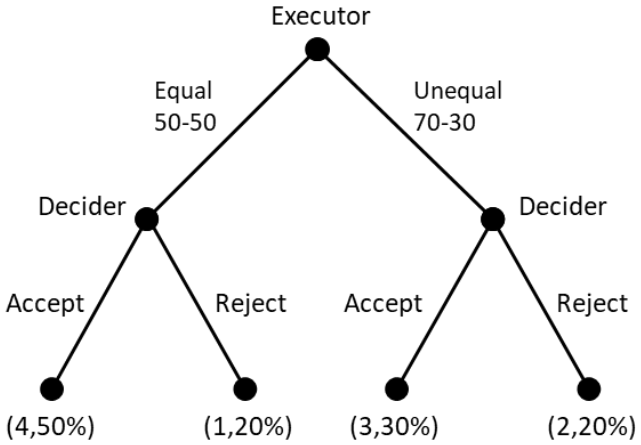


Figure 23. Ultimatum game – modified payoffs

The critical point here is that the payoff should reflect the value of the various options as perceived by the players in the game.

. . .

The following game (known as the Pirates' Treasure Puzzle) has a very large number of possible alternatives at each decision point. However, as we shall see, the puzzle can be solved without enumerating all the possible alternatives. The puzzle goes as follows:

> Five pirates (all men except the captain) of different ranks (numbered 1 to 5) have a stolen treasure of 100 gold coins, which they need to split amongst themselves. Pirate #1 (the captain) makes the first proposal concerning how the coins are to be divided. All the pirates (including the captain) vote for or against the proposal. If 50% or more of the pirates vote for the proposal, the coins will be shared as specified by the captain; otherwise, the captain is thrown overboard, and the process is repeated with the remaining pirates (according to rank) until agreement is reached. Further, if a pirate can get the same number of coins by voting for or against a proposal, he or she will vote against the proposal. Assume all the pirates are rational and seek the highest number of coins, and do not wish to be thrown overboard. What is the subgame perfect equilibrium?

By starting from the bottom (i.e., Pirate #5), we can solve the problem without explicitly drawing the entire game tree. A summary of the analysis is shown in Table 45. The table is explained in the following text.

Table 45. Backward analysis for Pirates' Treasure Puzzle

		Proposals received				
		Pirate 1	Pirate 2	Pirate 3	Pirate 4	Pirate 5
Proposals offered	Pirate 1	98	0	1	0	1
	Pirate 2	Overboard	99	0	1	0
	Pirate 3	Overboard	Overboard	99	0	1
	Pirate 4	Overboard	Overboard	Overboard	100	0
	Pirate 5	Overboard	Overboard	Overboard	Overboard	100

Pirate #5 only gets a chance to make a proposal if Pirates #1-4 have all made proposals that have been rejected and thus thrown overboard. Pirate #5 would be the only pirate left onboard, and has no choice but to offer 100 coins to himself, and of course, he accepts. See the bottom row Table 45.

Working backward one step, assume Pirate #4 is to make a proposal, Pirates #1-3 have been thrown overboard and Pirate #5 remains onboard. Again, the decision is easy for Pirate #4 since only 50% of the votes are required to accept a proposal. Knowing this, Pirate #4 proposed to take all 100 coins and give no coins to Pirate #5. Pirate #4 votes to accept and this ensures 50% of the vote. Note that we have discounted the possibility of Pirate #4 offering himself 99, 98, … coins, with the rest going to Pirate #5.

Working back yet another step, assume Pirate #3 is to make a proposal, Pirates #1 and #2 have been thrown overboard, and Pirates #4 and #5 are still onboard.

- If Pirate #3 takes all 100 coins, the other two (getting 0) will reject the offer and throw #3 overboard. So, this is not an option.
- Pirate #5 knows that if Pirate #3's offer gets rejected, the problem reduces to Pirate #4 making an offer. In this case, Pirate #5 knows that he will get nothing. So, Pirate #3 only needs to offer Pirate #5 one coin to get his vote. It is not a great deal for Pirate #5 but it is better than 0. So, Pirate #3 makes an offer of 0 for Pirate #4, 1 for Pirate #5 and 99 coins for himself. Pirates #3 and #5 vote to accept.

Going back one more step, Pirate #1 has been ejected from the ship, and Pirate #2 needs to decide on an offer.

- If Pirate #2 offers himself all the coins, his offer will be rejected by a vote of 3-1.
- Noting the Pirate #5 gets 1 in the event Pirate #2 gets rejects, Pirate #5 would need to be offered at least 2 coins to accept Pirate #2's offer.
- However, Pirate #2 can do even better by offering but 1 coin to Pirate #4, since Pirate #4 would get 0 coins if Pirate #2 is rejected. In this case, Pirates #2 and #4 accept the offer, with a rejection from Pirates #3 and #5.

Finally, we get back to the pirate captain. She needs 3 pirates (including herself) to accept her offer, or she will get ejected from the ship. If the captain gets ejected, Pirates #3 and #5 know they can be forced to take 0 by Pirate #2 in the next stage. So, the captain only needs to offer 1 coin each to Pirates #3 and #5 to get their vote, and this is the final solution.

. . .

The centipede game [72] tests the limits of the backward induction approach. The game involves two players (X and Y) who, at each turn, are given the option of taking $2 and splitting the existing pot of money, or putting the $2 into the pot, with the other player then given the same option. The pot starts with $0. To be clear, the $2 comes from a source outside of the two players, i.e., it is gifted at each round of play. The game consists of 100 rounds. On the last possible round, if Player Y puts the $2 gift into the pot, the game ends by splitting the money evenly between the two players. The game tree is depicted in Figure 24, the Player X's payoff listed first in each of the payoff pairs.

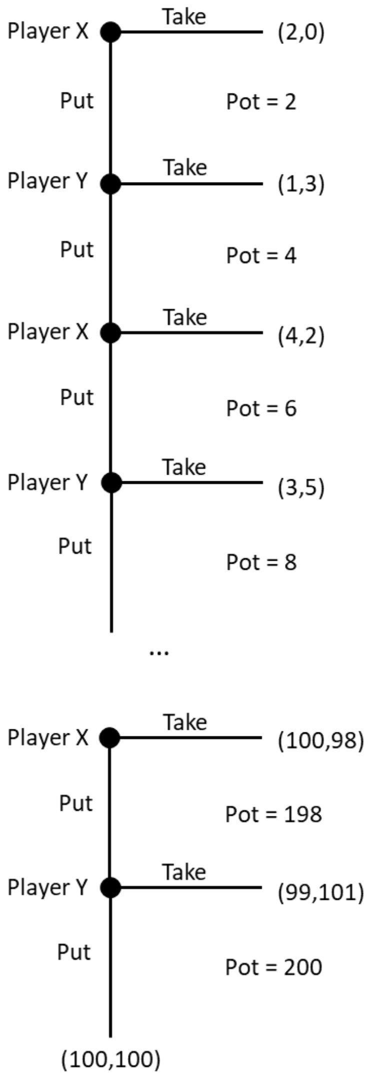


Figure 24. Centipede game

Starting at the bottom of the game tree, Player Y gets a higher payoff ($101) by taking the $2 and splitting the existing pot of $198. Moving back one round, Player X gets a higher payoff ($100) by taking the $2 and splitting the existing pot of $196. The same result (taking versus putting) follows all the way back to the first round, where X does better by taking the $2 and the game ends right there. So, the optimal solution (based on backward induction) is for X to take the initial $2 gift and stop the game.

However, the result appears to violate common sense. If the two players understand the game correctly, it makes most sense to keep taking the $2 at each round and build the pot to the maximum possible value. While it is true that there is a small advantage in taking at each round, this is outweighed by the goal of getting a bigger payout in the final round.

One possible explanation for the non-intuitive result is that the game payoffs are directly linked to the immediate financial payouts in each round. However, once the game is well understood by each player, both players are likely to see the value in cooperation and will want to extend the game to the last round. This can be reflected in the payoffs, as shown in Figure 25. In the alternate payoff scheme, each player is shown to prefer putting the $2 in the pot and extending the game. The key point here is that games should reflect what the modeler thinks are the preferences of the players. This aspect of game theory (i.e., considering social preferences, social utility and other psychological factors) is known as behavioral game theory [74].

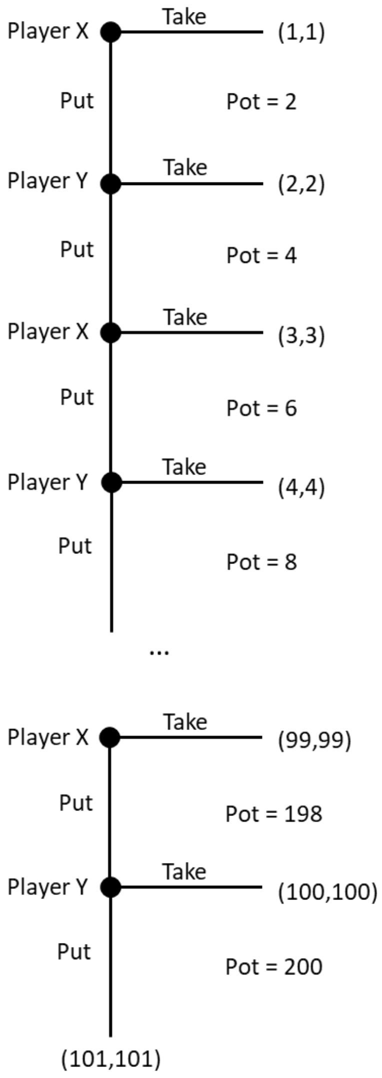


Figure 25. Centipede game – alternate payoff scheme

10 Linear Programming

10.1 Overview

Prerequisites: algebra, summation notation, inequalities

Linear Programming (LP) is an approach for optimizing a function (known as the **objective function**) under a set of constraints that define a **feasible region** to be considered. LP falls within a branch of mathematics known as operations research. The term "linear" refers to the stipulation that the function to be optimized is a linear combination of variables, and the constraints are linear equations and inequalities. The term "programming" is not used in the computer programming sense but rather in the sense of planning a best approach to a problem.

For example, consider the problem of finding the maximum value of

$$y - \frac{x}{4}$$

subject to the constraints

$$y - \frac{x}{2} \leq 3$$

$$y + \frac{x}{3} \leq 8$$

$$x \geq 0, y \geq 0$$

The area confined by the constraints is the gray polygon in Figure 26. This is known as the feasible region for the problem. The dashed line is $y - \frac{x}{4} = 7$ and the dotted line is $y - \frac{x}{4} = \frac{9}{2}$ (which turns out to be the maximum value of the objective function within the feasible region for the problem). All other points within the feasible region will give a smaller value for the objective function. For example, the objective function evaluated at the point $(6,4)$ is $4 - \frac{6}{4} = \frac{5}{2} < \frac{9}{2}$. The points on the dashed line result in higher values for the objective function but none of the points on the dashed line is in the feasible region of the problem.

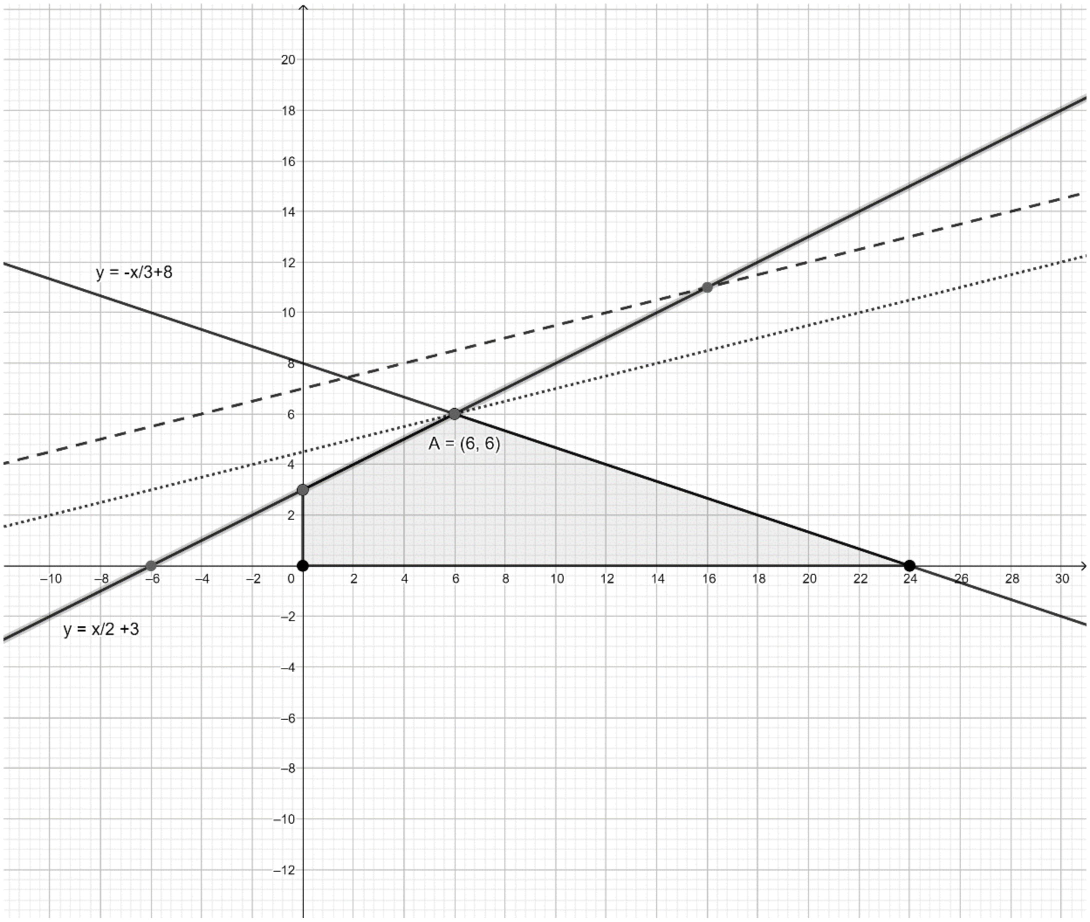


Figure 26. Linear Programming example in 2-dimensions

In general, a linear programming problem in two dimensions has a feasible region in the shape of a polygon (possibly unbounded above or below) and an objective function of the form $ax + by$ with a and b being constants, and x and y being variables. The goal is to find the point (x', y') in the feasible region such that $ax' + by' = c$ is maximal (or minimal), where $ax' + by' = c$ is the equation for a line.

In three dimensions, the feasible region is typically a convex polyhedron, i.e., a flat-faced solid such that the line segment between any two points of the solid is completely within the solid. The objective function is of the form $ax + by + cz$. Again, we seek a point (x', y', z') in the feasible region such that $ax' + by' + cz' = d$ is maximal (or minimal), where $ax' + by' + cz' = d$ is the equation for a plane.

In n dimensions, the feasible region is a convex polytope [75], i.e., a convex flat-faced solid where the faces are of dimension $n - 1$. The objective function is of the form $a_1x_1 + a_2x_2 + \cdots + a_nx_n$. Again, we seek a point $(x_1', x_2', \dots, x_n')$ in the feasible region such that $a_1x_1' + a_2x_2' + \cdots + a_nx_n' = c$ is maximal (or minimal), where $a_1x_1' + a_2x_2' + \cdots + a_nx_n' = c$ is the equation for a hyperplane (i.e., a subspace of dimension $n - 1$).

In all cases, point(s) yielding the maximum (or minimum) value for the objective function are on the boundary of the feasible region. This follows by something known as the maximum principle for convex functions [76].

This section focuses on various applications of LP and on problem formulation. For readers interested in the various methods of solution and the theory behind LP, a good starting point is the book by Brickman [77].

10.2 General Problem

In the general LP problem, there are n variables, the objective function can be minimized or maximized, and the constraints can consist of equalities and inequalities (combinations of less than or greater than). In terms of notation, we have

Maximize (or Minimize) the objective function:

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n + d, \text{ with the } c_i \text{ and } d \text{ being constants}$$

Subject to a finite number of constraints of the form:

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n \ (\leq, = \text{ or } \geq)\ b$$

such that

$$x_i \geq 0, i = 1,2, \dots, n$$

For example, we could have

Minimize:

$$3x + 4y - 2z$$

Subject to

$$3x + 2y = 4$$

$$-2x + 7y - 3z \geq 8$$

$$x \geq 0, \qquad y \geq 0, \qquad z \geq 0$$

Using an online LP application (e.g., https://linprog.com/en/main-simplex-method), we find the minimum value of the objective function (within the feasible region) to be 4, with that value occurring at point $(0, 2, 2)$.

10.3 Applications

10.3.1 Production Problem

In the production problem, a manufacturer can make n different products $(P_1, P_2, \dots, P_n)$ where one unit of P_i yields a profit of c_i when sold. The manufacturer would like to maximize its profit, i.e., maximize the objective function

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

However, there are resource constraints. The production of each product requires some combination of m resources $(R_1, R_2, \dots, R_n)$. In particular, one unit of product P_i requires a_{ji} units of resource R_j. Further, the amount of resource R_j is limited to an amount b_j over a given time period.

In term of equations, we have

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leq b_1 \qquad \text{(constraint on } R_1\text{)}$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \leq b_2 \qquad \text{(constraint on } R_2\text{)}$$
$$\ldots$$
$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leq b_m \quad \text{(constraint on } R_m\text{)}$$

with $x_i \geq 0, i = 1,2, \ldots, n$.

The above can be written more succinctly using summation notation as

Maximize: $\sum_{i=1}^{n} c_i x_i$

Subject to:

$\sum_{i=1}^{n} a_{ji} x_i \leq b_j$, for $j = 1,2, \ldots, m$

$x_i \geq 0, i = 1,2, \ldots, n$

This can also be written in matrix form

Maximize: $\boldsymbol{c}^t \boldsymbol{x}$

Subject to:

$A\boldsymbol{x} \leq \boldsymbol{b}$

$\boldsymbol{x} \geq \boldsymbol{0}$

In the above, $\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{b}$ and $\boldsymbol{0}$ are column vectors, $\boldsymbol{c^t}$ is the transpose of **c** (and thus a row vector), and A is an $m\ x\ n$ matrix.

. . .

A barbell manufacturer produces two types of barbell sets. The barbell sets (plates and bar) are molded from iron by means of a large pressing machine and then finished by hand labor.

- Barbell set B1 requires 120 pounds of iron, 6 minutes of machine time, and 3 minutes of finishing labor.
- Barbell set B2 requires 80 pounds of iron, 5 minutes of machine time, and 5 minutes of finishing labor.

Over the next month, the barbell manufacturer can dedicate up to 60,000 pounds of iron, 3,000 minutes of machine time, and 2,000 minutes of labor for the production of barbell sets.

The manufacturer makes a profit of \$50 on the sale barbell set B1, and a \$60 profit on the sale barbell set B2. Assuming that all the barbell sets made can be sold, how many of each type should be manufactured in the next month in order to maximize profits?

If we let x be the number of barbell sets of type B1 produced (and sold), and y be the number of barbell sets of type B2 produced (and sold), then we have the following constraints:

$$120x + 80y \leq 60{,}000$$
$$6x + 5y \leq 3{,}000$$
$$3x + 5y \leq 2{,}000$$

$$x \geq 0, \qquad y \geq 0$$

The goal here is to maximize the following objective function:

$$50x + 60y$$

Using the online LP application at https://linprog.com/en/main-simplex-method, we determine a maximum profit $28,666.67 when

$$x = 333\frac{1}{3}, \qquad y = 200$$

Since x is not a whole number, the manufacturer will need to round-down the number. This type of result (i.e., non-integer solutions) is typical in real-life examples. If there is a hard requirement to get integer solutions, there is something called integer programming (which we discuss in Section 11).

If we plug $x = 333\frac{1}{3}$ and $y = 200$ back into the constraint expressions, we get

$$120x + 80y = 56{,}000$$

$$6x + 5y = 3{,}000$$

$$3x + 5y = 2000$$

All the machine time and finishing labor is being used to the maximum but not all of the (potential) iron is being used. This type of analysis can be helpful to determine which of the constraints are limiting factors. For the problem at hand, the manufacturer might add another shift (at a time when the machines would otherwise be idle), and thus, be able to produce additional products while still remaining within the 60,000 ton limit.

. . .

A company, that packages almonds, operates two processing plants (A and B). The company has three suppliers (S1, S2 and S3) who can supply almonds in the following amounts (per week):

- S1: 200 tons at $11/ton
- S2: 310 tons at $10/ton
- S3: 420 tons at $9/ton

Shipping costs per ton of almonds are summarized in the following table:

		To	
		Processing Plant A	Processing Plant B
From	Supplier S1	$3	$3.5
	Supplier S2	$2	$2.5
	Supplier S3	$6	$4

Plant capacities and labor costs are summarized in the following table:

	Plant A	Plant B
Capacity	460 tons (per week)	560 tons (per week)
Labor Cost	\$26/ton	\$21/ton

The packaged almonds are sold at \$50/ton to distributors. The almond packaging company can sell all it can produce at this price.

The objective is to find the best mixture of the quantities supplied by the three suppliers to the two processing plants so that the company maximizes its profit. All elements of the problem are on a per weekly basis.

We need to determine how much should be purchased from each of the three suppliers and targeted to each of the two processing plants. To this end, we let x_{ij} be the number of tons purchased from grower i ($i = 1,2,3$ for S1, S2 and S3, respectively) and targeted at plant j ($j = 1$ for Plant A and $j = 2$ for Plant B) where $x_{ij} \geq 0, i = 1,2,3;\ j = 1,2.$

In terms of capacity, we have the following constraints:

$$x_{11} + x_{21} + x_{31} \leq 460$$

$$x_{12} + x_{22} + x_{32} \leq 560$$

In terms of supply, we have the following constraints:

$$x_{11} + x_{12} \leq 200$$

$$x_{21} + x_{22} \leq 310$$

$$x_{31} + x_{32} \leq 420$$

The cost is

$$x_{11}(11 + 3 + 26) + x_{21}(10 + 2 + 26) + x_{31}(9 + 6 + 26) +$$

$$x_{12}(11 + 3.5 + 21) + x_{22}(10 + 2.5 + 21) + x_{32}(9 + 4 + 21)$$

$$= 40\,x_{11} + 38\,x_{21} + 41\,x_{31} + 35.5\,x_{12} + 33.5\,x_{22} + 34\,x_{32}$$

The revenue is

$$50 \sum_{i=1}^{3} \sum_{j=1}^{2} x_{ij}$$

The profit (and objective function to be maximized) is the revenue minus the cost, i.e.,

$$10\,x_{11} + 12\,x_{21} + 9\,x_{31} + 14.5\,x_{12} + 16.5\,x_{22} + 16\,x_{32}$$

Using the online LP application at https://linprog.com/en/main-simplex-method, we determine a maximum profit of \$13,070 per week when

$$x_{11} = 200, \qquad x_{21} = 170, \qquad x_{31} = 0, \qquad x_{12} = 0, \qquad x_{22} = 140, \qquad x_{32} = 420$$

10.3.2 Blending Problem

The blending problem entails the combining of various ingredients ($A_1, A_2, \dots, A_n$) into a mixture. Each ingredient has varying amounts of the components $B_1, B_2, \dots, B_m$. Each unit of the mixture requires a specific amount of each component. The cost of each ingredient is known and the amount of each component per ingredient is also known. The task is to minimize the cost of producing the mixture.

We make the assignments:

- x_i is the number of units of ingredient A_i used in the mixture
- c_i is the cost of one unit of ingredient A_i
- b_i is the number of units of component B_i required in the mixture
- a_{ij} is the number of units of component B_i in ingredient A_j.

The task is to minimize the cost of producing a unit of the mixture. So, the objective function to be minimized is

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

subject to the constraints

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \geq b_1$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \geq b_2$$
$$\dots$$
$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \geq b_m$$
$$. . .$$

As an example, consider a diet that is to include all 9 of the Essential Amino Acids (EAAs) for humans. The EAAs are to come from 5 different protein-rich foods. The unit cost of each food is known and the amount of each EAA in each food is also known. The task is to minimize the cost of the foods in the diet of a 75 kg person while meeting EAA requirements (as described below).

According to the World Health Organization (WHO) the following amounts of each EAA are required per kilogram (kg) of body weight. Although methionine, but not cysteine, is considered an essential amino acid, the addition of dietary cysteine “spares” and reduces the metabolic requirement for methionine, which is why the WHO requirements lists these two amino acids together. For similar reasons, phenylalanine and tyrosine are combined.

Essential Amino Acid	**milligrams (mg) per kg of body weight**	**mg required for 75 kg person**
Histidine	10	750
Isoleucine	20	1500
Leucine	39	2925
Lysine	30	2250
Methionine (plus Cysteine)	15	1125
Phenylalanine (plus Tyrosine)	25	1875
Threonine	15	1125
Tryptophan	4	300
Valine	26	1950

The following table was constructed from data provided by the United States Food and Drug Administration (FDA), see https://fdc.nal.usda.gov/fdc-app.html#/. The number below each food item is the FDC ID (not important here unless the reader wants to trace back to the exact food item in the FDA food database).

		mg of EAA per 100 grams				
		Eggs 748967	**Ground Beef** 174036	**Chicken** 171477	**Lentils** 172420	**Rice** 169711
EAA	Histidine	283	558	963	693	47
	Isoleucine	616	759	1640	1060	87
	Leucine	1050	1340	2330	1790	167
	Lysine	832	1420	2640	1720	73
	Methionine (plus Cysteine)	803	619	1256	532	88
	Phenylalanine (plus Tyrosine)	1172	1198	2280	1878	175
	Threonine	594	665	1310	882	72
	Tryptophan	166	87	362	221	23
	Valine	734	844	1540	1220	123

The following are prices for the five food items (based on an informal survey of supermarket prices in March of 2022). Two eggs weigh about 100 grams. The cost of the rice is based on a 20 pound (9.072 kg) bag of white rice.

	Eggs	**Ground Beef**	**Chicken**	**Lentils**	**Rice**
Cost per 100 grams	$0.58	$1.32	$1.43	$0.55	$0.11

Let x_1, x_2, x_3, x_4, x_5 be the number of units (100 grams) of eggs, ground beef, chicken, lentils and rice, respectively. Our task is to minimize

$$.58\ x_1 + 1.32\ x_2 + 1.43\ x_3 + .55\ x_4 + .33\ x_5$$

subject to the following constraints

$$283\ x_1 + 558\ x_2 + 963\ x_3 + 693\ x_4 + 47\ x_5 \geq 750$$
$$616\ x_1 + 759\ x_2 + 1640\ x_3 + 1060\ x_4 + 87\ x_5 \geq 1500$$
$$1050\ x_1 + 1340\ x_2 + 2330\ x_3 + 1790\ x_4 + 167\ x_5 \geq 2925$$
$$832\ x_1 + 1420\ x_2 + 2640\ x_3 + 1720\ x_4 + 73\ x_5 \geq 2250$$
$$803\ x_1 + 619\ x_2 + 1256\ x_3 + 532\ x_4 + 88\ x_5 \geq 1125$$
$$1172\ x_1 + 1198\ x_2 + 2280\ x_3 + 1878\ x_4 + 175\ x_5 \geq 1875$$
$$594\ x_1 + 665\ x_2 + 1310\ x_3 + 882\ x_4 + 72\ x_5 \geq 1125$$
$$166\ x_1 + 87\ x_2 + 362\ x_3 + 221\ x_4 + 23\ x_5 \geq 300$$
$$734\ x_1 + 844\ x_2 + 1540\ x_3 + 1220\ x_4 + 123\ x_5 \geq 1950$$

Using the online LP application at https://linprog.com/en/main-simplex-method, the minimum cost per 100 grams is $1.03, with

$$x_1 = .52, \qquad x_2 = 0, \qquad x_3 = 0, \qquad x_4 = 1.33, \qquad x_5 = 0$$

10.3.3 Transportation Problem

The transportation problem entails a market consisting of a given number of providers and consumers of a commodity, and a network of routes between the providers and the consumers. The transportation network is given by a set A of routes, where $(i, j) \in A$ means there is a route connecting the provider i and the consumer j.

We make the following assignments:

- let c_{ij} be the unit shipment cost over route (i, j),
- let s_i be the amount of the given commodity that is available from provider i
- let d_j be the demand of consumer j
- let x_{ij} be number of units of the commodity shipped over route $(i, j) \in A$

The task here is to minimize the transportation costs subject to the supply and demand constraints, i.e.,

Minimize

$$\sum_{(i,j)\in A} c_{ij}\ x_{ij}$$

Subject to

$$\sum_{j} x_{ij} \leq s_i, \forall i$$

$$\sum_{i} x_{ij} \geq d_j, \forall j$$

The first constraint says that the supply s_i should be greater than the sum of the demands on provider i. The second constraint says that the demand d_j should be less than what can be supplied by the providers to consumer j.

. . .

A producer of wood chips has two sites (S1 and S2). S1 can produce a maximum of 350 tons of wood chips per week and S2 can produce a maximum of 550 tons of wood chips per week. The producer has three customers, with the following requirements

- Customer 1 (C1) requires at least 275 tons of wood chips per week
- Customer 2 (C2) requires at least 325 tons of wood chips per week
- Customer 3 (C3) requires at least 300 tons of wood chips per week.

The following table shows the transportation costs (in dollars per ton of wood chips) between supplier and customer locations.

	C1	**C2**	**C3**
S1	17	22	15
S2	18	16	12

Let $x_{ij} \geq 0$ represent the amount in tons to be shipped weekly from Si to Cj , for $i = 1{,}2$ and $j = 1{,}2{,}3$.

The task is to minimize the shipping costs

$$17x_{11} + 22x_{12} + 15x_{13} + 18x_{21} + 16x_{22} + 12x_{23}$$

subject to the supply constraints

$$x_{11} + x_{12} + x_{13} \leq 350$$
$$x_{21} + x_{22} + x_{23} \leq 550$$

and the demand constraints

$$x_{11} + x_{21} \geq 275$$
$$x_{12} + x_{22} \geq 325$$
$$x_{13} + x_{23} \geq 300$$

Using the previously mentioned online LP app, we get a minimum weekly shipping cost of $13,700 with

$$x_{11} = 275, \quad x_{12} = 0, \quad x_{13} = 75, \quad x_{21} = 0, \quad x_{22} = 325, \quad x_{23} = 225$$

10.3.4 Dynamic Planning Problem

All the examples that we have seen so far involve a single time period. It is also possible to formulate LP problems that cover several time periods, with interrelationships among the periods.

Consider the example of a garden shed seller. In the current month (April), the seller has 25 sheds in inventory. In each of the next 3 months, the seller can buy from the shed manufacturer up to 65 sheds, and can sell up to 100 sheds to customers. The buy (from manufacturer) and sell (to customer) prices are as shown in the following table:

	Buy	**Sell**
May	60	90
June	65	110
July	68	105

Further, the shed seller can store up to 45 sheds at a cost of $7/shed/month. Assume the seller can sell all sheds that are offered to customers in a given month. The task is to determine the optimal plan for buying, selling, and storing the sheds.

Make the following assignments:

- let b_i be the number of shed purchased from the manufacturer in month i, where $i = 1$ corresponds to May, $i = 2$ corresponds to June and $i = 3$ corresponds to July
- let s_i be the number of sheds sold to customers in month i
- let h_i be the number of sheds held in storage during month i

With the above notations in hand, we want to maximize

$$(90s_1 + 110s_2 + 105s_3) - (60b_1 + 65b_2 + 68b_3) - 7(h_1 + h_2 + h_3)$$

subject to following input/output constraints

$$25 + b_1 = s_1 + h_1$$
$$h_1 + b_2 = s_2 + h_2$$
$$h_2 + b_3 = s_3 + h_3$$

and the additional constraints

$$0 \leq b_i \leq 65, \quad 0 \leq s_i \leq 100, \quad 0 \leq h_i \leq 45, \text{ for } i = 1,2,3$$

Using a different online LP solver this time (see http://online-optimizer.appspot.com/), we get the optimal solution of $9995 with

$$b_1 = b_2 = b_3 = 65, s_1 = 45, s_2 = 100, s_3 = 75, h_1 = 45, h_2 = 10, h_3 = 0$$

10.4 Methods of Solution

[Author's Remark: In this book, I made the decision to focus on LP problem formulation as this will be of most immediate use to the reader. However, there is a significant body of knowledge concerning solution methods for LP problems. In what follows, we give a very brief summary.**]**

Various methods for solving linear programming problems have been developed. One set of methods are known as basis exchange algorithms where the exterior of the feasible is searched for an optimal solution. The best known of the basic exchange algorithms is the simplex method. The other set of methods are known as interior point algorithms where one starts with an interior point of the feasible solution (which is not optimal) and then iterates until an optimal solution is found.

10.4.1 Simplex Method

As noted in Section 10.1, a system of linear inequalities defines a polytope as a feasible region. The simplex method begins at a starting vertex and moves along the edges of the polytope until it reaches the vertex of the optimal solution (as illustrated in Figure 27).

Credit to Victor Treushchenko for Figure 27, see https://commons.wikimedia.org/wiki/File:Simplex-description-en.svg.

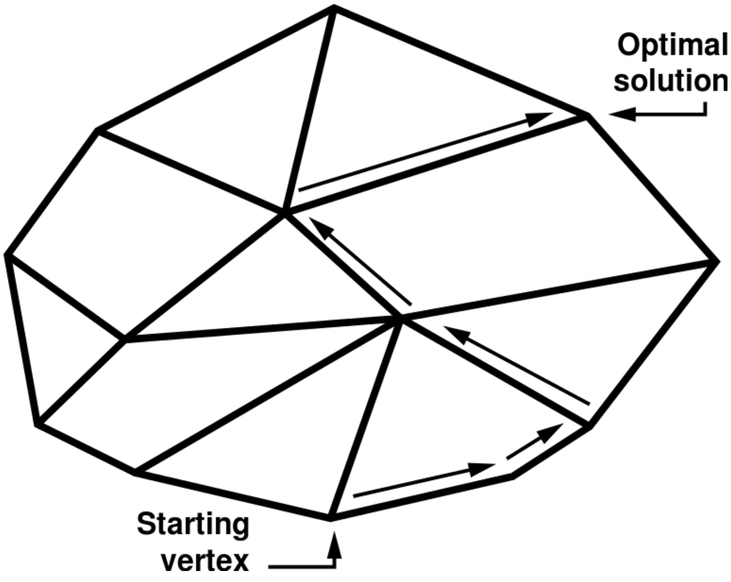


Figure 27. Graphic for simplex method in 3 dimensions

The simplex method was first developed by George Dantzig. The Wikipedia article on this topic [78] provides the following history:

> George Dantzig worked on planning methods for the US Army Air Force during World War II using a desk calculator. During 1946 his colleague challenged him to mechanize the

> planning process to distract him from taking another job. Dantzig formulated the problem as linear inequalities inspired by the work of Wassily Leontief, however, at that time he didn't include an objective as part of his formulation. Without an objective, a vast number of solutions can be feasible, and therefore to find the "best" feasible solution, military-specified "ground rules" must be used that describe how goals can be achieved as opposed to specifying a goal itself. Dantzig's core insight was to realize that most such ground rules can be translated into a linear objective function that needs to be maximized. Development of the simplex method was evolutionary and happened over a period of about a year.

10.4.2 Interior Point Methods

In contrast to the simplex method, which finds an optimal solution by traversing the edges between vertices on the polytope comprising the feasible region for an LP problem, interior-point methods move through the interior of the feasible region.

From the Wikipedia article entitled “Interior-point method” [79]:

> An interior point method was discovered by Soviet mathematician I. I. Dikin in 1967 and reinvented in the U.S. in the mid-1980s. In 1984, Narendra Karmarkar developed a method for linear programming called Karmarkar's algorithm, which runs in provably polynomial time and is also very efficient in practice. It enabled solutions of linear programming problems that were beyond the capabilities of the simplex method. Contrary to the simplex method, it reaches a best solution by traversing the interior of the feasible region. The method can be generalized to convex programming based on a self-concordant barrier function used to encode the convex set.

The YouTube video “Interior Point Method Demonstration” [80] provides a short (but somewhat technical) presentation of the differences between the simplex method and interior point methods.

10.5 LP Software

There are many free online applications that can solve relatively small LP problems. For example, see the online sites used to solve the examples in this section (i.e., https://linprog.com/en/main-simplex-method and http://online-optimizer.appspot.com/).

The CVXOPT package for the Python programming language can solve a variety of optimization problems, including linear programming (see https://cvxopt.org/userguide/coneprog.html#linear-programming).

An extensive list of free and commercial (for fee) optimization software (not restricted to LP) can be found at https://en.wikipedia.org/wiki/List_of_optimization_software.

11 Integer Programming

11.1 Overview

Prerequisites: algebra, summation notation, inequalities, linear programming (from previous section)

Integer programming is similar to linear programming in the sense that a linear function in several variables is to be optimized over some feasible region defined by linear constraints. The difference is that, in the case of integer programming, the variables are restricted to integer values. If there is a mixture of integer and real-valued variables, the problem is known as **mixed-integer programming**. The case where the integer variables are restricted to be 0 or 1 arises often. Such problems are called pure (mixed) 0-1 programming problems or pure (mixed) binary integer programming problems.

Recall the barbell production problem from the previous section. Our LP solution was to produce $333\frac{1}{3}$ of the smaller barbell set. We had no issue in just rounding the number to 333. The small excess of iron from the $\frac{1}{3}$ of a barbell set could easily be set aside for the next month. On the other hand, suppose the problem entailed the production of a more expensive product such as a house and we got a solution of 12.5 houses of one type and 5.7 houses of another type. In this case, the decision to round up or down is more difficult, and it would be better to solve the problem for whole number solutions only.

This restriction to nonnegative integers may seem harmless, but in reality it has significant implications. On the one hand, modeling with integer variables has shown to be useful well beyond restrictions to integral production quantities. For example, the use of integer variables allows for the modeling of logical requirements, fixed costs, sequencing and scheduling requirements.

The statement of an integer programming problem looks very similar to that of an LP problem. For example, consider the following integer programming problem:

Maximize

$$2y + x$$

Subject to

$$-x + y \leq 1$$

$$2x + 3y \leq 12$$

$$3x + 2y \leq 12$$

$$x \geq 0, \qquad y \geq 0$$

$$x, y \in \mathbb{Z}$$

The last constraint is the only difference from an LP problem. The constraint is shorthand for saying that x and y are elements of the set of integers.

The black dots in Figure 28 are the feasible integer points, and the gray area is the smallest convex polyhedron that contains all the feasible integer points. The convex polyhedron is within the feasible region defined by the problem minus the integer constraint. The goal is to find the feasible point which yields the largest value for the objective function. Since we only have 11 feasible points, it is easy to check each one (rather than using an algorithm). In this case, the maximum value of $2y + x$ is 6 and it occurs at point $(2,2)$.

If we remove the integer constraint (in effect, leaving us with a LP problem), the maximum value of the object function is $7\frac{2}{5}$ and it occurs at point $(\frac{9}{5}, \frac{14}{5})$.

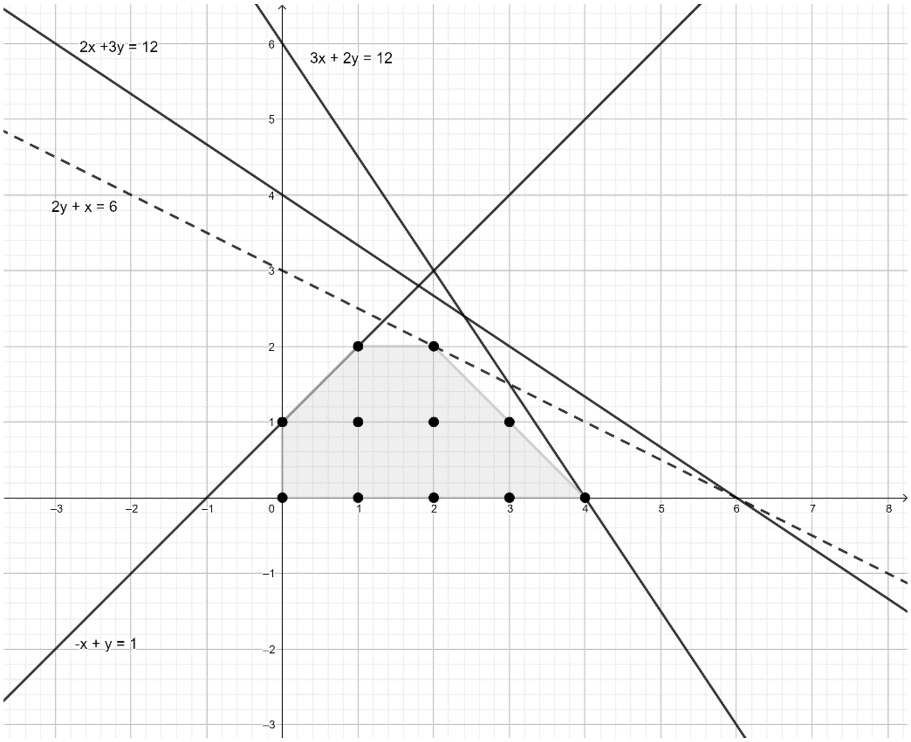


Figure 28. Integer programming example

11.2 Applications

11.2.1 Knapsack Problem

In the knapsack problem, a hiker has a collection of items that could be placed in a knapsack. However, there is a weight issue and the hiker cannot carry the full set of items. Further, each item has a value, e.g., water and a fire starter may have relatively high values compared to a digital music player which is nice to have but not necessary for the hike. The goal is to stay within a given weight constraint for the knapsack, and to optimize the value of the included items.

As noted in the Wikipedia article on this topic [81]:

> Knapsack problems appear in real-world decision-making processes in a wide variety of fields, such as finding the least wasteful way to cut raw materials, selection of investments and portfolios, selection of assets for asset-backed securitization, and generating keys for the Merkle–Hellman and other knapsack cryptosystems.

In what follows, we formulate the knapsack problem as a 0-1 integer programming problem. Assume that we have n possible items to place in a given knapsack.

- For item i, let x_i represent whether the item is or is not included in the knapsack. If $x_i = 1$, the item is included. If $x_i = 0$, the item is not included.
- Let c_i be the value of item i with respect to the given hike.
- Let a_i be the weight of item i.
- Let b be the maximum amount of weight for all the items included in the knapsack.

Our problem can be stated as

maximize

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

subject to

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n \leq b$$

$$x_i \in \{0,1\}, \qquad i = 1,2,\ldots,n$$

. . .

A hiker has five items that could be included in her knapsack for a hike. She has decided that the maximum weight for the items to be included in the knapsack is 15 pounds. The details for each potential item are listed in the following table.

Item	Weight	Value
1	5	5
2	4	3
3	7	6
4	6	6
5	2	2

Our task is to

maximize

$$5x_1 + 3x_2 + 6x_3 + 6x_4 + 2x_5$$

subject to the constraints

$$5x_1 + 4x_2 + 7x_3 + 6x_4 + 2x_5 \leq 15$$

Using the 0-1 integer programming application at atozmath.com, we find that inclusion of items 1,2 and 4 yields the maximum value of 14.

11.2.2 Capital Budgeting

11.2.2.1 Concepts

Present value refers to the sum of money which if invested now at a given rate of interest will accumulate to a specified amount at a particular future date.

For example, what is the present value of \$25,000 three years from now assuming 5% interest compounded monthly? In other words, how much money needs to be invested now at 5% interest compounded monthly (with no additional investments or withdraws) to have \$25,000 in three years? In terms of the compound interest formula, the following needs to be solved for P

$$\$25{,}000 = P(1 + \frac{.05}{12})^{(12)(3)}$$

Solving this equation, we get $P = \$21{,}524.41$. Someone might do such a calculation if he or she wants to buy a \$25,000 car three years from now and wants to know how much to put into a given fixed interest investment (such as a government bond) right now.

For a given investment, let PV be the present value, FV be the future value, and assume there are s compounding periods with interest rate i per compounding period. We then have the following formula related PV and FV:

$$FV = PV(1 + i)^s$$

We can also solve the above equation for PV in terms of FV to get

$$\boldsymbol{PV = FV(1 + i)^{-s}}$$

With regard to capital budgeting, present value is used to make comparisons at the same focal date. For example, consider a restaurateur who has \$50,000 to invest. She can buy a pizza oven that she estimates will entail new revenue of \$15,000, \$20,000 and \$22,000 in the first three years of operation, or she can invest the money in a high-quality corporate bond that pays 4% per year (compounded annually). To make a comparison between the two options, we need to compute the present value of the profit amounts for the business supported by the pizza oven and add the three results. The present value for each profit amount is shown in the table below (assuming 4% per year, compounded annually). The sum of the present values of the anticipated profits from the pizza business is \$52,472.12. So, given the stated assumptions, the restaurateur would make more money with the investment in the pizza oven (and that is only considering 3 years into the future).

	15,000	20,000	22,000
Present Value	14,423.08	18,491.12	19,557.92

11.2.2.2 Capital Budgeting as an Integer Programming Model

The capital budgeting problem can be considered over various time periods, e.g., months, quarters, years. The goal is to determine the maximum profit an investor can attain using a combination of various investment options. Since the profits occur at different points in time, it is necessary to calculate the profits based on a common focal date (typically, but not necessarily, the present time).

The general problem is as follows. Over a prescribed planning period T, business has n possible projects in which it can invest, but not all can be selected because of budget limitations in each time period. We make the following assumptions:

- Project i has a present value of c_i and requires an investment of a_{ti} in time period t (for $t = 1,2, \dots, T$).
- The capital available in time period t is b_t.

- A project is selected in total or not, i.e., cannot do part of a project.
- The variable x_i represent the decision of project i.

The goal is to maximize the total present value subject to the budgetary constraints in each time period over the planning period T. In terms of mathematical notation, the problem is stated as follows:

Maximize

$$z = \sum_i c_i x_i$$

subject to

$$\sum_i a_{ti} x_i \le b_t, \qquad t = 1,2,\dots,T$$

$$x_i \in \{0,1\}, \qquad i = 1,2,\dots,n$$

The constraint $x_i \in \{0,1\}$ implies that a project is either selected or not. This is a pure binary integer programming problem.

11.2.2.3 Capital Budgeting Examples

A specialty steel company has $140,000 to invest in new equipment that would allow the company to offer additional products. The investment opportunities are as follows:

- Opportunity 1 requires an investment of $50,000 and has a present value of $80,000.
- Opportunity 2 requires $70,000 and has a present value of $110,000.
- Opportunity 3 requires $40,000 and has a present value of $60,000; and
- Opportunity 4 requires $30,000 and has a present value of $40,000.

Each of the present values are based on the expected additional profit gained by making and selling new products (made possible by the new equipment purchased with the investment money) one year from the present. So, there is only one time period in this example. Each opportunity is all or nothing, e.g., cannot do $\frac{1}{3}$ of Investment 2.

What selection of opportunities maximize total present value?

Let x_i be a binary variable (for $i = 1,2,3,4$) such that $x_i = 1$ if opportunity i is selected and $x_i = 0$ if investment i is not selected. Also, the numbers below are in units of $10,000. Our task is to

maximize

$$8x_1 + 11x_2 + 6x_3 + 4x_4$$

subject to the constraints

$$5x_1 + 7x_2 + 4x_3 + 3x_4 \le 14$$

$$x_i \in \{0,1\}, \qquad i = 1,2,3,4$$

There are only 16 possible points in the feasible region. Checking these point manually, we find that the maximum present value is 21, which is attained when $x_1 = 0$, $x_2 = 1$, $x_3 = 1$ and $x_4 = 1$.

We can add some complications to the problem, e.g.,

There are a number of additional constraints we might want to add. For instance, consider the following constraints:

- Only two opportunities are allowed.
- Opportunities 2 and 4 must be selected together, or neither shall be selected.
- Opportunities 1 and 3 cannot both be selected.

The additional constraints can be represented as

$$x_1 + x_2 + x_3 + x_4 \leq 2$$

$$x_2 - x_4 = 0$$

$$x_1 + x_3 \leq 1$$

For the revised problem, the maximum present value of 15 is attained when $x_1 = 0$, $x_2 = 1$, $x_3 = 0$ and $x_4 = 1$.

. . .

As an example of a capital budgeting problem over several periods, we consider a restaurateur who has four options for the purchase of capital items that are expected to enhance business. The investments associated with the opportunities span a period of 3 months. The restaurateur has $14,000, $12,000, and $15,000 to invest in months 1, 2, and 3, respectively. The opportunities are as follows:

- Opportunity 1 (low humidity refrigerators) requires investments of $5,000, $8,000, and $2,000 in months 1, 2, and 3, respectively, and has a present value of $8,000.
- Opportunity 2 (a pizza oven) requires an investment of $7,000 in month 1 and $10,000 in month 3, and has a present value of $11,000.
- Opportunity 3 (tables with large umbrellas for outdoor dining) requires an investment of $4,000 in month 2 and $6,000 in month 3, and has a present value of $6,000.
- Opportunity 4 (several food processing machines) requires investments of $3,000, $ 4,000, and $5,000 in months 1, 2 and 3, respectively, and has a present value of $4,000.

For each opportunity, the expected profits can be realized (anticipated) at various points during the planning horizon which could be longer than the 3 months investment periods. For each opportunity, the present value for each anticipated profit needs to be calculated and then added together. For this example, we provide the final result of the present value calculations.

Our task is to determine which of the opportunities should be selected to maximize present value, subject to the various constraints stated in the problem.

Let x_i be a binary variable (for $i = 1,2,3,4$) such that $x_i = 1$ if opportunity i is selected and $x_i = 0$ if investment i is not selected.

Stated in mathematical notation, our problem is to

maximize

$$8x_1 + 11x_2 + 6x_3 + 4x_4$$

subject to the constraints

$$5x_1 + 7x_2 + 0x_3 + 3x_4 \leq 14$$

$$8x_1 + 0x_2 + 4x_3 + 4x_4 \leq 12$$

$$2x_1 + 10x_2 + 6x_3 + 5x_4 \leq 15$$

$$x_i \in \{0,1\}, i = 1,2,3,4$$

Under the stated constraints, the maximum value of the objective function is 19, and that occurs when $x_1 = 1$, $x_2 = 1$, $x_3 = 0$ and $x_4 = 0$. The solution was computed with the 0-1 integer programming application at atozmath.com.

11.2.3 Set Covering

Take the set of numbers $A = \{1,2,3,4,5,6,7,8,9,10,11\}$ and the following collection of subsets of A

$$S = \{\{\mathbf{1},2,3,4\},\{1,\mathbf{2},3,5\},\{1,2,\mathbf{3},4,5,6\},\{1,3,\mathbf{4},6,7\},\{2,3,\mathbf{5},6,8,9\},\{3,4,5,\mathbf{6},7,9,10\},$$
$$\{4,6,\mathbf{7},10\},\{5,\mathbf{8},9,11\},\{5,6,8,\mathbf{9},10,11\},\{6,7,9,\mathbf{10},11\},\{8,9,10,\mathbf{11}\}\}$$

Set S was derived by creating the set of all adjacent blocks for each numbered block in Figure 29 below and shown in bold above. The union of all the elements (sets) in set S is necessarily A. What is the least number of elements from S that covers (includes) every element of A?

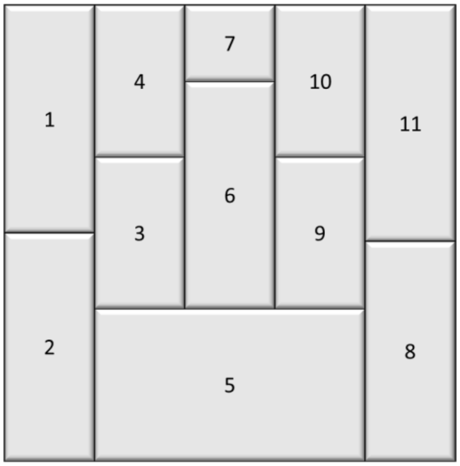


Figure 29. Set covering example

Let s_i represent the i^th^ element listed in set S, e.g., $s_3 = \{1,2,\mathbf{3},4,5,6\}$. Further, let x_i denote whether s_i is included in the covering set A (value of 1 means "included" and 0 means "not included").

We again have a 0-1 integer programming problem. Our task is to minimize

$$\sum_{i=1}^{11} x_i$$

subject to the constraints (with explanation below)

$$x_1 + x_2 + x_3 + x_4 + 0 \cdot x_5 + 0 \cdot x_6 + 0 \cdot x_7 + 0 \cdot x_8 + 0 \cdot x_9 + 0 \cdot x_{10} + 0 \cdot x_{11} \geq 1$$

$$x_1 + x_2 + x_3 + 0 \cdot x_4 + x_5 + 0 \cdot x_6 + 0 \cdot x_7 + 0 \cdot x_8 + 0 \cdot x_9 + 0 \cdot x_{10} + 0 \cdot x_{11} \geq 1$$

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + 0 \cdot x_7 + 0 \cdot x_8 + 0 \cdot x_9 + 0 \cdot x_{10} + 0 \cdot x_{11} \geq 1$$

$$x_1 + 0 \cdot x_2 + x_3 + x_4 + 0 \cdot x_5 + x_6 + x_7 + 0 \cdot x_8 + 0 \cdot x_9 + 0 \cdot x_{10} + 0 \cdot x_{11} \geq 1$$

$$0 \cdot x_1 + x_2 + x_3 + 0 \cdot x_4 + x_5 + x_6 + 0 \cdot x_7 + x_8 + x_9 + 0 \cdot x_{10} + 0 \cdot x_{11} \geq 1$$

$$0 \cdot x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + 0 \cdot x_8 + x_9 + x_{10} + 0 \cdot x_{11} \geq 1$$

$$0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + x_4 + 0 \cdot x_5 + x_6 + x_7 + 0 \cdot x_8 + 0 \cdot x_9 + x_{10} + 0 \cdot x_{11} \geq 1$$

$$0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 0 \cdot x_4 + x_5 + 0 \cdot x_6 + 0 \cdot x_7 + x_8 + x_9 + 0 \cdot x_{10} + x_{11} \geq 1$$

$$0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 0 \cdot x_4 + x_5 + x_6 + 0 \cdot x_7 + x_8 + x_9 + x_{10} + x_{11} \geq 1$$

$$0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 0 \cdot x_4 + 0 \cdot x_5 + x_6 + x_7 + 0 \cdot x_8 + x_9 + x_{10} + x_{11} \geq 1$$

$$0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 0 \cdot x_4 + 0 \cdot x_5 + 0 \cdot x_6 + 0 \cdot x_7 + x_8 + x_9 + x_{10} + x_{11} \geq 1$$

$$x_i \in \{0,1\}, i = 1,2, \dots ,11$$

If the first constraint does not hold, i.e., $x_1 = x_2 = x_3 = x_4 = 0$, then there is no way for element 1 of A to be covered. If the second constraint does not hold, i.e., $x_1 = x_2 = x_3 = x_5 = 0$, then there is no way for element 2 of A to be covered. Similar arguments hold for the other constraints.

The minimum solution is 3, and there are multiple ways to achieve this, e.g.,

$x_3 = x_7 = x_9 = 1$, with the other variables equal to 0

$x_1 = x_6 = x_8 = 1$, with the other variables equal to 0

$x_2 = x_6 = x_{11} = 1$, with the other variables equal to 0

$x_1 = x_5 = x_{10} = 1$, with the other variables equal to 0

The "problem" element is 7, which forces a third set to be included.

. . .

The set covering problem can be stated more generally, and does not need to be based on the geometric adjacency approach used in the previous example.

Let $A = \{1,2,\dots,n\}$ and let S be a collection (set) of subsets of A such that the union of the sets in $S = \{s_1, s_2, \dots, s_m\}$ equals A, i.e.,

$$A = \bigcup_{i=1}^{m} s_i$$

The problem is to find the minimum number of elements of S that cover (include) all of A. Define the variables $x_i,\ i = 1,2,\dots,m$ such that $x_i = 1$ if s_i is included in the covering of A, and $x_i = 0$ otherwise. The problem is to minimize

$\sum_{i=1}^{m} x_i$ (minimize the number of elements of S required to cover A)

subject to the constraints

$$\sum_{s:1\in s} x_s \geq 1$$

$$\sum_{s:2\in s} x_s \geq 1$$

$$\dots$$

$$\sum_{s:n\in s} x_s \geq 1$$

$$x_i \in \{0,1\}, \qquad i = 1,2,\dots,m$$

The first constraint says that for element 1 at least one element of S must include 1. The notation $s\colon 1 \in s$ means "summation over all $s \in S$ such at $1 \in s$. Similar statements can be made for the other constraints involving a summation.

11.2.4 Traveling Salesperson Problem

In the traveling salesperson problem, one is given a list of locations and the costs between each pair of locations. The costs are in terms of money, time, distance or some other measure of interest. The task is to determine the "shortest" possible (i.e., least cost) route that visits each location exactly once.

There are several formulations of the problem. For the discussion here, we make use of the Miller–Tucker–Zemlin (MTZ) formulation.

- The locations are labeled with the numbers $1,2,\dots,n$.
- The cost between location i and j is represented by c_{ij}.
- The binary variables x_{ij} are defined such $x_{ij} = 1$, if the route includes a link between locations i and j; otherwise, $x_{ij} = 0$.

The problem is to minimize

$$\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n} c_{ij}x_{ij}$$

The above expression represents the sum of the costs over all the links included in a route.

The first set of constraints below ensure that each location j has exactly one incoming link, and the second constraint ensure that each location i has exactly one outgoing link.

$$\sum_{i=1,i\neq j}^{n} x_{ij} = 1, \qquad j = 1,2,\dots,n$$

$$\sum_{j=1,j\neq i}^{n} x_{ij} = 1, \qquad i = 1,2,\dots,n$$

. . .

The following table shows the costs between 5 different locations in a traveling salesperson problem.

		To Location				
		1	**2**	**3**	**4**	**5**
From Location	**1**	-	4	2	5	1
	2	1	-	7	2	3
	3	3	5	-	3	1
	4	4	2	3	-	1
	5	3	5	6	4	-

Using the traveling salesperson application at https://www.easycalculation.com/operations-research/traveling-salesman-problem.php, we find the route of minimum cost to be

$$1 \to 3 \to 4 \to 2 \to 5$$

with the cost being $2 + 3 + 2 + 3 = 10$.

11.3 Relationship to Linear Programming

Given an integer programming problem of the form

Maximize (or Minimize) the objective function:

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

Subject to a finite number of constraints of the form:

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n \ (\leq, = \text{ or } \geq)\ b$$

$$x_i \in \{0,1\}, \qquad i = 1,2,\dots,n$$

there is a corresponding linear programming problem, known as the **linear relaxation** of the integer programming problem, which is formed by removing the integer restrictions, i.e., Maximize (or Minimize) the objective function:

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

Subject to a finite number of constraints of the form:

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n \ (\leq, = \text{ or } \geq) \ b$$

Since the linear relaxation problem has a subset of the constraints imposed on the original integer programming problem, the following statements are true:

- If the integer programming problem is a minimization problem, the minimum value of the objective function for the linear relaxation problem is less than or equal to the minimum value of the objective function for the integer programming problem.
- If the integer programming problem is a maximization problem, the maximum value of the objective function for the linear relaxation problem is greater than or equal to the maximum value of the objective functions for the integer programming problem.
- If the linear relaxation problem is infeasible, then so is the corresponding integer programming problem.
- If the linear relaxation problem is optimized by integer-valued variables, then that solution is feasible and optimal for the corresponding integer programming problem.

The various methods of solution for integer programming problems make use of the above relationships between the integer programming problem and its linear relaxation. A discussion of solutions methods for integer programming problems is beyond the scope of this book. Readers interested in such methods are referred to the following sources:

- The online article “Integer Programming” from ScienceDirect [82] contains several book excerpts that cover algorithms for the solution of integer programming problems.
- Part III of the book “Applied Integer Programming” [83] has extensive coverage of methods for the solution of integer programming problems.

12 Number Theory

12.1 Overview

Prerequisites: arithmetic, algebra, mathematical induction (Section 2), Diophantine equations (see Section 7)

The Wikipedia article on number theory provides the following definition:

> Number theory (or arithmetic or higher arithmetic in older usage) is a branch of pure mathematics devoted primarily to the study of the integers and integer-valued functions.

The adjective "pure" is arguable since there are many applications of number theory, e.g., number theory is used extensively in cryptography (see Section 13).

We've already covered one aspect of number theory, i.e., Diophantine equations (Section 7). In this section, we will focus on aspects of number theory that are used in cryptography.

12.2 Divisibility

Given any integer a and positive integer b, such that $b \neq 0$, there exists unique integers q (quotient) and r (remainder) such that

$$a = qb + r, \qquad 0 \leq r < b$$

where $|b|$ is the absolute value of b.

If a and b are negative integers, we solve the problem for $-a$ and $-b$ and then multiply the resulting equation by -1 (see the third and fourth examples below). In such cases, $-b < r \leq 0$.

The above is known as the **Quotient Remainder Theorem**. Some examples

- if $a = 8, b = 2$, we can write $8 = 4 \cdot 2 + 0$
- if $a = 7, b = 2$, we can write $7 = 3 \cdot 2 + 1$
- if $a = 3, b = 11$, we can write $3 = 0 \cdot 11 + 3$
- if $a = -9, b = -3$, then the conditions of the theorem do not hold, i.e., b is not a positive integer. However, note that $\frac{-9}{-3} = \frac{9}{3}$. We apply the theorem and get $9 = 3 \cdot 3 + 0$. If we multiply by -1, we get a solution to the original problem, i.e., $-9 = 3(-3) + 0$.
- if $a = -9, b = -7$, we first solve the problem $a = 9, b = 7$ to get $9 = 1 \cdot 7 + 2$ and then multiply by -1 to get a solution to the original problem, i.e., $-9 = 1(-7) - 2$
- if $a = -8, b = 7$, we can write $-8 = (-2)7 + 6$. This one is a bit tricky. Keep in mind that there is no restriction on q but r needs to be a positive integer or zero, and less than b.

An integer a is said to divide integer b, if there exists an integer k such that $a = kb$. This is represented as $a|b$. If a does not divide b, then we write $a \nmid b$.

Clearly, for any integer $a \neq 0$, we have $a|0$ and $a|a$. For any integer a (including 0), we have $1|a$.

Theorem 41. Division is transitive, i.e., if $a|b$ and $b|c$, then $a|c$.

> **Proof**: $a|b$ implies that $b = ka$ for some integer k, $b|c$ implies that $c = mb$ for some integer m. So, $c = mb = (mk)a$ which implies $a|c$.

Theorem 42. If $a|b$ and $a|c$, then $a|(xb + yc)$ for any integers x and y.

> **Proof**: $a|b$ and $a|c$ implies there exist integers k and m such that $b = ka$ and $c = ma$. So, we can write $xb + yc = xka + yma = (xk + ym)a$ and thus, $a|(xb + yc)$.

12.3 Prime Numbers

A **prime number** is a positive integer that is only divisible by 1 and itself. The first few primes are 2,3,5,7,11,13,17,19,23, …

Non-prime positive integers are known as **composite numbers**, e.g., $4,6,8,9,10,12,14,15$.

The next theorem plays a critical role in number theory. It says that every positive integer can be represented uniquely as the product of powers of prime numbers. For example,

- $782 = 2 \cdot 17 \cdot 23$
- $162{,}450 = 2 \cdot 3^2 \cdot 5^2 \cdot 19^2$

Theorem 43 (Fundamental Theorem of Arithmetic) Every positive integer $n > 1$ can be represented uniquely as a product of prime powers, i.e.,

$$n = p_1^{k_1} p_2^{k_2} \dots p_m^{k_m}$$

where $p_1, p_2, \dots, p_m$ are prime numbers.

Proof:

(Existence) We first prove that every integer greater than 1 is either prime or a product of primes. The proof is by strong induction [118]. The initial (or base) case is $n = 2$, which we know to be prime. By the strong induction hypothesis, assume the existence part of the theorem is true for all numbers greater than 1 and less than n. If n is prime, we are done. If n is composite, then by definition of composite there exists integers a and b, such $n = ab$, and $1 < a \le b < n$. By the induction hypothesis, a and b are each the products of primes, i.e., $a = p_1 p_2 \dots p_j$ and $b = q_1 q_2 \dots q_k$. Thus, we can write n product of primes, i.e., $n = p_1 p_2 \dots p_j q_1 q_2 \dots q_k$, which completes the induction.

(Uniqueness) We will use proof by contraction, i.e., assume s is the smallest positive integer that can be represented as the product of primes in more than one way. More formally, this means

$$s = p_1 p_2 \dots p_u = q_1 q_2 \dots q_v$$

Every p_i must be different from every q_j. If not (e.g., $p_i = q_j$), then $t = \frac{s}{p_i} = \frac{s}{q_j}$ would be an integer smaller than s that has two distinct prime factorizations (contradicting our initial assumption). Further, and without loss of generality (wlog), we can assume $p_1 < q_1$.

Let $P = p_2 p_3 \dots p_u = s/p_1$ and $Q = q_2 q_3 \dots q_v = s/q_1$. Since $p_1 < q_1$, we have that $Q < P$. Further,

$$s > s - p_1 Q = (q_1 - p_1)Q = p_1(P - Q)$$

Since we assumed that all positive integers less than s have unique prime factorizations (and the above expression is less than s), it must be that p_1 appears in the prime factorization of either $(q_1 - p_1)$ or Q.

- Since Q is less than s, it must have a unique factorization (i.e., $Q = q_2 q_3 \dots q_v$) and thus, p_1 cannot appear in the factorization of Q.
- If p_1 appears in the prime factorization of either $q_1 - p_1$ (i.e., it is a divisor of $q_1 - p_1$), then it must also be a divisor of $(q_1 - p_1) + p_1 = q_1$ which cannot be since p_1 and q_1 are distinct primes.

Either way, we arrive at a contradiction, and thus, our original assumption must be false. So, there cannot exist a smallest positive integer with more than one distinct prime factorization. ■

It has been known for several millennia that there are an infinite number of prime numbers. The earliest recorded proof is attributed to Euclid (circa 300 BC), as described in the following theorem.

Theorem 44. For any finite list of primes, there is at least one prime not in that list.

Proof: Take any finite set of primes (call the set A). Let P be the product of all the elements of set A and consider the number $P + 1$.

If $P + 1$ is a prime number, we are done since $P + 1$ would be a prime number not in A.

If $P + 1$ is not a prime number, then it must be divisible by some smaller prime p. If p is in A, then p divides both $P + 1$ and P. By Theorem 42, we have that $p|(P + 1) - P = 1$, but this is a contradiction and so, p is not in A. ■

Sometime later (circa 1860), Charles Hermite provided the following variation on the statement (and proof) concerning an infinite number of prime numbers.

Theorem 45, For an integer $n > 1$, if p is a prime divisor of $(n! + 1)$ then $p > n$ (which implies there are an infinite number of primes).

Proof: If $p \leq n$, then $p|n!$ but the condition of the theorem gives us that $p|(n! + 1)$. So, by Theorem 42, $p|(n! + 1) - n! = 1$, which is a contradiction. So, it must be that $p > n$. ■

There are other proofs concerning the infinite number of prime numbers. The two above are among the simplest.

The largest known prime number (at time of this writing) is $2^{82,589,933} - 1$, a number which has 24,862,048 digits when written in base 10 [119].

The following theorem gives us a way to determine the approximate number of prime numbers less than an given positive number x.

Theorem 46 (Prime Number Theorem) The number of primes less than x is approximately $f(x) = \frac{x}{\ln x}$ or more precisely,

$$\lim_{x \longrightarrow \infty} \frac{f(x)}{\frac{x}{\ln x}} = 1$$

The proof is fairly complex. Several proofs are described in the Wikipedia article “Prime number theorem” [123].

The theorem is useful if one is looking for large prime numbers in the context of cryptography. For example, let’s say that we needed to find 500-digit prime numbers for a given cipher. How many are there (if any)? In fact, there are an astronomical number of 500-digit primes. We can compute an approximation using Theorem 46, i.e.,

$$\frac{10^{500}}{\ln 10^{500}} - \frac{10^{499}}{\ln 10^{499}} \cong 7.81556 \times 10^{496}$$

The above calculation was done using Wolfram Alpha (at https://www.wolframalpha.com/). If you want to try this, the input is [10^500/ln(10^500)] - [10^499/ln(10^499)].

A **palindromic prime** is a prime number whose decimal expansion is a palindrome, see the definition at reference [120]. The first few palindromic primes are $2, 3, 5, 7, 11, 101, 131, 151, 181, 191, 313, 353, 373, 383$. The largest known palindromic prime is $10^{1888529} - 10^{944264} - 1$. However, it is not known whether there are an infinite number of palindromic primes [121].

An **emirp** (prime spelled backward) is a prime whose digits in reverse order also form a prime number. Palindromic primes are not considered to be emirps. The first few emirps are $13, 17, 31, 37, 71, 73, 79, 97, 107, 113, 149, 157, 167, 179, 199$. The largest known emirp is $10^{10006} + (941992101 \times 10^{4999}) + 1$. However, it is not known if there are an infinite number of emirps [122].

12.4 GCD and LCM

The **Greatest Common Divisor** (GCD) of two integers is the largest positive number that evenly divides (i.e., no remainder) both. For example, the GCD of 6 and 15 is 3 since 3 is the greatest number that divides both 6 and 15. This is written as $\gcd(6,15) = 3$.

Some additional examples:

$$\gcd(3,7) = 1, \qquad \gcd(6,21) = 3, \qquad \gcd(-5,30) = 5$$

In the example on the right, note that the GCD is a positive number even though one of the numbers is negative. The GCD is always a positive number even if both numbers are negative, e.g., $\gcd(-49, -7) = 7$. When the GCD of two numbers is 1 (such as for the example on the left), the two numbers are said to be **relatively prime** (or **coprime**).

There are several methods for finding the GCD of two numbers. In the "brute force" approach, one tests all combinations (multiples) of divisors of each number. However, there is a better (more efficient) way to eliminate some of the possibilities, i.e., factor each number into primes and then look for common factors between the two numbers. For example, take 24 and 112. We can write 24 as $2^3 \cdot 3$ and 112 as $2^4 \cdot 7$. Given the factorization, it is easy to see that $\gcd(24,112) = 2^3 = 8$.

Some examples to try:

$$\gcd(36,42)\,, \qquad \gcd(1155, 1225)\,, \qquad \gcd(400,1100)$$

Answers can be checked using the Microsoft Excel GCD function or the Google Sheets GCD function.

One more example, which suggests a general approach:

$$\gcd(2^{17}5^{11}7^{47}, 5^{13}7^{43}17^{12})$$

The two numbers are already represented in their unique prime factorizations. The numbers are too big for most calculators or spreadsheets but you should be able to easily read off the answer, i.e., just take the smallest power of each factor present in both numbers. In this case, the answer is $5^{11}7^{43}$. Note that neither powers of 2 nor powers of 17 appear in both numbers and so, they are not included in the GCD. In general, to calculate the GCD of two numbers, we first determine the prime factorization of two numbers, and then select the smallest power of each factor present in both numbers. More formally, if the unique prime factorizations of integers a and b are

$$a \ = p_1^{k_1} p_2^{k_2} \dots p_m^{k_m}$$

and

$$b = p_1^{l_1} p_2^{l_2} \dots p_m^{l_m}$$

where $k_i \geq 0$ and $l_i \geq 0$, then the GCD of a and b is

$$p_1^{\min{(k_1,l_1)}} p_2^{\min{(k_2,l_2)}} \dots p_m^{\min{(k_m,l_m)}}$$

The Euclidean algorithm [124] is another approach used to find the GCD. This approach entails an iterative use of the quotient remainder theorem. While the approach above (using prime factorization) is more intuitive, the Euclidean algorithm is better suited for automation via a computer program.

Theorem 47. GCDs exhibit the following properties:

i. $\gcd(a,b,c) \ = \ \gcd(a, \gcd(b,c))$ and with recursive application of this property, we can find the GCD of any number of integers.

ii. If m is a positive integer, then $\gcd(ma, mb) \ = \ m \cdot \gcd(a,b)$.

iii. If m is a positive common divisor of a and b, then $\gcd\left(\frac{a}{m}, \frac{b}{m}\right) = \frac{\gcd(a,b)}{m}$.

iv. If $a|b$, then $\gcd(a,b) = a$.

For an extensive list of GCD properties, see “Greatest common divisor – properties” [125].

. . .

The **Least Common Multiple** (LCM) of two integers is the smallest number that is the product of both numbers. For example, the LCM of 7 and 11 is 77 since 77 is the smallest number which is a multiple of both numbers. This is written as $\text{lcm}(7,11) = 77$.

Some additional examples:

$$\text{lcm}(4,24) = 24, \qquad \text{lcm}(6,21) = 42, \qquad \text{lcm}(24,60) = 120$$

The GCD and LCM are related by the following formula,

$$\text{lcm}(x,y) = \frac{x \cdot y}{\gcd(x,y)}$$

For example, the $\gcd(24,60) = 12$ and $24 \cdot 60 = 1440$. So, from the equation above, we have $\text{lcm}(24,60) = \frac{1440}{12} = 120$.

A key principle in mathematics is the reduction of a problem to one or more simpler problems. For the task at hand, i.e., finding the LCM of two numbers, one approach is to reduce the problem to one we already know how to solve. The idea is to first compute the GCD and then divide into $x \cdot y$.

Another approach is to use a method similar to the one we developed for the GCD. Again, the idea is to first determine the prime factorizations for the two numbers and then use this information to determine the LCM.

Consider the problem of finding the LCM of $x = 2^{17}5^{11}7^{47}$ and $y = 5^{13}7^{43}17^{12}$. The answer is $2^{17}5^{13}7^{47}17^{12}$. If we included any of the prime factors at a smaller power (exponent), e.g., $2^{\mathbf{16}}5^{13}7^{47}17^{12}$, then we would no longer have a common multiple. The general answer is to take the largest power of each prime factor of either number.

Try using the factorization method to solve the following:

$$\operatorname{lcm}(800, 560), \qquad \operatorname{lcm}(1265, 3025), \qquad \operatorname{lcm}(187, 51)$$

Check your answer with the LCM function in either Microsoft Excel or Google Sheets.

12.5 Modular Arithmetic

12.5.1 Concept

For a given positive integer n, it is possible to divide all integers into n partitions known as **congruence** classes. Relative to the selection of n, two integers a and b are in the same congruence class if $a - b = n$. In this case, we say that a is congruent to b modulo n and write this as

$$a \equiv b \pmod{n}$$

If we take $n = 5$, there are five congruence classes, i.e.,

$$\overline{0} = \{\dots, -10, -5, 0, 5, 10, \dots\}$$
$$\overline{1} = \{\dots, -9, -4, 1, 6, 11, \dots\}$$
$$\overline{2} = \{\dots, -8, -3, 2, 7, 12, \dots\}$$
$$\overline{3} = \{\dots, -7, -2, 3, 8, 13, \dots\}$$
$$\overline{4} = \{\dots, -6, -1, 4, 9, 14, \dots\}$$

In the above example and in general (for any value of n), each integer is in one and only one congruence class.

The set of congruence class modulo n, i.e., $\{\overline{0}, \overline{1}, \overline{2}, \dots, \overline{n-1}\}$ is denoted by the symbol $\mathbb{Z}_n$.

12.5.2 Basic Operations

Addition is modulo n. For example, if $n = 9$, then $\overline{4} + \overline{7} = \overline{2}$ since

$$4 + 7 \equiv 11 \equiv 2 \pmod{9}$$

In the above equation and in general, we usually only write the modulo once at the end.

Subtraction works in an analogous way. For example,

$$17 - 24 \equiv -7 \equiv 3 \pmod{10}$$

For multiplication, just multiply the integers and then reduce relative to the associated modulo. For example,

$$6 \cdot 19 \equiv 114 \equiv 2 \pmod{7}$$

Note that the remainder of 114 divided by 7 is 2 and thus, the above result.

The following theorem is useful in combining modular equations.

Theorem 48. If $a \equiv b\ (mod\ n)$ and $c \equiv d\ (mod\ n)$, then $ac \equiv bd\ (mod\ n)$.

Proof: a ≡ b $(mod\ \mathrm{n})$ implies $a = b + km$ for some integer k, and c ≡ d $(mod\ \mathrm{n})$ implies $c = d + lm$ for some integer l. Multiplying the two equations gives us $ac = bd + blm + dkm + klm^2 = bd + m(bl + dk + klm)$ which implies ac ≡ bd $(mod\ \mathrm{n})$.

Division is not as straightforward as the previously discussed operations.

For example, solve for x in the equation $7x \equiv 2 \pmod{10}$.

- One approach is to note that $7x \equiv 42 \pmod{10} \equiv 6 \cdot 7 \pmod{10}$ is equivalent to the original equation since $42 \equiv 2 \pmod{10}$. We can then divide both sides of the equation 7 to get the solution, i.e., $x \equiv 6 \pmod{10}$.
- Another approach is to determine the multiplicative inverse of $7 \pmod{10}$, i.e., the number when multiplied times 7 is $1 \pmod{10}$. By trying a few values, we see that $3 \cdot 7 \equiv 21\ \equiv 1 \pmod{10}$. Thus, 3 is the multiplicative inverse of 7 modulo 10, and we can multiple both sides of $7x \equiv 2 \pmod{10}$ to get $3 \cdot 7x \equiv x \equiv 2 \cdot 3 \equiv 6 \pmod{10}$.
- We cannot directly divide 2 by 7 to get the repeating decimal .285714 285714 285714 … since this number is not an element of $\mathbb{Z}_{10}$.

In general, we would like to know the conditions under which equations of the form $ax \equiv b \pmod{m}$ have solutions, and the form of the solutions when such exist. In basic algebra (with real numbers), this is a simple division problem, i.e., if $a \neq 0$, then the equation $ax = bm$ has the unique solution $x = \frac{b}{a}m$. In modular arithmetic, the problem is not so simple. To solve this problem, we need the following three results.

Theorem 49. The integer x is a solution of $ax \equiv b\ (mod\ m)$ if and only if there exists an integer y such that $ax - my = b$.

Proof: If x is a solution to $ax \equiv b \pmod{m}$, then there exists y such that $ax = b + my$ (by definition of modulo m).

Going in the other direction, if there exists y such that $ax - my = b$, then $ax = b + my$ which implies (by definition of modulo m) that $ax \equiv b \pmod{m}$. ■

Theorem 50. If a, b and c are positive integers such that $gcd(a, b) = 1$ and $a|bc$, then $a|c$.

Proof: By Theorem 14, there exists integers x and y that satisfy the equation $ax + by = 1$. Multiplying both sides of the equation by c, we get

$$acx + bcy = c$$

Clearly, $a|acx$. Further, $a|bcy$ since we are given that $a|bc$. Thus, $a|\ acx + bcy = c$. ■

Theorem 51. If a, b, c, and m are integers such that $m > 0, d = gcd(c, m)$, and $ac \equiv bc(mod\ m)$, then $a \equiv b(mod\,\frac{m}{d})$.

Proof:

If $ac \equiv bc(\text{mod } m)$, then we have $m|(ac - bc) = c(a - b)$, i.e., there exist an integer k such that $c(a - b) = km$. Next, divide both sides by d to get

$$\left(\frac{c}{d}\right)(a - b) = k\left(\frac{m}{d}\right)$$

By Theorem 47 (iii), $\gcd\left(\frac{m}{d}, \frac{c}{d}\right) = \frac{1}{d}\gcd(m, c) = 1$. So, by Theorem 50, we have that $\frac{m}{d}|(a - b)$. Thus, $a \equiv b(\text{mod}\,\frac{m}{d})$. ■

We are now in a position to prove the conditions under which the equation $ax \equiv b\ (mod\ m)$ has solutions.

Theorem 52. Let a, b and m be integers such that $m > 0$, and $gcd(a, m) = d$. If $d \nmid b$ (i.e., d does not divide b), then $ax \equiv b(mod\ m)$ has no solutions. If $d|b$, then $ax \equiv b(mod\ m)$ has exactly d incongruent solutions modulo m.

Proof: By Theorem 49, we know that x is a solution of $ax \equiv b(\text{mod } m)$ if and only if there is an integer y that satisfies the equation $ax - my = b$.

By Theorem 14, we can conclude that the equation has no solution if d ∤ b.

If $d|b$, then by Theorem 15, the equation $ax - my = b$ has infinitely many solutions of the form

$$x = x_0 + \left(\frac{m}{d}\right)t,\ \ y = y_0 + \left(\frac{m}{d}\right)t$$

where (x_0, y_0) is a particular solution of the equation. However, we seek incongruent solutions modulo m (i.e., not in the same congruence class).

If two solutions $x_1 = x_0 + (\frac{m}{d})t_1$ and $x_2 = x_0 + (\frac{m}{d})t_2$ are in the same congruence class, then

$$x_0 + \left(\frac{m}{d}\right)t_1 \equiv x_0 + \left(\frac{m}{d}\right)t_2\ (\text{mod } m)$$

which implies

$$\left(\frac{m}{d}\right)t_1 \equiv \left(\frac{m}{d}\right)t_2\ (\text{mod } m) \qquad (1)$$

Since $\gcd(a, m) = d$, it must be that $d|m$ which implies there exist k such that $m = kd$. Noting that $k = \frac{m}{d}$ is an integer, we can also say that $k = \frac{m}{d}|m$ which implies $\gcd\left(m, \frac{m}{d}\right) = \frac{m}{d}$ (follows from Theorem 47 iv). We can now apply Theorem 51 to equation (1) above to get

$$t_1 \equiv t_2\ (\text{mod } d)$$

In other words, two solutions are incongruent if $t_1 \not\equiv t_2 \ (\mathrm{mod}\ d)$. Thus, for each $t = 0,1, \dots , d - 1$ we get a different solution to $ax \equiv b(\mathrm{mod}\ m)$. ■

If $\gcd(a, m) = 1$, then Theorem 52 implies that $ax \equiv b(\mathrm{mod}\ m)$ has exactly one solution. We've already seen an example of this case, i.e., $7x \equiv 2 \ (\mathrm{mod}\ 10)$ which has the one solution $x \equiv 6 \ (\mathrm{mod}\ 10)$.

Consider the equation $6x \equiv 2 \ (\mathrm{mod}\ 9)$. We have that $\gcd(6,9) = 3 \nmid 2$ and by Theorem 52, the equation $6x \equiv 2 \ (\mathrm{mod}\ 9)$ has no solutions. This is easy to see since the only values (mod 9) that one can generate using different values of x in $6x$ are $0, 3$ and 6 (no way to get 2).

By Theorem 52, we know that the equation $5x \equiv 10 \ (\mathrm{mod}\ 25)$ has 5 solutions, since $\gcd(5,25) = 5$ and $5|10$. It is easy to see one solution, i.e., $x_0 = 2$. The complete set of solutions is given by the formula $x = x_0 + \left(\frac{m}{d}\right) t = 2 + 5t$, for $t = 0,1,2,3,4$. Thus, the solutions are 2,7,12,17,22.

12.5.3 Inverses and Fractions

To find the inverse of a number a modulo some integer m, we need find a number (call it a^{-1}) which when multiplied times a equals 1 modulo m, i.e.,

$$aa^{-1} \equiv 1 \ (\mathrm{mod}\ m)$$

By Theorem 49, this problem is equivalent to solving the Diophantine equation

$$aa^{-1} - my = 1$$

We already know how to solve such problems (see the discussion and examples following Theorem 15). The idea is to compute the continued fraction for, in this case, $\frac{a}{m}$.

If the continued fraction has an even number of terms n, then a solution to $aa^{-1} - my = 1$ is given by the penultimate convergent, i.e., (q_{n-1}, p_{n-1}).

If the continued fraction has an odd number of terms n, then a solution to $aa^{-1} - my = 1$ is given by the negative of penultimate convergent, i.e., $(-q_{n-1}, -p_{n-1})$.

. . .

The number a has an inverse modulo m when there is a solution to the equation $ax \equiv 1 \ (\mathrm{mod}\ m)$. By Theorem 52, we know that $ax \equiv 1 \ (\mathrm{mod}\ m)$ has exactly one solution when $gcd(a, m)\,|1$, i.e., when $gcd(a, m) = 1$.

. . .

For example, to find the inverse of 2324 modulo 7963, we solve the equation

$$2324a^{-1} - 7963y = 1$$

Using the calculator at https://www.alpertron.com.ar/CONTFRAC.HTM, we find that the continued fraction expansion of $\frac{2324}{7963}$ is [0,3,2,2,1,8,1,3,2,1,2] and the penultimate convergent is $\frac{863}{2957}$. Since there are an odd number of terms in the continued fraction expansion, the solution is $(-863, -2957)$, i.e.,

$$(2324)(-2957) - (7963)(-863) = -6872068 + 6872069 = 1$$

So, $a^{-1} \equiv (-2957)(\text{mod } 7963) \equiv 5006(\text{mod } 7963)$. As a check, we see that $2324 * 5006 \equiv 1 \ (\text{mod } 7963)$.

. . .

Let's try an example where the continued fraction expansion has an even number of terms. To find the inverse of 951 modulo 4241, we solve the equation

$$951a^{-1} - 4241y = 1$$

Using the calculator at https://www.alpertron.com.ar/CONTFRAC.HTM, we find that the continued fraction expansion of $\frac{951}{4241}$ to be $[0,4,2,5,1,2,12,2]$ with a penultimate convergent of $\frac{457}{2038}$. Since the number of terms in the continued fraction is even in this case, the solution is $(2038, 457)$, i.e.,

$$(951)(2038) - (4241)(457) = 1$$

So, $a^{-1} \equiv 2038(\text{mod } 4241)$. As a check, we see that $951 * 2038 \equiv 1 \ (\text{mod } 4241)$.

. . .

When finding the inverse of a number modulo n, first reduce the number modulo n (if the number is larger than n) and then find the inverse. For example, if we want to find the inverse of $30 \ (\text{mod } 7)$, first note that $2 \equiv 30 \ (\text{mod } 7)$. By trial and error, it is easy to see that $2^{-1} \equiv 4 \ (\text{mod } 7)$. Also, note that $4 \cdot 30 = 120 \equiv 1 \ (\text{mod } 7)$. In shorthand notation, we can write

$$30^{-1} \equiv 2^{-1} \equiv 4 \ (\text{mod } 7)$$

To be clear, 30^{-1} is the inverse of 30 modulo 7 and it is **not** (in this context) $\frac{1}{30}$.

. . .

The set of numbers $\mathbb{Z}_n$ has exactly n elements, i.e., $\{\overline{0}, \overline{1}, \overline{2}, \dots, \overline{n-1}\}$. However, we can state a convention concerning the meaning of fractional notation in $\mathbb{Z}_n$. The expression $\frac{a}{b} \ (\text{mod } n)$ is understood to mean $ab^{-1}(\text{mod } n)$ when b^{-1} exists, i.e., when $\gcd(b, n) = 1$.

For example, $\frac{73}{951} \ (\text{mod } 4241)$ is to be interpreted as $(951^{-1})(73) \ \equiv 2038 \cdot 73 \equiv 339(\text{mod } 4241)$.

12.5.4 Powers

Using a brute force method to compute high powers of a number in modular arithmetic can be prohibitive in terms of computational resources. However, there is an easier way using the following theorem.

Theorem 53. If $a \equiv b \ (\text{mod } n)$, then $a^k \equiv b^k \ (\text{mod } n)$ for any positive integer k.

Proof: $a \equiv b \ (\text{mod } n)$ if and only there exists an integer m such that $a = b + mn$. Applying the binomial theorem, we have

$$a^k = (b+mn)^k = \sum_{j=0}^{k} \binom{k}{j} b^k (mn)^{k-j}$$

$$a^k = b^k + \sum_{j=0}^{k-1} \binom{k}{j} b^k (mn)^{k-j}$$

$$a^k = b^k + n\left[\sum_{j=0}^{k-1} \binom{k}{j} b^k m^{k-j} n^{k-j-1}\right]$$

The last equation implies that $a^k \equiv b^k \pmod{n}$. ■

The above theorem can be applied multiple times to determine a very large power of a number modulo another number. For example, consider $7^{258} \pmod{235}$. If we first try to compute 7^{256}, we get an immense number, but we know the final solution is less than 235. Another approach is to start with

$$7^4 \equiv 2401 \equiv 51 \pmod{235}$$

Square both sides and continue the process

$$7^8 \equiv 51^2 \equiv 2601 \equiv 16 \pmod{235}$$
$$7^{16} \equiv 16^2 \equiv 256 \equiv 21 \pmod{235}$$
$$7^{32} \equiv 21^2 \equiv 441 \equiv 206 \pmod{235}$$
$$7^{64} \equiv 206^2 \equiv 42436 \equiv 136 \pmod{235}$$
$$7^{128} \equiv 136^2 \equiv 18496 \equiv 166 \pmod{235}$$
$$7^{256} \equiv 166^2 \equiv 27556 \equiv 61 \pmod{235}$$

We are almost there – just need to multiple the last equation above by $7^2 \equiv 49 \pmod{235}$ to get

$$7^{258} \equiv 61 \cdot 49 \equiv 2989 \equiv 169 \pmod{235}$$

This can be checked with Wolfram Alpha. If one enters "7^258 mod 235", the answer of 169 is returned very quickly.

12.5.5 Some Basic Theorems

The Chinese remainder theorem provides a mechanism for finding the solution to a system of modular arithmetic equations. The earliest known statement of the theorem is by the Chinese mathematician Sun-tzu in Chapter 3 of the Sunzi Suanjing [126] from the 3rd century AD.

Theorem 54 (Chinese Remainder Theorem) Let $n_1, n_2, \dots, n_k$ be a set of pairwise relatively prime integers such that each integer is greater than 1. If $a_1, a_2, \dots, a_k$ are integers such that $0 \le a_i < n_i$ for $i = 1,2,\dots,k$, then there exists a unique integer $x \ (mod\ N)$where $0 \le x < N = n_1 n_2 \dots n_k$ and

$$x \equiv a_1 \pmod{n_1}$$
$$x \equiv a_2 \pmod{n_2}$$
$$\dots$$
$$x \equiv a_k \pmod{n_k}$$

Proof: The proof is by construction, i.e., we show how to determine x.

(**Existence**) In notation, the condition about the n_i integers being pairwise relatively prime translates to

$$\gcd(n_i, n_j) = 1, \qquad i \neq j$$

Let $N_i = \frac{N}{n_i}$ for $i = 1,2, \dots, k$. In words, N_i is the product of all the n_j terms except for n_i. So, $\gcd(N_i\,, n_i) = 1$ by repeated application of Theorem 50.

By Theorem 14, there exists x_i and y_i such that

$$N_i x_i + n_i y_i = 1, \ \ i = 1,2, \dots, k$$

which implies

$$N_i x_i \equiv 1 \ (\text{mod}\ n_i), \ \ i = 1,2, \dots, k$$

Multiplying both sides of the by a_i, we get

$$N_i a_i x_i \equiv a_i \ (\text{mod}\ n_i), \ \ i = 1,2, \dots, k$$

Let $x = N_1 a_1 x_1 + N_2 a_2 x_2 + \cdots + N_k a_k x_k$.

Take a particular $j \in \{1,2, \dots, k\}$. Since $N_i \equiv 0 \ (\text{mod}\ n_j)$ for $i \neq j$, we have that

$$x \equiv N_j a_j x_j \ (\text{mod}\ n_j)$$

Since $N_j a_j x_j \equiv a_j \ (\text{mod}\ n_j)$, the above equation is equivalent to

$$x \equiv a_j \ (\text{mod}\ n_j)$$

and thus, we have constructed a solution to the system of equations stated in the theorem.

(**Uniqueness**) Assume that x and y are solutions to the system of equations stated in the theorem, i.e.,

$$\begin{array}{cc} x \equiv a_1 \ (\text{mod}\ n_1) & y \equiv a_1 \ (\text{mod}\ n_1) \\ x \equiv a_2 \ (\text{mod}\ n_2) & y \equiv a_2 \ (\text{mod}\ n_2) \\ & \dots \\ x \equiv a_k \ (\text{mod}\ n_k) & y \equiv a_k \ (\text{mod}\ n_k) \end{array}$$

Subtracting term by term, we get

$$x - y \equiv 0 \ (\text{mod}\ n_1)$$

$$x - y \equiv 0 \ (\mathrm{mod}\ n_2)$$

$$\dots$$

$$x - y \equiv 0 \ (\mathrm{mod}\ n_k)$$

The above implies that $n_i | (x - y)$, for $i = 1,2, \dots, k$ which, in turn, implies that

$$N = n_1 n_2 \dots n_k | (x - y)$$

and so,

$$x - y \equiv 0 \ (mod\ N)$$

which was to be proved. ■

As an example of the Chinese remainder theorem, consider the set of equations

$$x \equiv 2 \ (\mathrm{mod}\ 5)$$

$$x \equiv 3 \ (\mathrm{mod}\ 6)$$

$$x \equiv 4 \ (\mathrm{mod}\ 7)$$

We first note that 5,6 and 7 are pairwise relatively prime.

Next, do the following computations

- $N = n_1 n_2 n_3 = 5 \cdot 6 \cdot 7 = 210$
- $N_1 = 42, N_2 = 35, N_3 = 30$

The x_i terms are computed using the formula $x_i \equiv N_i^{-1} \ (\mathrm{mod}\ n_i)$

- $x_1 \equiv 42^{-1} \ (\mathrm{mod}\ 5) \equiv 2^{-1} \ (\mathrm{mod}\ 5) \equiv 3$
- $x_2 \equiv 35^{-1} \ (\mathrm{mod}\ 6) \equiv 5^{-1} \ (\mathrm{mod}\ 6) \equiv 5$
- $x_3 \equiv 30^{-1} \ (\mathrm{mod}\ 7) \equiv 2^{-1} \ (\mathrm{mod}\ 7) \equiv 4$

Now we have all the information required to compute x:

$$x \equiv N_1 a_1 x_1 + N_2 a_2 x_2 + N_3 a_3 x_3 \equiv 42 \cdot 2 \cdot 3 + 35 \cdot 3 \cdot 5 + 30 \cdot 4 \cdot 4 \equiv 1257 \equiv 207 \ (\mathrm{mod}\ 210)$$

Additional examples can be computed using the application at https://www.dcode.fr/chinese-remainder.

. . .

Fermat's Little Theorem and Euler's Theorem are important fundamental theorems in number theory with applicability to cryptography.

We need a preliminary result in order to prove Fermat's Little Theorem, i.e.,

Theorem 55. Let a, b, c, n be integers such that $n \neq 0$ and $gcd(a, n) = 1$. If $ab \equiv ac \ (mod\ n)$, then $b \equiv c \ (mod\ n)$.

Proof: Since $\gcd(a, n) = 1$, there exists integers x and y such that $ax + by = 1$ (by Theorem 14). Multiply both sides of the previous equation by $(b - c)$ to get

$$(ab - ac)x + n(b - c)y = b - c$$

By assumption, $ab - ac$ is a multiple of n and clearly, $n(b - c)y$ is a multiple of n. Thus, $b - c$ is a multiple of n, i.e., $b \equiv c \pmod{n}$. ■

Theorem 56 (Fermat's Little Theorem) If p is a prime number and p does not divide the integer a, then $a^{p-1} \equiv 1\ (mod\ p)$.

Proof: Let $S = \{1,2, \dots, p - 1\}$ and define the mapping $f: S \to S$ by $f(x) \equiv ax \pmod{p}$. For example, if $p = 11$ and $a = 4$, then $f(5) \equiv 20 \equiv 9 \pmod{11}$.

We first verify that, as defined above, $f: S \to S$. This could only be false if $f(x) = 0$ for some $x \in S$. Suppose $f(x) = 0$ for some $x \in S$. This implies that $(x) \equiv ax \pmod{p} \equiv 0 \pmod{p}$. The condition that p is prime and does not divide a implies $\gcd(a, p) = 1$ and by Theorem 55 we can divide both sides of the previous equation by a to get $x \equiv 0 \pmod{p}$, which implies $x \notin S$. This contradicts our assumption and thus, $f(x) \neq 0$ for any $x \in S$ and so, it is true that $f: S \to S$.

Next we show that $f(x)$ generates a distinct element for each value of x. Assume $f(x) = f(y)$ for $x, y \in S$. This implies that $ax \equiv ay \pmod{p}$ and by Theorem 55, we can cancel a on both sides of the equation to get $x \equiv y \pmod{p}$. So, $f(1), f(2), \dots, f(p - 1)$ are distinct elements in S. Since S has exactly $p - 1$ elements, it must be that $S = \{f(1), f(2), \dots, f(p - 1)\}$, not necessarily in the same order as $\{1,2,3, \dots, p - 1\}$. Thus, we have

$$1 \cdot 2 \cdot 3 \cdot \dots \cdot p - 1 \equiv f(1) \cdot f(2) \cdot f(3) \cdot \dots \cdot f(p - 1) \equiv a \cdot 2a \cdot 3a \cdot \dots \cdot (p - 1)a$$
$$\equiv a^{p-1}\ (1 \cdot 2 \cdot 3 \cdot \dots \cdot p - 1) \pmod{p}$$

Since $\gcd(i, p) = 1$ for $i \in S$, we can divide both sides of the above equation by $1 \cdot 2 \cdot 3 \cdot \dots \cdot p - 1$ (repeatedly using Theorem 55) to get $1 \equiv a^{p-1} (\mathrm{mod}\ p)$ which is equivalent to the result that we wanted to prove. ■

For example, we know by Fermat's Little Theorem that $49467^{12} \equiv 1 \pmod{13}$ since $13 \nmid 49467$.

The Fermat primality test [127] is a probabilistic approach for determining if a number is prime. The test goes as follows:

> Given an integer n, choose another integer a such that $\gcd(a, n) = 1$. Compute $a^{n-1} \pmod{n}$. If the result does not equal 1, then n is composite (by Fermat's Little Theorem). If the result is 1, then n may be prime and further testing is required.

The test is probabilistic since there are exceptions. For example, $2^{560} \equiv 1 \pmod{561}$ but 561 is not prime since $561 = 3 \cdot 11 \cdot 17$.

. . .

Euler's theorem (based on something called Euler's phi function or sometimes Euler's totient function) is an extension of Fermat's Little Theorem.

For a given integer n, Euler's function returns the number of integers between 1 and n that are relatively prime to n. More formally, $\phi(n)$ counts the number of integers a such that $1 \leq a \leq n$ and $\gcd(a, n) = 1$. For example, $\phi(18) = 6$, since 1,5,7,11,13,17 are relatively prime with 18.

A closed form (formula) for $\phi(n)$ exists, but first we need some preliminary results.

Theorem 57. If $gcd(a, n) = 1$ and $gcd(b, n) = 1$, then $gcd(ab, n) = 1$.

Proof: Apply Theorem 14 several times, we have

$\gcd(a, n) = 1$ implies there exists x and y such that $ax + ny = 1$

$\gcd(b, n) = 1$ implies there exists u and v such that $bu + nv = 1$

Multiplying the two equations above, we get

$$ab(xu) + n(buy + vax + nvy) = 1$$

Applying Theorem 14 again, we have that $\gcd(ab, n) = 1$. ■

Theorem 58. If m and n are relatively prime positive integers, then $\phi(mn) = \phi(m)\phi(n)$.

Proof: We first list the integers from 1 to mn in the tabular form below. An analysis of the rows and columns will give us the desired result.

1	$m+1$	$2m+1$	…	$(n-1)m+1$
2	$m+2$	$2m+2$	…	$(n-1)m+2$
3	$m+3$	$2m+3$	…	$(n-1)m+3$
…	…	…		…
r	$m+r$	$2m+r$	…	$(n-1)m+r$
…	…	…		…
m	$2m$	$3m$	…	nm

Assume $\gcd(m, r) = d > 1$. Each element in row r is of the form $km + r$, with k being an integer between 1 and $n - 1$. Since $d|m$ and $d|r$, we have that $d|(km + r)$. Further, $d|m$ implies $d|mn$. Thus, no number in row r is relatively prime to mn. Consequently, to find the integers (in the table above) that are relatively prime to mn, we only need to consider a row s if $\gcd(m, s) = 1$.

Consider a row s such that $\gcd(s, m) = 1$ (there are $\phi(m)$ such rows). Each element in row s is relatively prime to m. To see this, note that each element in row s is of the form $km + s, 1 \leq k \leq n - 1$. Assume $\gcd(km + s, m) = d > 1$. This implies that $d|m$ which means $d|km$. But if $d|km$ and $d|km + s$, then $d|[(km + s) - km] = s$. So, $d|m$ and $d|s$ with $d > 1$, which is a contradiction to our assumption that $\gcd(s, m) = 1$.

The n integers in row s are distinct modulo n. To see this, note that $km + r \equiv jm + r \pmod{n}$ implies $km \equiv jm \pmod{n}$ which, in turn, implies $k \equiv j \pmod{n}$ since we can cancel m on both sides of the equation by Theorem 55. Thus, exactly $\phi(n)$ of the integers in row s are relatively prime to n. Since the $\phi(n)$ integers in row s are relatively prime to m and n, they are also relatively prime to mn by Theorem 57.

Finally, since there are $\phi(m)$ rows, each of which have $\phi(n)$ integers that are relatively prime to mn, we have that $\phi(mn) = \phi(m)\phi(n)$. ■

Theorem 59. If p is a prime number and a is a positive integer, then $\phi(p^a) = p^a - p^{a-1}$.

Proof: The set of integers that are less than or equal to p^a that are **not** relatively prime to p^a is the same set as the integers divisible by p, i.e., the set $kp, 1 \leq k \leq p^{a-1}$. Thus, there are p^{a-1} integers less than or equal to p^a that are not relatively prime to p^a. If we subtract p^{a-1} from the total number of integers less than or equal to p^a, we get the number of integers relatively prime to p^{a}, i.e., $p^a - p^{a-1}$. ■

We are now ready to derive a closed formula for $\phi(n)$.

Theorem 60. If the prime factorization of a positive integer n is $p_1^{a_1}p_2^{a_2}\dots p_k^{a_k}$ then

$$\phi(n) = n\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\dots\left(1-\frac{1}{p_k}\right)$$

Proof: By Theorem 58 and mathematical induction, we have that

$$\phi(n) = \phi\left(p_1^{a_1}\right)\phi\left(p_2^{a_2}\right)\dots\phi\left(p_k^{a_k}\right)$$

By Theorem 59,

$$\phi\left(p_i^{a_i}\right) = p_i^{a_i} - p_i^{a_i-1} = p_i^{a_i}\left(1-\frac{1}{p_j}\right), \quad i = 1,2,\dots,k$$

Put the above two result together, we get

$$\phi(n) = p_1^{a_1}\left(1-\frac{1}{p_1}\right)p_2^{a_2}\left(1-\frac{1}{p_2}\right)\dots p_k^{a_k}\left(1-\frac{1}{p_k}\right)$$

$$= p_1^{a_1}p_2^{a_2}\dots p_k^{a_k}\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\dots\left(1-\frac{1}{p_k}\right)$$

$$= n\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\dots\left(1-\frac{1}{p_k}\right)$$

which is the result we set out to prove. ■

For example, $\phi(799) = \phi(17)\phi(47) = 16\cdot 46 = 736$. The values of $\phi(n)$ for $1 \le n \le 100$ are shown in Table 46.

Table 46. First 100 values for Euler Phi Function

+	1	2	3	4	5	6	7	8	9	10
0	1	1	2	2	4	2	6	4	6	4
10	10	4	12	6	8	8	16	6	18	8
20	12	10	22	8	20	12	18	12	28	8
30	30	16	20	16	24	12	36	18	24	16
40	40	12	42	20	24	22	46	16	42	20
50	32	24	52	18	40	24	36	28	58	16
60	60	30	36	32	48	20	66	32	44	24
70	70	24	72	36	40	36	60	24	78	32
80	54	40	82	24	64	42	56	40	88	24
90	72	44	60	46	72	32	96	42	60	40

In Table 46, $\phi(n)$ is even for all $n \geq 3$. This is true in general. We capture this result in the following theorem (stated without proof).

Theorem 61. For $n \geq 3$, $\phi(n)$ is an even number.

Proof: Assume the prime factorization of n is $p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}$. From the details of the proof for Theorem 60, we have that

$$\phi(n) = p_1^{a_1}\left(1 - \frac{1}{p_1}\right) p_2^{a_2}\left(1 - \frac{1}{p_2}\right) \dots p_k^{a_k}\left(1 - \frac{1}{p_k}\right)$$

$$= p_1^{a_1 - 1}(p_1 - 1) p_2^{a_2 - 1}(p_2 - 1) \dots p_k^{a_k - 1}(p_k - 1)$$

For each of the k terms in the above expression, there are two cases:

- If $p_i > 2$, then $p_i - 1$ is even and $p_i^{a_i - 1}$ is odd (being the product of an odd number). Thus, in this case, $p_i^{a_i - 1}(p_i - 1)$ is odd, since the product of an even and odd number is an even number.
- If $p_i = 2$ and $a_i > 1$, then $p_i^{a_i - 1}(p_i - 1) = p_i^{a_i - 1}$ which is an even number.

Given that $n \geq 3$, at least one of the two cases holds true. So, $p_i^{a_i - 1}(p_i - 1)$ is an even number for at least one of the terms and thus $\phi(n)$ is an even number. ■

Pick any positive integer and compute the sum of the values of the Euler phi function for all its divisors. For example, the divisors of 10 are 1,2,5 and 10. Taking the Euler phi function for each divisor of 10, we get $\phi(1) + \phi(2) + \phi(5) + \phi(10) = 1 + 1 + 4 + 4 = 10$. It turns out that this property is always true, i.e.,

Theorem 62. For each positive integer n, it is true that

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

Proof: See Section 7.1 of “Elementary Number Theory and Its Applications” [128].

We now state and prove the main result concerning Euler’s phi function.

Theorem 63. (Euler’s Theorem) For relatively prime integers a and n, $a^{\phi(n)} \equiv 1 \ (mod\ n)$.

Proof: Let R be the set of integers less than some given integer n and relatively prime to n, i.e.,

$$R = \{x_1, x_2, \dots, x_{\phi(n)}\}$$

Multiple each member of set R by $a \ (\text{mod } n)$ to get the set S, i.e.,

$$S = \{(ax_1 \bmod n), (ax_2 \bmod n), \dots, \left(ax_{\phi(n)} \bmod n\right)\}$$

S is a permutation of R since

- The elements of S are distinct. To see this, assume $ax_i \pmod{n} \equiv ax_j \pmod{n}$. Since $\gcd(a,n) = 1$, we can cancel terms to get $x_i \pmod{n} \equiv x_j \pmod{n}$ and thus, $x_i = x_j$.
- Further, since $\gcd(a,n) = 1$ and $\gcd(x_i,n) = 1$, we have by Theorem 57 that $\gcd(ax_i,n) = 1$ for $i = 1,2,\dots,\phi(n)$. So, all members of S are less than n and relatively prime to n.

Hence, we can write

$$\prod_{i=1}^{\phi(n)} ax_i \pmod{n} \equiv \prod_{i=1}^{\phi(n)} x_i$$

or equivalently,

$$\prod_{i=1}^{\phi(n)} ax_i \equiv \prod_{i=1}^{\phi(n)} x_i \pmod{n}$$

which implies

$$a^{\phi(n)} \prod_{i=1}^{\phi(n)} x_i \equiv \prod_{i=1}^{\phi(n)} x_i \pmod{n}$$

$$a^{\phi(n)} \equiv 1 \pmod{n}$$

This completes the proof. ■

One application of Euler's theorem is primality testing, e.g., see the Miller–Rabin primality test [129] and the Solovay–Strassen primality test [130].

Euler's theorem can be used to reduce modular arithmetic equations with large exponents. This is done by making use of the following theorem.

Theorem 64. Let a,n,x,y be integers with $n \geq 1$ and $gcd(a,n) = 1$. If $x \equiv y \ (mod\ \phi(n))$, then $a^x \equiv a^y \ (mod\ n)$.

Proof: We are given that $x \equiv y \pmod{\phi(n)}$ which implies there exist an integer k such that $x = y + \phi(n)k$. Thus, we have

$$a^x = a^{y+\phi(n)k} = a^y\left(a^{\phi(n)}\right)^k \equiv a^y 1^k \equiv a^y \pmod{n}$$

In the above, we make use of Euler's theorem, i.e., $a^{\phi(n)} \equiv 1 \pmod{n}$. ■

We can use Theorem 64 to compute $2^{84757} \pmod{131}$. Since $84757 = 651(130) + 127$, we have $84757 \equiv 127 \pmod{130}$, noting that $\phi(131) = 130$ since 131 is a prime number. Applying Theorem 64, we have that

$$2^{84757} \equiv 2^{127} \pmod{131}$$

which is easier to solve than the original problem. Using the technique described in Section 12.5.4, we have

$$2^8 \equiv 256 \equiv 125 \pmod{131}$$

$$2^{16} \equiv 15625 \equiv 36 \pmod{131}$$

$$2^{32} \equiv 1296 \equiv 117 \pmod{131}$$

$$2^{64} \equiv 13689 \equiv 65 \pmod{131}$$

$$2^{128} \equiv 4225 \equiv 33 \pmod{131}$$

We are almost there. Using Theorem 48, multiple both sides of the above equation by $2^{-1} \equiv 66 \pmod{131}$ to get

$$2^{127} \equiv 2^{-1}2^{128} \equiv 66 \cdot 33 \equiv 2178 \equiv 82 \pmod{131}$$

The answer can be checked using Wolfram Alpha. Just enter

2^84757 mod 131

and the answer 82 comes back in a few seconds.

12.5.6 Primitive Roots

A number g is a **primitive root** modulo n if for every integer a, that is relatively prime to n, there exists an integer k such $g^k \equiv a \pmod{n}$.

For example, 3 and 5 are primitive roots of 7. To see this, note that

$$3^0 \equiv 1, \quad 3, \quad 3^2 \equiv 2, \quad 3^3 \equiv 6, \quad 3^4 \equiv 4, \quad 3^5 \equiv 5, \quad 3^6 \equiv 1$$
$$5^0 \equiv 1, \quad 5, \quad 5^2 \equiv 4, \quad 5^3 \equiv 6, \quad 5^4 \equiv 2, \quad 5^5 \equiv 3, \quad 5^6 \equiv 1$$

So, the powers of 3 or 5 both generate all number relatively prime to 7 (modulo 7). Note that, example, this is not true of 2

$$2^0 \equiv 1, \quad 2, \quad 2^2 \equiv 4, \quad 2^3 \equiv 1$$

If n is not prime, then a primitive root only needs to generate the number relatively prime to n. For example, take $n = 10$ and $a = 3$. The numbers relatively prime to 10 are 1,3,7 and 9, allow of which are generated by powers of 3, as shown below.

$$3^0 \equiv 1, \quad 3, \quad 3^2 \equiv 9, \quad 3^3 \equiv 7, \quad 3^4 \equiv 1$$

Primitive roots (typically modulo a prime number) are used in cryptography, e.g., see the Diffie-Hellman Key Exchange in Section 13.3.2.

Primitive roots do not exist for all numbers, e.g., there are no primitive roots modulo 8 or 12.

The numbers which have primitive roots are $1, 2, 4$ and numbers of the form p^i and $2p^i$, where p is an odd prime and $i \geq 1$. Further, it can be shown that there are $\phi(p-1)$ primitive roots modulo p.

12.6 Unsolved Problems in Number Theory

Number theory is famous for having a large number of problems that are easy to state but very hard to prove. For example, Fermat's Last Theorem states that no non-trivial positive integer solutions to the equation $x^n + y^n = z^n$ for any integer value of $n > 2$. The conjecture was first stated by Pierre de Fermat around 1637 in the margin of a copy of Arithmetica. After 358 years of effort by mathematicians, the first successful proof was released in 1994 by Andrew Wiles, and formally published in 1995.

The following are a small sample of the unsolved problems in number theory. For an extensive list of such problems, see the Wikipedia category entitled "Unsolved problems in number theory" [132].

12.6.1 Goldbach's Conjecture

Goldbach's conjecture states that every even positive integer greater than two is the sum of two prime numbers. Christian Goldbach first proposed the conjecture in 1742. In spite of numerous attempts, the conjecture still remains unproven.

12.6.2 Twin Prime Conjecture

Twin primes have a difference of two, e.g., (11,13), (17,19) and (29,31).

It is not known if there are an infinite number of twin primes.

On 14 September 2016, Prime Grid's Sophie Germain found the current world record for twin primes, i.e.,

$$2996863034895 \times 2^{1290000} \pm 1$$

12.6.3 Prime Triplets Conjecture

A prime triplet is a set of three prime numbers in which the smallest and largest differ by 6. The triplets must have the form $(p, p+2, p+6)$ or $(p, p+4, p+6)$ where p is a prime. With the exceptions of $(2,3,5)$ and $(3,5,7)$, the prime triplet formats are the closest possible groupings of three prime numbers, since one of every three sequential odd numbers is a multiple of three, and thus, not prime.

It is not known if there are an infinite number of prime triplets.

The first known gigantic prime triplet was found in 2008 by Norman Luhn and François Morain. The primes are $(p, p+2, p+6)$ with $p = 2072644824759 \times 2^{33333} - 1$. As of October 2020, the largest known prime triplet is of the form $(p, p+2, p+6)$ with $p = 4111286921397 \times 2^{66420} - 1$.

12.6.4 Infinite Number of Perfect Numbers Conjecture

A perfect number is a positive integer that is equal to the sum of its positive divisors, excluding the number itself. For example, $28 = 14 + 7 + 4 + 2 + 1$. The distance between perfect numbers grows very fast and yet it is not known if there are an infinite number of perfect numbers. The first few perfect numbers are

$$6$$

$$28$$

$$496$$

$$8128$$

$$33550336$$

$$8589869056$$

$$137438691328$$

$$2305843008139952128$$

$$2658455991569831744654692615953842176$$

$$191561942608236107294793378084303638130997321548169216$$

12.6.5 Collatz conjecture

The Collatz conjecture concerns the convergence of a particular sequence which is defined as follows.

Define the function

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

Next, define the sequence

$$a_i = \begin{cases} n, & \text{for } i = 0 \\ f(a_{i-1}), & \text{for } i > 0 \end{cases}$$

The Collatz conjecture (yet unproven true or false) is that the sequence a_i will converge to 1 regardless of which positive integer n is chosen. For example, the number 33 generates the following sequence:

$$100, 50, 25, 76, 38, 19, 58, 29, 88, 44, 22, 11, 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1$$

A Collatz conjecture calculator is available at https://goodcalculators.com/collatz-conjecture-calculator/.

13 Cryptography

13.1 Overview

Prerequisites: number theory (see Sections 7 and 12), algebra, matrices

13.1.1 Definitions

Cryptography is the study of methods for the encryption (disguising) and decryption of information. The following are some definitions of cryptography:

- From the National Institute of Standards and Technology (NIST) [84]: “Cryptography uses mathematical techniques to transform data and prevent it from being read or tampered with by unauthorized parties. That enables exchanging secure messages even in the presence of adversaries. Cryptography is a continually evolving field that drives research and innovation.”
- From Wikipedia [85]: “Cryptography, or cryptology (from Ancient Greek: κρυπτός, romanized: kryptós ‘hidden, secret’; and γράφειν graphein, ‘to write’, or -λογία -logia, ‘study’, respectively), is the practice and study of techniques for secure communication in the presence of adversarial behavior. More generally, cryptography is about constructing and analyzing protocols that prevent third parties or the public from reading private messages; various aspects of information security such as data confidentiality, data integrity, authentication, and non-repudiation are central to modern cryptography. Modern cryptography exists at the intersection of the disciplines of mathematics, computer science, electrical engineering, communication science, and physics. Applications of cryptography include electronic commerce, chip-based payment cards, digital currencies, computer passwords, and military communications.”
- From The Economic Times [86]: “Cryptography is associated with the process of converting ordinary plain text into unintelligible text and vice-versa. It is a method of storing and transmitting data in a particular form so that only those for whom it is intended can read and process it. Cryptography not only protects data from theft or alteration, but can also be used for user authentication.”

The term **plaintext** is used to describe text that is not in encrypted form, i.e., readable without any conversion. The term **ciphertext** is used to describe text after it has been encrypted.

In the context of cryptography, a **key** is information (e.g., a binary string) that allows one to encrypt text or decrypt encrypted text. In some methods, different keys are used for encryption and decryption.

There are three basic types of cryptographic methods:

- **Symmetric-key cryptography**: The sender and receiver use the same key. The sender uses the key to encrypt plaintext which is then sent to the receiver. The receiver applies the same key to decrypt the message and recover the plaintext.
- **Public-key cryptography**: In this approach, two related keys (public and private) are used. Public keys are freely distributed, while its paired private key remains a secret. The public key is used for encryption and private key is used for decryption.

- **Cryptographic hash function [87]**: A cryptographic hash function (CHF) is a mathematical algorithm that maps data of an arbitrary size (often called the "message") to a bit array of a fixed size (the "hash value", "hash", or "message digest"). It is a one-way function, that is, a function for which it is practically infeasible to invert or reverse the computation. Ideally, the only way to find a message that produces a given hash is to attempt a brute-force search of possible inputs to see if they produce a match, or use a rainbow table of matched hashes. Cryptographic hash functions are a basic tool of modern cryptography.

. . .

Cryptanalysis is the process of finding weaknesses in cryptographic schemes, and using such weaknesses to decipher or alter the ciphertext without knowledge of the key or keys.

In the **brute force attack**, the attacker (usually with the aid of a computer) attempts to break the cipher by testing all possible values of the secret key. One goal of the cipher designer is to make brute force attacks practically infeasible in a given amount of time.

Attacks on encrypted information can be classified based on the information available to the attacker:

- In a ciphertext-only attack, the attacker only has access to the ciphertext. This, of course, implies the attacker is able to intercept and listen-in on communications between the sender and receiver of the encrypted information.
- In a known-plaintext attack, the attacker has access to some ciphertext and its corresponding plaintext.
- In a chosen-plaintext attack (sometimes known-plaintext attack) [88], the attacker has temporary access to the encryption mechanism. The attacker does not have access to the key. However, the attacker can choose various plaintext passages, generate the associated ciphertexts and attempt to deduce the key. The idea here is for the attacker to pose as a friendly associate of the receiver, while sending encrypted messages to the receiver. The receiver responds with an encrypted message which the receiver can not decrypt but each returned message gives the attacker more clues. For example, this approach was used by the British Government Code and Cypher School at Bletchley Park, England, during World War II, for schemes to entice the Germans to include particular words in responses to encrypted messages from the British [89].
- In a chosen-ciphertext attack [90], the attacker has temporary access to the decryption mechanism. The attacker does not have access to the key. However, the attacker can choose various ciphertext strings of symbols, generate the associated plaintexts and attempt to deduce the key.
- In a man-in-the-middle attack, the attacker is able to intercept encrypted messages from the sender, modify the messages and then send them along to the receiver (without the sender or receiver knowing). From the Wikipedia article on this topic [91]: “In cryptography and computer security, a man-in-the-middle, monster-in-the-middle, machine-in-the-middle, monkey-in-the-middle, meddler-in-the-middle (MITM) or person-in-the-middle (PITM) attack is a cyberattack where the attacker secretly relays and possibly alters the communications between two parties who believe that they are directly communicating with each other, as the attacker has inserted themselves between the two parties. One example of a MITM attack is active eavesdropping, in which the attacker makes

> independent connections with the victims and relays messages between them to make them believe they are talking directly to each other over a private connection, when in fact the entire conversation is controlled by the attacker. The attacker must be able to intercept all relevant messages passing between the two victims and inject new ones. This is straightforward in many circumstances; for example, an attacker within the reception range of an unencrypted Wi-Fi access point could insert themselves as a man-in-the-middle. As it aims to circumvent mutual authentication, a MITM attack can succeed only when the attacker impersonates each endpoint sufficiently well to satisfy their expectations. Most cryptographic protocols include some form of endpoint authentication specifically to prevent MITM attacks. For example, TLS can authenticate one or both parties using a mutually trusted certificate authority."

. . .

Kerckhoffs's principle [92] states that the security of a cryptographic system must depend on the secrecy of its keys only and everything else, including the algorithm itself, should be considered public knowledge. A reformulation of Kerckhoffs's principle by Claude Shannon goes as follows: "one ought to design systems under the assumption that the enemy will immediately gain full familiarity with them" (this is known as **Shannon's maxim**).

13.2 Symmetric Ciphers

13.2.1 Basic Concepts

Symmetric encryption (also known as conventional encryption or single-key encryption) was the only type of encryption in use prior to the advent of public key encryption in the 1970s. As noted, symmetric ciphers use the same cryptographic keys for both the encryption of plaintext and the decryption of ciphertext. The key is necessarily a shared secret between two or more parties that can be used to keep exchanged information private. The requirement that all parties (in the secure exchange) have knowledge of the secret key is a critical drawback of symmetric-key encryption, in comparison to public key encryption (also known as asymmetric ciphers). This drawback can be addressed by using public key encryption to exchange the secret key for symmetric key encryption.

Symmetric ciphers operate in either stream or block mode, or both.

- **Stream ciphers** encrypt the digits (usually bits or bytes), or letters (in the case of substitution ciphers) of a message one at a time.
- **Block ciphers** encrypt fixed-length groups of bits known as blocks.

13.2.2 Classical Encryption Methods

13.2.2.1 Substitution Ciphers

In a **substitution cipher**, units of plaintext characters (e.g., letters, numbers, bits) are replaced with other characters. The units of plaintext may be single characters or collections of characters, with the latter case being a block cipher. The identity of each unit of plaintext is changed while its position remains unchanged.

On the other hand, a **transposition cipher** (to be discussed on Section 13.2.2.2) rearranges the position of the units of plaintext. In a transposition cipher approach, the position of the plaintext units is changed but the identity of the plaintext unit is not changed. A transposition cipher is a permutation of plaintext units (which can be single characters or blocks of characters).

As our first example of a substation cipher, we consider what are called Caesar ciphers. In this approach, each letter of plaintext is replaced by a letter some fixed number of positions down in the alphabet. The method is named after Julius Caesar, who used it in his private communications.

The table below shows the mapping for a Caesar cipher with a shift of 3 places to the right, or equivalently, 23 places to the left. The text “I am Steve.” would get mapped to “FXJPQBSB”. Spaces and punctuation are typically removed from the ciphertext. For this type of cipher, the key is the number of places to shift the text (3 to the right in this example).

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

The Caesar cipher can be represented using modular arithmetic where the letters of the alphabet are mapped to integers starting with 0, i.e., A → 0, B → 1, ..., Z → 25. A shift to the right by n units is represented as

$$E_n(x) \equiv (x+n)(\mathrm{mod}\ 26)$$

and the associated deciphering is represented as

$$D_n(x) \equiv (x-n)(\mathrm{mod}\ 26)$$

This type of cipher clearly does not satisfy Kerckhoffs's principle. If a potential code breaker knows that a Caesar cipher is being used but does not know the key (i.e., the value of n), there are only 25 possible shifts to check (noting that a shift of 0 leaves the plaintext as-is).

. . .

The Caesar cipher barely scratches the surface with respect to single letter to single letter substitution ciphers. The number of possible substitution ciphers is immense, i.e., $26! = 403291461126605635584000000 \cong 4.032914611 \times 10^{26}$ possibilities (assuming a 26-letter alphabet). The cryptograms posed in newspapers, magazines and now online (e.g., https://api.razzlepuzzles.com/cryptogram) are usually substitution ciphers.

For the general case of a substitution cipher, a brute force (test all possibilities) is infeasible. However, there are other code breaking techniques. When nothing more is known about the specific type of monoalphabetic substitution cipher being used, a typical first step is to do a frequency analysis on the ciphertext. The frequency of each letter for many languages is known, see the Wikipedia article on this topic [93]. If given a sufficiently large amount of ciphertext, one can try to map the ciphertext letters to their corresponding plaintext letters. For example, if the Z appears most often in ciphertext, it is likely to be the letter E in plaintext since E is the most common letter (at least for English). After looking at single letter frequencies, one can also check on the frequencies of digrams (or bigrams) [94], i.e., two-letter combinations that appear often, e.g., “TH” or “HE”.

Substitution ciphers are relatively easy to break because they reflect the letter and letter combination frequency profile of the original alphabet. One remedy to this problem is to use multiple possible substitutions (known as **homophonic substitutions**) for some or all of the plaintext letters. The idea is to map (encrypt) letters that appear more often (such as E, A and N) to several symbols in the ciphertext with the intent of smoothing out the frequencies of the ciphertext symbols.

In Table 47, several of the more common plaintext letters are mapped to several two-digit numbers. When encrypting a plaintext letter with several ciphertext options, one can either randomly choose which ciphertext option to use or take each option in turn (round-robin fashion).

Table 47. Example homophonic substitution table

Plaintext	**Ciphertext**
A	21,27,31,40
B	15
C	01,33
D	20,34
E	22,28,32,36,37
F	05
G	17
H	14
I	02,29,38,41
J	19
K	03
L	07,39,42
M	09,43
N	12,48,97
O	18,60,85
P	26,44
Q	25
R	24,49
S	10,30,45,99
T	06,96,55
U	16,94
V	23
W	13
X	11
Y	08
Z	04

Using Table 47, the plaintext

WE HOLD THESE TRUTHS TO BE SELF-EVIDENT.

is encrypted to the following ciphertext (using the round-robin approach)

1322 14180720 0614281032 962416551430 0660 1536 4537390522230234281296

Note that the period was omitted from the encryption (this is typically of all punctuation). Also, the spaces in the ciphertext are always omitted in real applications, but are left here to make it easier to see how each word of plaintext is encrypted.

Credits: The above table and example encryption come from a website known as dCode, see https://www.dcode.fr/homophonic-cipher.

. . .

Affine ciphers can be seen as an extension of the Caesar cipher. For an affine cipher the encryption mapping is given by

$$E(x) \equiv ax + b \pmod{26}$$

where $\gcd(a, 26) = 1$ and $0 \leq b \leq 26$. The condition $\gcd(a, 26) = 1$ ensures that a has an inverse in $\mathbb{Z}_{26}$.

Decryption is determined by solving $y = ax + b \pmod{26}$ for x, i.e.,

$$D(y) \equiv a^{-1}y - a^{-1}b \pmod{26}$$

In this scheme, the key is (a, b). Since we require $\gcd(a, 26) = 1$, there are only 12 possible choices for a, i.e., 0,3,5,7,9,11,15,17,19,21,23,25). Given the 26 possible choices for b, we get a total of $12 \cdot 26 - 1 = 311$ possible choices for the key (we subtract one to exclude the case of $a = b = 0$ which does not transform the plaintext at all). With only 311 possible keys, this cipher is easy to break with a computer.

As an example, take $a = 7$ and $b = 2$. This gives the following encryption mapping

$$E(x) \equiv 7x + 2 \pmod{26}$$

Using the stated values for a and b, the plaintext ELEPHANT gets encrypted as EBEDZCPF. The calculation is as follows:

plaintext	E	L	E	P	H	A	N	T
x	4	11	4	15	7	0	13	19
7x + 2	30	79	30	107	51	2	93	135
(7x + 2) mod 26	4	1	4	3	25	2	15	5
ciphertext	E	B	E	D	Z	C	P	F

(Note that the plaintext is first converted to numbers between 0 and 26, i.e., $A = 0, B = 1, C = 2, \ldots$)

Since $7 \cdot 15 \equiv 1 \pmod{26}$, $a^{-1} = 15$. Using the general formula that we compute above, the decryption mapping for this example is

$$D(y) \equiv 15y - 15 \cdot 2 \pmod{26} \equiv 15y + 22 (\text{mod } 26)$$

Applying the above to EBEDZCPF, we do get back the original plaintext ELEPHANT, see the calculations below:

plaintext	E	B	E	D	Z	C	P	F
x	4	1	4	3	25	2	15	5
15x + 22	82	37	82	67	397	52	247	97
(15x + 22) mod 26	4	11	4	15	7	0	13	19
ciphertext	E	L	E	P	H	A	N	T

. . .

A **monoalphabetic cipher** uses the same substitution scheme over the entire message, whereas a **polyalphabetic cipher** uses a different substitution scheme at different positions in the message. All the ciphers that we have seen thus far are of the monoalphabetic type.

As an example of a polyalphabetic cipher, we consider the Vigenère cipher. From the Wikipedia article on the topic [95]:

> The Vigenère cipher is a method of encrypting alphabetic text by using a series of interwoven Caesar ciphers, based on the letters of a keyword. It employs a form of polyalphabetic substitution.
>
> First described by Giovan Battista Bellaso in 1553, the cipher is easy to understand and implement, but it resisted all attempts to break it until 1863, three centuries later. This earned it the description le chiffrage indéchiffrable (French for 'the indecipherable cipher'). Many people have tried to implement encryption schemes that are essentially Vigenère ciphers. In 1863, Friedrich Kasiski was the first to publish a general method of deciphering Vigenère ciphers.

For example, take the keyword “pinelands”. We first write the keyword based on the position of each letter of the keyword in the alphabet, i.e., “15 8 13 4 11 0 13 3 18”. Each number in the key represents a Caesar shift cipher. The first letter in the message is shifted 15 letters to the right, the second letter in the plaintext message is shifted 8 letters to the right, and so on. When we get to the 10th letter in the plaintext message (i.e., when we use all the numbers in the key), we return to the first number in the key.

Using our keyword, the plaintext

> when in the course of human events

gets converted to

> LPRR TN GKW RWHVDE BI ZJUNR PVRQLH

Credits: The above encryption was done using the Vigenere tool at https://www.boxentriq.com/code-breaking/vigenere-cipher.

An extension of the Vigenère cipher known as One Time Pad (OTP) is considered to be unbreakable. The OTP cipher makes use of a single-use keyword that is at least as long as the plaintext message. In the same manner as the Vigenère cipher, the keyword is converted to a string of numbers.

According to the Wikipedia article on this topic [96], the OTP cipher is unbreakable if the following conditions are met:

- “The numbers in the key are used to shift corresponding letters in the message.
- The key must be at least as long as the plaintext.
- The key must be random (uniformly distributed in the set of all possible keys and independent of the plaintext), entirely sampled from a non-algorithmic, chaotic source such as a hardware random number generator. It is not sufficient for OTP keys to pass statistical randomness tests as such tests cannot measure entropy, and the number of bits of entropy must be at least equal to the number of bits in the plaintext. For example, using cryptographic hashes or mathematical functions (such as logarithm or square root) to generate keys from fewer bits of entropy would break the uniform distribution requirement, and therefore would not provide perfect secrecy.
- The key must never be reused in whole or in part.
- The key must be kept completely secret by the communicating parties.”

The main issue with this type of cipher is key distribution between the communicating parties. One approach is to randomly generate a long string which represents many keys. Each time a message is sent, a different segment of the string is used as the key. Both sides need to keep track of the part of the string that has been used. The string could be put on a memory stick and physically transported to the other party. See “One-time pad: Key distribution” [97] for further thoughts on this issue.

13.2.2.2 Transposition Ciphers

Transposition ciphers permute the positions of plaintext units (which are typically letters, numbers or groups thereof), i.e., the order of the units is changed (the plaintext is reordered) but the units are not replaced with different characters (as is done with substitution ciphers).

As an initial example of a transposition cipher, we consider the **rail fence cipher**. In this cipher, plaintext characters are placed diagonally only in the gaps in a chain-link fence (as shown in Figure 30) and the ciphertext is read off of the rows, starting from the top. The plaintext HELLO WORLD is enciphered as HL ERD LO LW O (spaces are included to separate the rows for readability but would not be included in the actual ciphertext).

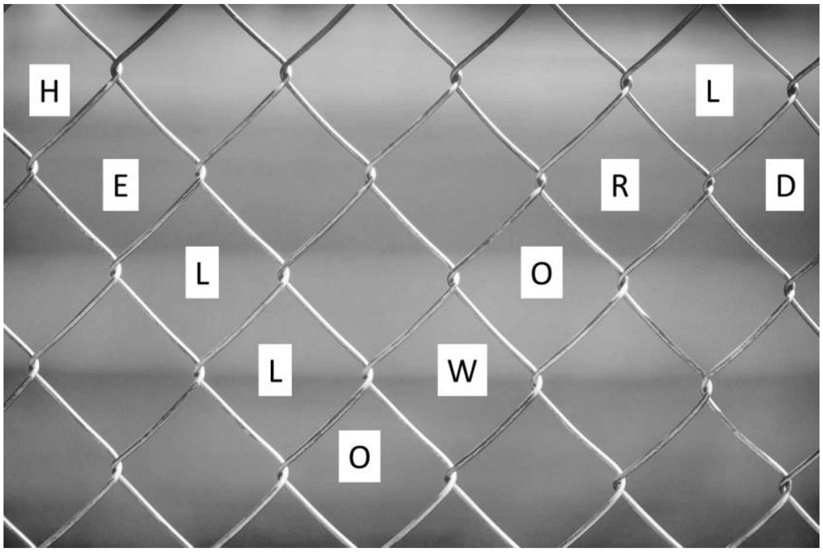


Figure 30. Rail Fence Cipher – Hello World

The rail fence cipher differs based on the number of rows. In fact, the number of rows is the key.

For example, if we encrypt the Shakespeare quote "Wisely and slow, they stumble that run fast." get encrypted as follows if we use 3 rows

```
WLDWYMTRA IEYNSOTESUBEHTUFS SALHTLANT
```

The corresponding layout is

```
W   L   D   W   Y   M   T   R   A
 I E Y N S O T E S U B E H T U F S
  S   A   L   H   T   L   A   N   T
```

If we apply 5 rows to the same plaintext, we get

```
WDYTA INSESEHFS SALHTLANT EYOTUBTU LWMR
```

The corresponding layout is

```
W       D       Y       T       A
 I     N S     E S     E H     F S
  S   A   L   H   T   L   A   N   T
   E Y     O T     U B     T U
    L       W       M       R
```

Credits: The two encryptions above and the associated layouts were generated using the "Rail Fence Cipher - Decoder and Encoder" at https://www.boxentriq.com/code-breaking/rail-fence-cipher.

There are online applications that attempt to determine the type of cipher used, given a sample of ciphertext. Using the Cipher Identifier at https://www.dcode.fr/cipher-identifier with input

```
WLDWYMTRAIEYNSOTESUBEHTUFSSALHTLANT
```

the rail fence cipher is returned as the second most likely cipher.

Similar in concept to rail fence ciphers is something called **columnar transposition** where plaintext is written out along the rows of a matrix with a fixed number of columns and with the number of rows depending on the size of the plaintext message. The ciphertext is generated by reading the matrix column by column, where the columns are chosen in some scrambled order based on a keyword. For example, the keyword GERMANY is of length 7 (so the rows are of length 7), and the permutation is defined by the alphabetical order of the letters in the keyword. In this case, the order would be "3 2 6 4 1 5 7".

For example, take the plaintext “In the beginning, God created the heaven and the earth.” Without spaces and punctuation, we have

```
INTHEBEGINNINGGODCREATEDTHEHEAVENANDTHEEARTH
```

Applying the columnar transposition, using GERMANY as the keyword, we get the following ciphertext. Each group of text is a column from the matrix below. Spaces would not be added in a real application of the cipher.

```
EIRHNE NIOEATH IGGTEDT HNCTEE BNEEAA TNDDVH EGAHNR
```

The associated matrix is

G	**E**	**R**	**M**	**A**	**N**	**Y**
3	**2**	**6**	**4**	**1**	**5**	**7**
I	N	T	H	E	B	E
G	I	N	N	I	N	G
G	O	D	C	R	E	A
T	E	D	T	H	E	H
E	A	V	E	N	A	N
D	T	H	E	E	A	R
T	H					

Credits: The above cipher and associated matrix was generated by the “Columnar Transposition Cipher Decoder and Encoder” at https://www.boxentriq.com/code-breaking/columnar-transposition-cipher.

A single columnar transposition could be broken by guessing possible column lengths, writing the message out in its columns (but in the wrong order, since the key is not yet known), and then looking for possible anagrams that are actual English words. To make this type of attack more difficult, columnar transposition can be applied twice, using keys of different length. For example, we could take the output from the example above and apply the key WOMBAT to get

```
NATCEDAHEGNNNGROGHBTEIIITEAHRETETEDHENHDEAVN
```

The associated matrix is

```
W  O  M  B  A  T
6  4  3  2  1  5
E  I  R  H  N  E
N  I  O  E  A  T
H  I  G  G  T  E
D  T  H  N  C  T
E  E  B  N  E  E
A  A  T  N  D  D
V  H  E  G  A  H
N  R
```

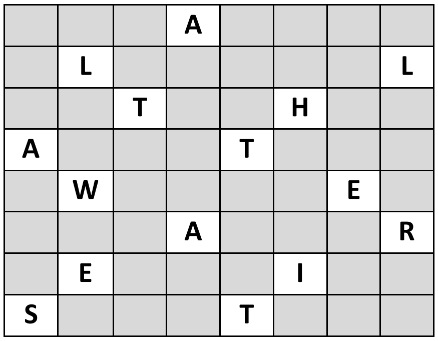
. . .

A **grille cipher** [98] is a method for encrypting a plaintext by writing it onto the open holes (grid) over a matrix. The earliest known description is due to the mathematician Girolamo Cardano in work "De Subtilitate libri XXI" (1550).

As an example, take the quote from Buddha

All that we are is the result of what we have thought

Place the quote in the open (white) squares in the grid shown in Figure 31 (starting from the top of the grille and going left to right).



			A				
	L						L
		T			H		
A				T			
	W					E	
			A				R
	E				I		
S				T			

Figure 31. Grille example

The second step is to remove the grille (mask) and fill-in the remaining squares with random characters, as shown in Figure 32.

2	R	5	**A**	G	B	S	P
X	**L**	Q	Y	4	N	L	**L**
V	5	**T**	U	Z	**H**	M	K
A	8	B	N	**T**	L	U	P
1	**W**	N	C	Q	A	**E**	Z
E	D	B	**A**	M	L	9	**R**
Q	**E**	T	Y	B	**I**	K	7
S	W	4	S	**T**	O	I	D

Figure 32. Insertion of random characters to fill grille

Next, the ciphertext in the matrix is sent row by row, starting from the top.

Since only part of the plaintext was encrypted in the initial grille, we need to repeat the process for the remaining characters in the plaintext quote.

In a variation of the grille method the grille (mask) is rotated. This is known as the **Fleissner grille** cipher (or turning grille cipher). To illustrate the technique, we will consider a 4 by 4 grille (noting that other sizes are possible). In Figure 33, consider the location of the number 1 as the grille is successively rotated 90, 180 and 270 degrees counterclockwise (or clockwise). Similarly, the location of the numbers 2, 3 and 4 are shown as the grille is rotated.

2	4	1	2
1	3	3	4
4	3	3	1
2	1	4	2

Figure 33. Turning grille structure

Next, we cut-out four squares (one for each number), shown as gray squares in Figure 34. The other squares shown in white are covered (think of a 4 by 4 cardboard template with 4 squares cut-out and placed over a matrix).

2	4	1	2
1	3	3	4
4	3	3	1
2	1	4	2

Figure 34. Turning grille with cut-outs

The critical point here is that as we rotate the template 90, 180 and 270, all of the underlying squares are exposed. This only works if we choose to remove (make a cut-out for) one of the squares with a 1, one of the squares with a 2, one of the squares with a 3 and one of the squares with a 4.

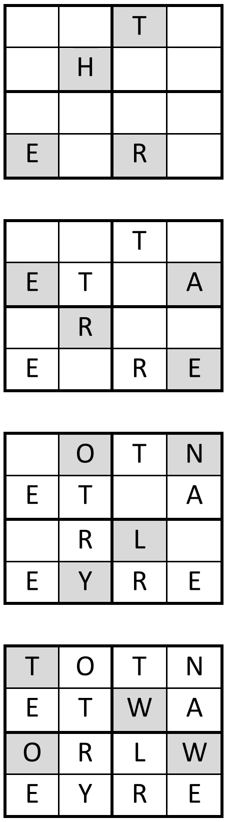
This gives us a way to encrypt text. For example, take the following quote from Albert Einstein

> There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.

The plaintext is encoded as shown in Figure 35. We put the first four characters of the plaintext into the open cells exposed by the unrotated grille (going from top to bottom, and left to right). Next, rotate the grille 90 degrees counterclockwise to expose four additional open cells. Put the next four characters of the plaintext into the open cells. Do the same for 180 and 270 degree rotations. At this point the matrix is fully populated with characters from the plaintext. The bottom matrix in the figure provides the following encryption from the first part of the quote, i.e.,

TOTN ETWA ORLW EYRE

Since we have only encrypted part of the plaintext, we repeat the process until all the plaintext has been encrypted.



		T	
	H		
E		R	

		T	
E	T		A
	R		
E		R	E

	O	T	N
E	T		A
	R	L	
E	Y	R	E

T	O	T	N
E	T	W	A
O	R	L	W
E	Y	R	E

Figure 35. Turning grille example

The paper “Encryption Using a Variant of the Turning-Grille Method” [99] provides additional background and examples on this topic.

13.2.3 Block Ciphers

With the exception of the columnar transposition cipher, all the ciphers that we have studied so far are stream ciphers. With stream ciphers, changing one letter in the plaintext changes exactly one letter in the ciphertext. This makes it easier to find the key using frequency analysis. Block ciphers make analysis (code breaking) more difficult by encrypting blocks of plaintext simultaneously. With most block ciphers, a change in one character of the plaintext results in many changes to the corresponding ciphertext block.

As an example of a block cipher, we consider the **Hill cipher**. The Hill cipher is a polygraphic block substitution cipher based on modular arithmetic and linear algebra. It was invented by Lester S. Hill in 1929. To encrypt a message, each block of n plaintext characters (represented as an n-component column vector) is multiplied by an invertible $n \times n$ matrix modulo 26. To decrypt the message, each block of ciphertext (also represented as a column vector) is multiplied by the inverse of the matrix used for encryption.

For example, consider the following invertible matrix

$$A = \begin{bmatrix} 6 & 24 & 5 & 3 \\ 12 & 16 & 10 & 21 \\ 20 & 17 & 15 & 19 \\ 7 & 15 & 23 & 2 \end{bmatrix}$$

Using the modular arithmetic inverse calculator at https://www.dcode.fr/matrix-inverse, we find the inverse of A to be

$$A^{-1} = \begin{bmatrix} 20 & 9 & 25 & 15 \\ 19 & 16 & 3 & 22 \\ 23 & 25 & 10 & 24 \\ 4 & 3 & 22 & 14 \end{bmatrix}$$

In summary, to encrypt plaintext, we convert the plaintext to a column vector of numbers (call it x) and the multiple by A, i.e., $Ax = y$. The result y is the ciphertext (numbers) which is converted to text before sending to the receiver. If the matrix A is $n \times n$, then only n characters of plaintext can be converted at a time, and the process needs to be repeated.

Going in the other direction, we convert ciphertext to numbers, place in a column vector y and then multiply by A inverse, i.e., $A^{-1}y = x$. The result x (when converted from numbers to text) is the original plaintext.

Alternately, one can use row vectors and multiple on the right, i.e., $xA = y$ for encryption and $yA^{-1} = x$. For the same matrix A, this leads to a different encryption.

To see how this type of encryption works, consider the plaintext

THE TIME IS NOW

The first step is to convert the plaintext to numbers modulo 26, i.e.,

$$19\ 7\ 4\ 19\ 8\ 12\ 4\ 8\ 18\ 13\ 14\ 22$$

Next, we multiple matrix A times the first 4 numbers in the plaintext, i.e.,

$$\begin{bmatrix} 6 & 24 & 5 & 3 \\ 12 & 16 & 10 & 21 \\ 20 & 17 & 15 & 19 \\ 7 & 15 & 23 & 2 \end{bmatrix} \begin{bmatrix} 19 \\ 7 \\ 4 \\ 19 \end{bmatrix} \equiv \begin{bmatrix} 359 \\ 779 \\ 920 \\ 368 \end{bmatrix} (\text{mod } 26) \equiv \begin{bmatrix} 21 \\ 25 \\ 10 \\ 4 \end{bmatrix}$$

We then convert $(21, 25, 10{,}4)$ to letters, i.e., VZKE.

The same process is applied to each of the next two sets of four plaintext numbers, with results being $(16{,}2{,}4{,}6)$ and $(10{,}12{,}13{,}11)$. Covering to letters, we get the following ciphertext which is then sent to the receiving side of the interaction

VZK EQCE GK MNL

This can be checked using the Hill Cipher online application at https://www.dcode.fr/hill-cipher.

If the number of plaintext characters is not a multiple of n, then we need to pad the plaintext with some random characters.

On the receiving end, VZKE is converted $(21, 25, 10{,}4)$. We then multiple by A^{-1} to get

$$\begin{bmatrix} 20 & 9 & 25 & 15 \\ 19 & 16 & 3 & 22 \\ 23 & 25 & 10 & 24 \\ 4 & 3 & 22 & 14 \end{bmatrix}\begin{bmatrix} 21 \\ 25 \\ 10 \\ 4 \end{bmatrix} \equiv \begin{bmatrix} 955 \\ 917 \\ 1304 \\ 435 \end{bmatrix} (\text{mod } 26) \equiv \begin{bmatrix} 19 \\ 7 \\ 4 \\ 19 \end{bmatrix}$$

which maps to the first four letters of the plaintext, i.e., “THE T”.

We need to apply A^{-1} twice more to recover the rest of the plaintext.

. . .

With the advent of electronic computers, computational demanding block ciphers were developed. One example is the **Feistel cipher**, designed by Horst Feistel and Don Coppersmith in 1973. The Feistel cipher is a component of the Data Encryption Standard (DES), a symmetric-key block cipher published by the National Institute of Standards and Technology (NIST) [100].

The following steps are used to encrypt plaintext using the Feistel cipher.

1. Convert the plaintext characters to a bit stream. Both sides of the communication need to agree on how this is done, e.g., using a specific variant of the American Standard Code for Information Interchange (ASCII).

2. Divide the bit stream into fixed-length blocks. For DES, the size of a block is 64 bits.

3. Take the left-half of the first block and put it into variable L_0 and put the right-half into R_0.

4. R_0 and a key K_0 are used as input to a function F. The output $F(R_0, K_0)$ is added to L_0 modulo 2 to get R_1, i.e., $R_1 = F(R_0, K_0) \oplus L_0$ where $\oplus$ is the exclusive OR symbol (equivalent to addition mod 2 for bits). We define $L_1 = R_0$. [In general, there is an overall key for the Feistel cipher, and a subkey is derived from the overall key at each step. The function F is defined as a component for each particular encryption scheme that uses the Feistel cipher. For DES, the definition of the subkeys and the function F are complex multi-step procedures, see DES [100] for the details.]

5. The second step of the encryption takes L_1, R_1 as input, generates another subkey K_1, and then computes $R_2 = F(R_1, K_1) \oplus L_1$ and then makes the assignment $L_2 = R_1$. This process continues for n steps, where at step i, we have $R_i = F(R_{i-1}, K_{i-}) \oplus L_{i-1}$, and $L_i = R_{i-1}$. The general flow is shown on the left of Figure 36. For DES, $n = 16$.

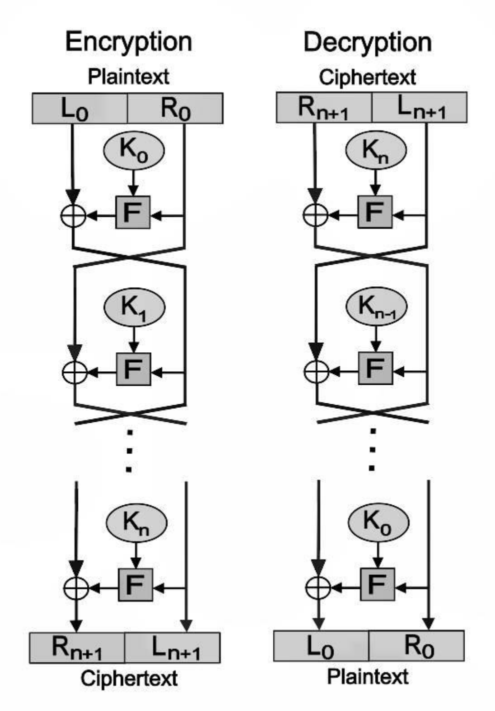


Figure 36. Feistel cipher flow diagram

Credit for Figure 36 goes to Amirki, see
https://commons.wikimedia.org/wiki/File:Feistel_cipher_diagram_en.svg.

Decryption of a Feistel cipher follows the same process as encryption except that the initial inputs are final outputs from the encryption process, as shown on the right of Figure 36.

The Feistel cipher is only part of DES. DES does a permutation of the bits in each block before and after the 16 rounds of the Feistel cipher. Even with all its complexity, DES is no longer considered the gold standard for encryption. From the Wikipedia article on DES [101]

> DES is insecure due to the relatively short 56-bit key size. In January 1999, distributed.net and the Electronic Frontier Foundation collaborated to publicly break a DES key in 22 hours and 15 minutes. There are also some analytical results which demonstrate theoretical weaknesses in the cipher, although they are infeasible in practice. The algorithm is believed to be practically secure in the form of Triple DES *[i.e., the application of DES three times to each block]*, although there are theoretical attacks. This cipher has been superseded by the Advanced Encryption Standard (AES). DES has been withdrawn *[19 May 2005]* as a standard by the National Institute of Standards and Technology.

The replacement for DES, i.e., AES [102], uses a block size of 128 bits and key sizes of 128, 192, or 256 bits. The larger the key the more secure the cipher (at least for AES). The Wikipedia article on this topic provides an overview of AES encryption and decryption processes.

13.3 Asymmetric Ciphers

13.3.1 Basic Concepts

In the previous section, we studied symmetric ciphers, i.e., ciphers that use the same key(s) for encryption and decryption. Also, the decryption process is typically the reverse of the encryption process. **Asymmetric ciphers** use different keys for encryption and decryption. The term "public-key cipher" is often used interchangeably with "asymmetric cipher". Asymmetric ciphers use a public key to encrypt messages and a private key to decrypt the message. The advantage of asymmetric ciphers is that the public key can be openly published, thus allowing parties to establish secure communication without having a shared secret key. Public key ciphers are used by DES and AES to exchange their keys.

The general idea is that each entity involved in an asymmetric cipher ecosystem publishes its public key. For Entity A to send a secure message to Entity B, Entity A encrypts the message using Entity B's public key. Only Entity B has the private key to decrypt messages that have been encrypted with its public key. Entity B can respond to Entity A by encrypting a response using Entity A's public key (which only Entity A can decrypt using its private key). In summary, the key for encrypting messages directed to a given entity is public (hence "public key") but the key for decrypting messages sent to a given entity is private (known only to the intended recipient of a message).

13.3.2 Diffie-Hellman Key Exchange

The Diffie–Hellman key exchange (named after Whitfield Diffie and Martin Hellman) is a method of securely exchanging cryptographic keys over a public (insecure) channel. The concept was published in 1976 [104] and, this is the earliest publicly known work that proposed the idea of a private key and a corresponding public key. The Diffie–Hellman approach allows two parties with no prior knowledge of each other to jointly establish a shared secret key over an insecure channel. This key can then be used to encrypt subsequent communications using a symmetric key cipher.

As an example, assume two parties (Angus and Bonnie) wish to communicate securely using DES. The first step is to agree on a common key for DES, using the Diffie-Hellman key exchange. The steps for the Diffie-Hellman key exchange are as follows:

1. Agree on a prime number p and a primitive root modulo p. Angus and Bonnie will use $p = 1033$ and primitive root $g = 47$ (this information can be made public).
   [Recall that g is a primitive root modulo p if for every integer x relatively prime with p, there is some integer k for which $g^k \equiv x \pmod{p}$. The calculator at http://bluetulip.org/2014/programs/primitive.html lists all the primitive roots modulo a given prime number.]
2. Angus chooses a secret integer $a = 7$, computes $A \equiv g^a \pmod{p} \equiv 47^7 \pmod{1033} \equiv 178$. [Wolfram Alpha does a good job of computing this number.] Angus sends the number 178 to Bonnie.
3. Bonnie chooses a secret integer $b = 11$, computes $B \equiv g^b \pmod{p} \equiv 47^{11} \pmod{1033} \equiv 663$. Bonnie sends the number 663 to Angus.
4. Angus computes $s \equiv B^a \pmod{p} \equiv 663^7 \pmod{1033} \equiv 748$, and Bonnie computes $s \equiv A^b \pmod{p} \equiv 178^{11} \pmod{1033} \equiv 748$. The secret key s is typically used in a subsequent symmetric cipher.

That the two computations have the same result is no coincidence, since

$$A^b \pmod p \equiv g^{ab} \pmod p \equiv g^{ba} \pmod p \equiv B^a \pmod p$$

In the above transaction, a and b are kept secret, and the values for p, g, A and B are sent in the clear. For large values of p, a and b, it is computational very difficult to determine the secret key s. Several sources on the Internet suggest selecting p with 600 digits or more, e.g., see "How did they break Diffie-Hellman?" [105].

If Angus and Bonnie want to use their secret key as input to DES, they have a problem since DES requires a 56-bit key and 748 converted to binary is 1011101100 (only 10 bits). One approach (recommended) would be to use a much larger value of p. This would likely result in a secret key that is more than 56 bits long. In this case, the problem would be to agree on a method for selecting 56 bits from the secret key to be used as input to DES.

13.3.3 RSA

RSA (Rivest–Shamir–Adleman) is a public-key cipher invented by Ron Rivest, Adi Shamir and Leonard Adleman [106]. Similar to the Diffie-Hellman, RSA depends on a number theory problem that is easy to do in the forward direction by very hard to invert (and thus break the code). Neither Diffie-Hellman nor RSA depend on transpositions or substitutions.

In a public-key cipher, the encryption key is made public and distinct from the decryption key, which is known only to the recipient. RSA is based on public keys that are the product of two very large primes numbers. To break the cipher, one needs to factor the resulting product of the two primes (very hard to do computationally). The prime numbers are kept secret. Messages can be encrypted by anyone, via the public key, but can only be decoded by someone who knows the prime numbers.

Since RSA is a slow algorithm, it is not typically used to directly encrypt user data but rather as an exchange mechanism for keys to be used in symmetric ciphers.

The algorithm goes as follows:

1. Choose distinct prime numbers p and q. In practice, p and q are typically over 1024 bits (i.e., about 308 digits). For our example, we choose $p = 89$ and $q = 131$.
2. Next, compute the product of the two primes selected in Step 1. This is part of the public key. In our example, $n = pq = 89 \cdot 131 = 11{,}659$.
3. Compute the Carmichael totient function $\lambda(n)$, where $\lambda(n)$ is the smallest positive integer m such that $a^m \equiv 1 \pmod n$ for every integer a between 1 and n that is relatively prime with n. The computation of $\lambda(n)$ is simplified since for RSA, n is the product of two primes. Using known formulas for Carmichael's totient function, we have $\lambda(n) = \mathrm{lcm}(\lambda(p), \lambda(q))$. Further, since p and q are prime, $\lambda(p) = \varphi(p) = p - 1$ and $\lambda(q) = \varphi(q) = q - 1$ where $\varphi(x)$ is Euler's totient function (i.e., the number of integers less than or equal to x and relatively prime to x). Thus, our computation reduces to determining $\lambda(n) = \mathrm{lcm}(p - 1, q - 1)$. The value of $\lambda(n)$ is kept secret. For our example, $\lambda(n) = \mathrm{lcm}(88{,}130) = 5720$.
4. Choose an integer e such that $\gcd\big(e, \lambda(n)\big) = 1$ and $1 < e < \lambda(n)$. The value of e is also part of the public key. For our example, we choose $e = 7$.

5. Lastly, determine the multiplicative inverse d of e modulo $\lambda(n)$, i.e., $de \equiv 1 \pmod{\lambda(n)}$. The value of d is used as the exponent of the private key. For our example, $d = e^{-1} \equiv 4903 \pmod{5720}$.

In summary, the public key is the pair (n, e) and the private key is d. Think e for “encrypt” and d for “decrypt”.

Now, for the punchline in our example. Assume that Entity A wants to send message $m = 33$ to Entity B.

1. Entity A obtains the public key of Entity B, i.e., $(n = 5720, e = 7)$.
2. Entity A encrypts the message $m = 33$ as follows: $E(33) \equiv 33^7 (\text{mod } 5720) \equiv 4257$
3. Entity A sends the encrypted message (i.e., the number 4257) to Entity B.
4. Entity B decrypts the message as follows: $D(4257) \equiv 4257^{4903} (\text{mod } 5720) \equiv 33$. Success!

In general, this works for message m, since

$$(m^e)^d \equiv m^{ed} \equiv m^{de} \equiv m \pmod{n}$$

The above calculations were done using Wolfram Alpha, e.g., just enter "4257^4903 mod 5720" (without the quotes) at the Wolfram Alpha prompt and the result 33 will be returned.

13.4 Cryptographic Hashes

13.4.1 Overview

In general (not specific to cryptography), a **hash** is a function that can be used to map data of arbitrary size to fixed-size values. The values returned by a hash function are called hash values, hash codes, digests, or simply hashes.

From the Wikipedia article entitled “Hash function” [107]:

> Hash functions and their associated hash tables are used in data storage and retrieval applications to access data in a small and nearly constant time per retrieval. They require an amount of storage space only fractionally greater than the total space required for the data or records themselves. Hashing is a computationally and storage space-efficient form of data access that avoids the non-constant access time of ordered and unordered lists and structured trees, and the often exponential storage requirements of direct access of state spaces of large or variable-length keys.

The Wikipedia article on “Cryptographic hash function” provides the following definition :

> A **cryptographic hash** function (CHF) is a mathematical algorithm that maps data of an arbitrary size (often called the "message") to a bit array of a fixed size (the "hash value", "hash", or "message digest"). It is a one-way function, that is, a function for which it is practically infeasible to invert or reverse the computation. Ideally, the only way to find a message that produces a given hash is to attempt a brute-force search of possible inputs to see if they produce a match, or use a rainbow table of matched hashes. Cryptographic hash functions are a basic tool of modern cryptography.

The emphasis with a cryptographic hash is on the integrity of the input data.

The following requirements apply to cryptographic hash functions:

1. (Speed) Cryptographic hash functions should be computationally efficient, i.e., fast.
2. (Avalanche Effect) A minor change in the input message to the hash function should result in a major change in the resulting hash value.
3. (Deterministic) A given input message should always result in exactly the same hash value.
4. (One-Way Function) It should be computationally infeasible to determine the input given the hash value. In other words, given a hash value h, it should be infeasible to find any message *m* such that *h = hash(m)*.
5. (Weak Collision Resistance) Given m and $hash(m)$, it should be computationally infeasible to find a different input message m^* such that $hash(m) = hash(m^*)$.
6. (Strong Collision Resistance) It should be computationally infeasible to find two different messages m_1 and m_2 such that $hash(m_1) = hash(m_2)$.

In the above, the term “infeasible” is used since it is not possible to guarantee “impossible.” Consider that we are mapping a larger set of messages into smaller (fixed length) hashes, and thus collisions are definitely possible.

13.4.2 Usages

13.4.2.1 Message Authentication

Message authentication entails verifying whether or not a message has been modified while in transit (data integrity), while also ensuring the identity of the source of the message to the receiving party.

A common approach to message authentication is known as Message Authentication Code (MAC). In this approach, both sides of the communication have a secret key. On the sending side, the secret key and message are used as input to the MAC function which, in turn, produces a hash value, referred to as the MAC. The message recipient can check for message integrity by applying the MAC function (plus secret key) to the received message. If the hash does not match, the message is very likely to have been altered. An attacker who intercepts the message will not be able to alter the hash value (in an undetectable manner) without knowledge of the secret key. Further, by checking the hash value with the MAC function (plus secret key), the receiving party can be confident that the message has come from the identified sending party since no other entity knows the secret key.

[This is not to be confused with the Media Access Control (MAC) addresses assigned to a Network Interface Controller (NICs).]

Figure 37 is a high-level depiction of how a MAC works. The sender of a message uses a MAC algorithm along with a secret key to generate a MAC. The message and MAC are sent to the intended recipient of the message. The receiver (who also knows the secret key) applies the MAC algorithm to the received message. If the generated MAC matches the MAC sent from the sender, then the receiver has good reason to assume the message is authentic and has not been altered.

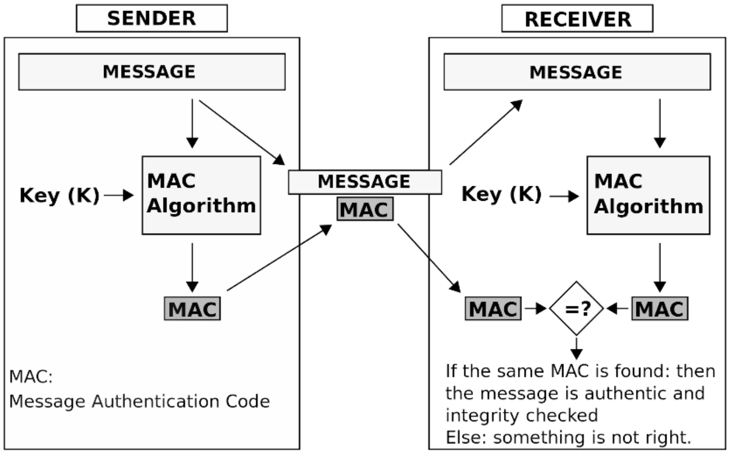

SENDER
MESSAGE
Key (K)
MAC Algorithm
MAC
MAC:
Message Authentication Code
MESSAGE
MAC
RECEIVER
MESSAGE
Key (K)
MAC Algorithm
MAC
=?
MAC
If the same MAC is found: then the message is authentic and integrity checked
Else: something is not right.

Figure 37. MAC example

Many different MACs are in use today. A small sampling is listed below:

- ChaCha20-Poly1305 [113]
- VMAC [114]
- SipHash [115]

13.4.2.2 Password Verification

Cryptographic hashes are typically used in support of the password verification process. Storing user passwords as plaintext (unencrypted) allows for a security breach if the password file is compromised. One method to mitigate this threat is to store the hash value for each password (but not the password itself). When authenticating a user, the presented password is hashed and compared with the stored hash of the password.

Multi-Factor Authentication (MFA), usually two-step verification, is an authentication technique in which a user is granted access to a website or application only after successfully presenting two or more pieces of evidence (or factors) to an authentication mechanism, e.g.,

- knowledge (something only the user knows), e.g., answers to previous recorded security questions (the classic “mother’s maiden name” or “name of your first pet”)
- possession (something only the user has), e.g., sending a code to the user’s cell phone which the user needs to enter along with his or her password
- inherence (something only the user is), e.g., biometrics such as a fingerprint, eye iris pattern, or voice pattern.

For added security, the server-side of an interaction could just store the hashed version of the MFA related information.

13.4.2.3 Blockchain

Blockchain is a secure, distributed, immutable ledger technology. Some definitions

- From Investopedia [109]:

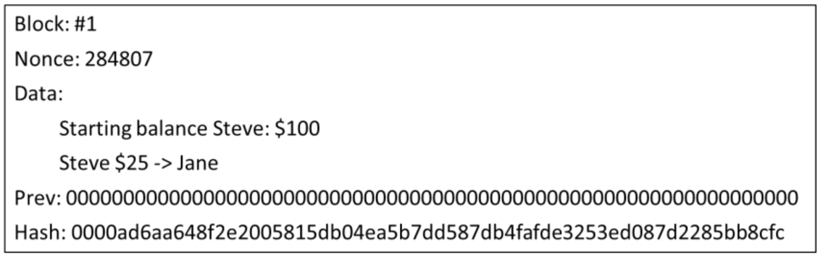
  “A blockchain is a distributed database or ledger that is shared among the nodes of a computer network. As a database, a blockchain stores information electronically in digital format. Blockchains are best known for their crucial role in cryptocurrency systems, such as Bitcoin, for maintaining a secure and decentralized record of transactions. The innovation with a blockchain is that it guarantees the fidelity and security of a record of data and generates trust without the need for a trusted third party.

One key difference between a typical database and a blockchain is how the data is structured. A blockchain collects information together in groups, known as blocks, that hold sets of information. Blocks have certain storage capacities and, when filled, are closed and linked to the previously filled block, forming a chain of data known as the blockchain. All new information that follows the freshly added block is compiled into a newly formed block that will then also be added to the chain once filled."

- From NIST [110]: "A blockchain is a collaborative, tamper-resistant ledger that maintains transactional records. The transactional records (data) are grouped into blocks. A block is connected to the previous one by including a unique identifier that is based on the previous block's data. As a result, if the data is changed in one block, its unique identifier changes, which can be seen in every subsequent block (providing tamper evidence). This domino effect allows all users within the blockchain to know if a previous block's data has been tampered with. Since a blockchain network is difficult to alter or destroy, it provides a resilient method of collaborative record keeping."
- From IBM : "Blockchain is a shared, immutable ledger that facilitates the process of recording transactions and tracking assets in a business network. An asset can be tangible (a house, car, cash, land) or intangible (intellectual property, patents, copyrights, branding). Virtually anything of value can be tracked and traded on a blockchain network, reducing risk and cutting costs for all involved."

A critical point in all the above definitions is the immutability of each block of data. The immutability is accomplished by the chaining of hashes.

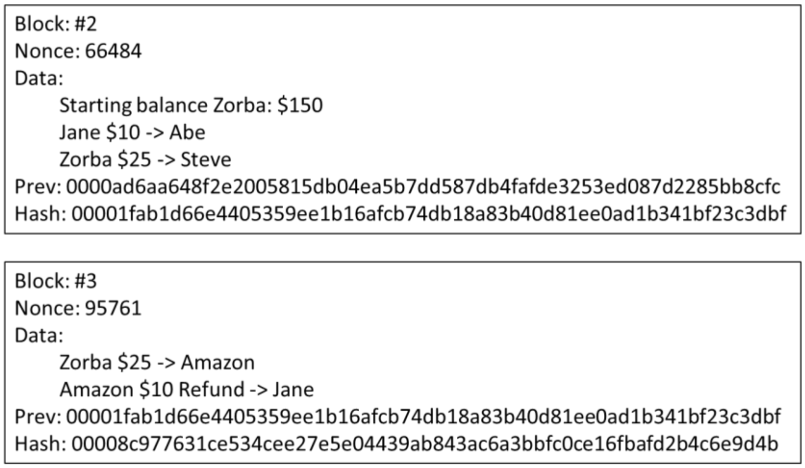
For example, we consider a very short blockchain that records financial transactions. The first block in a blockchain is known as the genesis block (example shown in Figure 38). As with all blocks, the genesis block has a block number, nonce, data, and hash. The second block and beyond will also include the hash for the previous block. Each blockchain will have some requirement on the hash. For the example at hand, the requirement is that the first four hex digits be zero. In order to generate a hash with four zeros at the beginning, the nonce is used. Multiple values of the nonce are tried until the hash requirement is met (in this case, 4 zeros at the beginning of the hash). The process of meeting the hash requirement is known as mining. Once the hash is appropriately generated, the block is considered to be signed (closed to any modification).



Block: #1
Nonce: 284807
Data:
    Starting balance Steve: $100
    Steve $25 -> Jane
Prev: 0000000000000000000000000000000000000000000000000000000000000000
Hash: 0000ad6aa648f2e2005815db04ea5b7dd587db4fafde3253ed087d2285bb8cfc

Figure 38. Genesis block in example blockchain

Figure 39 shows two more signed blocks in our example. Notice that the Prev field in Block #2 is the Hash from the genesis block, and the Prev field in Block #3 is the hash from Block #2.



```
Block: #2
Nonce: 66484
Data:
      Starting balance Zorba: $150
      Jane $10 -> Abe
      Zorba $25 -> Steve
Prev: 0000ad6aa648f2e2005815db04ea5b7dd587db4fafde3253ed087d2285bb8cfc
Hash: 00001fab1d66e4405359ee1b16afcb74db18a83b40d81ee0ad1b341bf23c3dbf
```

```
Block: #3
Nonce: 95761
Data:
      Zorba $25 -> Amazon
      Amazon $10 Refund -> Jane
Prev: 00001fab1d66e4405359ee1b16afcb74db18a83b40d81ee0ad1b341bf23c3dbf
Hash: 00008c977631ce534cee27e5e04439ab843ac6a3bbfc0ce16fbafd2b4c6e9d4b
```

Figure 39. Blocks #2 and #3 in example blockchain

If (for example) someone tried to change the data in the genesis block, the hashes in all subsequent blocks would change and more importantly, the subsequent hashes would no longer satisfy the requirement of having 4 leading zeros. This can only be corrected by re-mining the subsequent blocks. Further, the blockchain is distributed in many places (all with the same information). So, if the intruder modifies one instance of the blockchain, the other instances will be out of sync with the modified instance. Thus, allowing for detection of the bogus modification.

The YouTube video “Blockchain 101 - A Visual Demo” provides an excellent demonstration of blockchain. There are also several websites that allow one to experiment with the operation of blockchain, e.g., see https://demoblockchain.org/blockchain.

The mostly commonly used hash functions for blockchain come from the Secure Hashing Algorithm (SHA) family of hash functions. The SHA family of cryptographic hash functions are published by the National Institute of Standards and Technology (NIST) as a U.S. Federal Information Processing Standard (FIPS). For example, SHA-256 and SHA-512 are hash functions from the SHA-2 set of hashing functions. These are complex, multi-round hashing functions that make use of bit padding, transpositions, modular arithmetic and bit rotations. The YouTube video “SHA 512 - Secure Hash Algorithm - Step by Step Explanation” provides a clear and detailed presentation of every step in SHA-512 [116].

14 Epilogue

"Live as if you were to die tomorrow. Learn as if you were to live forever."

— Mahatma Gandhi

In this book, I have attempted to provide the reader with a very small selection of the many interesting topics from mathematics and associated fields such as cryptography. The body of knowledge for any one of these topics is large. So, these short vignettes can only serve (at best) as inspirations for the reader to learn more. If I have encouraged the reader to delve into any of the topics covered in this book, I have accomplished my intended task. Happy learning!

"Education is the kindling of a flame, not the filling of a vessel."

— Socrates

"There is no absolute knowledge. And those who claim it, whether they are scientists or dogmatists, open the door to tragedy."

— Jacob Bronowski, The Ascent of Man

Acronyms

AES – Advanced Encryption Standard

aka – also known as

ASCII – American Standard Code for Information Interchange

CHF – Cryptographic Hash Function

DES – Data Encryption Standard

EAA – Essential Amino Acid

e.g. – “exempli gratia” in Latin and “for example” in English

GCD – Greatest Common Divisor

i.e. – “id est” in Latin and “that is” in English

IESDS – Iterated Elimination of Strictly Dominated Strategies

IETF – Internet Engineering Task Force

LCM – Least Common Multiple

LP – Linear Programming

MAC – Message Authentication Code

MFA – Multi-Factor Authentication

MTZ – Miller Tucker Zemlin

NIST – National Institute of Standards and Technology

OTP – One Time Pad

RSA – Rivest–Shamir–Adleman

SHA – Secure Hashing Algorithm

wlog – without loss of generality

References

[1] Graham, R., Knuth, D., Patashnik, O., 1994, *Concrete Mathematics: A Foundation for Computer Science*, 2nd edition, Addison-Wesley.

[2] Kramer, J., von Pippich, A., *From Natural Numbers to Quaternions*, Springer, 2017.

[3] Aczel, A., *The Origin of the Number Zero*, Smithsonian Magazine, December 2014.

[4] Fratini, S., *Mathematical Thinking*, self-published on Amazon, https://www.amazon.com/dp/B08F75CDD6, August 2020.

[5] *Fraction*, Wikipedia, https://en.wikipedia.org/wiki/Fraction#Converting_between_decimals_and_fractions, accessed on 5 August 2021.

[6] *Transcendental number*, Wikipedia, https://en.wikipedia.org/wiki/Transcendental_number, accessed on 5 August 2021.

[7] Juan Carlos Ponce Campuzano, *Complex Analysis: A Visual and Interactive Introduction*, https://complex-analysis.com/, accessed on 18 June 2022.

[8] Brooks, M., *Octonions: The strange maths that could unite the laws of nature*, NewScientist, Issue 3400 , published 20 August 2022.

[9] *Euler's Formula*, Wikipedia, https://en.wikipedia.org/wiki/Euler%27s_formula, accessed on 7 August 2021.

[10] Elaydi, S., *An Introduction to Difference Equations** (3rd Edition), Springer, 2006.

[11] *Tribonacci numbers*, The On-line Encyclopedia of Integer Sequences®, http://oeis.org/A000213, accessed on 23 August 2021.

[12] *Young Gauss and the sum of the first n positive integers*, Math and Multimedia, http://mathandmultimedia.com/2010/09/15/sum-first-n-positive-integers/, accessed on 27 August 2021.

[13] *Lazy caterer's sequence*, Wikipedia, https://en.wikipedia.org/wiki/Lazy_caterer's_sequence, accessed on 28 August 2021.

[14] *Number of Regions N Lines Divide Plane*, "Cut the Knot" website, https://www.cut-the-knot.org/proofs/LinesDividePlane.shtml, accessed on 29 August 2021.

[15] Spiegel, M. R., *Theory and Problems of Calculus of Finite Differences and Difference Equations*, Schaum's Outlines, McGraw-Hill, Inc., 1971.

[16] Elaydi, S., *An Introduction to Difference Equations**, Springer, 1996.

[17] *Logistic map*, Wikipedia, https://en.wikipedia.org/wiki/Logistic_map, accessed on 5 September 2021.

[18] *Euclidean division*, Wikipedia, https://en.wikipedia.org/wiki/Euclidean_division, accessed on 16 September 2021.

[19] Khinchin, A.Y., *Continued Fractions**, University of Chicago Press, 1964.

[20] Olds, C.D., *Continued Fractions**, Random House, 1963.

[21] *Nested radicals*, Wikipedia, https://en.wikipedia.org/wiki/Nested_radical, accessed on 25 September 2021.

[22] Zimmerman, S., & Ho, C. (2008). *On Infinitely Nested Radicals*. Mathematics Magazine, 81(1), 3–15. http://www.jstor.org/stable/27643075.

[23] *Nested Radical*, Wolfram MathWorld, https://mathworld.wolfram.com/NestedRadical.html, accessed on 25 June 2022.

[24] *Wiles's proof of Fermat's Last Theorem*, Wikipedia, https://en.wikipedia.org/wiki/Wiles%27s_proof_of_Fermat%27s_Last_Theorem, accessed on 30 September 2021.

[25] *Taxicab number*, Wikipedia, https://en.wikipedia.org/wiki/Taxicab_number, accessed on 30 September 2021.

[26] Boyer, C., *New Upper Bounds for Taxicab and Cabtaxi Numbers*, Journal of Integer Sequences, Vol. 11 (2008), Article 08.1.6, http://www.christianboyer.com/taxicab/TaxicabUpperBounds.pdf.

[27] Lander, L.J., Parkin, T.R., *Counterexamples to Euler's conjecture on sums of like powers,* Bulletin of the American Mathematical Society, Volume 72, 1966, p. 1079. https://projecteuclid.org/journals/bulletin-of-the-american-mathematical-society-new-series/volume-72/issue-6/Counterexample-to-Eulers-conjecture-on-sums-of-like-powers/bams/1183528522.full

[28] Elkies, N.D., *On* $A^4 + B^4 + C^4 = D^4$, Mathematics of Computation, Volume 51, Number 184, Pages 825-835, 1988. https://www.ams.org/journals/mcom/1988-51-184/S0025-5718-1988-0930224-9/S0025-5718-1988-0930224-9.pdf

[29] Frye, Roger E. (1988), *Finding 958004 + 2175194 + 4145604 = 4224814 on the Connection Machine*, Proceedings of Supercomputing 88, Vol. II: Science and Applications, pp. 106–116.

[30] Andreescu, T., Andrica, D., Cucurezeanu, I., *An Introduction to Diophantine Equations: A Problem-Based Approach*, Birkhauser, 2010.

[31] *Bézout's identity*, Wikipedia, https://en.wikipedia.org/wiki/B%C3%A9zout's_identity, accessed on 3 October 2021.

[32] *Euclid's lemma*, Wikipedia, https://en.wikipedia.org/wiki/Euclid's_lemma, accessed on 3 October 2021.

[33] Gilbert, W., Pathria, A., *Linear Diophantine Equations*, preprint 1990, ncatlab.org/nlab/files/GilbertPathria90.pdf.

[34] *Euclidean algorithm*, Wikipedia, https://en.wikipedia.org/wiki/Euclidean_algorithm, accessed on 8 October 2021.

[35] *Equivalence relation*, Wikipedia, https://en.wikipedia.org/wiki/Equivalence_relation, accessed on 28 November 2021.

[36] *Bienaymé's identity*, Wikipedia, https://en.wikipedia.org/wiki/Bienaym%C3%A9%27s_identity, accessed on 1 July 2022.

[37] *Weisstein, Eric W. "Pólya's Random Walk Constants." From MathWorld--A Wolfram Web Resource. https://mathworld.wolfram.com/PolyasRandomWalkConstants.html*, accessed on 17 December 2021.

[38] L.H. Liyanage, C.M. Gulati, J.M. Hill, *A bibliography on applications of random walks in theoretical chemistry and physics*, Advances in Molecular Relaxation and Interaction Processes, Volume 22, Issue 1, 1982, Pages 53-72, ISSN 0378-4487, https://doi.org/10.1016/0378-4487(82)80019-8.

[39] Naoki Masuda, Mason A. Porter, Renaud Lambiotte, *Random walks and diffusion on networks*, Physics Reports, Volumes 716–717, 2017, Pages 1-58, ISSN 0370-1573, https://doi.org/10.1016/j.physrep.2017.07.007.

[40] Weiss, George H. "Random Walks and Their Applications: Widely Used as Mathematical Models, Random Walks Play an Important Role in Several Areas of Physics, Chemistry, and Biology." American Scientist, vol. 71, no. 1, Sigma Xi, The Scientific Research Society, 1983, pp. 65–71, http://www.jstor.org/stable/27851819.

[41] Doyle, P.G., Snell, J.L., *Random Walks and Electric Networks**, The Mathematical Association of America (MAA) Press, 1984.

[42] Ross, S.M., *Introduction to Probability Models**, 12th Edition, Academic Press, 2019. Earlier editions available at http://mitran-lab.amath.unc.edu/courses/MATH768/biblio/introduction-to-prob-models-11th-edition.PDF and https://faculty.ksu.edu.sa/sites/default/files/introduction-to-probability-model-s.ross-math-cs.blog_.ir_.pdf.

[43] Smith, P., Jones, P. W., *Stochastic Processes: An Introduction*, Third Edition, United Kingdom: CRC Press, 2007.

[44] *Survival function: properties*, Wikipedia, https://en.wikipedia.org/wiki/Survival_function#Properties, accessed on 6 July 2022.

[45] *Processes with Stationary, Independent Increments*, Random (online book), https://www.randomservices.org/random/processes/Increments.html, accessed on 6 July 2022.

[46] *Lévy process*, Wikipedia, https://en.wikipedia.org/wiki/L%C3%A9vy_process, accessed on 5 January 2022.

[47] Stephen Fratini (https://math.stackexchange.com/users/545634/stephen-fratini), Can a stochastic process have stationary increments which are not independent increments?, URL (version: 2022-01-05): https://math.stackexchange.com/q/4349593.

[48] "Combining and Splitting Poisson Processes." 2021. September 19, 2021. https://eng.libretexts.org/@go/page/44607.

[49] *Wald's equation*, Wikipedia, https://en.wikipedia.org/wiki/Wald%27s_equation, accessed on 9 January 2022.

[50] *Yule Process*, Science Direct, https://www.sciencedirect.com/topics/mathematics/yule-process, accessed on 7 July 2022.

[51] *Exponential distribution*, section entitled "Distribution of the minimum of exponential random variables", Wikipedia,

https://en.wikipedia.org/wiki/Exponential_distribution#Distribution_of_the_minimum_of_exponential_random_variables, accessed on 10 January 2022.

[52] Karlin, S., McGregor, J., *The classification of birth and death processes*, Trans. Amer. Math. Soc. 86 (1957), 366-400, https://www.ams.org/journals/tran/1957-086-02/S0002-9947-1957-0094854-8/S0002-9947-1957-0094854-8.pdf.

[53] *Birth and death processes*, section on *Stationary Solutions*, Wikipedia, https://en.wikipedia.org/wiki/Birth%E2%80%93death_process#Stationary_solution, accessed on 12 January 2022.

[54] *Queueing Theory*, Investopedia, https://www.investopedia.com/terms/q/queuing-theory.asp, accessed on 12 January 2022.

[55] *Queueing theory*, Wikipedia, https://en.wikipedia.org/wiki/Queueing_theory, accessed on 12 January 2022.

[56] Kleinrock, Leonard. *Queueing Systems, Volume I**. United Kingdom: Wiley, 1974.

[57] Kleinrock, Leonard. *Queueing Systems, Volume 2*: Computer Applications*. United Kingdom: Wiley, 1976.

[58] *M/M/1 queue*, Wikipedia, https://en.wikipedia.org/wiki/M/M/1_queue, 8 July 2022.

[59] *M/M/c queue*, Wikipedia, https://en.wikipedia.org/wiki/M/M/c_queue, 8 July 2022.

[60] *Kendall's notation*, Wikipedia, https://en.wikipedia.org/wiki/Kendall's_notation, accessed on 13 January 2022.

[61] Norris, J.R., *Markov Chains*, Cambridge Series on Statistical and Probabilistic Mathematics, Cambridge University Press, 1997.

[62] Durrett, R., *Essentials of Stochastic Processes*, Third Edition, Springer Texts in Statistics, 2016.

[63] Spaniel, W., *Game Theory 101: The Complete Textbook*, self-published by the author, https://www.amazon.com/Game-Theory-101-Complete-Textbook/dp/1492728152, 2011.

[64] Spaniel, W., *Game Theory 101* (a collection of videos and some that complement the associated textbook of the same name), http://gametheory101.com/courses/game-theory-101/, accessed on 12 March 2022.

[65] *Economic Applications of Game Theory*, MIT Open Courseware, https://ocw.mit.edu/courses/economics/14-12-economic-applications-of-game-theory-fall-2012/index.htm, accessed 12 March 2022.

[66] I. Gilboa, E. Kalai, E. Zemel, *On the order of eliminating dominated strategies*, Operations Research Letters, Volume 9, Issue 2, 1990, Pages 85-89.

[67] *Nash equilibrium (proof of existence)*, Wikipedia, https://en.wikipedia.org/wiki/Nash_equilibrium#Proof_of_existence, accessed on 15 March 2022.

[68] *Minimax*, Wikipedia, https://en.wikipedia.org/wiki/Minimax, accessed on 19 March 2022.

[69] Maschler, M., Solan, E., Zamir, S., *Game Theory**, Cambridge University Press, 2013.

[70] Osborne, M.J., *An Introduction to Game Theory,* Oxford University Press, 2003.

[71] John C. Harsanyi, 1961. "On the rationality postulates underlying the theory of cooperative games," Journal of Conflict Resolution, Peace Science Society (International), vol. 5(2), pages 179-196, June.

[72] *Centipede game*, Wikipedia, https://en.wikipedia.org/wiki/Centipede_game, accessed on 27 March 2022.

[73] Nagel R, Tang FF. Experimental Results on the Centipede Game in Normal Form: An Investigation on Learning. J Math Psychol. 1998 Jun;42(2/3):356-84. doi:10.1006/jmps.1998.1225. PMID: 9710555.

[74] Behavioral game theory, Wikipedia, https://en.wikipedia.org/wiki/Behavioral_game_theory, accessed 11 June 2022.

[75] *Polytope*, Wikipedia, https://en.wikipedia.org/wiki/Polytope, accessed on 4 April 2022.

[76] *Maximum principle*, Wikipedia, https://en.wikipedia.org/wiki/Maximum_principle, accessed on 5 April 2022.

[77] Brickman, L., *Mathematical Introduction to Linear Programming and Game Theory**, Springer-Verlag, 1989.

[78] *Simplex algorithm*, Wikipedia, https://en.wikipedia.org/wiki/Simplex_algorithm, accessed on 10 April 2022.

[79] *Interior-point method*, Wikipedia, https://en.wikipedia.org/wiki/Interior-point_method, accessed on 10 April 2022.

[80] *Interior Point Method Demonstration*, YouTube video by sjbaran, https://youtu.be/MsgpSl5JRbI, accessed on 10 April 2022.

[81] *Knapsack Problem*, Wikipedia, https://en.wikipedia.org/wiki/Knapsack_problem, accessed on 15 April 2022.

[82] Integer Programming, ScienceDirect, https://www.sciencedirect.com/topics/mathematics/integer-programming, accessed on 18 April 2022.

[83] Chen, D., Batson, R., Dang, Y., *Applied Integer Programming: Modeling and Solution*, Wiley, 2010.

[84] *Cryptography*, National Institute of Standards and Technology (NIST), https://www.nist.gov/cryptography, accessed on 15 May 2022.

[85] *Cryptography*, Wikipedia, https://en.wikipedia.org/wiki/Cryptography, accessed on 15 May 2022.

[86] *What is Cryptography*, The Economic Times, https://economictimes.indiatimes.com/definition/cryptography, accessed on 15 May 2022.

[87] *Cryptographic hash function*, Wikipedia, https://en.wikipedia.org/wiki/Cryptographic_hash_function,

[88] *Known-plaintext attack*, Wikipedia, https://en.wikipedia.org/wiki/Known-plaintext_attack, accessed on 16 May 2022.

[89] *Gardening (cryptanalysis)*, Wikipedia, https://en.wikipedia.org/wiki/Gardening_(cryptanalysis), accessed on 16 May 2022.

[90] *Chosen-ciphertext attack*, Wikipedia, https://en.wikipedia.org/wiki/Chosen-ciphertext_attack, accessed on 16 May 2022.

[91] *Man-in-the-middle attack*, Wikipedia, https://en.wikipedia.org/wiki/Man-in-the-middle_attack, accessed on 16 May 2022.

[92] *Kerckhoffs's principle*, Wikipedia, https://en.wikipedia.org/wiki/Kerckhoffs%27s_principle, accessed on 16 May 2022.

[93] *Letter frequency*, Wikipedia, https://en.wikipedia.org/wiki/Letter_frequency, accessed on 17 May 2022.

[94] *Bigram*, Wikipedia, https://en.wikipedia.org/wiki/Bigram, accessed on 17 May 2022.

[95] *Vigenère cipher*, Wikipedia, https://en.wikipedia.org/wiki/Vigen%C3%A8re_cipher, accessed on 18 May 2022.

[96] One-time pad, Wikipedia, https://en.wikipedia.org/wiki/One-time_pad, accessed on 19 May 2022.

[97] One-time pad: Key distribution, Wikipedia, https://en.wikipedia.org/wiki/One-time_pad#Key_distribution, accessed on 19 May 2022.

[98] *Grille (cryptography)*, Wikipedia, https://en.wikipedia.org/wiki/Grille_(cryptography), accessed on 21 May 2022.

[99] Stephen Fratini (2002), Encryption Using a Variant of the Turning-Grille Method, Mathematics Magazine, 75:5, 389-396, DOI: 10.1080/0025570X.2002.11953934.

[100] *Data Encryption Standard (DES)*, NIST, FIPS 46-1, https://csrc.nist.gov/publications/detail/fips/46/1/archive/1988-01-22.

[101] *Data Encryption Standard*, Wikipedia, https://en.wikipedia.org/wiki/Data_Encryption_Standard, accessed on 24 May 2022.

[102] *Advanced Encryption Standard (AES)*, NIST, FIPS 197, https://www.nist.gov/publications/advanced-encryption-standard-aes.

[103] *Advanced Encryption Standard*, Wikipedia, https://en.wikipedia.org/wiki/Advanced_Encryption_Standard, accessed on 24 May 2022.

[104] W. Diffie and M. Hellman, "New directions in cryptography," in IEEE Transactions on Information Theory, vol. 22, no. 6, pp. 644-654, November 1976, doi: 10.1109/TIT.1976.1055638. Available at https://ee.stanford.edu/~hellman/publications/24.pdf.

[105] Grooten, M., *How did they break Diffie-Hellman?*, ArsTechnica, https://arstechnica.com/information-technology/2015/11/op-ed-how-did-they-break-diffie-hellman/, accessed on 26 May 2022.

[106] R. L. Rivest, A. Shamir, and L. Adleman. 1978. A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM 21, 2 (Feb. 1978), 120–126. https://doi.org/10.1145/359340.359342.

[107] *Hash function*, Wikipedia, https://en.wikipedia.org/wiki/Hash_function, accessed on 29 May 2022.

[108] *Cryptographic hash function*, Wikipedia, https://en.wikipedia.org/wiki/Cryptographic_hash_function, accessed on 29 May 2022.

[109] Hayes, A., *What is a Blockchain*, Investopedia, https://www.investopedia.com/terms/b/blockchain.asp, accessed on 22 July 2022.

[110] *Blockchain*, NIST, https://www.nist.gov/blockchain, accessed on 30 May 2022.

[111] *What is blockchain technology?*, IBM, https://www.ibm.com/topics/what-is-blockchain, accessed on 30 May 2022.

[112] Brownworth, A., *Blockchain 101 - A Visual Demo*, YouTube, https://youtu.be/_160oMzblY8, accessed on 31 May 2022.

[113] RFC 8439, *ChaCha20 and Poly1305 for IETF Protocols*, IETF, https://datatracker.ietf.org/doc/html/rfc8439, accessed on 1 June 2022.

[114] Krovetz, T., Dai, W., *VMAC: Message Authentication Code using Universal Hashing*, IETF, https://datatracker.ietf.org/doc/html/draft-krovetz-vmac-01, accessed on 1 June 2022.

[115] *SipHash*, Wikipedia, https://en.wikipedia.org/wiki/SipHash, accessed on 1 June 2022.

[116] Satish, C.J., *SHA 512 - Secure Hash Algorithm - Step by Step Explanation - Cryptography - Cyber Security - CSE4003*, YouTube, https://youtu.be/JViXozmJnSk, accessed on 2 June 2022.

[117] *Number theory*, Wikipedia, https://en.wikipedia.org/wiki/Number_theory, accessed on 3 June 2022.

[118] *Complete (strong) induction*, Wikipedia, https://en.wikipedia.org/wiki/Mathematical_induction#Complete_(strong)_induction, accessed on 4 June 2022.

[119] *Largest known prime number*, Wikipedia, https://en.wikipedia.org/wiki/Largest_known_prime_number, accessed on 5 June 2022.

[120] *Palindromic primes*, The On-Line Encyclopedia of Integer Sequences (OEIS), https://oeis.org/A002385, accessed on 5 June 2022.

[121] *Palindromic prime*, Wikipedia, https://en.wikipedia.org/wiki/Palindromic_prime, accessed on 5 June 2022.

[122] *Emirp*, Wikipedia, https://en.wikipedia.org/wiki/Emirp, accessed on 5 June 2022.

[123] *Prime number theorem*, Wikipedia, https://en.wikipedia.org/wiki/Prime_number_theorem, accessed on 5 June 2022.

[124] *Euclidean algorithm – Procedure*, Wikipedia, https://en.wikipedia.org/wiki/Euclidean_algorithm#Procedure, accessed on 5 June 2022.

[125] *Greatest Common Divisor – Properties*, Wikipedia, https://en.wikipedia.org/wiki/Greatest_common_divisor#Properties, accessed on 6 June 2022.

[126] Sunzi Suanjing, Wikipedia, https://en.wikipedia.org/wiki/Sunzi_Suanjing, accessed on 11 June 2022.

[127] *Fermat primality test*, Wikipedia, https://en.wikipedia.org/wiki/Fermat_primality_test, accessed on 13 June 2022.

[128] Rosen, K.H., *Elementary Number Theory and Its Application**, 6th Edition, Pearson, 2010.

[129] *Miller–Rabin primality test*, Wikipedia, https://en.wikipedia.org/wiki/Miller%E2%80%93Rabin_primality_test, accessed on 15 June 2022.

[130] *Solovay–Strassen primality test*, Wikipedia, https://en.wikipedia.org/wiki/Solovay%E2%80%93Strassen_primality_test, accessed on 15 June 2022.

[131] *Three-pass protocol*, Wikipedia, https://en.wikipedia.org/wiki/Three-pass_protocol, accessed on 15 June 2022.

[132] *Category: Unsolved problems in number theory*, Wikipedia, https://en.wikipedia.org/wiki/Category:Unsolved_problems_in_number_theory, accessed on 17 June 2022.

* Indicates the book or article is available for borrowing from the Internet Achieve at https://archive.org/.

Index of Terms