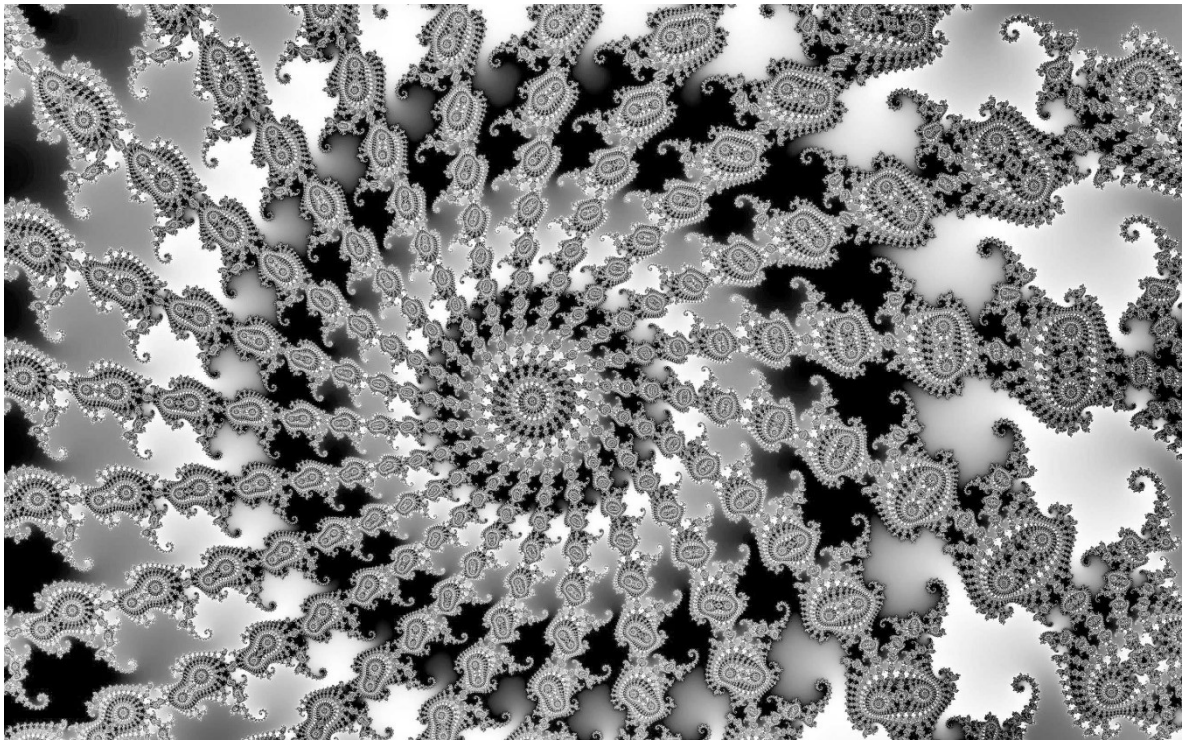
Mathematical Vignettes II



by **Stephen Fratini**

Table of Contents

List of Figures

List of Tables

Preface

"Only a few know, how much one must know to know how little one knows."
Werner Heisenberg

"The more I read, the more I acquire, the more certain I am that I know nothing."
Voltaire

“A good book is like a good friend. It will stay with you for the rest of your life.”
Charlie Lovett

“One friend in a storm is worth more than a thousand friends in the sunshine.”
Matshona Dhiliwayo

This book is a sequel to my previous book, Mathematical Vignettes [1]. It is a collection of short introductions to various topics in mathematics, each of which is about 10-30 pages long. The topics in this book are different from those in the first volume. My goal is to expose the reader to the wonderful world of mathematics and encourage further study. To this end, I have included over 180 references to additional material.

The topics covered in this book include:

- Combinatorial designs (e.g., Latin squares)
- Magic squares
- Finite geometries
- Abstract algebra
- Error correcting codes
- Geometric packing problems
- Knot theory
- Voting theory (including Arrow's impossibility theorem)
- The application of probability to genetics

Some proofs of theorems are included in this book to illustrate the particular topic at hand. However, the proofs of more complex theorems are often omitted, with reference to where the proof can be found. This is because including long and complex proofs would make the book too long and defeat the goal of providing short introductions to various aspects of mathematics..

In places where I state an opinion, I start my comment with “**Author’s Remark**”; otherwise, I’ve tried to stick to the facts.

“A man can fail many times, but he isn't a failure until he begins to blame somebody else.”
John Burroughs

Acknowledgements

The author would like to thank Tony Clark for his review and comments on Section 10 “Application of Probability to Genetics.”

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 (2nd Edition)*, self-published on Amazon, https://www.amazon.com/dp/B0CL34FRP1, October 2023.
- *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.
- *Mathematical Vignettes: Number theory, stochastic processes, game theory, cryptography, linear programming and more*, self-published on Amazon, https://www.amazon.com/Mathematical-Vignettes-stochastic-cryptography-programming-ebook/dp/B0BBP1PBJQ/, August 2022.
- *Learning Math through Puzzles: Number properties, counting, sequences and series, algebra, functions, and mathematical reasoning*, self-published on Amazon, https://www.amazon.com/dp/B0BZFRZP5B, March 2023.

Electronic versions of my books are available (free of charge) at https://www.artofmanagingthings.com/home/my-books.

1 Introduction

"A man who reads too much and uses his own brain too little falls into lazy habits of thinking."
Albert Einstein

“Nature is written in mathematical language.”
Galileo Galilei

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

In terms of prerequisites, Sections 2,3, 4, 7, 8 and 9 only require an understanding of high school algebra (including series and sequences), basic geometry and the fundamentals of matrices. Knowledge of abstract algebra (as summarized in Section 5) is required for an understanding of Section 6 concerning error correcting codes and to a lesser degree Section 7 concerning geometric packing problems. The required background in probability for Section 10 is provided at the beginning of the section.

1.4 Outline

The outline for the book is as follows:

- Section 1 is this introduction.
- Section 2 is about combinatorial designs (including Latin squares). Combinatorial design theory deals with the existence, construction and properties of systems of finite sets whose arrangements satisfy generalized concepts of balance and/or symmetry. Combinatorial designs are used in experimental designs, finite geometry (discussed in Section 4) and tournament scheduling.
- Magic squares (a popular topic in recreational mathematics) are discussed in Section 3.
- Section 4 covers finite geometries (i.e., geometries with a finite number of points). The various geometries are based on a relatively small number of axioms (assumptions) from which theorem are proven.
- Section 5 provides an short overview of abstract algebra. This is the most technically demanding section of the book.
- Section 6 is about error detecting and correcting codes. This section requires a basic understanding of abstract algebra (as covered in the previous section).

- Section 7 is about geometric packing problems, e.g., what is largest radius that allows one to pack (without overlap) 10 smaller circles into a unit circle?
- Mathematical knots (closed loops with several crossovers) are discussed in Section 8.
- Voting theory (including Arrows impossibility theorem) is covered in Section 9.
- Section 10 is about the application of probability to genetics (with summaries of the required prerequisites in probability and genetics).
- A list of acronyms, references and an index of terms are provided at the end of the book.

2 Combinatorial Designs

"Don't just teach your children to read. Teach them to question what they read. Teach them to question everything." — George Carlin

2.1 Overview

Combinatorial design theory addresses problems concerning the arrangement of elements from a finite set into subsets that have various balance properties. From the Wikipedia article on combinatorial design:

> Combinatorial design theory is the part of combinatorial mathematics that deals with the existence, construction and properties of systems of finite sets whose arrangements satisfy generalized concepts of balance and/or symmetry. These concepts are not made precise so that a wide range of objects can be thought of as being under the same umbrella. At times this might involve the numerical sizes of set intersections as in block designs, while at other times it could involve the spatial arrangement of entries in an array as in sudoku grids.

For example, consider the problem of assigning 7 people to 7 sets of size 3 such that each pair is assigned to exactly one group. Number the people from 1 to 7. Seven sets that satisfy the problem are

$$\{1,2,3\},\{1,4,5\},\{1,6,7\},\{2,4,6\},\{2,5,7\},\{3,4,7\},\{3,5,6\}$$

The solution is an example of something called a Balanced Incomplete Block Design (BIBD). The solution can also be represented as a graph (known as The Fane Plane) where each set is on a line or the circle, see Figure 1. BIBDs are discussed further in Section 2.3.2.

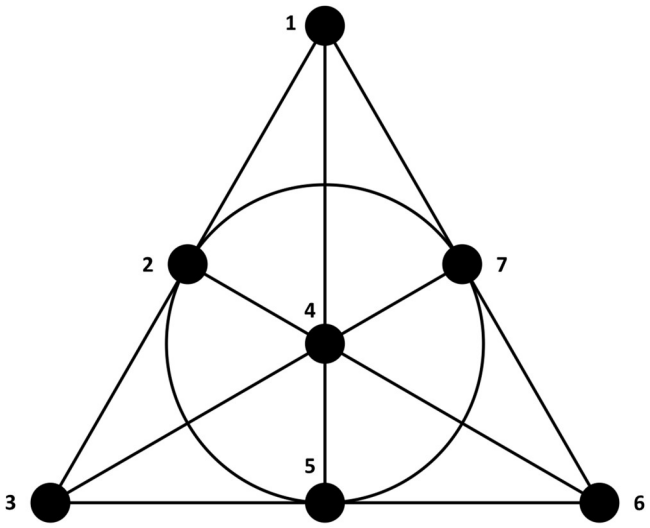


Figure 1. The Fano Plane

Combinatorial design theory has its beginnings in recreational mathematics. Combinatorial designs have many applications beyond recreational mathematics, e.g., tournament scheduling, experimental design, and coding theory (error detection and correction). In this section, we cover just a small sampling of the many types of designs and their applications. For more extensive coverage of the topic, see the following books:

- Combinatorial Design: Constructions and Analysis [3]
- Handbook of Combinatorial Design [4].

2.2 Latin Squares and Rectangles

Prerequisites: equivalence relationships [5], modular arithmetic (just the very basics) [6]

2.2.1 Definitions and Concepts

A **Latin square** is an $n \times n$ matrix populated with n different symbols such that each type of symbol occurs exactly once in each row and exactly once in each column. The term "Latin" is a reference to Latin letters which are sometimes used as the symbols in a Latin square. However, any collection of symbols can be used, e.g., positive integers or colors.

Figure 2 is an example of a $4\ x\ 4$ Latin square using symbols from the set $\{1,2,3,4\}$.

Figure 2. 4 x 4 Latin square using integers

4	2	1	3
1	3	4	2
3	1	2	4
2	4	3	1

Assuming there is some agreed order to the set of symbols being used (which is the case for integers and letters, but not for colors), a Latin square is said to be in **reduced (normal or standard) form** if both its first row and its first column are in the order ascribed to the corresponding set of symbols. The Latin square in Figure 2 is not in reduced form. It is always possible to rearrange a Latin square into reduced form using column and row operations. In Figure 3, the columns of the Latin square in Figure 2 have been permutated to put the entries in the top row in ascending order.

Figure 3. Reducing a Latin square – column rearrangement

1	2	3	4
4	3	2	1
2	1	4	3
3	4	1	2

In Figure 4, the rows of the Latin square in Figure 3 have been permutated to put the entries in the left column in ascending order. The resulting Latin square is now in reduced form.

Figure 4. Reducing a Latin square – row rearrangement

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

If the rows, columns, or the names of the symbols in a Latin square are permutated, a new Latin square is obtained. If a Latin square can be obtained from another using one or more row, column or symbol permutations, the two Latin squares are said to be **isotopic**. Isotopism is an equivalence relation [5], and thus, the set of all Latin squares is divided into equivalence classes (referred to as isotopy classes). For example, the Latin squares in Figure 2, Figure 3 and Figure 4 are in the same isotopy class. As an example of symbol permutation, we switch each appearance of 1 and 4 in Figure 4 to get the isotopic Latin square in Figure 5.

Figure 5. Symbol permutation in a Latin square

4	2	3	1
2	4	1	3
3	1	4	2
1	3	2	4

An **intercalate** is a $2 \; x \; 2$ subset of a Latin square which is also a Latin square. The elements of the intercalate do not need to be adjacent within the containing Latin square. For example, the four shaded elements in Figure 6 are an intercalate. The Latin square in Figure 6 has a total of 12 intercalate. See if you can find the other 11.

Figure 6. Intercalate in a Latin square

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

Since the number of intercalates in a Latin square is not changed by row permutation, column permutation or symbol permutation, isotopic Latin squares have the same number of intercalates. The two Latin squares in Figure 7 are not isotropic, since the Latin square on the left has 12 intercalates and the one on the right has only 4 (one of which is shown in gray).

Figure 7. Non-isotropic Latin squares

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

1	2	3	4
2	3	4	1
3	4	1	2
4	1	2	3

There are only two equivalence classes of $4\ x\ 4$ Latin squares. The Latin squares in Figure 7 are representatives from each equivalence class. For $5\ x\ 5$ Latin squares there are also only two equivalence classes. However, as the size of the Latin square increases, the number of equivalence classes increases very quickly, as shown in Table 1 (from OEIS sequence A040082 [8]).

Table 1. Number of n x n Latin square equivalence classes

n	Number of Latin square equivalence classes
1	1
2	1
3	1
4	2
5	2
6	22
7	564
8	1676267
9	115618721533
10	208904371354363006
11	12216177315369229261482540

Since each equivalence class contains several Latin squares the total number Latin squares of a given sizes is even larger, as shown in Table 2 (from OEIS sequence A002860 [7]).

Table 2. Total number of n x n Latin squares

n	Total number of Latin squares
1	1
2	2
3	12
4	576
5	161280
6	812851200
7	61479419904000
8	108776032459082956800
9	5524751496156892842531225600
10	9982437658213039871725064756920320000
11	776966836171770144107444346734230682311065600000

. . .

If each entry of an $n \times n$ Latin square is written as a triple (r, c, s), where r is the row, c is the column, and s is the symbol, we obtain what is called the **orthogonal array representation** of the Latin square. Based on the orthogonal array representation, we can reformulate the definition of a Latin square as follows:

> A Latin square is a set of n^2 triples (r, c, s), where $1 \leq r, c, s \leq n$, such that all ordered pairs (r, c) are distinct, all ordered pairs (r, s) are distinct, and all ordered pairs (c, s) are distinct.

The orthogonal array representation of the Latin square in Figure 4 is shown in Figure 8.

Figure 8. Orthogonal array representation of a Latin square

row	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
column	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
symbol	1	2	3	4	2	1	4	3	3	4	1	2	4	3	2	1

The orthogonal array representation gives us another way to describe rearrangements of a Latin square that, in turn, produce another Latin square. The rearrangements are achieved by systematically and consistently permutating the three items in each triple in the representation. For example, we can replace each triple (r, c, s) by (c, r, s) which corresponds to transposing the square (reflecting about its main diagonal). Surprisingly, permutation of the symbols with either the column or row entries yields another Latin square. There are 6 possible permutations including "do nothing", giving us 6 Latin squares called the **conjugates** (also **parastrophes**) of the original Latin square.

. . .

A **transversal** of an $n \; x \; n$ Latin square is a set S of n cells, such that each row contains one element (i.e., cell) from S, each column contains one element (i.e., cell) from set S, and each symbol of the Latin square appears in exactly on cell of S.

In Figure 9, one of the four transversals of the given Latin square is shown (see the gray cells).

Figure 9. Transversal of a Latin square

1	2	3	4
3	4	1	2
4	3	2	1
2	1	4	3

Not all Latin squares have transversals. For example, the Latin square in Figure 10 does not have any transversals.

Figure 10. Latin square with no transversals

1	2	3	4
2	3	4	1
3	4	1	2
4	1	2	3

. . .

Two Latin squares are said to be **orthogonal** if when superimposed the ordered paired entries in the positions are all distinct. In such cases, the Latin squares are said to be orthogonal mates.

Figure 11. Orthogonal Latin squares

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

1	2	3	4
4	3	2	1
2	1	4	3
3	4	1	2

In Figure 12, we've superimposed the two Latin squares from Figure 11. In each cell of Figure 12, the first number in the ordered pair is from the Latin square on the left in Figure 11 and the second number in the ordered pair is from the Latin square on the right. The key requirement is that each pair is different (which is the case in this example).

Figure 12. Overlap of orthogonal Latin squares

(1,1)	(2,2)	(3,3)	(4,4)
(2,4)	(1,3)	(4,2)	(3,1)
(3,2)	(4,1)	(1,4)	(2,3)
(4,3)	(3,4)	(2,1)	(1,2)

The following is a puzzle whose solution entails orthogonal Latin squares:

> Take all the aces, kings, queens and jacks from a standard deck of 52 playing cards, and arrange them in a $4\ x\ 4$ grid such that each row and each column contains all four suits as well as one of each face value.

One solution is essentially the grid in Figure 12, if we replace $\{1,2,3,4\}$ in the first Latin square with {Ace, King, Queen, Jack}, and replace $\{1,2,3,4\}$ in the second Latin square with {Spade, Heart, Diamond, Club}, as shown below.

(Ace, Spade)	(King, Heart)	(Queen, Diamond)	(Jack, Club)
(King, Club)	(Ace, Diamond)	(Jack, Heart)	(Queen, Spade)
(Queen, Heart)	(Jack, Spade)	(Ace, Club)	(King, Diamond)
(Jack, Diamond)	(Queen, Club)	(King, Spade)	(Ace, Heart)

The "Thirty-six officers' problem" is another puzzle that involves orthogonal Latin squares. The puzzle goes as follows:

> There are 6 ranks of officers in a military unit (e.g., 1st Lieutenant, 2nd Lieutenant, Captain, Major, Lieutenant Colonel, Colonel) and 6 regiments within the unit (number them from 1 to 6). There is one officer for each rank/regiment pair for a total of 36 officers. Arrange the 36 officers in a square grid such that in each line (both horizontal and vertical) there are 6 officers of different ranks and different regiments.

A pair of $6\ x\ 6$ orthogonal Latin squares would solve the problem if such existed. However, it has been proven in 1901 by Gaston Tarry that no pairs of $6\ x\ 6$ orthogonal Latin squares exist [9].

. . .

Assume that L1 and L2 are $n\ x\ n$ orthogonal Latin squares. Consider all the cells of L2 that contain the same symbol s, then the entries in the corresponding cells of the L1 must all be different; otherwise, the squares would not be orthogonal. For example, L2 is the Latin square on the right of Figure 13 and L1 is on the left. We choose symbol 1 (i.e., $s = 1$) in L2 and note that the symbols in the same position in L1 are all different.

Figure 13. Orthogonal Latin squares and transversals

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

1	2	3	4
4	3	2	1
2	1	4	3
3	4	1	2

Since the symbol s occurs exactly once in each row and once in each column of L2, the n entries in L1 corresponding to the n appearances of the symbol s in L2, have the property of all being different, and occurring once in each row and once in each column of the square L1, i.e., they are a transversal of L1. Thus, L1 has n transversals (one for each symbol in L2). Similarly, L2 has n transversals (one for each symbol in L1). Thus, we have the following theorem.

Theorem 1. *A Latin square of order* $\boldsymbol{n}$ *has an orthogonal mate if and only if it has* $\boldsymbol{n}$ *disjoint transversals.*

The Wikipedia article entitled "Mutually orthogonal Latin squares" [10] provides additional concepts concerning orthogonal squares, e.g., there exist pairs of orthogonal Latin squares for every odd order.

. . .

A **Latin rectangle** is an $m \times n$ matrix (with $m \leq n$), populated with n different symbols, e.g., $\{1,2,3, \ldots n\}$, such that no symbol occurs more than once in any row or column. The definition implies there are exactly m appearances of each of the n symbols. An example of a Latin rectangle is shown in Figure 14.

Figure 14. Example of a 3 x 7 Latin rectangle

1	2	3	4	5	6	7
2	3	4	5	6	7	1
3	4	5	6	7	1	2

. . .

It is also possible to place several symbols in each cell. Such an entity is known as a **semi-Latin rectangle**. An example of a $(5 \times 10)/2$ semi-Latin rectangle is shown in Figure 15. The notation $(5 \times 10)/2$ means the dimension of the rectangle is 5×10 and each cell has 2 symbols. Further, each symbol appears the same number of times in each row, i.e., 2 times, and each symbol appears the same number of times in each column, i.e., once. The order of the two symbols in each cell does not matter, which is why they are represented as a set and not an ordered pair.

Figure 15. Semi-Latin rectangle example

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

More formally, we define a semi-Latin rectangle as follows:

> Let v, r, c, k, x, y be positive integers such that $r > 1, c > 1$, $yv = kr$, and $xv = kc$. Then, an $(r \times c) / k$ semi-Latin rectangle is a row-column design in which v symbols are arranged into r rows and c columns, where each row-column intersection (cell) contains k symbols such that each symbol appearing at most once in each cell. Further, each symbol appears x times in each row, and y times in each column.

In our example, $v = 10, r = 5, c = 10, k = 2, x = 2, y = 1$.

The example in Figure 16 was constructed using something called a **starter**. In this case, the starter is the first column of the rectangle. To create the second column, we add $1\ (mod\ 9)$ to each number in the first column. When then add $1\ (mod\ 9)$ to each number second column to create the third column, and so on. The proper selection of the starter ensures that the rectangle will meet the requirements for a semi-Latin rectangle. For more details on constructing semi-Latin rectangles, see the article entitled “Constructions for regular-graph semi-Latin rectangles with block size two” [11].

2.2.2 Experimental Design

Latin squares are used in experimental designs to average out what are called nuisance factors, i.e., factors which are not of interest but may influence the results of an experiment. The assumption is that nuisance factors are known and can be controlled (at least in the context of an experiment).

For example, assume that a company wants to determine the effectiveness of a new fertilizer on leafy house plants. Further, the fertilizer comes in 3 different strengths which are referred to as treatments 1, 2 and 3. The nuisance factors are determined to be the type of potting soil used and the type of house plant. The experiment designer decides there are 3 types of potting soils (S1, S2 and S3) and 3 types of leafy plants (P1, P2 and P3) to be considered in the design. The Latin square in Figure 16 shows the protocol to be used for the experiment.

Figure 16. Latin square design for fertilizer experiment

	P1	***P2***	***P3***
S1	1	2	3
S2	2	3	1
S3	3	1	2

In this design, each treatment is tested against each of the soil types and each of the plant types **exactly once**. Each of the nine cells in the shaded region of the Latin square represents a test. For example, treatment 1 is tested against the combination P1 and S1 (top-left shaded cell).

In this example and in general, for experiments based on Latin squares, each treatment is **not** tested against all combinations of nuisance factors. In our example, each treatment is only tested against 3 of the 9 possible combination of nuisance factors.

In this type of experiment, the typical hypothesis is that there is no difference in the treatments. The intent (hope) is to reject the hypothesis with a high degree of certainty and thus, conclude that there is a difference in the treatments. The next step would be to determine which treatment is best (this could involve another experiment).

. . .

A shortcoming of Latin square experiment design is that there needs to be the same number of levels (types) for each nuisance factor as there are treatments. If, in the above experiment, we only had two potting soil types (say S1 and S2), then a Latin square design could not be applied directly. Similarly, a Latin square design would not work if we had more levels for a given nuisance type than the number of treatments. If we had plant types P1, P2, P3 and P4 in our example, Latin square design would not apply.

On approach, is to test all combinations of nuisance factors with each treatment. For the example suggested above, that would be $2 \times 4 \times 3 = 24$ tests.

The exhaustive approach can become very expensive depending on the cost of each test, the number of nuisance factor levels and number of treatments. For example, consider an experiment with nuisance factor A having 5 levels, nuisance factor B having 9 levels, and 3 types of treatments. This could be modeled with the unbalanced Latin rectangle design shown in Figure 17. The rows represent the levels of nuisance factor A, and the columns represent the levels of nuisance factor B. The treatments (1,2 and 3) are distributed cyclically (with wrap-around) starting from the top-left of the rectangle. Each treatment is used three times for each level of nuisance factor B (this is balanced part of the design). However, the columns are unbalanced (using two treatments twice and one treatment once).

Figure 17. Latin rectangle unbalance design

1	3	2	1	3	2	1	3	2
2	1	3	2	1	3	2	1	3
3	2	1	3	2	1	3	2	1
1	3	2	1	3	2	1	3	2
2	1	3	2	1	3	2	1	3

Another issue occurs when there are more treatments than the number of levels of either nuisance factor. For example, consider an experiment with nuisance factor A having 5 levels, nuisance factor B having 5 levels, and 8 treatments (numbered 1 through 8). One approach would be to first create a Latin square using only 5 of the treatments, as shown below.

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

The next step is to add in treatments 6,7,8 while keeping the designed balanced, see below. Each treatment appears once in each row, and once in each column. However, some cells (blocks) have 2 treatments and the other cells have just one treatment. An exhaustive design would require $5 \times 5 \times 8 = 200$ tests, but the design below requires only 40 tests.

{1,6}	2	3	{4,8}	{5,7}
{2,7}	{3,6}	4	5	{1,8}
{3,8}	{4,7}	{5,6}	1	2
4	{5,8}	{1,7}	{2,6}	3
5	1	{2,8}	{3,7}	{4,6}

We can recast the above as a semi-Latin rectangle (square in this case) if we add two "do nothing" treatments (call them treatments 9 and 10). This gives us the equivalent design shown below:

{1,6}	{2,***10***}	{3,***9***}	{4,8}	{5,7}
{2,7}	{3,6}	{4,***10***}	{5,***9***}	{1,8}
{3,8}	{4,7}	{5,6}	{1,***10***}	{2,***9***}
{4,***9***)	{5,8}	{1,7}	{2,6}	{3,***10***}
{5,***10***}	{1,**9**}	{2,8}	{3,7}	{4,6}

. . .

Another shortcoming of Latin squares is that they can only handle 2 nuisance factors. However, one can use Mutually Orthogonal Latin Squares (MOLS) to handle 3 nuisance factors. For our fertilizer experiment, assume that humidity is another nuisance factor. Divide humidity into the following levels:

- H1: humidity between 0 and 33 percent
- H2: humidity between 34 and 67 percent
- H3: humidity between 68-100 percent.

We next select a Latin square that is orthogonal to the one in Figure 16, see Figure 18.

Figure 18. Second Latin square for enhanced fertilizer experiment

H1	H2	H3
H3	H1	2
H2	H3	H1

Next, we superimpose the two orthogonal Latin squares, see Figure 19. The first element in each pair is from the Latin square in Figure 16, and the second element is from the Latin square in Figure 18. For example, the element (3,1) means to use treatment 3 with potting soil S2 on plant type P2 and with humidity level H1. We have effectively created a 3-dimensional table. In terms of balance, each humidity level is used with each type of potting soil, and each humidity level is used with each type of plant.

Figure 19. Latin square design for enhanced fertilizer experiment

	P1	***P2***	***P3***
S1	(1,H1)	(2,H2)	(3,H3)
S2	(2,H3)	(3,H1)	(1,H2)
S3	(3,H2)	(1,H3)	(2,H1)

2.2.3 Error Detecting and Correcting Codes

Error detecting and correcting codes can be used to enhance the quality of electronic communications. The applicability of Latin squares to this topic is covered in Section 4.

2.2.4 Mathematical Puzzles

Several mathematical puzzles are based on Latin squares. The most famous of these puzzles is **Sudoku**. A Sudoku puzzle is a 9×9 Latin square with several entries missing. To solve the puzzle, one must use logical reasoning to determine the values of the missing entries. Further, there is one additional set of conditions, i.e., each 3×3 sub-square within the larger square must also be a Latin square. There are 9 of these 3×3 sub-squares (separated by heavy lines in Figure 20). Sudoku puzzles are constructed to have a unique solution.

From the Wikipedia article on Sudoku [12]:

> The fewest clues possible for a proper Sudoku is 17 (proven January 2012, and confirmed September 2013). Over 49,000 Sudokus with 17 clues have been found, many by Japanese enthusiasts.
>
> The number of classic 9×9 Sudoku solution grids is 6,670,903,752,021,072,936,960 (sequence A107739 in the OEIS [13]), or around 6.67×10^{21}. This is roughly 1.2×10^{-6} times the number of 9×9 Latin squares. Various other grid sizes have also been enumerated—see the main article for details. The number of essentially different solutions, when symmetries such as rotation, reflection, permutation, and relabeling are taken into account, was shown to be just 5,472,730,538 (sequence A109741 in the OEIS [14]).

The book by Rosenhouse and Taalman [15] goes into the detailed mathematics behind Sudoku.

Figure 20. Sudoku puzzle

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

The unique solution to the puzzle in Figure 20 is shown in Figure 21.

Figure 21. Solution to Sudoku puzzle

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

. . .

KenKen® is another puzzle game based on Latin squares. In KenKen, a Latin square is divided into various regions (often called “cages”) with an arithmetic clue given for each cage. For example, in the puzzle shown in Figure 22, "$24\times$" means that the product of the three cells in the associated cage must multiply to equal 24 (nothing is implied about the order of the numbers within the region). The expression "$2\div$" means that the quotient of the two cells in that cage must equal 2. The value of only one cell is given, i.e., the 3 in the second row and second column. Many KenKen puzzles don’t provide the value for any of the cells.

From the Wikipedia article on KenKen®:

> KenKen and KenDoku are trademarked names for a style of arithmetic and logic puzzle invented in 2004 by Japanese math teacher Tetsuya Miyamoto, who intended the puzzles to be an instruction-free method of training the brain. The name derives from the Japanese

word for cleverness. The names Calcudoku and Mathdoku are sometimes used by those who do not have the rights to use the KenKen or KenDoku trademarks.

Figure 22. KenKen puzzle

$24\times$		$8+$	
	3		
$2\div$	$3-$		$12+$

The solution to the KenKen puzzle in Figure 22 is shown in Figure 23. For additional KenKen puzzles, see https://www.kenkenpuzzle.com/.

Figure 23. Solution to KenKen puzzle

3	2	1	4
4	3	2	1
2	1	4	3
1	4	3	2

2.3 Block Designs

2.3.1 Overview

Latin squares are an example of something called a block design. In this section, we will formally define "block design" and then analyze some examples.

A **block design** is a pair (X, A) such that

- X is a set of elements called points, and
- A is a multiset consisting of nonempty subsets of X called blocks.

[**Note**: According to accepted definitions in mathematics, the elements of a set are distinct and unordered, whereas a multiset can have repeated elements. For example, $A = \{1,2,3\}$ is a set and $B = \{1,2,2,3\}$ is a multiset but not a set.]

Example 1: The orthogonal array representation of the Latin square shown in Figure 8 is an example of a block design with

$$X = \{1,2,3,4\}$$

$$A = \{\{1,1,1\},\{1,2,2\},\{1,3,3\},\{1,4,4\},\{2,1,2\},\{2,2,1\},\{2,3,4\},\{2,4,3\},$$
$$\{3,1,3\},\{3,2,4\},\{3,3,1\},\{3,4,2\},\{4,1,4\},\{4,2,3\},\{4,3,2\},\{4,4,1\}\}$$

In this example, there are no repeated blocks. Such block designs are referred to as simple block designs. In what follows and when there is no potential for confusion, we will use a shorthand notation for the elements of A, e.g., we will write 313 instead of $\{3,1,3\}$.

A block design is said to be **balanced** (up to t) if all t-subsets (i.e., subsets size t) of the original set X occur in equally many blocks. Such a block design is known as a **t-design**.

Example 2: The block design with

$$X = \{0,1,2,3,4,5,6,7\}$$

$$A = \{0123, 0124, 0156, 0257, 0345, 0367, 0467, 1267, 1346, 1357, 1457, 2347, 2356, 2456\}$$

is a 2-design since each set of 2 elements from X appear in exactly 3 blocks.

The Latin square in Figure 8 is not a 2-design since not all pairs appear the same number of times, e.g., $\{1,2\}$ appears in 3 blocks but $\{3,4\}$ appears in 6 blocks. It is not even a 1-design since 1 appears in 10 blocks, and 2,3, and 4 each appear in 9 blocks,.

Example 3: Some Latin squares are 2-designs. For example, the Latin square in Figure 24 is a 2-design, with each pair of distinct symbols appearing in exactly 4 blocks.

Figure 24. Latin square that is a 2-design

row	1	1	1	2	2	2	3	3	3
column	1	2	3	1	2	3	1	2	3
symbol	1	2	3	2	3	1	3	1	2

If no block contains all the elements of X, then the design is said to be **incomplete**. Examples 1 and 2 above are incomplete block designs.

If each element of a design occurs in the same number of blocks (the **replication number**, denoted by r), then the design is said to be **regular**.

If all the blocks in a design are of the same size (denoted by k), then the design is said to be **uniform**.

Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are two designs where X and Y have the same number of elements. The two designs are said to be **isomorphic** if there exists a bijection $\alpha: X \to Y$ such that

$$\{\{\alpha(x) : x \in A\} : A \in \mathcal{A}\} = \mathcal{B}$$

In words, if we rename every point $x \in X$ by $\alpha(x)$, then the collection of blocks $\mathcal{A}$ is transformed into the collection of blocks $\mathcal{B}$.

[**Note**: A **bijection** is a function between the elements of two sets, where each element of one set is paired with exactly one element of the other set, and each element of the other set is paired with exactly one element of the first set.]

2.3.2 Balanced Incomplete Block Designs

A **Balanced Incomplete Block Design** (BIBD) is an incomplete block design where each pairs of elements occur in the same number of blocks. BIBDs are used extensively in experimental design. The Fano plane in Figure 1 is an example of a BIBD. More formally, we define a BIBD as follows:

> Let v, k, and λ be positive integers such that $v > k \geq 2$. A (v, k, λ) balanced incomplete block design (abbreviated as $(v, k, \lambda) - BIBD$) is a design (X, A) such that:
>
> - the number of elements in X is v

- (uniform design) each block contains exactly k elements
- (regular) every elements appears in r blocks where r is the replication number
- (balanced) every pair of distinct elements is contained in exactly λ blocks.

The number of blocks b can be calculated from the given parameters, i.e., $b = \frac{vr}{k}$.

. . .

As an example, consider a baking contest with 7 bakers (each making their favorite pastry) and 14 judges where each judge is to evaluate 3 pasties. This would require each baker to make $\frac{3 \cdot 14}{7} = 6$ instances of their pastry (so the replication number is 6). Further, we have the requirement that each pair of pastries (from different bakers) is to be evaluated by 2 judges. What we have described is a (7,3,2)-BIBD where the 7 pastry types are the elements of the BIBD, and each judge is a block. The block design is shown in Figure 25. The top row has the block number.

Figure 25. (7,3,2)-BIBD

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
2	1	2	2	1	3	1	1	3	3	1	1	4	2
6	3	6	4	2	5	4	2	4	5	4	5	5	3
7	6	7	5	3	7	7	5	6	7	7	6	6	4

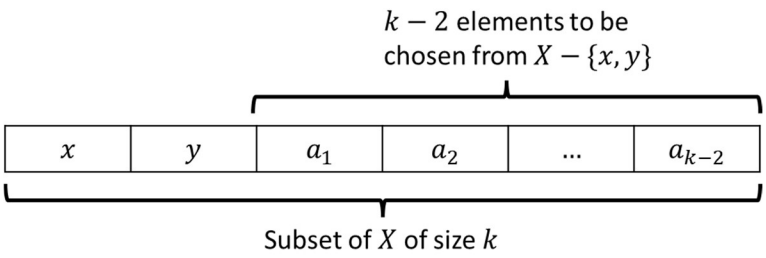
Figure 26 shows an alternate solution to the problem. The BIBD was calculated using the online capability at https://rdrr.io/cran/ibd/man/bibd.html using input bibd(7,14,6,3,2,pbar=FALSE).

Figure 26. Alternate solution for (7,3,2)-BIBD

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	1	5	5	1	1	3	1	3	2	1	2	2	2
3	4	6	6	4	3	4	2	4	4	2	4	3	3
6	5	7	7	5	7	7	7	6	7	6	6	5	5

. . .

In a block design, let X be of size v and let A be the collection of all subsets of X each of size k. Given any pair $\{x, y\}$ from X there are $\binom{v-2}{k-2}$ subsets of size k that contain $\{x, y\}$, see the figure below. Thus, (X, A) is a $\left(v, k, \binom{v-2}{k-2}\right)$-BIBD. Keep in mind that the order of the elements in the subsets of X does not matter.



. . .

Given the number of elements in X (i.e., v), the size of the blocks (i.e., k), and λ, it is possible to compute r (the number of blocks in which each element of X appears).

Theorem 2. In a (v,k,λ)-BIBD, every element occurs in exactly $r = \frac{\lambda(v-1)}{(k-1)}$ blocks.

Proof: Let (X,A) be the (v,k,λ)-BIBD given in the statement of the theorem.

Let $x \in X$, and let r_x be the number of blocks containing x.

Define the set

$$S = \{(y,B) : y \in X, y \neq x, B \in A, \{x,y\} \subseteq B\}$$

In words, S is the set of ordered pairs (y,B) such that y is an element of X, $y \neq x$, B is an element of A (i.e., B is a block in the BIBD), and $\{x,y\}$ is a pair in B.

Next, we compute the number of elements in S in two different ways, and then equate the results (which will give us the desired result).

First, there are $v-1$ ways to choose $y \in X$ such that $y \neq x$. For each such y, there are λ blocks B such that $\{x,y\} \subseteq B$ (this follows from the definition of λ in a (v,k,λ)-BIBD). Thus, the number of elements in S is $\lambda(v-1)$.

Alternately, there are r_x ways to choose a block B such that $x \in B$. For each choice of B, there are $k-1$ ways to choose $y \in B$ where $y \neq x$. So, using this approach, the number of elements in S is $r_x(k-1)$.

Combining the two results, we have

$$\lambda(v-1) = r_x(k-1)$$

which implies

$$r_x = \frac{\lambda(v-1)}{(k-1)}$$

Noting that above equation is independent of x, we get the desired result. ■

Applying Theorem 2 to the baking contest example, we determine that each pastry is evaluated by $r = \frac{2(7-1)}{(3-1)} = 6$ judges.

Applying Theorem 2 to the example concerning all subsets of size k, we have that each element in the BIBD appears in the following number of blocks:

$$\binom{v-2}{k-2}\cdot\frac{(v-1)}{(k-1)}=\frac{(v-2)!}{(v-k)!\,(k-2)!}\cdot\frac{(v-1)}{(k-1)}=\frac{(v-1)!}{(v-k)!\,(k-1)!}=\binom{v-1}{k-1}$$

It is possible to compute the number of blocks in a design given other information as described in the following theorem.

Theorem 3. The number of blocks b in a (v,k,λ)-BIBD with replication number r is given by

$$b=\frac{vr}{k}=\frac{\lambda(v^2-v)}{k^2-k}$$

Proof: Let the BIBD be (X,A). Using a technique similar to the proof of the previous theorem, we define the following set and compute its number of elements in two different ways.

$$S\ =\ \{(x,B):\ x\in\ X,B\in A\}$$

There are v ways to choose $x\in X$. For each such x, there are r blocks B such that $x\in B$ (this is given in the statement of the theorem). Thus, S has vr elements.

Alternately, there are b ways to choose a block $B\in A$ (also given in the statement of the theorem). For each choice of B, there are k ways to choose $x\in B$. So, the S has bk elements.

Combining these two results, we have $bk=vr$ which implies $b=\frac{vr}{k}$. Further, from Theorem 2, we have $r=\frac{\lambda(v-1)}{(k-1)}$ which when substituted into the formula for b, we get $b=\frac{\lambda v(v-1)}{k(k-1)}=\frac{\lambda(v^2-v)}{k^2-k}$. ■

The previous two theorem can be used to prove the non-existence of certain BIBDs.

- Theorem 2 implies that $k-1$ must exactly divide $\lambda(v-1)$. For example, a BIBD with parameters $v=9,\ k=6,\ \lambda=3$ is not possible since $k-1=5$ does not divide $\lambda(v-1)=3\cdot 8=24$.
- Theorem 3 implies that $k(k-1)$ must exactly divide $\lambda v(v-1)$. For example, a BIBD with parameters $v=11,k=4,\ \lambda=5$ is not possible since $k(k-1)=12$ does not divide $\lambda v(v-1)=5\cdot 110=550$.

It is sometimes convenient to use the notation $(\boldsymbol{v},\boldsymbol{b},\boldsymbol{r},\boldsymbol{k},\boldsymbol{\lambda})$**-BIBD** to represent a (v,k,λ)-BIBD with b blocks and replication number r even though b and r can be derived from the other 3 variables.

In addition to Theorem 2 and Theorem 3, the following theorem provides a necessary condition for the existence of a BIBD with given parameters.

Theorem 4. (Fisher's Inequality) Given a (v, b, r, k, λ)-BIBD, then it must be that b ≥ v, i.e., the number of blocks must be greater than or equal to the number of elements in the BIBD.

Note: From Theorem 3, the condition $r\geq k$ is equivalent to $b\geq v$.

Proof: See the Wikipedia article entitled "Fisher's inequality" [17].

Is a BIBD with $v=17$, $k=9$ and $\lambda=1$ possible? Using the formula from Theorem 3, we have that

$$r=\frac{\lambda(v-1)}{(k-1)}=\frac{1(17-1)}{(9-1)}=\frac{16}{8}=2$$

Since $r = 2 < 9 = k$, we know that such a BIBD is not possible by Theorem 4.

. . .

There is an alternative representation of a block design (not just BIBDs) known as an **incidence matrix**. When representing a block design with an incidence matrix, a 1 is used to indicate when an element is in a given block and 0 is used to indicate an element is not in a given block. More formally, we define an incidence matrix as follows.

Let (X, A) be a block design where $X = \{x_1, x_2, \dots, x_v\}$ and $A = \{A_1, A_2, \dots, A_b\}$. The incidence matrix of (X, A) is the $v \times b$ matrix $M = (m_{ij})$ where

$$m_{ij} = \begin{cases} 1, & x_i \in A_j \\ 0, & x_i \notin A_j \end{cases}$$

The following properties hold true for the incidence matrix M of a (v, b, r, k, λ)-BIBD:

- every column of M contains exactly k ones
- every row of M contains exactly r ones
- two distinct rows of M each contain ones in exactly λ columns.

For example, take the (7,3,1)-BIBD with elements $\{1, 2, 3, 4, 5, 6, 7\}$, as depicted in Figure 27. Using Theorem 3, we can compute the number of blocks, i.e.,

$$b = \frac{\lambda(v^2 - v)}{k^2 - k} = \frac{1 \cdot (49 - 7)}{9 - 3} = \frac{42}{6} = 7$$

The top row in the figure lists the block numbers.

Figure 27. (7,3,1)-BIBD

1	**2**	**3**	**4**	**5**	**6**	**7**
1	1	1	2	2	3	4
2	3	5	3	6	4	5
4	7	6	5	7	6	7

The associated incidence matrix is shown in Figure 28. Each row represents elements within a block, and each column indicates the blocks in which a given element appears.

Figure 28. Incidence matrix for (7,3,1)-BIBD

1	1	0	1	0	0	0
1	0	1	0	0	0	1
1	0	0	0	1	1	0
0	1	1	0	1	0	0
0	1	0	0	0	1	1
0	0	1	1	0	1	0
0	0	0	1	1	0	1

. . .

Experiment design using a BIBD: Suppose that a small research company has developed 7 potential treatments (drugs) for a rare disease. Based on the ingredients in the treatments and on some animal studies, the treatments are considered to be safe for administration to humans. The researchers think that the treatments work well individually or in combination, and would like to try out every possible combination, i.e., $2^7 - 1 = 127$. However, administration of the treatment combinations take several weeks, and it will be costly to employ 127 test subjects. In fact, the company only has funds (from a government grant) for 7 test subjects. The researchers think that only small gains will be achieve by using more than 2 treatments, but even trying every pair of treatments on a different person would require $\binom{7}{2} = 21$ people.

Given the limitations on the number of test subjects, one strategy would be to use a (7,3,1)-BIBD, with $X = \{1,2,3,4,5,6,7\}$ and $A = \{123,145,167,246,257,347,356\}$. The elements of X are the treatments, and each block corresponds to a test subject. For example, test subject #3 (block #3) is given treatments 1, 6 and 7. In this scheme, every pair of treatments is used together on precisely one of the test subjects.

2.3.3 Symmetric BIBDs

A **Symmetric Balanced Incomplete Block Design** (SBIBD) is a BIBD in which the number of elements equals the number of blocks, i.e., $v = b$. SBIBDs are an important and well-studied subclass of BIBDs.

The (7,3,1)-BIBD in the experimental design from the previous example is also an SBIBD since we have $v = b = 7$.

As another example, consider the collection of $(v, v-1, v-2)$-BIBDs with $v = k + 1$. From Theorem 2, we have

$$r = \frac{(v-2)(v-1)}{k-1} = \frac{(k-1)k}{k-1} = k$$

and from Theorem 3, we have

$$b = \frac{vr}{k} = \frac{(k+1)k}{k} = k + 1$$

Thus, $v = b$ and $(v, v-1, v-2)$-BIBD is a family of SBIBDs.

Theorem 5. In a symmetric (v, k, λ)-BIBD, any two blocks have λ elements in common.

Proof: For a proof, see Theorem 2.2 in Stinson [3].

One can verify that each pair of blocks in the symmetric (7,3,1)-BIBD example (Figure 27) has $\lambda = 1$ element in common.

2.3.4 Resolvable BIBDs

Suppose (X, A) is a (v, k, λ)-BIBD. A **parallel class** in (X, A) is a subset of disjoint blocks from A whose union is X. A partition of A into c parallel classes is called a resolution. A BIBD is said to be a **resolvable BIBD** if it has at least one resolution.

A BIBD is resolvable if and only if $b \geq v + c - 1$ (see Theorem 5.4 in the book "Design Theory" [18]).

We can partition the $(6,2,1)$-BIDB with $X = \{1,2,3,4,5,6\}$ and A consisting of all subsets of X of size two as follows:

$$P_1 = \{12, 36, 45\}$$
$$P_2 = \{13, 42, 56\}$$
$$P_3 = \{14, 53, 62\}$$
$$P_4 = \{15, 64, 23\}$$
$$P_5 = \{16, 25, 34\}$$
$$\ldots$$

The $(9,3,1)$-BIBD with $X = \{1,2,3,4,5,6,7,8,9\}$ and

$$A = \{123, 145, 167, 189, 246, 257, 289, 347, 358, 369, 456, 478\}$$

is not resolvable. To see this, consider the block containing 123. To avoid having more than one instance of 1, 2 or 3, this block must have either 456 or 478 (so as to include the element 4). Whether we choose 456 or 478, we have no way of completing the block since selection of any remaining block will cause result is multiple instances of at least one element.

The $(13,3,1)$-BIBD with

$$X = \{1,2,3,4,5,6,7,8,9,10,11,12,13\}$$
$$A = \{(1,2,3), (2,3,4), (3,4,5), (4,5,6), (5,6,7), (6,7,8), (7,8,9), (8,9,10),$$
$$(9,10,11), 10,11,12), (11,12,13), (12,13,1), (13,1,2)\}$$

is also not resolvable. This and the previous example come from a class of BIBDs known as Steiner triple systems [19].

$$\ldots$$

Kirkman's schoolgirl problem is a puzzle whose solution involves a resolvable BIBD. The puzzle goes as follows:

> Fifteen schoolgirls take a daily walk for seven consecutive days. They walk in 5 rows of 3. It is required to arrange them daily so that no two shall walk twice in the same row.

One can solve the problem by constructing a resolvable $(15,3,1)$-BIBD. The 15 elements of the BIBD represent the schoolgirls (we label them from A to O). Each block represents a row of 3 schoolgirls on a given day, and $\lambda = 1$ implies that each pair of girls only appears in a row once.

Using the formulas that were previously stated, we have $r = \frac{\lambda(v-1)}{k-1} = \frac{14}{2} = 7$ (implying each girl appears in 7 blocks – one per day) and $b = \frac{vr}{k} = \frac{15(7)}{3} = 35$ blocks.

One possible solution in shown in Figure 29. Each day corresponds to a parallel class in the partition of the BIBD. In all, there are seven possible solutions (up to isomorphism), see the Wikipedia article "Kirkman's schoolgirl problem" [20].

Figure 29. One solution to the schoolgirl puzzle

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
ABC	ADG	AEO	AIM	AFJ	AHK	ALN
DEF	BEH	BIJ	BDL	BKO	BGN	BFM
GHI	CJM	CDN	CEK	CGL	CFI	CHO
JKL	FKN	FHL	FGO	DHM	DJO	DIK
MNO	ILO	GKM	HJN	EIN	ELM	EGJ

2.3.5 Hadamard Matrices

A **Hadamard matrix** H is an $n \times n$ matrix, all of whose elements are either -1 or 1 such that $HH^T = nI_n$ where H^T is the transpose of H (i.e., the columns of H^T are formed by the consecutive rows of H) and I_n is the $n \times n$ identity matrix (i.e., 1 for each entry on the main diagonal and 0 everywhere else).

Hadamard matrices are named after the mathematician Jacques Hadamard. However, such matrices were first discovered earlier (1867) by another mathematician named James Joseph Sylvester.

Let H be an $n \times n$ Hadamard matrix. Then the partitioned matrix

$$\begin{bmatrix} H & H \\ H & -H \end{bmatrix}$$

is a $2n \times 2n$ Hadamard matrix. We can use this fact to construct a collection of Hadamard matrices. First note that we have the trivial Hadamard matrices $[1]$ and $[-1]$. From these, we can define two more Hadamard matrices, i.e.,

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \text{ and } -H_1 = \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix}$$

From H_1, we can generate the following Hadamard matrices

$$H_2 = \begin{bmatrix} H_1 & H_1 \\ H_1 & -H_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}, \text{ and } -H_2 = \begin{bmatrix} -1 & -1 & -1 & -1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & -1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix}$$

Continuing in this manner, we can construct Hadamard matrices of size $2^n \times 2^n$ for every non-negative integer n.

A necessary condition for the existence of a Hadamard matrix of size $n \times n$ is $4|n$ (i.e., n is a positive integer and multiple of 4). The Hadamard conjecture [21] proposes that a Hadamard matrix of size $4k \times 4k$ exists for every positive integer k. At the time of this writing, 668 is the smallest integer for which it is unknown whether or not a Hadamard matrix exists.

Hadamard matrices are related to symmetric BIBD by the following theorem.

Theorem 6. For an integer $m > 1$, there exists a Hadamard matrix of order $4m$ if and only if there exists a symmetric $(4m - 1, 2m - 1, m - 1)$-BIBD.

The proof is constructive and thus gives one a way to create a symmetric BIBD given a Hadamard matrix (and vice versa). Details of the proof and associated construction are provided in Theorem 4.5 of Stinson [3].

2.4 Factorial Designs

A **factorial design** can be used to model an experiment whose result (response) depends on multiple factors (variables). Historically, experiments were designed to determine the effect of one variable upon one response to an experiment. Statistician R.A. Fisher demonstrated the advantages of combining the study of multiple variables in one experiment. Factorial design can reduce the number of experiments one has to perform by studying multiple factors simultaneously.

Wikipedia provides the following definition of a factorial experiment [22]:

> In statistics, a full factorial experiment is an experiment whose design consists of two or more factors, each with discrete possible values or "levels", and whose experimental units take on all possible combinations of these levels across all such factors. A full factorial design may also be called a fully crossed design. Such an experiment allows the investigator to study the effect of each factor on the response variable, as well as the effects of interactions between factors on the response variable.

[Note: In what follows, we use “factorial design” and “factorial experiment” interchangeably.]

In an $s_1 \times s_2 \times \ldots \times s_k$ factorial experiment, there are k factors such that factor i has s_i levels. If all the k factors have the same number of levels (say n), then we use the shorthand notation n^k. Most common are factorial experiments with k factors, each with 2 levels (this is called a 2^k factorial experiment).

From Section 14.2 of the LibreTexts book entitled “Chemical Process Dynamics and Controls” [23]:

> Factorial design tests all possible conditions. Because factorial design can lead to a large number of trials, which can become expensive and time-consuming, factorial design is best used for a small number of variables with few states (1 to 3). Factorial design works well when interactions between variables are strong and important, and where every variable contributes significantly.

As a simple example, consider an experiment to evaluate the effects of an Essential Amino Acid (EAA) supplement and/or resistance exercises on an illness known as sarcopenia (muscle loss that occurs with aging and/or immobility). The effectiveness of the different treatments will be measured based on grip strength (measured before and after administration of the treatments). This experiment fits nicely into a 2×2 factorial design, as shown in Figure 30. The numbers in the table represent the average grip strength increase (in pounds) based on administration of the various treatment combinations on multiple participants.

Figure 30. Factorial design for sarcopenia experiment

	Placebo	EAA Supplement
No exercise	1	4
Resistance exercises	3	5

3 Magic Squares

The magic square is the hammer that shatters the ice of our unconscious. — Qingfu Chuzhen

3.1 Overview

A **magic square** is an $n \times n$ matrix of (typically) non-negative integers such that each row, each column and each of the two main diagonals add up to the same number (referred to as the **magic number** and represented by $\mathcal{S}$). An $n \times n$ magic square is said to be of order n. Each element in a magic square is referred to as a **cell**.

When a magic square of order n is populated with integers from 1 to n^2, it is called a **pure magic square** (also known as a traditional, normal, simple or ordinary magic square).

There is only one pure magic square of order 1, i.e., [1]. Excluding rotations and reflections, there is only one pure magic square of order 3, see Figure 31. The magic number is 15.

Figure 31. Unique magic square of order 3

8	1	6
3	5	7
4	9	2

For a pure magic square of order n, the sum of all the numbers in the matrix is

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

and so, the magic number is $\mathcal{S} = \frac{n(n^2+1)}{2}$.

The number of pure magic squares (excluding rotations and reflections) increase rapidly, as described in article A006052 from the Online Encyclopedia of Integer Sequences [24] and shown in Figure 32. The number of pure magic squares of order 4 was published by Frenicle de Bessy in 1693. The number of pure magic squares of order 5 was computed by Richard C. Schroeppel in 1973.

Order	Number of pure magic squares
1	1
2	0
3	1
4	880
5	275,305,224

Figure 32. Number of pure magic squares

Figure 33 shows an example of a 4th order pure magic square. Using the formula for the magic number, we have that $\mathcal{S} = \frac{4(4^2+1)}{2} = 34$.

2	16	13	3
11	5	8	10
7	9	12	6
14	4	1	15

Figure 33. Example of a 4th order pure magic square

The center of a magic square is a cell if the order of the magic square is odd. If the order of the magic square is even, its center is the point of intersection between the vertical and horizontal lines that bisect the magic square. Cells placed at equal distances from the center of a magic square and on opposite ends of the line through the center are called **skew-related cells**. In Figure 33, a few examples of skew-related cells are indicated with similar shading, e.g., the cells containing 7 and 10 are skew-related. For this example, the sum of any two skew-related cells is the same (17 in this case). In general, a magic square of order n is said to be an **associative magic square** if each pair of skew-related cells adds to $n^2 + 1$.

. . .

It is also possible to start with a number $a \neq 1$ and increment by an integer $d \neq 1$, i.e., the cells of the $n \times n$ square are populated by the numbers

$$a, a + d, a + 2d, \dots, a + (n^2 - 1)d$$

In his book "The Zen of Magic Squares, Circles, and Stars An Exhibition of Surprising Structures across Dimensions" [25], Pickover refers to such entities as **imperfect magic squares**. Figure 34 depicts a 4th order imperfect magic square with $a = 17$ and $d = 3$.

Figure 34. Imperfect magic square of order 4

17	59	56	26
50	32	35	41
38	44	47	29
53	23	20	62

In general, the sum of the cells in an imperfect magic square with parameters a and d is

$$a + (a + d) + (a + 2d) + \cdots + (a + (n^2 - 1)d)$$

$$= n^2 a + d\big(1 + 2 + \cdots + (n^2 - 1)\big)$$

$$= n^2 a + \frac{d(n^2 - 1)n^2}{2}$$

$$= n^2 \left(\frac{2a + d(n^2 - 1)}{2}\right)$$

So, the magic number is the above expression divided by n, i.e.,

$$\mathcal{S} = n\left(\frac{2a + d(n^2 - 1)}{2}\right)$$

For the example in Figure 34, we have

$$\mathcal{S} = 4\left(\frac{2(17) + 3(4^2 - 1)}{2}\right) = 4\left(\frac{34 + 45}{2}\right) = 2 \cdot 79 = 158$$

. . .

Another approach for creating an imperfect magic square of order n is as follows:

- select a starting number a
- select an increment b to be used within each set of n consecutive numbers
- select an increment c between each set of n consecutive numbers.

For example, if we take $a = 5, b = 3, c = 2$ and $n = 3$, we get the following sequence of numbers:

$$5, 8, 11; 13, 16, 19; 21, 24, 27$$

These number can be used to create a 3rd order imperfect magic square with magic number 48, as shown in Figure 35.

Figure 35. Imperfect magic square of order 3

24	5	19
11	16	21
13	27	8

We can derive the general formula for the magic number of such imperfect squares as follows. First, let $a_i = a_{i-1} + c + (n-1)b$ for $i = 2, 3, \ldots, n$ and $a_1 = a + (n-1)b$. Next, write down the generated set of numbers in n rows, as shown in Figure 36.

a	$a + b$	$a + 2b$	…	$a_1 = a + (n-1)b$
$a_1 + c$	$a_1 + c + b$	$a_1 + c + 2b$	…	$a_2 = a_1 + c + (n-1)b$ $= a + c + 2(n-1)b$
$a_2 + c$	$a_2 + c + b$	$a_2 + c + 2b$	…	$a_3 = a_2 + c + (n-1)b$ $= a + 2c + 3(n-1)b$
…	…	…		…
$a_{n-1} + c$	$a_{n-1} + c + b$	$a_{n-1} + c + 2b$	…	$a_n = a_{n-1} + c + (n-1)b$ $= a + (n-2)c + (n-1)^2 b$

Figure 36. Arithmetic sequence for each set of n numbers with a jump between each set

The plan of attack is to sum all the cells in the above table, which would give us n times the magic number. Warning: the numbers in the above table are **not** in the form of a magic square but rather just a listing (in ascending order) of the numbers to be used in a magic square.

Observe that

$$a+\sum_{i=1}^{n-1} a_i = na + c\big(1+2+\cdots+(n-2)\big) + (n-1)b\big(1+2+\cdots+(n-1)\big)$$

$$= na + \frac{c(n-2)(n-1)}{2} + \frac{b(n-1)^2 n}{2}$$

The above sum appears in each of the n columns in the table in Figure 36, resulting in the sum

$$n^2 a + \frac{c(n-2)(n-1)n}{2} + \frac{b(n-1)^2 n^2}{2} \qquad (1)$$

The c terms in the table (outside of the a_i terms) sum to

$$n(n-1)c \qquad (2)$$

Finally, the b terms in the table (outside of the a_i terms) sum to

$$nb\big(1+2+\cdots+(n-1)\big) = \frac{b(n-1)n^2}{2} \qquad (3)$$

From expressions (1) and (3) above, we have the following b terms

$$\frac{b(n-1)^2 n^2}{2} + \frac{b(n-1)n^2}{2} = \frac{n^2(n-1)}{2}\big((n-1)+1\big)b = \frac{n^3(n-1)b}{2} \qquad (4)$$

From expressions (1) and (2) above, we have the following c terms

$$\frac{c(n-2)(n-1)n}{2} + (n-1)nc = (n-1)nc\left(\left(\frac{n-2}{2}\right)+1\right) = \frac{(n-1)n^2 c}{2} \qquad (5)$$

Adding the a term from expression (1), to the sum of expressions (4) and (5), and then dividing by n we get the following formula for the magic number

$$\boldsymbol{na + \frac{n^2(n-1)b}{2} + \frac{n(n-1)c}{2}}$$

Plugging the values from our previous example (i.e., $a=5, b=3, c=2$ and $n=3$) into the above formula, we get the magic number

$$3\cdot 5 + \frac{3^2\cdot 2\cdot 3}{2} + \frac{3\cdot 2\cdot 2}{2} = 15+27+6 = 48$$

Let's try another example with $a = 3, b = 2, c = 5$ and $n = 3$. The numbers to be used in the imperfect magic are

$$3, 5, 7; 12, 14, 16; 21, 23, 25$$

The imperfection magic square is shown below. The magic number is 42. We will describe the technique for generating this magic square in Section 3.4.

23	3	16
7	14	21
12	25	5

3.2 Some Historical Examples

3.2.1 The Luoshu Diagram

The Luoshu, Lo Shu , or Nine Halls Diagram is an ancient Chinese diagram, named after the Luo River near Luoyang, Henan. The Luoshu diagram is shown in Figure 37.

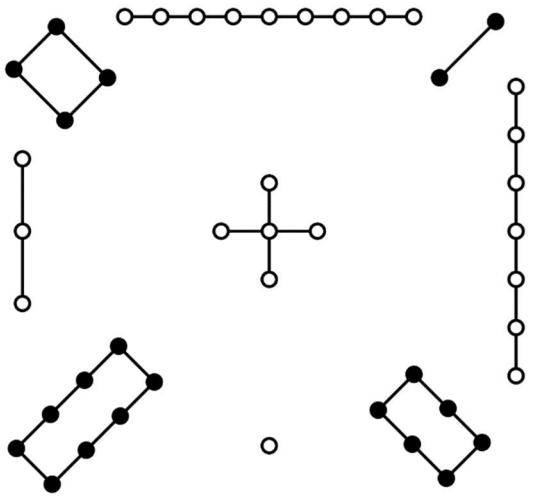


Figure 37. Luoshu diagram

Each of the 9 smaller configurations within the Luoshu diagram represent a number from 1 to 9. The mapping to numbers allows for the Luoshu diagram to be represented as the 3rd order magic square shown in Figure 38.

4	9	2
3	5	7
8	1	6

Figure 38. Luoshu magic square

The story of the Luoshu diagram goes back to the third millennia B.C.E. when there was a great flood in China. The legend goes that Emperor Yu (died 2197 B.C.E.), who was noted for his power over the water, directed people in the affected areas to build canals to control the flood waters. The canals where successful in mitigating the flood. According to legend, as Emperor Yu stood on the banks of the Luo River (one of the flooded rivers – a tributary of the Huang He or Yellow River) a tortoise emerged from the water bearing, on the underplate of its shell, an array of symbols, i.e., the Luoshu diagram as shown in Figure 37. As noted in the book "Legacy of the Luoshu" [26], the first textual reference to the Luoshu is found in the writings of Zhuang Zi (369–286 B.C.E.). This is not only the first reference to the Luoshu but also the first (albeit indirect) reference to magic squares.

3.2.2 Dürer's Magic Square

For our next historical example, we consider the 16th century painting *Melencolia I* by Albrecht Dürer. From the Wikipedia article on this topic [27]:

> *Melencolia I* is a large 1514 engraving by the German Renaissance artist Albrecht Dürer. The print's central subject is an enigmatic and gloomy winged female figure thought to be a personification of melancholia – melancholy. Holding her head in her hand, she stares past the busy scene in front of her. The area is strewn with symbols and tools associated with craft and carpentry, including an hourglass, weighing scales, a hand plane, a claw hammer, and a saw. Other objects relate to alchemy, geometry or numerology. Behind the figure is a structure with an embedded magic square, and a ladder leading beyond the frame. The sky contains a rainbow, a comet or planet, and a bat-like creature bearing the text that has become the print's title.

The drawing is available in the Wikipedia article noted above. The part of the drawing containing the magic square is shown in Figure 39.

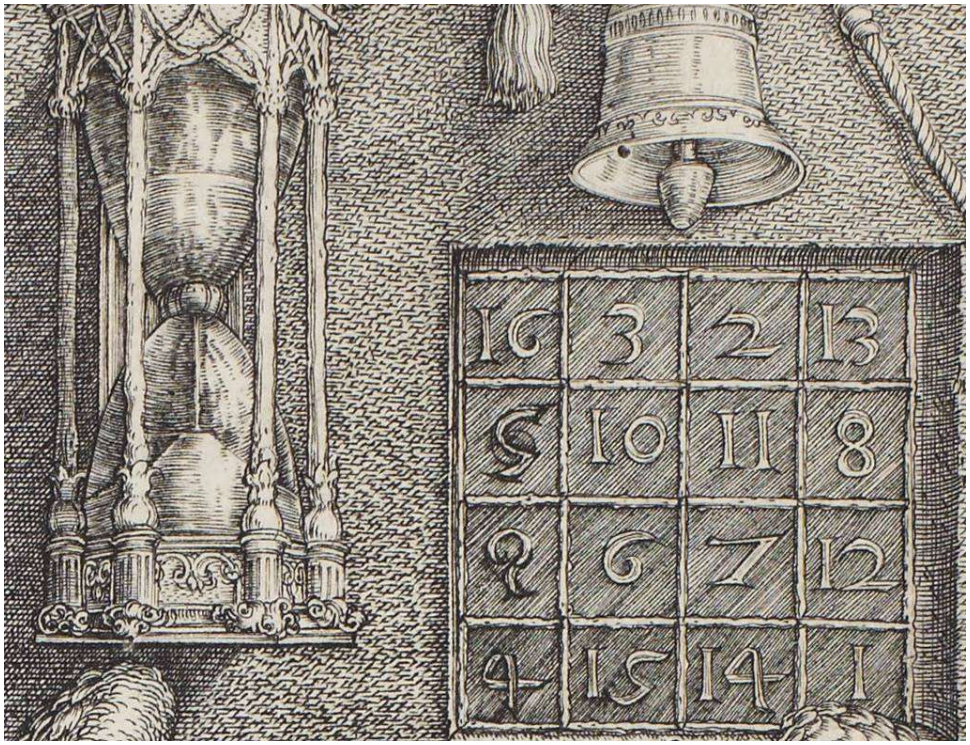


Figure 39. Magnification of magic square in Melencolia I

The pure magic square in *Melencolia I* is associative since all skew-related cells add to $n^2 + 1 = 17$. A few of the skew-related cells are indicated by the same shading in Figure 40. Each row, column and main diagonal add to the magic number for the magic square, i.e., 34. The four corners add to 34, and the middle 2×2 square adds to 34. Further, the year of the painting (1514) appears at the bottom-middle of the diagram.

16	3	2	13
5	10	11	8
9	6	7	12
4	15	14	1

Figure 40. Dürer's magic square

Many additional patterns exist in Dürer's magic square. For example, the like-shaded cells in each of the squares within Figure 41 add to the magic number 34.

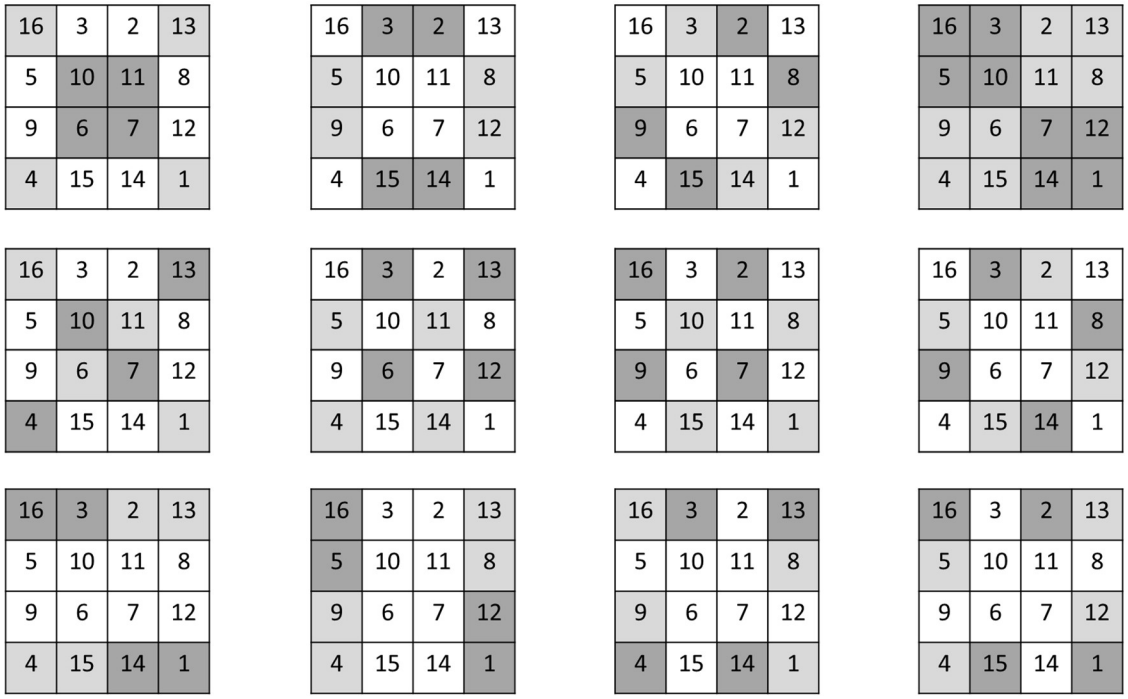


Figure 41. Additional patterns in the Dürer's magic square

Shading the cells with even numbers differently from the cells with odds numbers presents a pleasing pattern, as shown in the following figure.

16	3	2	13
5	10	11	8
9	6	7	12
4	15	14	1

3.2.3 Leonhard Euler

The famous mathematician Leonhard Euler (1707–1783) created a 4th order magic square consisting of integers squares (see Figure 42). Each row, each column and each of the main diagonals sum to 8515.

68^2	29^2	41^2	37^2
17^2	31^2	79^2	32^2
59^2	28^2	23^2	61^2
11^2	77^2	8^2	49^2

Figure 42. Euler's magic square of squares

The matrix in Figure 43 is known of *Euler's knight's square*. Starting from the cell with the number 1, a chess knight (using legal chess moves) lands on every square in the matrix (basically an 8×8 chess board). The amazing property of this square is that each row and each column add to 260. However, this is not a magic square since one diagonal sums to 264 and the other to 256. Squares having the property that the row sums and columns sums equal the same number but not the sums of the main diagonals are known as **semi-magic squares**.

The four 4×4 sub-squares (in each quadrant) of Euler's knight's square are also semi-magic squares with row and column sums equal to 130. The elements in most but not all of the 2×2 sub-square add to 130. For example, the elements in each of the lightly shaded 2×2 sub-squares add to 130, but the elements in the more heavily shaded sub-squares do not.

1	48	31	50	33	16	63	18
30	51	46	3	62	19	14	35
47	2	49	32	15	34	17	64
52	29	4	45	20	61	36	13
5	44	25	56	9	40	21	60
28	53	8	41	24	57	12	37
43	6	55	26	39	10	59	22
54	27	42	7	58	23	38	11

Figure 43. Euler's knight's square

If one connects the cells in sequential order, a symmetric pattern is revealed (as shown in Figure 44).

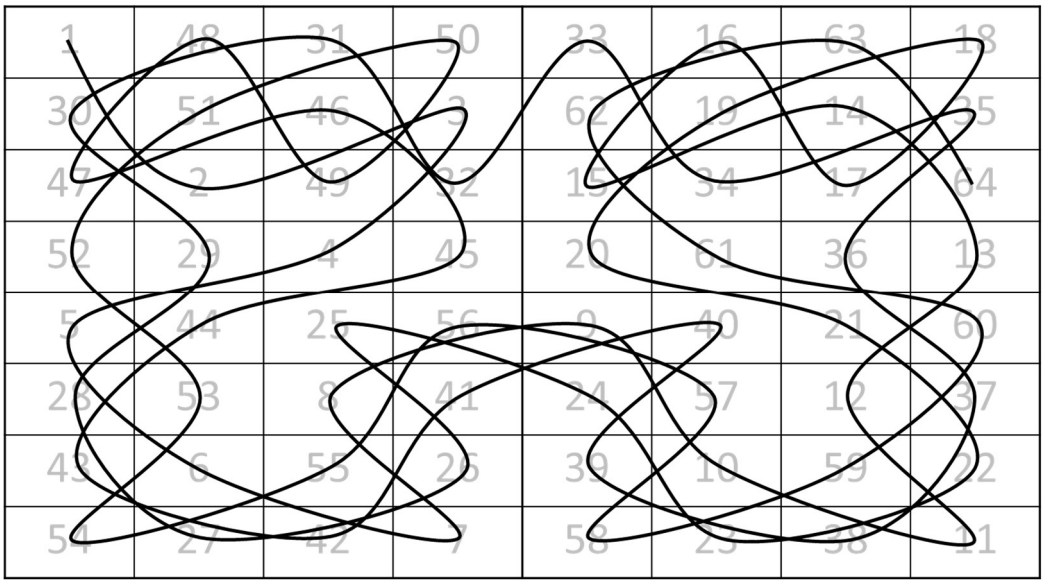


Figure 44. Underlying symmetric pattern in Euler's knight's square

3.2.4 Benjamin Franklin's Semi-magic Square

The 8×8 semi-magical square by Benjamin Franklin (1706–1790) is shown in Figure 45. Benjamin Franklin was an American polymath who was active as a writer, scientist, inventor, statesman, diplomat, printer, publisher, and political philosopher. Among the leading intellectuals of his time, Franklin was one of the Founding Fathers of the United States, a drafter and signer of the Declaration of Independence, and the first postmaster general. [29]

In Franklin's semi-magic square, the sum of each row and the sum of each column is 260. The sum of the diagonal elements from top left to bottom right is 228, and the sum of the diagonal elements from bottom left to top right is 292. Thus, we "only" have a semi-magic square. However, there are many other patterns in Franklin's creation, e.g., the sum of four corner numbers and the four center numbers add to 260 (see the 8 shaded cells in Figure 45). Further, any half-row starting from the right or left edge of the square (e.g., 14,3,62,51) and any half-column starting from the top or bottom of the square (e.g., 32,34,25,39) add to $\frac{260}{2} = 130$.

52	61	4	13	20	29	36	45
14	3	62	51	46	35	30	19
53	60	5	12	21	28	37	44
11	6	59	54	43	38	27	22
55	58	7	10	23	26	39	42
9	8	57	56	41	40	25	24
50	63	2	15	18	31	34	47
16	1	64	49	48	33	32	17

Figure 45. Benjamin Franklin's 8×8 semi-magic square

According to ChatGPT, Franklin used his semi-magic square to create a puzzle that he published in his "Pennsylvania Gazette" newspaper (1779). The puzzle consisted of removing eight numbers from the square so that the remaining numbers still formed a semi-magic square. This puzzle became very popular and was known as the "Franklin Magic Square" or "Ben Franklin's Missing Eight Puzzle."

[**Author's Remark**: As far as I can tell, the story from ChatGPT is fabricated by highly plausible, since Franklin did create an 8×8 semi-magic square and he did own the Pennsylvania Gazette. In any event, this inspired me to create my own 5×5 puzzle. For the magic square in Figure 46, each row, each column and the two main diagonals add to the same integer. The missing entries are positive integers. No number is repeated (including the numbers that go in the missing places). Find the numbers that go in the missing places. There are an infinite number of solutions.

51		4	23	42
	12	21	40	49
10	19		57	66
27	36	55	64	

34	53	72		25

Figure 46. Steve's 5×5 magic square puzzle

The answer is provided via an example in Section 3.4.]

Getting back to Franklin's 8×8 semi-magic square, Figure 47 shows 3 "bent rows" (each of which adds to 260). Bent rows can "wrap around" as is the case with the bent row with numbers 53,3,4,49,48,29,30 and 44 (the cells with heavy borders in the figure).

52	61	4	13	20	29	36	45
14	3	62	51	46	35	30	19
53	60	5	12	21	28	37	44
11	6	59	54	43	38	27	22
55	58	7	10	23	26	39	42
9	8	57	56	41	40	25	24
50	63	2	15	18	31	34	47
16	1	64	49	48	33	32	17

Figure 47. Bent rows

All the bent columns also add to 260. Three examples of bent columns are shown in Figure 48.

52	61	4	13	20	29	36	45
14	3	62	51	46	35	30	19
53	60	5	12	21	28	37	44
11	6	59	54	43	38	27	22
55	58	7	10	23	26	39	42
9	8	57	56	41	40	25	24
50	63	2	15	18	31	34	47
16	1	64	49	48	33	32	17

Figure 48. Bent columns

The gray and black checker patterns in Figure 49 and Figure 50 each add to 260. This works if one moves the pattern up or down (including wrap arounds).

52	61	4	13	20	29	36	45
14	3	62	51	46	35	30	19
53	60	5	12	21	28	37	44
11	6	59	54	43	38	27	22
55	58	7	10	23	26	39	42
9	8	57	56	41	40	25	24
50	63	2	15	18	31	34	47
16	1	64	49	48	33	32	17

Figure 49. Checker patterns – pointing down

52	61	4	13	20	29	36	45
14	3	62	51	46	35	30	19
53	60	5	12	21	28	37	44
11	6	59	54	43	38	27	22
55	58	7	10	23	26	39	42
9	8	57	56	41	40	25	24
50	63	2	15	18	31	34	47
16	1	64	49	48	33	32	17

Figure 50. Checker patterns – pointing up

The eight cells in Figure 51 add to 260. The pattern, when shifted up or down, or left or right, will still add to 260.

52	61	4	13	20	29	36	45
14	3	62	51	46	35	30	19
53	60	5	12	21	28	37	44
11	6	59	54	43	38	27	22
55	58	7	10	23	26	39	42
9	8	57	56	41	40	25	24
50	63	2	15	18	31	34	47
16	1	64	49	48	33	32	17

Figure 51. Eight-cell pattern

Finally, we mention that any 2×2 square within Franklin's semi-magic square adds to 130.

3.3 Classification

The following classification of magic squares is based on the scheme presented in the Wikipedia article on magic squares [28].

In a **semi-magic square**, the rows and columns sum to the same number, i.e., the magic number. However, the main diagonals do not necessarily add to the magic number. We discussed an example of a semi-magic square in Section 3.2.4.

In a **pure magic square**, each row, each column, and the two diagonals sum to the magic number. The Luoshu and Dürer square are examples of pure magic squares.

A self-complementary magic square is an $n \times n$ magic square which when complemented (i.e., each number subtracted from $n^2 + 1$) will give a rotated or reflected version of itself. On the left-side of Figure 52 is Dürer's magic square, and on the right-side is its complement. The complement of Dürer's magic square is a 180 degree counterclockwise rotation of the original magic square, and thus, Dürer's magic square is self-complementary.

16	3	2	13
5	10	11	8
9	6	7	12
4	15	14	1

1	14	15	4
12	7	6	9
8	11	10	5
13	2	3	16

Figure 52. Self-complementary magic square

An **associative magic square** is an $n \times n$ magic square with the property that every number added to the number equidistant, in a straight line, from the center equals $n^2 + 1$. They are also called symmetric magic squares. Associative magic squares do not exist for squares of singly even order (i.e., order divisible by 2 but not 4). All associative magic squares are self-complementary magic squares. Dürer's magic square is an associative magic square.

A **pandiagonal magic square** is a magic square with the property that the **broken diagonals** (i.e., set of n cells forming two parallel diagonal lines in a square matrix, see "Broken diagonal" [30]) sum to the magic constant. Pandiagonal magic squares do not exist for singly even orders. An example of a 5×5 pandiagonal magic square is shown in Figure 53. The gray cells represent one of the broken diagonals in the magic square. In the example, each row, each column, both of the main diagonals and each broken diagonal add to the magic number 65.

20	8	21	14	2
11	4	17	10	23
7	25	13	1	19
3	16	9	22	15
24	12	5	18	6

Figure 53. Pandiagonal magic square

One way to identify broken diagonals is to place two copies of a magic square next to each other, select a cell on the top row and then create a diagonal going down to the left or down to the right. In Figure 54, two broken diagonals are shown, i.e., 21,10,19,2,12 (cell in black with white numbers) and 20,23,1,9,12 (gray cells).

20	8	21	14	2	20	8	21	14	2
11	4	17	10	23	11	4	17	10	23
7	25	13	1	19	7	25	13	1	19
3	16	9	22	15	3	16	9	22	15
24	12	5	18	6	24	12	5	18	6

Figure 54. Broken diagonals

An ultra-magic square has the properties of both an associative and a pandiagonal magic square. Ultra-magic squares exist only for orders $n \geq 5$. The square in Figure 53 is an ultra-magic square - its complement is shown below.

6	18	5	12	24
15	22	9	16	3
19	1	13	25	7
23	10	17	4	11
2	14	21	8	20

A **bordered magic square** is a magic square, which remains magic when its borders are removed. For example, if the outer board (light gray) of the magic square in Figure 55 is removed, the remaining matrix is a magic square. If we remove both the light gray and darker gray borders, we again are left with a magic square. Finally, if we removed the light gray, darker gray and black borders, the remaining (white) matrix is a magic square. Thus, Figure 55 consists of three concentric bordered magic squares (noting that the white magic square is not a bordered magic square).

11	99	50	4	96	95	7	10	92	41
1	12	88	14	86	85	17	83	19	100
98	49	33	77	48	28	74	43	52	3
21	22	23	64	36	35	67	78	79	80
70	69	76	57	45	46	54	25	32	31
30	39	75	47	55	56	44	26	62	71
81	72	38	34	66	65	37	63	29	20
93	59	58	24	53	73	27	68	42	8
40	82	13	87	15	16	84	18	89	61
60	2	51	97	5	6	94	91	9	90

Figure 55. Concentric magic squares

A magic square of order mn is called **composite**, if it can be decomposed into m^2 magic sub-squares, each of order n. The minimum order for a composite magic square is 9, an example of which is shown in Figure 56 ($m = n = 3$ in this example).

71	66	67	20	25	24	29	34	33
64	68	72	27	23	19	36	32	28
69	70	65	22	21	26	31	30	35
8	3	4	40	39	44	74	79	78
1	5	9	45	41	37	81	77	73
6	7	2	38	43	42	76	75	80
47	54	49	56	63	58	11	16	15
52	50	48	61	59	57	18	14	10
51	46	53	60	55	62	13	12	17

Figure 56. Composite magic square of order 9

A composite magic square of order 12 is shown in Figure 57 ($m = 3$ and $n = 4$ in this example).

17	31	30	20	132	142	143	129	61	56	60	49
28	22	23	25	137	135	134	140	51	58	54	63
24	26	27	21	133	139	138	136	50	59	55	62
29	19	18	32	144	130	131	141	64	53	57	52
100	105	101	112	68	73	69	80	36	41	37	48
110	103	107	98	78	71	75	66	46	39	43	34
111	102	106	99	79	70	74	67	47	38	42	35
97	108	104	109	65	76	72	77	33	44	40	45
84	89	85	96	4	14	15	1	116	121	117	128
94	87	91	82	9	7	6	12	126	119	123	114
95	86	90	83	5	11	10	8	127	118	122	115
81	92	88	93	16	2	3	13	113	124	120	125

Figure 57. Composite magic square of order 12

A **most-perfect magic square** is a pandiagonal magic square with the following properties

- (**compactness**) each 2×2 sub-square adds to $\frac{1}{k}$ of the magic number where the order of the magic square is $4k$ (noting that most-perfect magic squares only exist for squares whose order is a multiple of 4)
- (**completeness**) all pairs of integers separated by distance $\frac{n}{2}$ along any diagonal (major or broken) are complementary (i.e., they sum to $n^2 + 1$).

Figure 58 shows an 8×8 most-perfect magic square, with $n = 8$ and $k = 2$. The magic number is $\mathcal{S} = \frac{n(n^2+1)}{2} = \frac{8(65)}{2} = 260$. Pick any 2×2 sub-square, and the sum will be $\frac{1}{2}(260) = 130$. Concerning the completeness property, all pairs at a distance $\frac{n}{2} = 4$ along any diagonal sum to $n^2 + 1 = 65$. For example, 56 and 9 (black cells with white numbers in the figure) are four cells apart along a main diagonal, and they sum to 65. In the broken diagonal (cells with heavy borders in the figure), 18 and 47 (gray cells) are four cells apart and add to 65.

1	16	17	32	53	60	37	44
63	50	47	34	11	6	27	22
3	14	19	30	55	58	39	42
61	52	45	36	9	8	25	24
12	5	28	21	64	49	48	33
54	59	38	43	2	15	18	31
10	7	26	23	62	51	46	35
56	57	40	41	4	13	20	29

Figure 58. Most-perfect magic square of order 8

A **Franklin magic square** is a semi-magic square with order a multiple of 4 such that

- every bent diagonal adds to the magic number
- every half row and half column, starting at an outside edge, sum to half the magic number
- the square is compact.

The 8×8 Franklin semi-magic square that we saw earlier fits this definition. Franklin also created a 16×16 semi-magic square with all the properties noted above (see Figure 59). Franklin described his creation in a letter to Peter Collinson Esq. of London [31]:

> Mr. Logan then showed me an old arithmetical book in quarto, wrote, I think, by one [Michel] Stifelius, which contained a square of 16 that he said he should imagine must have been a work of great labor; but I forget not, it had only the common properties of making the same sum, viz., 2056, in every row, horizontal, vertical, and diagonal. Not willing to be outdone by Mr. Stifelius, even in the size of my square, I went home and made that evening the following magical square of 16, which, besides having all the [special] properties of the 8×8 square (i.e., it would make 2056 in all the same rows and bent and broken bent rows), had this added: that a four-square hole being cut in a piece of paper of such a size as to take in and show through it just 16 of the little squares, when laid on the greater square, the sum of the 16 numbers so appearing through the hole, wherever it was placed on the greater square, should likewise make 2056.

An example of the 4×4 sub-square that Franklin mentions in his letter is shown in the figure (black cells with white numbers). The sum is 2056.

200	217	232	249	8	25	40	57	72	89	104	121	136	153	168	185
58	39	26	7	250	231	218	199	186	167	154	135	122	103	90	71
198	219	230	251	6	27	38	59	70	91	102	123	134	155	166	187
60	37	28	5	252	229	220	197	188	165	156	133	124	101	92	69
201	216	233	248	9	24	41	56	73	88	105	120	137	152	169	184
55	42	23	10	247	234	215	202	183	170	151	138	119	106	87	74
203	214	235	246	11	22	43	54	75	86	107	118	139	150	171	182
53	44	21	12	245	236	213	204	181	172	149	140	117	108	85	76
205	212	237	244	13	20	45	52	77	84	109	116	141	148	173	180
51	46	19	14	243	238	211	206	179	174	147	142	115	110	83	78
207	210	239	242	15	18	47	50	79	82	111	114	143	146	175	178
49	48	17	16	241	240	209	208	177	176	145	144	113	112	81	80
196	221	228	253	4	29	36	61	68	93	100	125	132	157	164	189
62	35	30	3	254	227	222	195	190	163	158	131	126	99	94	67
194	223	226	255	2	31	34	63	66	95	98	127	130	159	162	191
64	33	32	1	256	225	224	193	192	161	160	129	128	97	96	65

Figure 59. Franklin's 16×16 semi-magic square

A **multimagic square** is a magic square that remains magic even if all its numbers are replaced by their k^{th} power for $1 \leq k \leq p$ where p is an integer. The website http://www.multimagie.com/ provides extensive example of multimagic squares. For example, Figure 60 shows a tri-magic square (i.e., $p = 3$). The sums of the rows, columns and diagonals are equal to 870. The sums of the squares of the rows, columns and diagonals are equal to 83,810. The sums of the cubes of the rows, columns and diagonals are equal to 9,082,800. This creation was discovered by Walter Trump in June 2002. Trump's tri-magic square has the further property of being symmetrical, i.e., the i^{th} element of a row plus the $(13 - i)^{th}$ element of this same row equals $12^2 + 1 = 145$. See the two examples highlighted in black in the figure.

1	22	33	41	62	66	79	83	104	112	123	144
9	119	45	115	107	93	52	38	30	100	26	136
75	141	35	48	57	14	131	88	97	110	4	70
74	8	106	49	12	43	102	133	96	39	137	71
140	101	124	42	60	37	108	85	103	21	44	5
122	76	142	86	67	126	19	78	59	3	69	23
55	27	95	135	130	89	56	15	10	50	118	90
132	117	68	91	11	99	46	134	54	77	28	13
73	64	2	121	109	32	113	36	24	143	81	72
58	98	84	116	138	16	129	7	29	61	47	87
80	34	105	6	92	127	18	53	139	40	111	65
51	63	31	20	25	128	17	120	125	114	82	94

Figure 60. Tri-magic square

3.4 Construction of Magic Squares

3.4.1 Overview

Magic square and the associated methods for their construction fall into several categories. In this section, we provide one example construction scheme for magic squares in each of the following categories:

- Magic squares of odd order
- Magic squares of a doubly even order, i.e., order is divisible by 2 and 4
- Magic squares of a singly even order, i.e., order is divisible by 2 but not 4.

A comprehensive list of magic square constructions methods can be found at https://www.magic-squares.info/methods/intro.html.

The first step in constructing a magic square is the selection of a number set to be placed in the square. This is common to all the approaches described in this section. We have already seen several approaches for selecting the number set (listed below). In each case, we assume the magic square is of order n and thus, has n^2 cells to be populated with positive integers.

- consecutive positive integers, i.e., $a, a+1, a+2, \ldots, a+n^2-1$ where $a \geq 1$
- an arithmetic sequence, i.e., $a, a+d, a+2d, \ldots, a+(n^2-1)d$
- an arithmetic sequence for each set of n numbers with a jump between each sequence (recall the previous discussion related to Figure 36).

3.4.2 De la Loubère's Method

The De la Loubère method (also known as the Siamese or Up-Right method) is used for the creation of magic squares of odd order. The steps are as follows:

1. Place the first number in the middle cell of the top row.
2. To place the next number, go up one cell and to the right one cell.
3. Whenever up and right takes you off the top or right of the grid, continue to the opposite side of the square (i.e., wrap-around).
4. If you arrive at an already occupied cell, return to the last valid position and go one step down.

Let's try an example using the arithmetic sequence with a jump approach to select the set of numbers to be placed in the magic square. Select the starting number $a = 4$, the step $b = 2$, the jump between each sequence $c = 7$ and the order of the magic square $n = 5$. This yields the following set of numbers:

$$4, 6, 8, 10, 12$$
$$19, 21, 23, 25, 27$$
$$34, 36, 38, 40, 42$$
$$49, 51, 53, 55, 57$$
$$64, 66, 68, 70, 72$$

The figure below shows the first few steps in constructing a magic square using the numbers above.

- We start with a 4 in the middle of the top row.
- Next, we go up and to the right, but this is off the grid, and so, we wrap-around and place the 6 in the bottom row as shown in the figure. Placement of the next number (i.e., 8) follows the up and right rule.
- The next number (i.e., 10) goes off the grid, and so, we wrap-around. Placement of the next number (i.e., 12) follows the up and right rule.
- The next number cannot be placed right and up from 12 since that cell is occupied by 4. So, we follow rule #4, and place 19 below 12.

			6		
		4			
	12				
10	19				10
				8	
			6		

The figure below shows the next few steps.

- 21 and 23 are placed according to the right and up rule. 25 goes off the top of the gird, and so we wrap-around. Similarly, 27 goes off the right-side of the grid and we wrap-around.
- Going right and up from 27 lands us on an occupied cell, and so, we follow rule #4.
- Placement of 36, 38, 40 and 42 follow the right and up rule.
- Placement of 49 would take us off the gird to the right and over the top. In this special case, we use rule #4.

				25	
		4	23	42	
	12	21	40	49	
10	19	38			
27	36			8	27
34			6	25	

The final steps are shown in the following grid.

	53	72			
51	70	4	23	42	51
68	12	21	40	49	68
10	19	38	57	66	
27	36	55	64	8	
34	53	72	6	25	

This is one solution to the puzzle proposed in conjunction with Figure 46. Adding or subtracting the same integer value from the missing numbers 68, 70, 38, 6 and 8 leads to an infinite number of solutions to the puzzle.

3.4.3 Marking Diagonals

The method of creating magic squares known as “marking diagonals” is documented in Arabic sources dating from the 11th century AD. In this approach, the order n of the magic square must be a multiple of 4.

Using one of the approaches described in Section 3.4.1, we create a list of n^2 numbers to populate the magic square. The numbers are placed in the square, starting with the smallest number being placed at the top-left and proceeding down to the bottom-right (containing the largest number in the list).

As an example, we will create an 8×8 magic square using this method. Let’s select the set of numbers for the magic square using the arithmetic sequences with jumps approach, with $a = 5$,

$b = 2$ and $c = 4$. The numbers are placed in grid as shown below. Clearly, this is not yet a magic square and some rearrangement is necessary.

5	7	9	11	13	15	17	19
23	25	27	29	31	33	35	37
41	43	45	47	49	51	53	55
59	61	63	65	67	69	71	73
77	79	81	83	85	87	89	91
95	97	99	101	103	105	107	109
113	115	117	119	121	123	125	127
131	133	135	137	139	141	143	145

Next, divide the grid into 4 quadrants, and mark the diagonals in each quadrant (shown in gray in the following figure).

5	7	9	11	13	15	17	19
23	25	27	29	31	33	35	37
41	43	45	47	49	51	53	55
59	61	63	65	67	69	71	73
77	79	81	83	85	87	89	91
95	97	99	101	103	105	107	109
113	115	117	119	121	123	125	127
131	133	135	137	139	141	143	145

The complement of the i^{th} number in the ordered list of numbers used in the magic square is the $n - i + 1$ number in the list. For example, 25 is the 10^{th} number in our example and its complement is the 55th number in the list (since $n - 1 + 1 = 64 - 10 + 1 = 55$) where the 55th number in the list is 125.

Next, we replace each of the numbers in the shaded cells with their complement. This yields the magic square in the following figure. Each row, each column and the main diagonals sum to 600.

145	7	9	139	137	15	17	131
23	125	123	29	31	117	115	37
41	107	105	47	49	99	97	55
91	61	63	85	83	69	71	77
73	79	81	67	65	87	89	59
95	53	51	101	103	45	43	109
113	35	33	119	121	27	25	127
19	133	135	13	11	141	143	5

3.4.4 Philippe de la Hire's Method

Philippe de la Hire (1640–1719) devised a method of creating magic squares of singly even order (e.g., 6, 10, 14). His method makes use of two generating squares that get summed together. In what follows, we demonstrate de la Hire's method for a magic square of order 10.

Figure 61 depicts the first generating matrix to be used in our example of de la Hire's method. Given the order of the resulting magic square is to be 10, we use the integers 1 through 10 in the matrix.

- The first step is to populate the main diagonals of the matrix as shown in the figure.
- The next step is to populate the remaining (unfilled) cells in the 1st column with an equal number of 10s and 1s (we are free in terms of the arrangement). In the 2nd column we distribute an equal number of 9s and 2s (again, we are free in terms of the arrangement). We do the same for the 3rd, 4th and 5th columns.
- Once the left-side of the matrix is populated, the matrix is determined. The right-side is almost a mirror image of the right-side with the exception that we replace each number i with its complement, i.e., $n + 1 - i = 11 - i$. For example, 9 is replaced by $11 - 9 = 2$. The result for our example is shown in the figure below.

10	9	3	4	5	6	7	8	2	1
10	9	3	4	5	6	7	8	2	1
1	2	8	7	6	5	4	3	9	10
1	2	8	7	5	6	4	3	9	10
10	2	8	7	6	5	4	3	9	1
1	9	8	4	6	5	7	3	2	10
10	2	3	7	5	6	4	8	9	1
1	2	8	4	6	5	7	3	9	10
1	9	3	4	5	6	7	8	2	10
10	9	3	7	6	5	4	8	2	1

Figure 61. First generating matrix in de la Hire's method

Figure 62 shows the second generating matrix to be used in our example of de la Hire's method. Given that the order of the resulting magic square is to be 10, we use the integers $0, 10, 20, \ldots, 90$ in the matrix. The procedure for the second matrix is very similar to the first, except that we populate the top half of the matrix first.

- The first step is to populate the main diagonals of the matrix as shown in the figure.
- The next step is to populate the remaining cells in the 1st row with an equal number of 0s and 90s (we are free in terms of the arrangement). In the 2nd row we distribute an equal number of 10s and 80s (again, we are free in terms of the arrangement). We do the same for the 3rd, 4th and 5th rows.
- Once the top-half of the matrix is populated, the matrix is determined. The bottom-half is almost a mirror image of the top-half with the exception that we replace each number i with its complement, i.e., $100 - i$. The result for our example is shown in the figure below.

0	90	0	90	90	90	0	0	90	0
10	10	80	80	80	10	80	80	10	10
70	70	20	70	20	20	70	20	70	20
60	30	60	30	60	30	30	60	60	30
50	50	40	40	40	40	50	50	40	50
40	40	50	50	50	50	40	40	50	40
30	60	30	60	30	60	60	30	30	60
20	20	70	20	70	70	20	70	20	70
80	80	10	10	10	80	10	10	80	80
90	0	90	0	0	0	90	90	0	90

Figure 62. Second generating matrix in de la Hire's method

We are almost done. Just add the two matrices and we get the magic square shown in Figure 63. Each row, each column and both of the main diagonals sum up to 505.

10	99	3	94	95	96	7	8	92	1
20	19	83	84	85	16	87	88	12	11
71	72	28	77	26	25	74	23	79	30
61	32	68	37	65	36	34	63	69	40
60	52	48	47	46	45	54	53	49	51
41	49	58	54	56	55	47	43	52	50
40	62	33	67	35	66	64	38	39	61
21	22	78	24	76	75	27	73	29	80
81	89	13	14	15	86	17	18	82	90
100	9	93	7	6	5	94	98	2	91

Figure 63. Order 10 magic square constructed using de la Hire's method

4 Finite Geometries

"Anyone who stops learning is old, whether at 20 or 80. Anyone who keeps learning stays young. The greatest thing in life is to keep your mind young." — Henry Ford

4.1 Overview

A **finite geometry** is an axiomatic structure that has a finite number of points, some of which are grouped into sets which are called lines. The set of lines is also finite. In this section, we discuss various types of finite geometries. Each type of finite geometry is defined by a set of axioms (assumptions).

Each instance of a finite geometry is a **space** comprised of a finite set P whose elements are called points, and a finite set L of certain subsets of P, whose elements are called lines.

4.2 Near-linear Spaces

A finite **near-linear space** is a space $S = (P, L)$ consisting of a finite set of points P and a finite set of lines L such that

- Axiom NL1: every line has at least two points
- Axiom NL2: two points are on at most one line.

The number of points in a near-linear space are referred to as its **order**.

A near-linear space with no points or lines is allowed. It is called the empty space, denote ϕ.

In terms of its axioms, near-linear spaces are the simplest type of finite geometry.

Concerning notation, we will make use of the following:

- The number of points will be denoted by v.
- The number of lines will be denoted by b.
- The notation $v(\ell)$ denotes the number of points on a line ℓ. [In what follows, the expressions “points on a line”, “points in a line” and “points contained by a line” are meant to mean the same thing.]
    - If we label the lines as $\ell_1, \ell_2, \dots, \ell_b$, we use the shorthand v_i for $v(\ell_i)$.
- The notation $b(p)$ the number of lines containing point p.
    - If we label the points as $p_1, p_2, \dots, p_v$, we use the shorthand b_i for $b(p_i)$.

As an example, consider the space S with points $\{1, 2, 3, 4\}$ and lines $\{1,2\}, \{2,3\}, \{3,4\}, \{4,1\}, \{2,4\}, \{1,3\}$. This is a near-linear space since all the lines have at least 2 points (exactly 2 points in this case), and no two points are on more than one line. The example is depicted graphically in Figure 64. [**Warning**: The lines in the figure are only associations. There are only two points on each line (at either end). The two lines crossing in the center of the figure have no meaning with regard to the near-linear space being represented.]

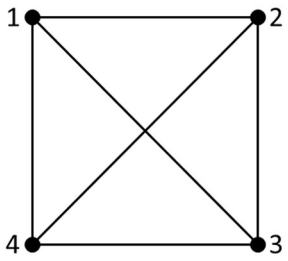


Figure 64. Near-linear space with 4 points and 6 lines

If we add the line $\{1,2,3\}$ to our example, then it is no longer a near-linear space since the points 1 and 2 belong to two lines (same issue with points 1 and 3, and point 2 and 3). See Figure 65 for a graphical representation of the modified example.

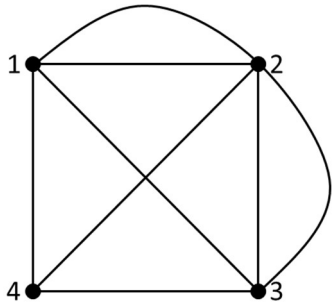


Figure 65. Space which is not a near-linear space

A space with no lines satisfies the axioms of a near-linear space. For example, the space with points $\{1,2,3,4,5\}$ but no lines, is a near-linear space. The associated graphical representation would just be 5 points with no interconnecting lines. If we added a line with one point, say $\{5\}$, then we would no longer have a near-linear space.

The points in common to two or more lines is known as their **intersection**. Because of Axiom NL2, two distinct lines in a near-linear space can have at most 1 point in their intersection.

If the set of points comprising line ℓ_1 is a subset of the set of points comprising line ℓ_2 (which we write as $\ell_1 \subseteq \ell_2$), then $\ell_1 = \ell_2$. This is true since, by Axiom NL1, ℓ_1 has a least two points (say a and b) and by the condition $\ell_1 \subseteq \ell_2$, a and b must also be on ℓ_2. However, Axiom NL2 tells us that a and b can be on at most one line and so, $\ell_1 = \ell_2$.

. . .

The **dual of the near-linear space** $S = (P, L)$ is defined to be the space $R = (P', L')$ such that

- $P' = L$, i.e., lines of S become the points of R.
- If we let $P' = L = \{\ell_1, \ell_2, \dots, \ell_k\}$ then a line of R is any subset of P' (with at least two members) such that each member of the subset contains a common point in S.

Figure 66 depicts the dual of the near-linear space from Figure 64. We used the lines of the near-linear space in Figure 64 to label the points in its dual. The label of each line in the dual contains one common point from the original near-linear space. For example, each point of line ℓ_1' has 1 as part of its label.

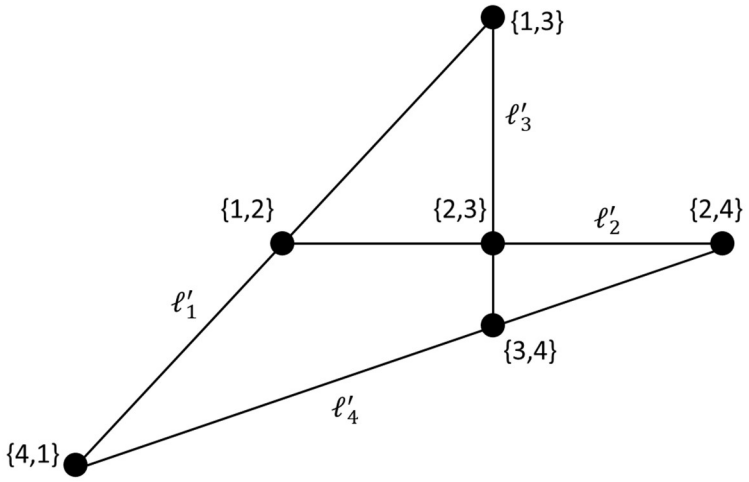


Figure 66. Example dual of a near-linear space

Theorem 7. The dual space of a near-linear space is a near-linear space.

Proof: Let S be a near-linear space and let R be its dual.

The definition of a dual space requires that each line have at least two points, and so, we have covered Axiom NL1.

Regarding Axiom NL2, assume two points in the dual space (say ℓ_1 and ℓ_2) are on two distinct lines of the dual space. Each line containing ℓ_1 and ℓ_2 in the dual space corresponds to a point of intersection of ℓ_1 and ℓ_2 in S. However, we showed earlier that there is at most one such point of intersection between two lines in a near-linear space. Thus, there is at most one line containing points ℓ_1 and ℓ_2 in the dual space, i.e., Axiom NL2 holds. ■

. . .

A **subspace** $R = (P', L')$ of a near-linear space $S = (P, L)$ is formed by a subset of P (call it Q) such that whenever $a, b \in Q$ and are both on a line $\ell \in L$, then the entire line ℓ is in L'.

The definition is a bit tricky since we start with $Q \subseteq P'$ but may need to add additional points to satisfy the condition about including all the points on any line the contains two points from Q. The adding of points can involve several iterations, as we illustrate in the example below.

A subspace of a near-linear space is also a near-linear space.

The empty set, any point, any line and the whole space itself are always subspaces of a given space.

Consider the near-linear space shown in Figure 67.

- If we take subset $Q_1 = \{1,2\}$, then we need to add the entire line containing points 1 and 2. This results in the subspace $R_1 = (\{1,2,7,10\}, \{\ell_1\})$.
- If we take subset $Q_2 = \{1,2,3,4\}$, then we need to add the all the other points on lines ℓ_1 and ℓ_2. Next, we notice that points 5 and 7 are in the subspace, and so we need to add line ℓ_3 (including point 6). This results in the $R_2 = (\{1,2,3,4,5,6,7,8,10\}, \{\ell_1, \ell_2, \ell_3\})$.

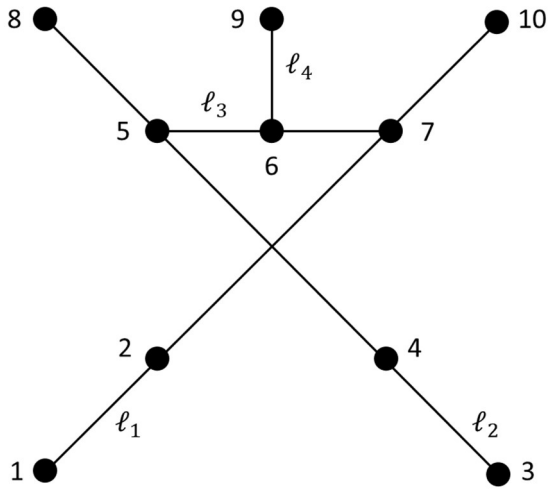


Figure 67. Near-linear space with interesting subspace

Theorem 8. The intersection of subspaces of a near-linear space $S = (P, L)$ is also a subspace of S.

Proof: Let $T = (P', L')$ be the intersection of subspaces of S. Formally, $T = S_1 \cap S_2 \cap \ldots \cap S_n$ where each $S_i = (P_i, L_i)$ is a subspace of S. We only need to show that if a and b are points in P', and a and b are on a line ℓ, then $\ell \in L'$. However, if $a, b \in P'$ and a and b are on ℓ, then it must be that $a, b \in P_i$ and $\ell \in L_i$, for $i = 1,2, \ldots, n$. Thus, $\ell \in L_1 \cap L_2 \cap \ldots \cap L_n = L'$. ■

. . .

Given a near-linear space $S = (P, L)$ and a subset Q of P, the **closure** of Q (written as $< Q >$) is the smallest subspace containing Q.

It follows from Theorem 8 that the closure of a subset of points Q in a near-linear space is the intersection of all subspaces containing Q.

For the near-linear space in Figure 67, we have

- The closure of $\{1,2\}$ is the subspace generated by $\{1,2,7,10\}$.
- The closure of $\{1,2,3,4\}$ is the subspace generated by $\{1,2,3,4,5,6,7,8,10\}$.
- The closure of $\{1,3\}$ is the entire space S since there is no smaller subspace that includes both 1 and 3. On the other hand, the subspace generated by 1 and 3 is the set of points $\{1,3\}$ with no lines.

. . .

An incidence matrix $R = [r_{ij}]$ is an alternate way of representing a near-linear space. The rows of an incidence matrix represent the points and the columns represent the lines. If a point (associated with row i) is on a line (associated with column j), then we set $r_{ij} = 1$ of the incidence matrix; otherwise, we put a 0 in cell $r_{ij} = 0$.

- Axiom NL1 implies that each column of an incidence matrix has at least two 1s.
- Axiom NL2 implies there is at most one k such that r_{ik} and r_{jk} are both 1 for distinct points i and j.

Table 3 is the incidence matrix for the near-linear space in Figure 67.

Table 3. Incident matrix for a near-linear space

	ℓ_1	ℓ_2	ℓ_3	ℓ_4
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	1	0	0
5	0	1	1	0
6	0	0	1	1
7	1	0	1	0
8	0	1	0	0
9	0	0	0	1
10	1	0	0	0

In general, if we sum the entries in a row (say row i), we get the number of lines containing the point p_i. This gives the formula

$$\sum_{j=1}^{b} r_{ij} = b(p_i) = b_i$$

If we sum the entries in a column (say row j), we get the number of points on the line ℓ_j. This gives the formula

$$\sum_{i=1}^{v} r_{ij} = v(\ell_j) = v_j$$

. . .

A near-linear space S is **line regular** if each line has the same number of points (assuming there are any lines in S, e.g., S could be a set of isolated points). This number is called the line regularity.

A near-linear space S is **point regular** if each point is on the same number of lines (assuming there are any points in S, e.g., S could be the empty set). This number is called the point regularity.

The near-linear space in Figure 64 has line regularity 2 and point regularity 3. The near-linear space in Figure 67 is neither line nor point regular.

Theorem 9. If a near-linear space S has line regularity x, point regularity y, v points and b lines, then $vy = bx$.

Proof: This follows from a consideration of the incidence matrix $R = [r_{ij}]$.

Since the line regularity is x, then each column of R has x ones, and since there are b columns in R, there are a total of bx ones in R.

Since the point regularity is y, then each row of R has y ones, and since there are v rows in R, there are a total of vy ones in R.

Thus, $vy = bx$. ■

. . .

For a point p_i **not** on the line ℓ_j, the **connection number** $c(p_i, \ell_j) = c_{ij}$ is defined to be the number of points on ℓ_j that are joined to p_i by a line. If p_i **is** on the line ℓ_j, then we define $c(p_i, \ell_j)$ to be 1. Thus, $c_{ij} = 1$ if $r_{ij} = 1$.

For any point $p_i$and line ℓ_j in a near-linear space, $c(p_i, \ell_j) \leq v_j$ (i.e., the number of points on line ℓ_j) since by Axiom NL2, the can be a most one line between p_i and any point on ℓ_j.

As an example of the connection number concept, consider the near-linear space shown in Figure 68.

- $c(p_1, \ell_1) = 5$ since the five points on ℓ_1 are joined to p_1 via lines the $\ell_2, \ell_3, \ell_4, \ell_5$ and ℓ_6, respectively.
- $c(p_4, \ell_2) = 2$ since two points on ℓ_2 are joined to p_4 via the lines ℓ_1 and ℓ_4.
- By definition, $c(p_4, \ell_1) = 1$.

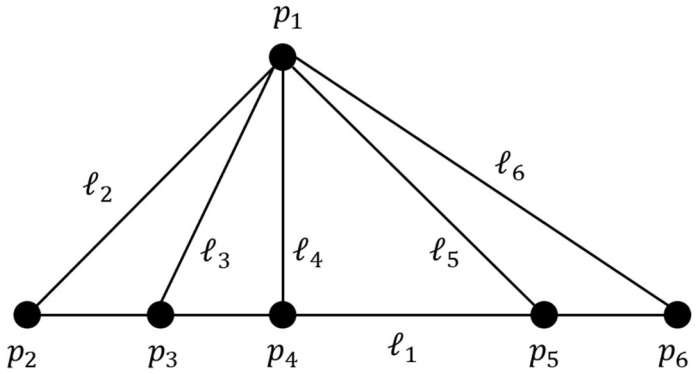


Figure 68. Connection number example

The near-linear space in Figure 69 has line regularity 2 and point regularity 2. Some example connection numbers are as follows: $c(p_1, \ell_1) = 1$ (by definition), $c(p_2, \ell_1) = 1$ and $c(p_3, \ell_1) = 0$.

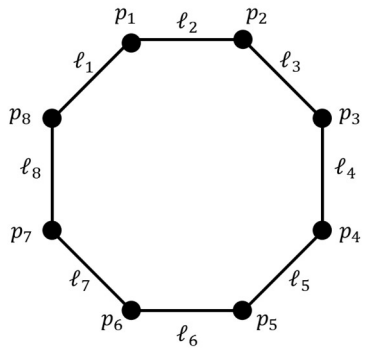


Figure 69. Near-linear space with equal line and point regularities

We have the following two theorems related to connection numbers.

Theorem 10. For point p and line ℓ in a near-linear space, $c(p,\ell) \leq v(\ell)$.

Proof: If p is on ℓ, then we know (by definition) that $c(p,\ell) = 1$, and by Axiom NL1, $v(\ell) \geq 2$. So, in this case, $c(p,\ell) = 1 < v(\ell)$.

If p is not on ℓ, then there can be at most one line connecting p to each point on line ℓ which implies $c(p,\ell) \leq v(\ell)$. ■

Theorem 11. In a near-linear space, if $r_{ij} = 0$, then the number of lines containing p_i which do not intersect ℓ_j is given by $b_i - c_{ij}$.

Proof: $r_{ij} = 0$ means that point p_i is not on line ℓ_j. By definition, the number of lines connecting p_i to a point on line ℓ_j is given by c_{ij}. The total number of lines containing p_i is b_i. Thus, the number of lines containing p_i but not intersecting ℓ_j is $b_i - c_{ij}$. ■

4.3 Linear Spaces

As one might expect, the definition of a linear space entails a small enhancement to the definition of a near-linear space. In particular, we require that any two points in a linear space are on exactly one line. More formally, a finite **linear space** is a space $S = (P, L)$ consisting of a finite set of points P and a finite set of lines L such that

- Axiom LS1: any line has at least two points (same as Axiom NL1)
- Axiom LS2: two points are on **exactly** one line.

Since a linear space satisfies the conditions of a near-linear space, all the results mentioned in the previous section about near-linear spaces apply to linear spaces.

For example, the spaces in Figure 68 and Figure 69 are linear spaces. The near-linear space in Figure 67 is not a linear space, since there is no line containing points 2 and 4 (as well as several other violations of Axiom LS2).

Theorem 12. A near-linear space S with at least one line is a linear space if $c_{ij} = v_j$ for every point p_i and line ℓ_j such that $r_{ij} = 0$ (i.e., point p_i is not on the line ℓ_j).

Proof: We are given that Axiom LS1 holds true. Assume the line given in the statement of the theorem is ℓ_k.

We need to show Axiom LS2 also holds true. Consider two points in S (call them p_n and p_m).

If $r_{nk} = r_{mk} = 1$, then the two points are on line ℓ_k and we are done since, by Axiom NL2, the two points cannot be on another line.

If $r_{nk} = 0$ (i.e., p_n is not on line ℓ_k) and if $r_{mk} = 1$ (i.e., p_m is on line ℓ_k), then, by assumption, we have $c_{nk} = v_k$, meaning the p_n is connected to every point on line ℓ_k by line. In particular, there is a line containing p_n and p_m.

If $r_{nk} = r_{mk} = 0$, then neither p_n nor p_m are on the line ℓ_k. Consider a point $p_i \in \ell_k$ (noting the ℓ_k must have a least two points by Axiom LS1). By hypothesis, $c_{nk} = v_k \geq 2$ and so, there must be a line connecting p_n to p_i (call this line ℓ). If p_m is on ℓ, we are done. If p_m is not on ℓ, we apply the hypothesis once again to show there must be a line connecting p_m to each point on the line ℓ (including the point p_n). ■

The following theorem provides an equivalent condition for a near-linear space to be a linear space.

Theorem 13. Let $S = (P, L)$ be a near-linear space with v points and b lines. S is a linear space if and only if

$$\sum_{i=1}^{b} v_i(v_i - 1) = v(v - 1)$$

Proof: A proof can be found in the book by Batten [33], see the corollary to Theorem 1.6.4. ■

For the linear space in Figure 68, we have $\sum_{i=1}^{b} v_i(v_i - 1) = 5(4) + 5 \cdot (2 \cdot 1) = 30$ and $v(v - 1) = 6(5) = 30$. So, this checks-out with what is expected.

For the near-linear space in Figure 69, we have $\sum_{i=1}^{b} v_i(v_i - 1) = 8 \cdot 2 \cdot 1 = 16$ which does not equal $v(v - 1) = 8(7) = 56$. This is as expected since this is not a linear space (there are multiple pairs of points which are not on a line).

A near-linear space can be extended to a linear space by adding a line between each pair of points that are not already on a line. For example, Figure 70 shows an extension of the near-linear space in Figure 69 to a linear space. The modified figure has $\frac{8(7)}{2} = 23$ lines. So, we now have $\sum_{i=1}^{b} v_i(v_i - 1) = 23 \cdot 2 \cdot 1 = 56 = v(v - 1)$.

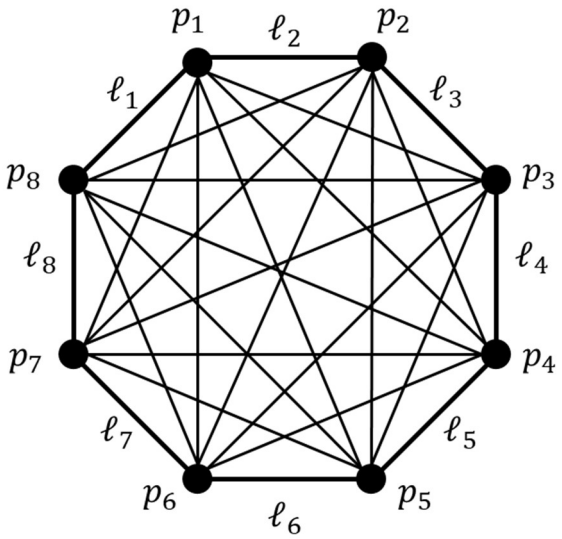


Figure 70. Extending a near-linear space to a linear space

The following theorem (known as the de Bruijn – Erdös theorem) is a key result concerning linear spaces.

Theorem 14. If S is a linear space with at least two lines (i.e., $b \geq 2$), then

i. $b \geq v$, or

ii. $b = v$. In this case, either

   a. one line has $v - 1$ points with all other lines having two points, or

b. every line has the same number of points (say x) and every point is on x lines.

Proof: See Theorem 2.2.2 in the book by Batten [33]. ■

Figure 70 is an example of the other alternative in Case i of Theorem 14, where $b = 23 > 8 = v$.

Figure 68 is an example of Case ii.a of Theorem 14, where one line (ℓ_1) has $v - 1 = 5$ points and the other lines each have 2 points.

The Fano plane (Figure 1) is an example of Case ii.a where each point is on 3 lines and each line has 3 points.

Another example of Case ii.a is shown in Figure 71. This is known as the projective plane PG(2,3). The numbers in the figure refer to the points. The lines are not labeled. Each line contains 4 points and each point is on 4 lines.

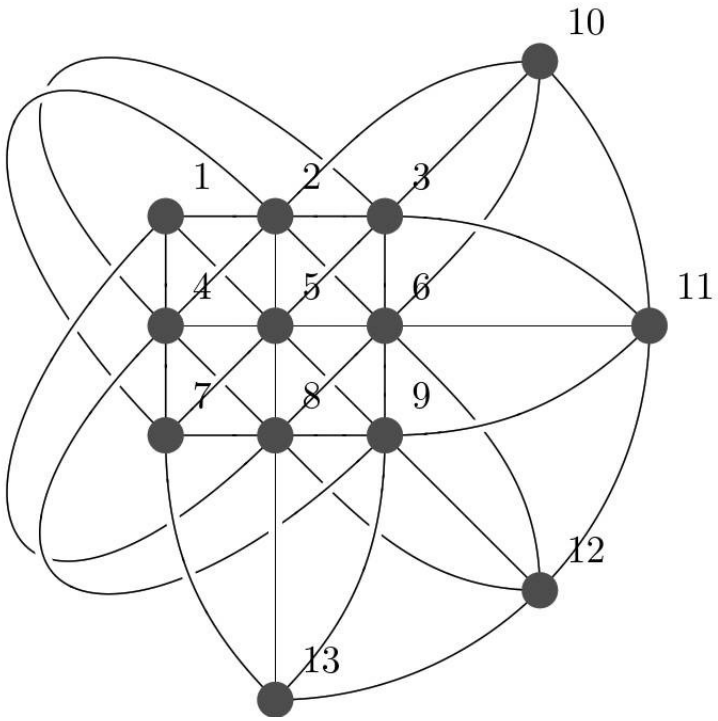


Figure 71. Projective plane PG(2,3)

A more succinct (but less visual) way to represent the linear space in Figure 71 is by listing the points on each line as follows:

1 2 3 11
1 5 9 12
1 4 7 13
1 6 8 10
2 4 9 10
2 7 6 12
2 5 8 13
3 6 9 13
3 5 7 10
3 4 7 12
4 5 6 11
7 8 9 11
10 11 12 13

. . .

There are many numerical properties of linear spaces. We list and prove a few here.

Theorem 15. If p_i is a point in a linear space S with b lines, v points and incidence matrix incidence matrix $R = [r_{ij}]$, then

$$\sum_{j=1}^{b} (v_j - 1) r_{ij} = v - 1$$

Proof: Recall the notation $v_j = v(\ell_j)$ represents the number of points on line ℓ_j.

By Axiom LS2, each point in S is on a unique line containing p_i. Thus, all points in S (excluding p_i) can be counted by tallying the points on each line containing p_i which is just the sum of the $v_j - 1$ term for each line ℓ_j containing p_i (i.e., when $r_{ij} = 1$). ■

The following theorem relates line regularity to point regularity.

Theorem 16. A line regular linear space S is also point regular.

Proof: Let the line regularity be x, i.e., each line contains x points. Thus, $v_j = x$ for every line $\ell_j \in S$. Also, assume S has v points and b lines.

Take any point $p_i \in S$. By Theorem 15, we have that

$$v - 1 = \sum_{j=1}^{b} (v_j - 1) r_{ij} = (x - 1) \sum_{j=1}^{b} r_{ij} = (x - 1) b_i$$

Thus, $b_i = \frac{v-1}{x-1}$ for each point in S, i.e., S is point regular with point regularity $\frac{v-1}{x-1}$. ■

The linear space in Figure 72 has point regularity 3, but is not line regular. Thus, the converse of Theorem 16 is not true.

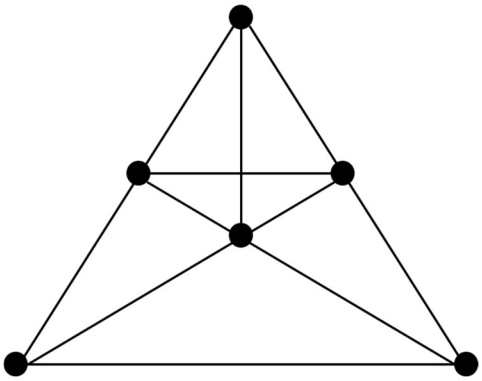


Figure 72. Linear space that is point regular but not line regular

It is also possible for a linear space to be both point and line regular where the point regularity does not equal the line regularity. For example, the linear space in Figure 73 has point regularity 3 and line regularity 2.

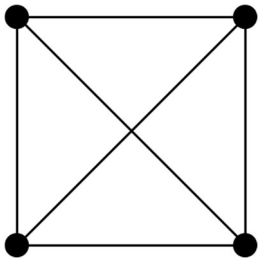


Figure 73. Linear space with line regularity not equal to point regularity

In the case that a linear space has the same line and point regularity, the following theorem relates the number of points and the number of lines to the line (and point) regularity.

Theorem 17. If S is a linear space with line and point regularity $x, x \geq 2$, then all lines intersect and $b = v = x^2 - x + 1$.

The conditions for S imply that it falls under case ii(b) in Theorem 14.

Assume S has v points and b lines.

Proof: Take lines $\ell_1, \ell_2 \in S$ and point $p \in \ell_2$.

If $p \in \ell_1$, then the two lines intersect and we are done (in terms of showing the two lines intersect).

If $p \notin \ell_1$, then every line containing p intersects ℓ_1 since (by hypothesis) there are x lines containing p, x points on line ℓ_1 and by Axiom LS2, two points are on exactly one line. Thus, ℓ_1 intersects ℓ_2.

Next, we determine v (the number of points in S) in terms of x. Take any point $p \in S$. From the details of the proof of Theorem 16, we have that the number of lines containing p is

$$b(p) = \frac{v-1}{x-1} = x$$

which implies $v = x^2 - x + 1$.

Finally, we show that $v = b$. Multiple the number of points in S (i.e., v) times the number of lines containing each point (i.e., x) and we get vx. In this calculation, each line is counted x times and thus,

$$b = \frac{vx}{x} = v$$

To visualize the previous statement, it may help to think in terms of the incidence matrix for S. ■

4.4 Projective Planes

A finite projective plane is a finite linear space with the following properties

- Axiom PP1 any two lines intersect.
- Axiom PP2 there exists a set of four points no three of which are collinear.

Axiom PP1 means there are no parallel lines in a projective space.

Axiom PP2 is meant to exclude the following degenerate cases:

i. the empty set

ii. a single point, with no lines

iii. a single line, with no points

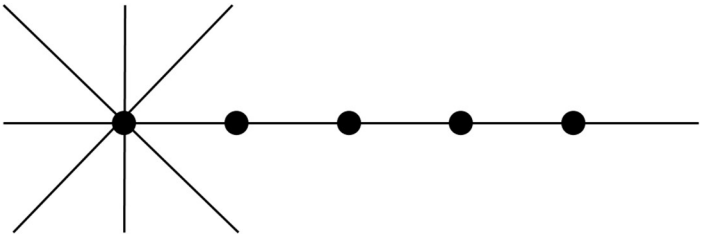
iv. a single point and a collection of lines, where all the lines intersect at the point

v. a single line and a collection of points, where all the points are on the line

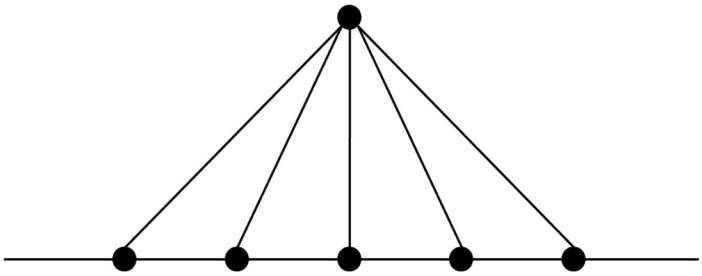
vi. a point p incident with a line ℓ, an arbitrary collection of lines all incident with p and an arbitrary collection of points all incident with ℓ

vii. a point p not incident with a line ℓ, an arbitrary (possibly empty) collection of lines all incident with p, and all the points of intersection of these lines with ℓ.

An example of the type of configuration excluded by case vi is shown in the figure below. Case vi is a combination of cases iv and v.



An example of the type of configuration excluded by case vii is shown in the figure below.



There are alternate but equivalent definitions of a projective plane. It can be shown that Axiom PP2 is equivalent to the following two axioms:

- Each line contains at least three points.
- Each point is on at least three lines.

It can also be proven that Axiom PP2 is equivalent to the following axiom:

> Two lines cannot contain all the points in a projective plane, i.e., for every two lines, there exists a point which is not on either line.

A proof of these equivalences is provided in the book *Finite Geometries* [34], see Lemma 1.4.

The following theorem is fundamental to the study of finite projective planes.

Theorem 18. All finite projective planes are point and line regular, with the point regularity equal to the line regularity. Further, if the point (line) regularity is x, then $b = v = x^2 - x + 1$.

Proof: See Theorem 1.8 in the book *Finite Geometries* [34].

The **order of a finite projective plane** is defined to be one less than the point (line) regularity.

If there is a bijective mapping [35] between the points and lines of two projective planes such that incidence is preserved, then the two projective planes are said to be **isomorphic**.

Much of the research into finite projective planes entails the determination of projective planes of various orders. Also, there are proofs that projective planes do not exist for certain orders. For example, it is known that projective planes of order 6 or 10 are not possible, see the journal article by Lam [36].

Figure 71 shows the finite projective plane of 3. Higher order projective planes are typically described by listing the points in each line since the figures become very difficult to draw. For example, the projective plane of order five (there is only one) is listed below. The points are numbered from 0 to 30, and each line in the list below represents a line. There are $6^2 - 6 + 1 = 31$ points (lines).

0 1 2 3 4 5
0 6 7 8 9 10
0 11 18 19 20 21
0 12 15 22 27 28
0 13 16 24 26 30
0 14 17 23 25 29
1 6 11 12 13 14
1 7 18 22 23 24
1 8 16 19 27 29
1 9 15 21 25 30
1 10 17 20 26 28
2 6 19 22 25 26
3 6 17 18 27 30
4 6 15 20 24 29
5 6 16 21 23 28
2 7 11 15 16 17
2 9 13 18 28 29
2 10 14 21 24 27
2 8 12 20 23 30
4 7 14 19 28 30
5 7 13 20 25 27
3 7 12 21 26 29
5 10 11 22 29 30
4 9 11 23 26 27
3 8 11 24 25 28
5 9 12 17 19 24
3 10 13 15 19 23
4 10 12 16 18 25
5 8 14 15 18 26
4 8 13 17 21 22
3 9 14 16 20 22

The webpage “Projective Planes of Small Order” provides a description of known projective planes of small order, see https://ericmoorhouse.org/pub/planes/. For some orders, there are multiple projective planes, e.g., there 193 known (non-isomorphic) projective planes of order 25.

. . .

The following theorem relates the existence projective planes of a given order to MOLS (which we discussed in Section 2.2.2). A set of $n-1$ MOLS (of dimension $n \times n$) is called a complete set of MOLS.

Theorem 19. There exists a projective plane of order n if and only if there exists a complete set of MOLS of order n.

Proof: See Theorem 1.29 in the book *Finite Geometries* [34].

Complete sets of MOLS of order n exist if n is a prime number or power of a prime number, so projective planes of such orders exist. Finite projective planes with an order different from these, and thus complete sets of MOLS of such orders, are not known to exist [10].

A key result concerning the non-existence of finite projective planes is the Bruck–Ryser theorem, which says that if a projective plane of order n exists and n is of the form $4k+1$ or $4k+2$, then n must be the sum of two perfect squares. For a proof of this result, see Theorem 1.32 in Finite Geometries [34].

. . .

The concept of a projective plane (two dimensions) can be extended to higher dimensions. Such spaces are known as projective spaces and are defined as follows:

> A **projective space** S is a linear space with the property that any two-dimensional subspace is a projective plane.

From the Wikipedia article on projective spaces [37]:

> In mathematics, the concept of a projective space originated from the visual effect of perspective, where parallel lines seem to meet at infinity. A projective space may thus be viewed as the extension of a Euclidean space, or, more generally, an affine space with points at infinity, in such a way that there is one point at infinity of each direction of parallel lines.

4.5 Affine Planes

An **affine plane** is a linear space with the following properties:

- Axiom AP1: any point p not on a line ℓ is on precisely one line not intersecting ℓ.
- Axiom AP2: there exists a set of three non-collinear points.

Two distinct lines in an affine plane are said to be **parallel** if they do not intersect (i.e., do not have a point in common). Also, a line is considered to be parallel to itself. If lines ℓ_1 and ℓ_2 are parallel, we use the notation $\ell_1 \parallel \ell_2$.

An example of an affine plane is shown in Figure 74. Each of the straight lines and the four curved lines (i.e., 267, 249, 168 and 348) are lines in this example.

- Axiom AP1 holds for all the points, e.g., point 8 is not on line 123 and there is one line (i.e., 789) that contains 8 and does not intersect line 123.
- Concerning Axiom AP2, there are several examples of three non-collinear points, e.g., the points 3, 4 and 5 are not on any one line.

There are many parallel lines in Figure 74, e.g., line 123 is parallel to line 789, and line 357 is parallel to line 168.

The configuration is not a projective plane since Axiom PP1 does not hold, e.g., lines 249 and 357 do not intersect.

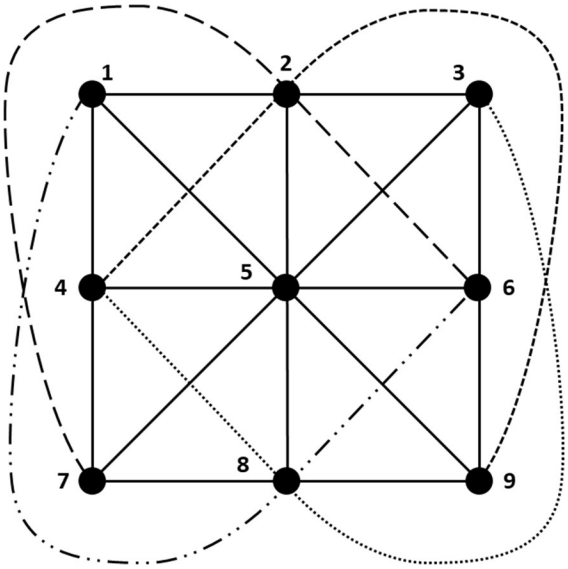


Figure 74. Affine plane with 9 points and 12 lines

As we saw in the previous example, line regularity is not necessarily equal to point regularity in an affine plane. However, we do have the following result that characterizes line and point regularity for all finite affine planes.

Theorem 20. In a finite affine plane $\mathcal{A}$, if a line contains (exactly) k points, then

i. every line of $\mathcal{A}$ contains exactly k points

ii. every point of $\mathcal{A}$ is on exactly $k + 1$ lines

iii. $\mathcal{A}$ has k^2 points and $k^2 + k$ lines.

Proof: See Theorem 1.21 of *Finite Geometries* [34].

The number k in the above theorem is known as the **order of the affine plane**.

The affine plane in Figure 74 has order 3. From Theorem 20, there are $3^2 = 9$ points, $3^2 + 3 = 12$ lines, and every point is on $3 + 1 = 4$ lines (all of which checks out with the actual figure).

In terms of line intersections, we have the following result.

Theorem 21. In an affine plane $\mathcal{A}$ of order k, each line intersects k^2 other lines.

Proof: Take any line $\ell \in \mathcal{A}$. By Theorem 20i, ℓ contains k points. At each point along ℓ, the line ℓ intersects k other lines by Theorem 20ii. This gives us a total of k^2 intersections of ℓ with other lines. If any of the lines of intersection were the same at two different points on ℓ, then we would have two points on two different lines which violates Axiom LS2. ■

In terms of parallel lines, we have the following result.

Theorem 22. In an affine plane $\mathcal{A}$ of order k, each line is parallel to k lines (including the line itself).

Proof: Let ℓ be any line in $\mathcal{A}$, and take any line g (distinct from ℓ) that intersects ℓ. The intersection can only be one point given Axiom LS2 (call this point p). Each line h that is parallel to ℓ does not contain p (by definition of parallel lines). Further, h intersects g; otherwise, there would be two lines containing p (i.e., lines ℓ and g) that are parallel to h, which contradicts Axiom AP1. Since **every** line parallel to ℓ intersects g in a unique point (i.e., there cannot be two lines through a point of g that are parallel to ℓ as this would violate Axiom AP1) and g has k points, then there are a total of k lines parallel to ℓ. ■

. . .

For a given affine plane $\mathcal{A}$, let $[\ell]$ represent all the lines that are parallel to line ℓ (noting that this set also includes ℓ). The set $[\ell]$ is an equivalence class [38]. Each line of $\mathcal{A}$ is in one and only one equivalence class, and thus, parallelism defines a partitioning of the lines in $\mathcal{A}$. Each equivalence class $[\ell]$ is known as a **point at infinity**. Further, we define a line ℓ_∞ that contains all the points of infinity for $\mathcal{A}$. This is known as the **line at infinity**.

Using points at infinity and the line at infinity for any finite affine plane $\mathcal{A}$, we can create a projective plane.

- The points of the projective plane are the points of $\mathcal{A}$ plus all the points at infinity
- Each line of $\mathcal{A}$ is modified to include the point at infinity, and then added to the projective plane.
- The line at infinity is also a line of the projective plane.

For a proof that this construction produces a valid projective plane, see Example 1.20 in Finite Geometries [34] or the article “Construction of projective planes from affine planes” [39].

Theorem 23. In an affine plane $\mathcal{A}$ of order k, there are $k+1$ equivalence classes (i.e., points at infinity).

Proof: By Theorem 22, each equivalence has k members. By Theorem 20, there are k^2+k lines. Putting the two facts together, there must be $\frac{k^2+k}{k} = k+1$ equivalence classes in $\mathcal{A}$. ■

As an example of the construction of a projective plane from an affine plane, we shall extend the affine plane in Figure 74 (call it $\mathcal{A}$) to a projective plane (call it $\mathcal{B}$). $\mathcal{A}$ has 4 points at infinity which become points of $\mathcal{B}$, i.e.,

$$p_1 = [123] = \{123, 456, 789\}$$
$$p_2 = [147] = \{147, 258, 369\}$$
$$p_3 = [357] = \{357, 249, 168\}$$
$$p_4 = [159] = \{159, 267, 348\}$$

The lines of $\mathcal{B}$ are as follows:

$$123p_1, 456p_1, 789p_1$$
$$147p_2, 258p_2, 369p_2$$
$$357p_3, 249p_3, 168p_3$$
$$159p_4, 267p_4, 348p_4$$

We also add the line $\ell_\infty = \{p_1, p_2, p_3, p_4\}$ to $\mathcal{B}$.

So, $\mathcal{B}$ has 13 lines and 13 points.

. . .

It is also possible to go in the other direction, i.e., modifying a projective plane to get an affine plane. The process is to remove one line and all its points from the projective plane. This means that points from the removed line need to be removed from the remaining lines. For a proof that this construction is valid, see Theorem 4.3.2 of *Combinatorics of Finite Geometries* [33].

For example, consider the projective plane in Figure 71 with the line 1, 2, 3, 11 and its points removed. The result is an affine plane with points {4, 5, 6, 7, 8, 9, 10 12, 14} and the lines listed below.

5 9 12
4 7 13
6 8 10
4 9 10
7 6 12
5 8 13
6 9 13
5 7 10
4 7 12
4 5 6
7 8 9
10 12 13

4.6 Relationship to BIBDs

Alternate but equivalent definitions of projective planes and affine planes can be stated in terms of BIBDs.

A projective plane of order $n \geq 2$ (if such exists) is an $(n^2 + n + 1, n + 1, 1)$-BIBD.

The $n^2 + n + 1$ elements are equivalent to points, and the $n + 1$ blocks are equivalent to lines.

Theorem 24. For every prime power $q \geq 2$, there exists a symmetric $(q^2 + q + 1, q + 1, 1)$-BIBD (i.e., a projective plane of order q).

Proof: See Theorem 2.10 in *Combinatorial Designs* [3].

We can also define affine planes in terms of BIBDs.

An affine plane of order $n \geq 2$ (if such exists) is a $(n^2, n, 1)$-BIBD.

Theorem 25. For every prime power $q \geq 2$, there exists a $(q^2, q, 1)$-BIBD (i.e., an affine plane of order q).

Proof: See Theorem 2.13 in Combinatorial Designs [3].

5 Abstract Algebra

"To attain knowledge, add things every day. To attain wisdom, remove things every day."
Lao Tzu

5.1 Overview

As the name suggests "abstract algebra" is an abstraction (and extension) of concepts from elementary algebra. Some definitions from various sources:

- (Wolfram MathWorld) "Abstract algebra is the set of advanced topics of algebra that deal with abstract algebraic structures rather than the usual number systems. The most important of these structures are groups, rings, and fields."
- (dictionary.com) "Abstract algebra: the branch of mathematics that deals with the extension of algebraic concepts usually associated with the real number system to other, more general systems."

From Google Bard (AI application), when asked for the definition of "abstract algebra":

"Abstract algebra is the branch of mathematics that studies the properties of algebraic structures, such as groups, rings, and fields. ***These structures are abstract in the sense that they are not tied to any particular set of numbers or objects. Instead, they are defined by their properties****, which can be studied independently of any specific implementation.*

Abstract algebra is a vast and complex subject, but it is also one of the most important branches of mathematics. It has applications in many other areas of mathematics, as well as in physics, chemistry, computer science, and other fields.

Some of the most important concepts in abstract algebra include:

- *Groups: A group is a set of elements together with an operation that combines any two elements of the set to produce another element of the set. The operation must satisfy certain properties, such as associativity, commutativity, and the existence of an identity element.*
- *Rings: A ring is a set of elements together with two operations, addition and multiplication, that satisfy certain properties. The properties of a ring are similar to those of a group, but with the additional requirement that multiplication is associative and distributes over addition.*
- *Fields: A field is a ring in which every nonzero element has an inverse under multiplication. Fields are the most important algebraic structures, and they have many applications in mathematics and other areas.*

Abstract algebra is a challenging but rewarding subject. It requires a high level of mathematical maturity and abstract thinking, but it can provide deep insights into the nature of mathematics."

In terms of preliminaries, it is assumed that the reader is familiar with the basics of set theory, number theory and permutations. Much of the assumed background material can be found, for example, in the textbook by Judson [40].

Group theory is a vast topic. This section is a very brief summary of some key concepts. A short but good video (from 1992, using old presentation technology) is recommended as a starting point before reading this section, see Part 1 of Neumann [41].

5.2 Groups

5.2.1 Definition

Consider the set of integers, i.e., $\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$.

If we add two integers, we necessarily get another integer, e.g., $(-3) + 7 = 4$. In general, a set having this property is said to be closed under addition. **Warning**: there is only one operation being considered here, i.e., addition. Subtraction is not an operation but rather a sign appended to a number.

The order of addition does not matter, e.g.,

$$2 + (-4) = (-4) + 2 = -2$$

This is known as the commutative property.

With respect to the addition of integers, grouping does not change the result, e.g.,

$$2 + (5 + (-11)) = (2 + 5) + (-11) = -4$$

This is known as the associative property.

The sum of any integer with 0 is the integer itself, e.g., $(-5) + 0 = -5$. With respect to integer addition, 0 is known as the identity element.

For every integer, there exists another number such that when the two numbers are added together, the result is 0, e.g., $2 + (-2) = 0$. This is the property of each number in a set having an inverse.

Many other sets (with an associated binary operation) follow the properties listed above for the integers with the binary operation $+$. In general, a **binary operation** * on a set G, is an operation that takes any two elements $a, b \in G$ and yields another element of G, i.e., $a * b \in G$. This property is known as **closure**.

A **group** is a set G and binary operation *, with the following properties:

- the associative law holds, i.e., for every $a, b, c \in G$, $(a * b) * c = a * (b * c)$
- there exists a special element e, called the identity, such that $e * a = a$ for every $a \in G$.
    - Sometimes 0 in the case of additive groups, and 1 in the case of multiplicative groups are used in lieu of e.
- for every $a \in G$, there exist an inverse element $a' \in G$ such that $a' * a = e$.
    - The alternate notation a^{-1} is often used to represent the inverse of a. In what follows, we sometimes use this alternate notation.
    - It is also true for groups that $a * a' = e$, and we prove this in Theorem 28.

Further, if $a * b = b * a$ for every $a, b \in G$, then G is said to be a **commutative** (or abelian) group.

A group is represented as the pair $(G, *)$.

A group with a finite number of elements is said to be a finite group. In this case, the number of elements in such a group is known as its **order**. The order of a finite group G is written as $|G|$.

5.2.2 Examples

As we saw, the set of integers (represented by the symbol $\mathbb{Z}$) under addition is a commutative group. The set of real numbers (represented by the symbol $\mathbb{R}$) and the set of rational numbers (represented by the symbol $\mathbb{Q}$) are also commutative groups under addition. These sets (with the missing 0) are written as $\mathbb{R}\backslash\{0\}$ (reads as the set of real number with 0 removed), $\mathbb{Q}\backslash\{0\}$ and $\mathbb{Z}\backslash\{0\}$. The backslash (in this context) should be interpreted as set subtraction.

If we consider the operation of multiplication, we have an issue with 0 since there is no multiplicative inverse of 0. However, if we remove 0, then the integers, rational numbers and real numbers are commutative groups under the operation of multiplication. In these cases, the identity element is the number 1.

Another example comes from modular arithmetic under addition [42]. The integers modulo n (represented by the symbol $\mathbb{Z}_n$) under modulo n addition form a commutative group. $\mathbb{Z}_n = \{\overline{0}, \overline{1}, \overline{2}, \dots, \overline{n-1}\}$ where addition on $\mathbb{Z}_n$ is defined as the remainder upon division by n. For example, consider $\mathbb{Z}_5 = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}\}$. If we add $\overline{3}$ and $\overline{4}$, the result is $\overline{2}$ since the remainder of $3 + 4 = 7$ is 2 when dividing by 5. In $\mathbb{Z}_5$, the inverse of $\overline{0}$ is itself, the inverse of $\overline{1}$ is $\overline{4}$, and the inverse of $\overline{2}$ is $\overline{3}$.

Modular arithmetic under multiplication works in a similar manner to modular arithmetic under addition, i.e., multiply two numbers and then take the remainder. For example, consider $\mathbb{Z}_6 = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}, \overline{5}\}$. If we multiply $\overline{2}$ times $\overline{4}$ we get $\overline{2}$ since the remainder of 8 when divided by 6 is 2. Is $\mathbb{Z}_6$ a multiplicative group? It is closed under multiplication and we do have a multiplicative identity, i.e., $\overline{1}$. The associative and commutative laws also hold true. However, there is a problem with inverses, i.e., $\overline{0}, \overline{2}, \overline{3}$, and $\overline{4}$ do not have multiplicative inverses. The general solution to this problem is to restrict $\mathbb{Z}_n$ to only those elements that are relatively prime to n. Under such a restriction $(\mathbb{Z}_n, \times)$ is known as the **multiplicative group of integers modulo n** [43]. (The symbol × is used to represent modular multiplication, e.g., we write $\overline{2} \times \overline{5} = \overline{4}$ in $\mathbb{Z}_6$.) For example, the multiplicative group $(\mathbb{Z}_{10}, \times)$ has elements $\{\overline{1}, \overline{3}, \overline{7}, \overline{9}\}$. The inverse of $\overline{1}$ is itself and the inverse of $\overline{9}$ is also itself. $\overline{3}$ and $\overline{7}$ are inverses of each other. If we take consecutive powers of $\overline{3}$, we get $\overline{3}, \overline{9}, \overline{7}, \overline{1}$. An element of a group whose powers cover every element in the group is known as a **generator** of the group. For $(\mathbb{Z}_{10}, \times)$, $\overline{3}$ is the generator.

As a shorthand, we sometime use the symbol $\mathbb{Z}_n^+$ to represent the additive group modulo n, and $\mathbb{Z}_n^\times$ to represent the multiplicative group modulo n.

. . .

The **symmetric group** [44], defined over any set, is the group whose elements are all the one-one functions from the given set onto itself (i.e., a bijective mapping [35]), and whose group operation is the composition of functions. In particular, the finite symmetric group S_n is defined as all permutations on the set $\{1, 2, \dots, n\}$. So, S_n has $n!$ elements.

All finite groups can be mapped to a subset of S_n for some positive integer n.

To check that the symmetric group on n elements is in fact a group, we need to verify the group axioms, i.e., closure, associativity, identity, and existence of an inverse for every element.

- The operation of function composition is closed in the set of permutations of the given set $\{1, 2, \dots, n\}$.
- Function composition (not just for permutations, but in general) is associative.
- The permutation that assigns each element to itself is the identity permutation.
- Every bijection has an inverse function that undoes its action, and thus each element of a symmetric group does have an inverse which is also permutation.

As an example, consider S_5. The following is one of the 120 elements in S_5

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 1 & 5 & 2 \end{pmatrix} = (1,3)(2,4,5)$$

The permutation is written in two different (by equivalent) ways in the above example. On the left, we explicitly list the mapping for each element, i.e., 1 to 3, 2 to 4, 3 to 1, 4 to 5 and 5 to 2. On the right, the permutation is divided into component cycles, i.e., $1 \to 3 \to 1$ and $2 \to 4 \to 5 \to 2$.

When written in the cyclic notation on the right, it is understood that one "wraps-around" at the end of the cycle. For example, $(2,4,5)$ represents the mapping $2 \to 4 \to 5 \to 2$.

The identity element for S_5 is

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} = (1)(2)(3)(4)(5)$$

The inverse of an element is the reverse of an element, e.g.,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 1 & 5 & 2 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 1 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

Note: we go from left to right when multiplying two permutation. Some books and articles on this topic do multiplication from right to left.

The following shows the multiplication of two elements of S_5, with the order of multiplication reversed in the second line. Since the result of multiplication depends on the order of the terms, S_5 is not a commutative group.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 1 & 5 & 2 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 5 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 1 & 4 & 2 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 5 & 1 & 2 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 1 & 5 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 1 & 2 & 3 & 4 \end{pmatrix}$$

For $n \geq 2$, S_n is a non-abelian (i.e., non-commutative) group.

Every permutation can be broken down into the product of one or more cycles. For example, consider the following permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 9 & 7 & 4 & 1 & 2 & 5 & 6 & 8 & 11 & 3 & 10 \end{pmatrix} = (1,9,11,10,3,4)(2,7,6,5)(8)$$

If the cycles are independent of each other (i.e., don't share any numbers), then the order does not matter. So, for example, the above permutation also equals $(2,7,6,5)(8)(1,9,11,10,3,4)$.

It is always possible to decompose a permutation into a product of two-cycles, and one-cycles. For example, the permutation $\sigma = (1,3,2,5,4,6)$ can be written as

$$(1,3)(1,2)(1,5)(1,4)(1,6)$$

This may be hard to processes upon seeing for the first time. So, let's expand the above multiplications in detail.

$$(1,3)(1,2) = \begin{pmatrix}1&2&3&4&5&6\\3&2&1&4&5&6\end{pmatrix} \circ \begin{pmatrix}1&2&3&4&5&6\\2&1&3&4&5&6\end{pmatrix} = \begin{pmatrix}1&2&3&4&5&6\\3&1&2&4&5&6\end{pmatrix}$$

$$(1,3)(1,2)(1,5) = \begin{pmatrix}1&2&3&4&5&6\\3&1&2&4&5&6\end{pmatrix} \circ \begin{pmatrix}1&2&3&4&5&6\\5&2&3&4&1&6\end{pmatrix} = \begin{pmatrix}1&2&3&4&5&6\\3&5&2&4&1&6\end{pmatrix}$$

$$(1,3)(1,2)(1,5)(1,4) = \begin{pmatrix}1&2&3&4&5&6\\3&5&2&4&1&6\end{pmatrix} \circ \begin{pmatrix}1&2&3&4&5&6\\4&2&3&1&5&6\end{pmatrix} = \begin{pmatrix}1&2&3&4&5&6\\3&5&2&1&4&6\end{pmatrix}$$

$$(1,3)(1,2)(1,5)(1,4)(1,6) = \begin{pmatrix}1&2&3&4&5&6\\3&5&2&1&4&6\end{pmatrix} \circ \begin{pmatrix}1&2&3&4&5&6\\6&2&3&4&5&1\end{pmatrix} = \begin{pmatrix}1&2&3&4&5&6\\3&5&2&6&4&1\end{pmatrix} = (1,3,2,5,4,6)$$

In general, any cycle can be decomposed as follows:

$$(a_1, a_2, a_3, \dots, a_n) = (a_1, a_2)(a_1, a_3) \dots (a_1, a_n) = (a_1, a_2)(a_2, a_3) \dots (a_{n-1}, a_n)$$

Other 2-cycle decompositions are possible, e.g., $(a_1, a_n)(a_2, a_n) \dots (a_{n-1}, a_n)$.

So, every permutation of a finite set can be expressed as the product of 2-cycles (known as **transpositions**). While several such expressions for a given permutation may exist, either all contain an even number of transpositions or they all contain an odd number of transpositions (for a proof of this fact, see Theorem 2.40 of Rotman [50]). Thus, all permutations can be classified as even or odd depending on the number of transpositions in its cyclic decomposition.

Some examples

- $(1,3,2,5,4,6) = (1,3)(1,2)(1,5)(1,4)(1,6)$ is an odd permutation.
- $\begin{pmatrix}1&2&3&4&5&6&7&8&9&10&11\\9&7&4&1&2&5&6&8&11&3&10\end{pmatrix} = (1,9,11,10,3,4)(2,7,6,5)(8) =$ $(1,9)(1,11)(1,10)(1,3)(1,4)(2,7)(2,6)(2,5)(8)$ is an even permutation. Cycles of length 1 (8 in this case) are not counted in the parity calculation.

. . .

The subset of S_n containing all the even permutations over a set of n elements is known as the **alternating group** of degree n, and is denoted by A_n [45].

Theorem 26. The number of even permutations in S_n is equal to the number of odd permutations and so, A_n has $\frac{n!}{2}$ elements.

Proof: See Proposition 5.17 of Judson [40]. ■

A_n is a group.

- If we multiple two even permutations, we get an even number of transpositions and so, closure holds true.
- The associative law is inherited from S_n.
- The identity permutation has 0 transpositions (i.e., is an even permutation) and therefore, is in A_n.
- We know that each element of A_n has an inverse, since A_n is a subset of S_n and S_n is a group. The question is whether the inverse of an element in A_n is also an even permutation. To see this, we decompose an element $\alpha \in A_n$ into a product of disjoint cycles and single permutations. Let $\alpha = \sigma_1\sigma_2 \dots \sigma_k(x_1)(x_2) \dots (x_h)$ where the σ_i terms are cycles and the x_i terms are single permutations (e.g., $4 \to 4$). There is no need to invert the single permutations. Let's look at a general cycle, i.e., $\beta = (a_1, a_2, a_3, \dots, a_n)$. Its inverse is $\beta' = (a_n, a_{n-1}, a_{n-2}, \dots, a_3, a_2, a_1)$. β' can be decomposed into the same number of transpositions as β. So, the inverse of a cycle has the same parity as the original permutation. The inverse of α is $\sigma_1'\sigma_2' \dots \sigma_k'(x_1)(x_2) \dots (x_h)$ and each of the σ_i' terms has the same parity as the corresponding σ_i terms. Thus, the parity of α' is the same as that of α (which as even since $\alpha \in A_n$).

. . .

A **dihedral group** [46] is the group of symmetries of a regular polygon, including rotations and reflections. A regular polygon with n sides (i.e., an n-gon) has $2n$ different symmetries: n rotational symmetries (in increments of $\frac{360}{n}$ degrees) and n reflection symmetries. The set of rotations and reflections make up the dihedral group D_n. (**Warning**: Since the group of symmetries on a regular n-gon is of order $2n$, some sources call this group D_{2n}.)

- If n is odd, each axis of symmetry (with respect to reflection) connects the midpoint of one side to the opposite vertex (for a total of n axes of symmetry).
- If n is even, there are $\frac{n}{2}$ axes of symmetry (with respect to reflection) connecting the midpoints of opposite sides, and $\frac{n}{2}$ axes of symmetry connecting opposite vertices.

In either case, there are n axis of symmetry (with respect to reflection).

The dihedral group D_3 consists of the following operations on an equilateral triangle:

- The 3 rotations of an equilateral triangle, i.e., 0 degrees rotation (the identity element e), 120 degrees clockwise rotation (denoted by a) and 240 degrees clockwise rotation (denoted by a^2; applying a twice to the triangle).
- The three reflections about the axes between a vertex and midpoint of the opposite side. If we let b be one of the 180 degrees reflections of the triangle, the other two reflections can be gotten by a combination of rotations and a reflection, i.e., ab and a^2b.

With the above conventions, we have 6 group elements, i.e., e, a, a^2, b, ab, a^2b. The various combinations of multiplying the elements (basically function composition) is shown in Table 4 (which is an example of what is called a Cayley table). The elements of the group are listed in the top row and left column. It works just like a multiplication table for numbers. The following identities are used in the table: $a^3 = e, a^4 = a, b^2 = e, ba = a^2b, ba^2 = ab$.

D_3 is a non-abelian group, e.g., $ab \neq ba$.

Table 4. Cayley table for D_3

∘	$\boldsymbol{e}$	$\boldsymbol{a}$	$\boldsymbol{a^2}$	$\boldsymbol{b}$	$\boldsymbol{ab}$	$\boldsymbol{a^2b}$
$\boldsymbol{e}$	e	a	a^2	b	ab	a^2b
$\boldsymbol{a}$	a	a^2	e	ab	a^2b	b
$\boldsymbol{a^2}$	a^2	e	a	a^2b	b	ab
$\boldsymbol{b}$	b	a^2b	ab	e	a^2	a
$\boldsymbol{ab}$	ab	b	a^2b	a	e	a^2
$\boldsymbol{a^2b}$	a^2b	ab	b	a^2	a	e

. . .

The set of all invertible matrices of size $n \times n$ with elements from the real numbers form a non-abelian group. This group is known as the **general linear group** [47] of degree n over the real numbers, and is denote by $GL_n(\mathbb{R})$. We can also define a general linear group over the rational numbers $GL_n(\mathbb{Q})$ or the complex numbers $GL_n(\mathbb{C})$. In general, one can define general linear groups over a field F. Such groups are denoted $GL_n(F)$. We will discuss fields in Section 5.4. The real numbers, rational numbers and complex numbers are examples of infinite fields. We have already seen an example of a finite field, i.e., $\mathbb{Z}_p$ the integers modulo p where p is a prime. Regardless of the underlying field F, all matrices in $GL_n(F)$ must have non-zero determinant (which implies the inverse of the matrix exists) [48]. For those familiar with linear algebra, the term "linear" in "general linear group" refers to the rows (columns) being linearly independent.

The **special linear group** [49], written $SL_n(F)$, is the set of matrices (with a determinant of 1) over a field F. $SL_n(F)$ is a subgroup within $GL_n(F)$.

5.2.3 Some Basic Theorems

The proof of the following theorem is included to illustrate usage of the properties of a group.

Theorem 27. If $a * a = a$ in a group $(G,*)$, then $a = e$.

Proof: Apply a' (i.e., the inverse of a) to both sides of the given equation to get

$$a' * (a * a) = a' * a$$

Apply the associative law to the left-side of the previous equation and the inverse property to the right side of the equation to get

$$(a' * a) * a = e$$

Applying the inverse property to the left-side of the equation, we get $e * a = e$. Finally, apply the identity property to the left-side of the equation to get the desired result of $a = e$. ■

Even in non-abelian groups, inverses commute, and the identity element commutes with all elements.

Theorem 28. In a group $(G,*)$, $a * a' = e$, and $a * e = a$ for every $a \in G$.

Proof: From the inverse property, we have $\mathrm{a}' * \mathrm{a} = \mathrm{e}$. Multiply both sides on the left by a to get

$$a * (a' * a) = a * e$$

Applying the associative law to the left-side of the above equation yields

$$(a * a') * a = a * e$$

Multiply both sides on the right by a' to get

$$(a * a') * (a * a') = a * (e * a') = (a * a')$$

Applying Theorem 27 to the above, we have that $a * a' = e$.

We use the first part of the theorem to prove the second part, i.e.,

$$a * e = a * (a' * a) = (a * a') * a = e * a = a$$

■

In the next theorem, we show that the identity element is unique for a given group, and that there is only one inverse for each element of a group.

Theorem 29. For a group $(G,*)$, the following statements hold true

i. If $f \in G$ satisfies $f * a = a$ for every $a \in G$, then $f = e$.

ii. For a given $a \in G$, if $b * a = e$, $b = a'$.

Proof: For the first part of the theorem, the given condition implies that $f * f = f$ and by Theorem 27, we have that $f = e$.

For the second part of the theorem, we use Theorem 28 and the associative law as follows:

$$b = b * e = b * (a * a')$$
$$= (b * a) * a' = e * a' = a'$$

■

Theorem 30. For a group $(G,*)$***, the following statements hold true***

i. If either $x * a = x * b$ or $a * x = b * x$, then $a = b$ (cancellation law).

ii. $(a')' = a$ (inverse of the inverse returns the original element)

iii. $(a_1 * a_2 * \ldots * a_n)' = a_n' * \ldots * a_2' * a_1'$

Proof:

(i) We have the following

$$a = e * a = (x' * x) * a = x' * (x * a)$$
$$= x' * (x * b) = (x' * x) * b = e * b = b$$

The proof for the other case follows in a similar manner.

(ii) By Theorem 29(ii), inverses are unique and so $(a')'$ is the unique inverse of a', i.e., $(a')'$ is the unique element of G such that $(a')' * a' = e$. By Theorem 28, $a * a' = e$ and thus, it must be that $(a')' = a$.

(iii) We show the result for $n = 2$. The general result follows by mathematical induction.

By multiple applications of the associative law, we have

$$(a * b) * b' * a' = \big(a * (b * b')\big) * a' = (a * e) * a' = a * a' = e$$

By Theorem 29(ii), $(a * b)' = b' * a'$. ■

[**Author's Remark**: The preceding theorems may seem a bit tedious and unnecessary (i.e., why not just assume these results as part of the definition of a group?). The point, however, is to define a group with the minimal set of properties. This approach (i.e., minimal definitions) is a general goal in mathematics.]

If one has a string of operations in a group, e.g., $a_1 * a_2 * a_3 * \ldots * a_{n-1} * a_n$, parentheses are needed. One can group the terms in any manner and the result will be the same. This property is known as **generalized associativity**. For a proof of this fact, see Theorem 2.49 in Rotman [50].

. . .

We can define **exponentiation for groups** in a very similar way to basic algebra for real numbers. For a group $(G,*)$ and $a \in G$, the n^{th} power of a is defined as $a^n = a * a * \ldots * a$ where a is repeated n times. The following laws of exponentiation hold:

- $a^{m+n} = a^m * a^n$
- $(a^m)^n = a^{mn}$
- If a and b commute, then $(ab)^n = a^n * b^n$

If we write the inverse of $a \in G$ as a^{-1} in lieu of our previous notation a', then the above expressions hold for negative as well as positive integers. For example,

$$(a^{-3})^2 = (a^{-1} * a^{-1} * a^{-1})^2 = a^{-1} * a^{-1} * a^{-1} * a^{-1} * a^{-1} * a^{-1} = a^{-6}$$

When talking about groups and exponents, it is more natural to use the alternate notation for the identity element, i.e., use 1 instead of e. For example, we can write $a^1 a^{-1} = a^0 = 1$.

The **order** of $a \in G$ is defined to be the smallest integer k such that $a^k = 1$.

5.2.4 Subgroups

As the name suggests, a **subgroup** H of a group G (written as $H \leq G$) is a subset of G that fulfills the group axioms. Since the associative law is inherited by all subsets of a group, we only need to show the closure, identity and inverse axioms.

A group always has at least two trivial subgroups, i.e., the group itself, and the subgroup consisting of only the identity element. Not very interesting, but important to note with regard to various proofs.

We have already seen a non-trivial example, i.e., the alternating group A_n is a subgroup of the symmetric group S_n.

Consider the group $\mathbb{Z}_8^+$ (integers modulo 8 under addition). The powers of $\overline{3}$, which generate the set $\{\overline{1}, \overline{3}\}$, form a subgroup. Also, the powers of $\overline{5}$ form the subgroup $\{\overline{1}, \overline{5}\}$.

The following theorem gives us another way of establishing that a given subset of a group is a subgroup.

Theorem 31. A subset H of a group G is a subgroup if and only if H is non-empty and, whenever $x, y \in H$, then $xy^{-1} \in H$.

Proof: If $H \leq G$, and $x, y \in H$, then $y^{-1} \in H$ and by the closure property, $xy^{-1} \in H$. Further, since $H \leq G$, $1 \in H$ and thus, H is non-empty.

Going in the other direction, assume $x, y \in H$ implies $xy^{-1} \in H$, and that H is non-empty.

Since H is non-empty, there exist $x \in H$. The assumption implies that $xx^{-1} = 1 \in H$. So, the identity property holds.

The associative law is inherited by H.

Since $1, x \in H$, our assumption implies $1x^{-1} = x^{-1} \in H$. Thus, inverse exist for each element in H.

If $x, y \in H$, then $x, y^{-1} \in H$ (since we just showed inverses exists). Using the assumption, we have that $x(y^{-1})^{-1} = xy \in H$, which establishes the closure property. ■

For finite groups, we have an even easier way of determining if a subset is a subgroup, but first we need the following intermediate (but important) result.

Theorem 32. If G is a finite group with element a, then $a^k = 1$ for some integer $k \geq 1$.

Proof: Consider the sequence

$$1, a, a^2, a^3, \ldots$$

Since G is finite the terms in the above (infinite) sequence, the terms cannot all be unique. So, there must exist integers m and n such that $m > n$ and $a^m = a^n$. Using the laws of exponents, we have

$$1 = a^{-n}a^n = a^{-n}a^m = a^{m-n}$$

So, $k = m - n$ fulfills the theorem. ■

Theorem 33. A non-empty subset of a finite group is a subgroup if it is closed under the binary operation of the group.

Proof: Let H be a subset of a finite group G.

H inherits associativity from G.

Since H is not empty, it has at least one element (call it a). Since H is closed, all powers of a are in H. From Theorem 32, there exist a positive integer k such that $a^k = 1$. Thus, $1 \in H$.

For any $h \in H$, we know by Theorem 32 that there exists a positive integer n such that $h^n = 1$. Since G is a group and h is also in G, we know that h^{-1} exists. Multiple h^{-1} to both sides of the equation $h^n = 1$ to get $h^{-1}h^n = h^{n-1} = h^{-1}$. So, $h^{-1} \in H$. ■

. . .

For a group G and $a \in G$, the set of all powers of a is written as

$$\langle a \rangle = \{a^n : a \in \mathbb{Z}\}$$

Since $a^0 = 1$, the identity is in $\langle a \rangle$. Since positive and negative powers of a are included, we have inverses. Closure and associativity clearly hold true. So, $\langle a \rangle \leq H$. In particular, $\langle a \rangle$ is referred to as the **cyclic subgroup** of G generated by a.

A group G is a **cyclic group** if there exists $a \in G$ such that $G = \langle a \rangle$. In this case, a is known as a generator of G (there could be more than one generator).

The next theorem characterizes the generators of a finite cyclic group. The expression $\gcd(a, b)$ means the greatest common divisor of a and b, e.g., $\gcd(14,49) = 7$.

Theorem 34. If $G = \langle a \rangle$ is a cyclic group of order n, then a^k is a generator of G if and only if $gcd\ (n, k) = 1$, i.e., n and k are relatively prime.

Proof: See Proposition 2.69 of Rotman [50].

Theorem 35. All cyclic groups are abelian (i.e., commutative).

Proof: Let a be the generator of group G. Take any $x, y \in G$. There exists integers m and n such that $x = a^m$ and $y = b^n$. Using the exponentiation rules for groups, we have $xy = a^m a^n = a^{m+n} = a^n a^m = yx$. ■

. . .

We have the following very powerful (and fundamental) theorem of group theory.

Theorem 36 (Lagrange's Theorem) If $H \leq G$, the $|H|$ divides $|G|$.

Clearly, this only applies to groups of finite order.

Proof: See Lagrange's theorem (group theory) [51].

Lagrange's theorem makes the proof of the following theorem rather simple.

Theorem 37. All groups of prime order are cyclic groups.

Proof: Assume that $|G| = p$ is a prime number. Choose $a \in G$, such that $a \neq 1$, and let $H = \langle a \rangle$ be the cyclic subgroup of G generated by a. By Lagrange's theorem, $|H|$ is a divisor of $|G| = p$. Since p is a prime and $|H| > 1$, it must be that $|H| = p$, and thus, $H = G$, i.e., G is a cyclic group generated by a. ■

5.2.5 Group Structure

A basic concept used in the study of group structure is that of a coset.

Given a subgroup H of a group G and $a \in G$, then the (left) **coset** aH is defined as

$$aH = \{ah : h \in H\}$$

It is common to use juxtaposition (i.e., aH) to indicate a coset rather than $a * H$. However, when the group operation is addition, the notation $a + H$ is more common. The coset aH is clearly a subset of G. Further, one can also define a right coset Ha in an analogous manner.

For example, consider the group of integers under addition and the subgroup consisting of multiples of three, i.e., $H = \{\dots, -9, -6, -3, 0, 3, 6, 9, \dots\}$. There are only three distinct cosets of H, i.e.,

$$0 + H = \{\dots, -9, -6, -3, 0, 3, 6, 9, \dots\}$$
$$1 + H = \{\dots, -8, -5, -2, 1, 4, 7, 10, \dots\}$$
$$2 + H = \{\dots, -7, -4, -1, 2, 5, 8, 11, \dots\}$$

Notice that the union of the cosets equals all of G. As we shall see, this is always true.

Our second example involves the group $S_3 = \{e, (1,2), (1,3), (2,3), (1,2,3), (1,3,2)\}$, and the subgroup generated by the powers of $(1,3)$, i.e., $H = \{e, (1,3)\}$. [Note that e is the identity permutation. In terms of single cycles, it is $(1)(2)(3)$.]

The left cosets of H are

$$eH = (1,3)H = \{e, (1,3)\}$$
$$(1,2)H = (1,2,3)H = \{(1,2), (1,2,3)\}$$
$$(2,3)H = (1,3,2)H = \{(2,3), (1,3,2)\}$$

The right cosets of H are

$$He = H(1,3) = \{e, (1,3)\}$$
$$H(1,2) = H(1,3,2) = \{(1,2), (1,3,2)\}$$
$$H(1,2,3) = H(2,3) = \{(2,3), (1,2,3)\}$$

So, we see that left and right cosets using the same element from G are not necessarily equal.

For our third example, consider the subset $H = \{e, b\}$ of the dihedral group D_3 (see Table 4). The left cosets of H are

$$eH = bH = \{e, b\}$$
$$aH = abH = \{a, ab\}$$
$$a^2H = a^2bH = \{a^2, a^2b\}$$

The right cosets of H are

$$He = Hb = \{e, b\}$$
$$Ha = Ha^2b = \{a, ba\} = \{a, a^2b\}$$
$$Ha^2 = Hab = \{a^2, ba^2\} = \{a^2, ab\}$$

Again, the left and right cosets are not equal in all cases, $aH \neq Ha$ and $a^2bH \neq Ha^2b$.

On the other hand, all the right and left cosets of $H = \{e, a, a^2\}$ are equal, i.e.,

$$eH = aH = a^2H = \{e, a, a^2\} = Ha^2 = Ha = He$$
$$bH = abH = a^2bH = \{b, a^2b, ab\} = Ha^2b = Hab = Hb$$

. . .

The index of a subgroup $H \leq G$, denoted $[G:H]$, is defined to be the number of cosets of H in G.

Theorem 38. If G is a finite group and $H \leq G$, then $[G:H] = \frac{|G|}{|H|}$.

Proof: See Corollary 2.82 of Rotman [50].

. . .

The case where all right and left cosets are equal for a given subgroup of a group is critical in the study of group structure.

A subgroup N of a group G is a **normal subgroup** of G if and only if for every $g \in G$ the corresponding left and right cosets are equal, i.e., $gN = Ng$. In terms of notation, we write $N \lhd G$ when N is a normal subgroup of G.

An equivalent definition states that N is a normal subgroup of G if for every $g \in G$ and $x \in N$, implies $gxg^{-1} \in N$. This is sometimes stated as $N = gNg^{-1}$ for every $g \in G$. If G is an abelian group, then every subgroup N is normal, since if $x \in N$ and $g \in G$, then $gxg^{-1} = xgg^{-1} = x \in N$.

For example, $H = \{e, a, a^2\}$ is a normal subgroup of D_3 (as we saw in the previous example).

The special linear group over field F of degree n, i.e., $SL_n(F)$, is a normal subgroup of $GL_n(F)$.

Proof (for those familiar with linear algebra): For any $A \in SL_n(F)$ and $B \in GL_n(F)$, we have that

$$\det(BAB^{-1}) = \det(B)\det(A)\det(B^{-1}) = \det(A) = 1$$

since the determinant of the product of matrices is the product of the determinant of each matrix, and $\det(C) = \frac{1}{\det(c)}$, when $\det(C)$ is non-zero, i.e., when C is invertible [52]. So, $BAB^{-1} \in SL_n(F)$ and thus, $SL_n(F)$ is a normal subgroup of $GL_n(F)$. ■

A_n is a normal subgroup of S_n.

Proof: The product of two odd or two even permutations is even. The product of an odd permutation and an even permutation is odd. By definition, A_n consists of all the even permutations.

- If $g \in A_n$, then g^{-1} (being an even permutation) is also in A_n. For $x \in A_n$, $g^{-1}xg$ is also even and thus, in A_n.
- If $g \in S_n \backslash A_n$ (i.e., in S_n but not in A_n, or in other words, g is an odd permutation), the g^{-1} is also an odd permutation. For $x \in A_n$, $g^{-1}xg$ (being the product of an odd times even times odd permutation) is an even permutation and thus, in A_n. ■

Theorem 39. If $H \leq G$ and $[G:H] = 2$, then $H \lhd G$.

Proof: We will show that if $h \in H$, then $ghg^{-1} \in H$ for every $g \in G$.

Since H is a subgroup, it is closed under the group operation. Thus, $xH = H$ for every $x \in H$. This gives us one coset. We are given that $[G:H] = 2$, and so, there is one other coset with respect to H. This coset must be of the form aH where a is any element in G but not in H.

- Case 1 (g is in coset H): This implies that $ghg^{-1} \in H$, since $H \leq G$ (in particular, the closure property is used here).
- Case 2 (g is in coset aH): In this case, we can write $g = ax$ for some $y \in H$. This gives us $ghg^{-1} = (ax)h(ax)^{-1} = a(xhx^{-1})a^{-1} = aka^{-1}$ where $k = xhx^{-1} \in H$ (noting that k is the product of three elements in H). If $ghg^{-1} \notin H$, then $ghg^{-1} = aka^{-1} \in aH$, i.e., $aka^{-1} = az$ for some $z \in H$. Multiple the previous equation by a^{-1} on the left yields $ka^{-1} = z$ which implies $a = z^{-1}k \in H$. This is a contradiction since we assumed $a \notin H$. Thus, our previous assumption of $ghg^{-1} \notin H$ is false, and $ghg^{-1} \in H$.

So, in either case $ghg^{-1} \in H$, and we have shown that $H \lhd G$. ■

Theorem 39 gives us another way of showing that $A_n \lhd S_n$ since we know that $[S_n:A_n] = 2$ by Theorem 26.

. . .

The set of cosets of N in G (when $N \lhd G$) form a group referred to as a **quotient group** or factor group (written as G/N). For quotient groups, the group operation * is defined as $aN * bN = abN$. The identity is eN. Since inverse exists for every $a \in G$, we have $aN * a^{-1}N = aa^{-1}N = eN$. If N is abelian (commutative), so is G/N.

Theorem 40. If G is a finite group and $N \lhd G$, then $|G/N| = |G|/|N|$.

Proof: This follows from Theorem 38 since the elements of G/N are the cosets of G with respect to N. ■

One can also form quotient groups from infinite groups. For example, consider the additive group of integers $\mathbb{Z}$, and the normal subgroup $n\mathbb{Z} = \{\dots, -3n, -2n, -n, 0, n, 2n, 3n, \dots\}$. The cosets are the collection $\{n\mathbb{Z}, 1 + n\mathbb{Z}, 2 + n\mathbb{Z}, \dots, (n-2) + n\mathbb{Z}, (n-1) + n\mathbb{Z}\}$. The quotient group $\mathbb{Z}/n\mathbb{Z}$ can be thought of as equivalent to the group of remainders modulo n, i.e., $\mathbb{Z}_n^+$.

The idea of equivalent groups is formalized by the concepts of group homomorphism and isomorphism.

Two groups $(G,*)$ and $(H,\star)$ are **homomorphic** if there exists a function $f: G \to H$ such that for all $x, y \in G$ the following holds true

$$f(x * y) = f(x) \star g(y)$$

The purpose of defining a group homomorphism is to create functions that preserve algebraic structure between groups.

An equivalent definition of group homomorphism is as follows:

> The function $f: G \to H$ is a group homomorphism if whenever $x * y = z$ for $x, y, z \in G$, we also have $f(x) \star f(y) = f(z)$.

Theorem 41. Let $f: G \to H$ be a homomorphism between groups $(G,*)$ and $(H,\star)$. Denote the identity of G by e_G, and the identity of H by e_H. The following is true

- ***f maps the identity of G to the identity of H***
- ***f maps inverses to inverses***

Proof: For any $g \in G$, $e_G * g = g$. By the alternative definition of group homomorphism, we have

$$f(e_G) \star f(g) = f(e_G * g) = f(g) = e_H \star f(g)$$

Thus, $f(e_G) = e_H$.

For any $g \in G$, $g * g^{-1} = e_G$. By the alternative definition of group homomorphism, we have

$$f(g) \star f(g^{-1}) = f(g * g^{-1}) = f(e_G) = e_H$$

Since $f(g) \star f(g^{-1}) = e_H$ and inverses are unique, it follows that $f(g^{-1}) = f(g)^{-1}$. ■

A key point regarding homomorphisms (from G to H) is that they are only required to map into a subgroup of H. For example, the following mapping (call it f) defines a homomorphism from $\mathbb{Z}_3^+$ to a subgroup of the dihedral group D_3 (see Table 4):

$$\overline{0} \to e$$

$$\overline{1} \to a$$

$$\overline{2} \to a^2$$

To prove that f is a homomorphism we need to show $f(x + y) = f(x)g(y)$ for every $x, y \in \mathbb{Z}_3^+$. We show three of the verifications below and leave the others to the reader.

$$a^2 = f(\overline{2}) = f(\overline{1} + \overline{1}) = f(\overline{1})f(\overline{1}) = aa = a^2$$

$$a = f(\overline{1}) = f(\overline{2} + \overline{2}) = f(\overline{2})f(\overline{2}) = a^2a^2 = a$$

$$e = f(\overline{0}) = f(\overline{1} + \overline{2}) = f(\overline{1})f(\overline{2}) = aa^2 = e$$

We note that D_3 has three additional elements for which there is no mapping from $\mathbb{Z}_3^+$. This is fine for a homomorphism.

A homomorphism $f: G \to H$ that is one-to-one (i.e., injective) is called an embedding, i.e., the group G "embeds" into H as a subgroup.

If $f(G) = H$, then f is onto (i.e., surjective).

A homomorphism that is both injective and surjective is an **isomorphism**.

The names of elements in two isomorphic groups may differ. They may also look different in terms of their Cayley diagrams, but the isomorphism guarantees that they have the same algebraic structure.

When two groups G and H are isomorphic, we write $G \cong H$.

. . .

A **simple group** is a group whose only normal subgroups are the trivial group (i.e., the subgroup containing only the identity element) and the group itself. A group that is not simple can be broken into two smaller groups, namely a nontrivial normal subgroup N and the corresponding quotient group G/N. This process can be repeated, and for finite groups one can determine a cascade of normal subgroups (known as a **composition series**). Stated more formally: If G is a group, then one can construction a series of the form

$$e = N_0 \lhd N_1 \lhd N_2 \lhd \cdots \lhd N_k = G$$

with strict inclusion (i.e., N_i is a proper subset of N_{i+1}) and each N_i is a maximal normal subgroup of N_{i+1}. Further, each quotient group N_{i+1}/N_i is a simple group.

The following theorem tell us that composition series for a finite group are equivalent.

Theorem 42 (Jordan–Hölder theorem). The composition quotient groups belonging to two composition series of a finite group G are, apart from their sequence, isomorphic in pairs.

In other words, if we have the following two composition series for finite group G

$$e \lhd N_1 \lhd N_2 \lhd \cdots \lhd N_s = G$$

$$e \lhd M_1 \lhd M_2 \lhd \cdots \lhd M_t \lhd G$$

then $t = s$, and all the corresponding quotient groups are isomorphic, i.e.,

$$N_{i+1}/N_i \cong M_{i+1}/M_i$$

Proof: See Baumslag [53].

5.2.6 Classification of Finite Simple Groups

The search for the classification of finite simple groups started (as best as one can tell) in 1892 with the now famous quote from Otto Hölder (translated from German to English) in Mathematische Annalen [54]:

> It would be of the greatest interest if it were possible to give an overview of the entire collection of finite simple groups.

The first paper classifying an infinite family of finite simple groups, starting from a hypothesis on the structure of certain proper subgroups, was published by Burnside in 1899 [55].

The comprehensive classification of finite simple groups is attributed to Daniel Gorenstein in 1983 [56] [57]. However, the classification was not declared complete until corrections of the proof were

made by Aschbacher and Smith in 2004 [58]. A detailed history of this work can be found in the Wikipedia article entitled "Classification of finite simple groups" [59].

The classification is divided among abelian and non-abelian finite simple groups, with the latter being vastly more complex.

5.2.6.1 Finite abelian groups

The following theorem describes the structure of all finite simple abelian groups.

Theorem 43. A finite abelian group is simple if and only if it has prime order p. In this case, it is isomorphic to the cyclic group $\mathbb{Z}_p$.

Proof: See "Abelian Group is Simple iff Prime" [60]

The **direct product** is an operation that takes two groups $(G,*)$ and $(H,\star)$ and constructs a new group, denoted $G \times H$.

- The underlying set of the new group $G \times H$ is simply all the ordered pairs (g,h) with $g \in G$ and $h \in H$.
- The binary operation (denoted with a dot) is $(g_1,h_1) \cdot (g_2,h_2) = (g_1 * g_2, h_1 \star h_2)$.

This definition can be extended to more than two groups.

Theorem 44. Every finite abelian group is isomorphic to the direct product of cyclic groups of prime power order, i.e., a group of the form

$$\mathbb{Z}_{p_1^{a_1}} \times \mathbb{Z}_{p_2^{a_2}} \times \ldots \times \mathbb{Z}_{p_n^{a_n}}$$

Proof: See "Fundamental Theorem of Finite Abelian Groups" [61].

5.2.6.2 Finite simple non-abelian groups

The finite simple non-abelian groups are of one of the following types:

- Alternating group $A_n \geq 5$
- Groups of the finite simple Lie type: these are an assortment of matrix groups with elements from a finite field such as $\mathbb{Z}_p$ for prime number p. We discuss fields further in Section 5.4.
- Sporadic groups: these are groups that don't fit into the first two types. There are only 26 sporadic groups; the largest of which (called the Monster or just M) has $808017424794512875886459904961710757005754368000000000$ elements.

A list of all the finite simple groups (abelian and non-abelian) along with some additional information such as the group order is available in the Wikipedia article entitled "List of finite simple groups" [62].

The finite simple groups can also be organized in a sort of periodic table, see https://irandrus.wordpress.com/2012/06/17/the-periodic-table-of-finite-simple-groups/.

5.3 Rings

5.3.1 Definitions and Basic Concepts

A ring is an algebraic structure with two binary operations (usually referred to as "addition" and "multiplication" even though the operations are not necessarily what we think of as addition and multiplication of real numbers). Ring elements may be numbers such as integers, real or complex numbers, but they may also be other types of objects such as polynomials, square matrices, functions, and power series.

More precisely, a **ring** R is algebraic structure, with operations referred to as addition $+$ and multiplication $*$ (or just juxtaposition in some cases, e.g., ab), having the following properties:

- With respect to addition, R is an abelian group. The additive identity is labeled as 0.
- The multiplicative operation is associative, and there is a multiplicative identity 1. There are no requirements for inverses, or for commutativity under multiplication.
- The following distributive laws must hold:
  - $a * (b + c) = (a * b) + (a * c)$ for all $a, b, c \in R$ (left distributivity).
  - $(b + c) * a = (b * a) + (c * a)$ for all $a, b, c \in R$ (right distributivity).

If $ab = ba$, for every $a, b \in R$, then R is said to be a **commutative ring**.

A **unit element of a ring** is an element that has a multiplicative inverse. So, an element u of a ring R is a unit if there exists $v \in R$ such that $uv = vu = 1$.

The following theorem establishes some basic properties of rings.

Theorem 45. Let R be a ring with additive identity 0 and multiplicative identity 1.

i. $0 * a = 0$ for every $a \in R$

ii. If $-a$ is the additive inverse of $a \in R$, then $(-1) * (-a) = a$

iii. $(-1) * a = -a$ for every $a \in R$.

Proof:

i) Since $0 + 0 = 0$, we have from the distributive law

$$0 * a = (0 + 0) * a = (0 * a) + (0 * a)$$

Since $0 * a \in R$, it has an additive inverse, i.e., $-(0 * a)$. Adding $-(0 * a)$ to both sides of the above gives us $0 = 0 * a$.

ii) Using the distributive law and part i) of this theorem, we have

$$0 = 0 * (-a) = (-1 + 1) * (-a) = (-1) * (-a) + (-a)$$

Adding a to both sides of the above yields $a = (-1) * (-a)$.

iii) Multiple both sides of the identity $(-1) * (-a) = a$ by -1 to get

$$(-1) * (-1) * (-a) = (-1) * a$$

Applying part ii) of this theorem, we have $(-1) * (-1) = 1$, and so, the above equation reduces to $-a = (-1) * a$. ■

The integers $\mathbb{Z}$ are a commutative ring. The integer 0 is the additive identity, and the integer 1 is the multiplicative identity. Other than 1, no number has a multiplicative inverse. The multiplicative inverse of $a \in \mathbb{Z}$ is $\frac{1}{a} \in \mathbb{Q}$ but $\frac{1}{a} \notin \mathbb{Z}$ for $a \neq 1$. However, notice that if $ca = cb$ and $c \neq 0$, then $a = b$. In general, this property is known as the **multiplicative cancellation law**.

The integers modulo a prime number p under both addition and multiplication is a commutative ring (actually a field). We denote this entity as $\mathbb{F}_p$.

If we take the integers modulo n (with n not being a prime number) under addition and multiplication, we still have a commutative ring. For example, $\mathbb{Z}_6$ is a commutative ring. In this case, we have what are called **zero divisors** (i.e., two non-zero elements that multiple to 0). For example, $\overline{2} \cdot \overline{3} \equiv 0 \bmod (6)$. In this case, the multiplicative cancellation law does not hold. For example, $\overline{2} \cdot \overline{3} \equiv \overline{0} \equiv \overline{2} \cdot \overline{0}$ but if we try to cancel the $\overline{2}$ on both sides of the equation, we get $\overline{3} \equiv \overline{0}$ which is false.

A nonzero commutative ring with no zero divisors is known as an **integral domain**. There are several equivalent definitions, see "Integral domain" [85].

Theorem 46. The property of not having zero divisors and the multiplicative cancellation law are equivalent.

Proof: Assume the cancellation law holds for a ring R. Suppose (by way of contradiction) that there are nonzero elements $a, b \in R$ such that $a * b = 0$. By Theorem 45(i), we have $0 * b = 0$ which implies $a * b = 0 = 0 * b$. Applying the cancellation law, we have $a = 0$ (a contradiction).

Conversely, assume R does not have any zero divisors. If $c * a = c * b$ with $c \neq 0$, then $0 = c * a - c * b = c * (a - b)$. Since $c \neq 0$ and we have assumed no zero divisors, it must be that $a - b = 0$ and thus, $a = b$ and the cancellation law holds. ■

. . .

Isomorphic rings are rings that have the same structural properties, even if their elements may be different. Formally, the rings $(R, +, *)$ and $(S, \oplus, \times)$ are **isomorphic** if there is a bijective function $f: R \to S$ such that

- $f(a + b) = f(a) \oplus f(b)\ for\ all\ a, b \in R$
- $f(a * b) = f(a) \otimes f(b)\ for\ all\ a, b \in R$
- Unit (multiplicative identity) preserving, i.e., $f(1_R) = 1_S$.

Several properties follow immediately from these assumptions, e.g., $f(0_R) = f(0_S)$ and $f(a) = -f(a)\ \forall a \in R$. For a comprehensive list of properties see "Ring homomorphism" [64].

An example of isomorphic rings is the ring of integers $\mathbb{Z}$ and the ring of even integers $2\mathbb{Z}$. A ring isomorphism between $\mathbb{Z}$ and $2\mathbb{Z}$ is the function $f: \mathbb{Z} \to 2\mathbb{Z}$ defined by $f(n) = 2n$. This function is bijective and preserves both addition and multiplication. For example, $f(1) = 2, f(2) = 4$ and $f(3) = 6$. Also, $f(1 + 2) = f(3) = 6 = f(1) + f(2)$ and $f(1 \cdot 2) = f(2) = 4 = f(1) \cdot f(2)$.

As an example of non-isomorphic rings consider the ring of polynomials with real coefficients $\mathbb{R}(X)$ and the ring of polynomials with complex coefficients $\mathbb{C}(X)$. To show that they are not isomorphic, assume there is an isomorphism $f: \mathbb{C}(X) \to \mathbb{R}(X)$. Then $f(i) = a$ for some real number a. By the multiplicative property of an isomorphism, we have $a^2 = f(i)^2 = f(i^2) = f(-1)$. From the properties of ring homomorphisms, $f(1) = 1$ and $f(-1) = -1$ as the additive inverse of 1. So,

$a^2 = -1$, but no real number when squared equals -1. Thus, we have a contraction and our initial assumption must be false, i.e., there is no isomorphism between $\mathbb{R}(X)$ and $\mathbb{C}(X)$.

5.3.2 Examples

The set of polynomials with coefficients from the real numbers, denoted $\mathbb{R}(X)$, is a commutative ring. For example, the following is an element of $\mathbb{R}(X)$

$$f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_kX^k$$

where each coefficient a_i is an element of $\mathbb{R}$.

Two polynomials are equal if their corresponding coefficients are equal.

The additive identity is the real number 0, and the multiplicative identity is the real number 1.

Addition is pairwise, i.e., add each coefficient of like terms.

The distributive laws for polynomials are defined as in the definition of a ring. For example, given $f(X), g(X), h(X) \in \mathbb{R}(X)$

$$f(X) * (g(X) + h(X)) = (f(X) * g(X)) + (f(X) * h(X))$$

Multiplication involves extensive use of the distributive law, e.g.,

$$(2 + 3X + X^2)(1 - 2X + 2X^2)$$

$$= 2 + (-4 + 3)X + (-6 + 4 + 1)X^2 + (-2 + 6)X^3 + 2X^4$$

$$= 2 - X - X^2 + 4X^3 + 2X^4$$

Additive inverses are formed by taking the negative of each coefficient.

Multiplicative inverses only exist for non-zero constant polynomials. For example, the multiplicative inverse of the polynomial $X^3 - 5X + 2$ is $\frac{1}{X^3-5X+2}$ which is not an element of $\mathbb{R}(X)$.

Polynomials can also be defined over fields other than the real numbers, see the Wikipedia article "Polynomial ring" [63].

. . .

Take any set S and form the power set $\mathbb{P}(S)$ of S (i.e., the set of all possible subsets including the empty set and the set S itself). Over the elements of $\mathbb{P}(S)$, define addition to be the symmetric difference of sets (aka exclusive OR, or just XOR), and multiplication to be intersection.

- The symmetric difference of sets A and B is the collection of elements in either A or B, but not in both. It is written as $A \triangle B$. The symmetric difference can also be expressed as the union of the two sets, minus their intersection, i.e., $A \triangle B = (A \cup B) - (A \cap B)$.
- The intersection of sets A and B is the collection of elements in both A and B. It is written $A \cap B$.

$\mathbb{P}(S)$ with the above operations is a commutative ring.

- Clearly, intersection and symmetric difference result in another subset of S, and thus, we have additive and multiplicative closure.
- The additive identity is the empty set $\phi = \{\}$.
- The additive inverse of set A is itself.
- Symmetric difference is commutative and associative.
- The multiplicative identity is S. (Only S has a multiplicative inverse.)
- Intersection is commutative and associative.
- Intersection distributes over symmetric difference [65].

Since the cancellation law does not hold for intersection, $\mathbb{P}(S)$ under the stated binary operations is not an integral domain.

. . .

Consider the set $\{a + b\sqrt{D}: a, b, D \in \mathbb{Z}\}$ which we denote by $\mathbb{Z}[\sqrt{D}]$. In order to reduce the elements of R to their simplest form, we require that D be square free, i.e., none of its factors is a square. For example, 24 is square free but $90 = 2 \cdot 5 \cdot 3^2$ is not. Numbers of the form $a + b\sqrt{D}$ are known as quadratic integers [67].

R is a commutative ring.

- $\left(a + b\sqrt{D}\right) + \left(a_1 + b_1\sqrt{D}\right) = (a + a_1) + (b + b_1)D$ (additive closure)
- The additive identity is 0.
- Additive associativity and commutativity are inherited from the real numbers.
- $-a - b\sqrt{D}$ is the additive inverse of $a + b\sqrt{D}$
- $\left(a + b\sqrt{D}\right)\left(a_1 + b_1\sqrt{D}\right) = (aa_1 + bb_1D) + (a_1b + ab_1)\sqrt{D} \in R$ (multiplicative closure)
- Multiplicative associativity and commutativity are inherited from the real numbers.
- The multiplicative identity is $1 = 1 + 0\sqrt{D}$.
- The distributive laws carry over from the real numbers.

The multiplicative inverse of $a + b\sqrt{D}$ is

$$\frac{1}{a + b\sqrt{D}} = \frac{1}{a + b\sqrt{D}} \cdot \frac{a - b\sqrt{D}}{a - b\sqrt{D}} = \frac{a - b\sqrt{D}}{a^2 - b^2D} \in \mathbb{Z}\left[\sqrt{D}\right]$$

Without further restrictions on D, there is an issue with our definition of $\mathbb{Z}[\sqrt{D}]$, i.e., we don't have unique factorization. For example, in $\mathbb{Z}[\sqrt{5}]$

$$\left(1 + \sqrt{5}\right)\left(1 - \sqrt{5}\right) = 2 \cdot (-2)$$

The solution to this problem is to define $\mathbb{Z}[\omega]$ as the set $\{a + b\omega\}$ with $a, b \in \mathbb{Z}$ and

$$\omega = \begin{cases} \sqrt{D}, & D \equiv 2,3 \pmod 4 \\ \dfrac{1+\sqrt{D}}{2}, & D \equiv 1 \pmod 4 \end{cases}$$

The rationale for this definition is far from obvious. An explanation can be found in the discussion thread entitled "Why is quadratic integer ring defined in that way?" [68].

For the example above, note that $D = 5 \equiv 1 \pmod 4$ which falls into the second case. Thus, to ensure unique factorization, we should define the ring $\mathbb{Z}\left[\frac{1+\sqrt{5}}{2}\right]$.

To represent $(1+\sqrt{5})(1-\sqrt{5})$ in $\mathbb{Z}\left[\frac{1+\sqrt{5}}{2}\right]$, we need to solve

$$(1+\sqrt{5})(1-\sqrt{5}) = x + y\left(\frac{1+\sqrt{5}}{2}\right)$$

which implies the equations

$$x + \frac{y}{2} = -4; \quad \frac{y}{2} = 0$$

The only solution is $y = 0$ and $x = -4$. So, in $\mathbb{Z}\left[\frac{1+\sqrt{5}}{2}\right]$, $(1+\sqrt{5})(1-\sqrt{5})$ is uniquely represented as $-4 + 0 \cdot \left(\frac{1+\sqrt{5}}{2}\right)$.

. . .

If R is a ring, the following sets of matrices are also rings:

- The set of all $n \times n$ matrices over R, denoted $M_n(R)$. This is sometimes referred to as the "full ring of $n \times n$ matrices".
- The set of all upper triangular matrices over R.
- The set of all lower triangular matrices over R.
- The set of all diagonal matrices over R.

5.3.3 Ideals

An **ideal** is a subset of a ring with special properties. Ideals generalize certain subsets of the integers, such as the even numbers or the multiples of a positive integer n.

An ideal has two defining properties:

- Subgroup property: An ideal is closed under addition and subtraction, which by Theorem 33, implies the ideal is a subgroup of the additive group of the associated ring.
- Ideal property: An ideal is closed under multiplication by any element of the ring.

An ideal can be used to construct a quotient ring in a way similar to how, in group theory, a normal subgroup can be used to construct a quotient group (discussed in the next section).

More formally, a subset I of a ring R is called a left ideal of R if it satisfies the following conditions:

- I is an additive subgroup of R
- For every $x \in I$ and $r \in R$, $rx \in I$.

A right ideal of R is defined in a similar manner. If I is both a left and right ideal of a ring R, then we just say that I is a "two-sided ideal" or just "an ideal" of R. In a commutative ring, all ideals are two-sided.

Keep in mind that an ideal is not required to have a multiplicative identity and thus, is not necessarily a subring.

. . .

For a given ring R, the ring itself is an ideal (known as the **unit ideal**). The subset of R consisting of only the 0 element is an ideal (this follows from Theorem 45i). This is referred to as the **zero ideal**. An ideal of a ring R this is neither the unit nor zero ideal is referred to as a **proper ideal** of R.

Theorem 47. If a left-ideal contains a unit (i.e., invertible) element, then it cannot be a proper ideal.

A similar theorem holds for right ideals.

Proof: Let I be a left-ideal in a ring R. Assume that $u \in I$ is a unit, i.e., u^{-1} exists. Take any $r \in R$. Since R (as a ring) is closed under multiplication, we have that $ru^{-1} \in R$. Since I is a left ideal, $r = (ru^{-1})u \in I$. Thus, $I = R$ and I is not a proper ideal. ■

In a commutative ring, the set $(a) = \{ar : r \in R\}$ is an ideal in R. It is referred to as the **principal ideal** generated by a.

Principal ideals can also be defined for non-commutative rings, see the Wikipedia article entitle "Principal ideal" [73].

. . .

The even integers form a principal ideal (denoted $2\mathbb{Z}$) in the ring of all integers $\mathbb{Z}$, since

- $2\mathbb{Z}$ is an additive subgroup of $\mathbb{Z}$
- For every $x \in 2\mathbb{Z}$ and $r \in \mathbb{Z}$, we have an even number times an integer (even or odd) which results in another even number, and so, $xr \in I$. Thus, I is a left ideal of R. Similarly , I is a right ideal of R.

The set $n\mathbb{Z} = \{\dots, -3n, -2n, -n, 0, n, 2n, 3n, \dots\}$ is an additive subgroup of $\mathbb{Z}$. Take any $kn \in n\mathbb{Z}$ and any $r \in \mathbb{Z}$. We have that $(kn)r = (kr)n \in n\mathbb{Z}$ and $r(kn) = (kr)n \in n\mathbb{Z}$. Thus, $n\mathbb{Z}$ is a principal ideal of $\mathbb{Z}$ with generator n.

. . .

The set of all polynomials with real coefficients which are divisible by $X^2 + 1$ (with no remainder) is an ideal (called it I) in the ring of all real-coefficient polynomials $\mathbb{R}(X)$. If $f(X), g(X) \in I$, the sum $f(X) + g(X)$ is also divisible by $X^2 + 1$, and so, $f(X) + g(X) \in I$ (closure). The polynomial 0 is

divisible by $X^2 + 1$ and so, $0 \in I$. If $f(X) \in I$, then so is $-f(X) \in I$ (inverses). Associativity and commutativity in I are inherited from R. So, I is an abelian subgroup with R.

For any $f(X) \in I$ and $r(X) \in R$, $f(X)r(X)$ is divisible by $X^2 + 1$ which implies $f(X)r(X) \in I$ and so, I is a left ideal. Similarly, we can show that I is a right ideal.

. . .

Some additional examples of ideals. We leave the verification to the reader.

- The ring of continuous function f from $\mathbb{R}$ to $\mathbb{R}$ under pointwise multiplication [69] such that $f(1) = 0$ is an ideal.
- Another ideal in the ring of continuous function f from $\mathbb{R}$ to $\mathbb{R}$ is the set of function that vanish for sufficiently large values of x, i.e., continuous functions f for which there exists a value L such that $f(x) = 0$ for all $|x| > L$.
- Consider the ring $M_n(R)$ of all $n \times n$ matrices over any ring R. The subset I_{row_i} comprised of all matrices in $M_n(R)$ whose i^{th} row consists of zeros is a right ideal of $M_n(R)$. Note that for any $A \in I_{row\_i}$ and $M \in M_n(R)$, the product AM has all zeros in row i and thus, $AM \in I_{row\_i}$. However, I_{row_i} is not a left ideal. On the other hand, the subset I_{col_j} comprised of all matrices in $M_n(R)$ whose j^{th} column consists of zeros is a left ideal of $M_n(R)$ but not a right ideal.

For even more examples of ideals, see "Ideal (ring theory)" [70].

5.3.4 Quotient Rings

A **quotient ring**, also known as factor ring, is a structure similar to a quotient group. Given a ring R and a two-sided ideal $I \subseteq R$, we can construct a quotient ring, denoted R/I, whose elements are the cosets of I in R, where the cosets of I in R are defined in a similar manner to the cosets of a subgroup within a group. Given any $r \in R$, $r + I = \{r + x : x \in I\}$ is a coset of I in R. The set of all cosets of I in R comprise the quotient ring of I with respect to the ring R.

A quotient ring is a ring that is constructed from another ring, called the parent ring, by dividing out by an ideal. Formally, the elements of R/I are the equivalence classes of R under the equivalence relation $\sim$ defined by $x \sim y$ if and only if $x + (-y) \in I$. The addition and multiplication operations on R/I are defined by

- $(x + I) + (y + I) = (x + y) + I$
- $(x + I)(y + I) = (xy) + I$

It can be shown that these operations are well-defined, and that R/I is a ring under these operations.

The key to understanding a particular quotient ring is to identify the cosets, i.e., the equivalence classes.

. . .

Given a ring R, the quotient ring $R/\{0\}$ is equivalent to R, since each coset $r + \{0\} = \{r + 0 : 0 \in I\}$ corresponds to an element of R. On the other hand, the quotient ring R/R has but one coset, i.e., $r + R = \{r + x : x \in R\}$ which is equivalent to all of R, regardless of the r we select. This one coset is basically the 0 element of R/R. In general, the larger the ideal I, the smaller the quotient ring R/I. If I is a proper ideal of R, i.e., $I \neq R$, then R/I is not the zero ring.

. . .

Consider the ring of integers $\mathbb{Z}$ and the ideal of even numbers, i.e., $2\mathbb{Z}$. The quotient ring $\mathbb{Z}/2\mathbb{Z}$ has only two elements, the coset $0 + 2\mathbb{Z}$ consisting of the even numbers and the coset $1 + 2\mathbb{Z}$ consisting of the odd numbers. $\mathbb{Z}/2\mathbb{Z}$ is isomorphic $\mathbb{Z}_2$ with addition and multiplication modulo 2. In fact, many text books will write $\mathbb{Z}/2\mathbb{Z}$ rather than $\mathbb{Z}_2$ (even before they introduce quotient rings).

. . .

Consider the polynomial ring $\mathbb{R}(X)$ and the ideal $\mathbb{R}(X)/I$, where I is the principal ideal generated by $(X^2 + 1)$, i.e., $I = \{a(X) \cdot (X^2 + 1) : a(X) \in \mathbb{R}(X)\}$. A coset of $\mathbb{R}(X)/I$ is of the form

$$g(X) + I = \{g(X) + a(X) \cdot (X^2 + 1)\}$$

The zero element of R/I is $X^2 + 1$. So, the coset $(X^2 + 1) + I$ consists of all multiples of $(X^2 + 1)$. Further, we have that $X^2 + 1 = 0$ with respect to the quotient ring, which implies $X^2 = -1$ or $X = \sqrt{-1} = i$ (with some abuse of notation).

Using the **division algorithm** for polynomials (also known as "polynomial long division") [71], any $g(X) \in \mathbb{R}(X)$ can be written in the form

$$g(X) = q(X) \cdot (X^2 + 1) + r(x)$$

where the remainder $r(X)$ is linear, i.e., of the form $aX + b$ for constants $a, b \in \mathbb{R}$, and $q(X) \in \mathbb{R}(X)$.

So, an arbitrary element of $\mathbb{R}(X)/I$ is of the form $aX + b$. The operations of the quotient ring are exactly the same as those for the complex numbers $\mathbb{C}$. More formally, $\mathbb{R}(X)/(X^2 + 1)$ is isomorphic to $\mathbb{C}$ under the mapping $f(aX + b) = ai + b$. To prove this using the definition of isomorphism, take two elements in $\mathbb{R}(X)/I$, i.e., $aX + b$ and $cX + d$.

For multiplication, we have $f\big((aX + b)(cX + d)\big) = f(acX^2 + (ad + bc)X + bd)$

and since $X^2 = -1$, the above reduces to

$$f\big((ad + bc)X - (bd - ac)\big) = (ad + bc)i + (bd - ac)$$

Further, we have that $f(aX + b)f(cX + d) = (ai + b)(ci + d) = (ad + bc)i + (bd - ac)$

and so, $f\big((aX + b)(cX + d)\big) = f(aX + b)f(cX + d)$.

For addition, we have $f\big((aX + b)(cX + d)\big) = f\big((a + c)X + (b + d)\big) = (a + c)i + (b + d)$ and $f(aX + b) + f(cX + d) = (ai + b) + (ci + d) = (a + c)i + (b + d)$.

Thus, $\mathbb{R}(X)/(X^2 + 1)$ and $\mathbb{C}$ are isomorphic rings.

. . .

In a generalization of the previous example, let R be any commutative ring and let $R(X)$ be the polynomial ring defined over R. Take an n^{th} degree polynomial in $R(X)$, i.e.,

$$f(X) = a_n X^n + \cdots + a_1 X + a_0$$

and form the principal ideal generated by $f(X)$, i.e.,

$$I = \{a(X)f(X) : a(X) \in R(X)\}$$

A coset of the quotient group $R(X)/I$ is of the form

$$g(X) + I = \{g(X) + a(X)f(X) : a(X) \in R(X)\}$$

However, by the division algorithm for polynomials, we have that any $g(x) \in R(x)$ can be written in the form

$$g(X) = q(X)f(X) + r(X)$$

where $\deg\big(r(x)\big) < \deg\big(f(x)\big)$, i.e., $r(X)$ is of the form $b_{n-1}X^{n-1} + \cdots + b_1 X + b_0$.

So, the cosets of $R(X)/I$ are polynomials of the form

$$g(X) + I = \{g(X) + a(X)f(X) : a(X) \in R(X)\}$$

$$= \{q(X)f(X) + r(X) + a(X)f(x) : a(X) \in R(X)\}$$

$$= \{r(X) + \big(q(X) + a(X)\big)f(X)\}$$

$$= r(X) + I$$

with $r(X)$ of the form noted above, i.e., polynomial over R of degree $n - 1$ or less.

5.4 Fields

5.4.1 Definitions and Basic Concepts

A **field** is a commutative ring (under both addition and multiplication) such that $0 \neq 1$ and all nonzero elements are invertible under multiplication. The Wikipedia article on fields offers the following (equivalent) definitions of a field [72]:

> A field has two operations, called addition and multiplication; it is an abelian group under addition with 0 as the additive identity; the nonzero elements are an abelian group under multiplication with 1 as the multiplicative identity; and multiplication distributes over addition.
>
> Even more summarized: a field is a commutative ring where $0 \neq 1$ and all nonzero elements are invertible under multiplication.

The following theorem relates rings and ideals to fields.

Theorem 48. A nonzero commutative ring R is a field if and only if its only ideals are $\{0\}$ and the ring itself.

Proof: Assume that R is a field. If $I \neq \{0\}$, it contains some nonzero element, and every nonzero element in a field is a unit. It follows by Theorem 47 that $I = R$.

Going the other way, assume that the only ideals of the ring R are $\{0\}$ and the ring itself. Since we are given that R is nonzero, it contains an element $a \neq 0$. The principal domain (a) must equal R, since we assumed the only ideals of R are $\{0\}$ and R, and $a \neq 0$. Since $R = (a)$, there exists an $r \in R$ such that $1 = ra$ which implies that a (which we picked arbitrarily) has an inverse. Thus, R is a field. ■

. . .

The rational numbers $\mathbb{Q}$, real numbers $\mathbb{R}$ and complex numbers $\mathbb{C}$ are examples of fields with an infinite number of elements.

The set of integers modulo p is a field if and only p is a prime number. In this case, we denote the field $\mathbb{F}_p$ or $GF(p)$ where GF stands for "Galois field."

The number of elements of a finite field is known as its order.

For two fields to be homomorphic (or isomorphic), their constituent groups need to be homomorphic (or isomorphic).

Finite fields (also called **Galois fields**) are fields with finitely many elements.

Theorem 49. The order of a finite field (i.e., its number of elements) is either a prime number or a prime power. For every prime number p and every positive integer k there are fields of order p^k, all of which are isomorphic.

Proof: See the Existence and Uniqueness section of the Wikipedia article on Finite Fields [74]. ■

5.4.2 Construction of Non-prime Finite Fields

As noted, finite fields are either of prime order p (isomorphic to $\mathbb{F}_p$) or of prime power order p^n. In this section we describe how to construct fields of prime power order.

Given a prime power $q = p^n$ with p prime and $n > 1$, the field $\mathbb{F}_q$ (also written as $GF(q)$) may be constructed as follows:

1. Choose an irreducible polynomial $P(X)$ in $\mathbb{F}_p(X)$ of degree n (such an irreducible polynomial always exists). By "irreducible polynomial", we mean a polynomial that has no roots in the given field $\mathbb{F}_p(X)$.
2. Construction the quotient ring $\mathbb{F}_p(X)/(P(X))$. The quotient ring is the field of order q that we seek, i.e., $\mathbb{F}_q \cong \mathbb{F}_p(X)/(P(X))$.
    a. $(P(X))$ is an ideal that consists of all multiples of $P(X)$, i.e., $\big(P(X)\big) = \{f(X) \cdot P(X) : f(X) \in \mathbb{F}_p(x)\}$.
    b. $P(X)$ is the additive zero of $\mathbb{F}_p(X)/(P(X))$, since the remainder when dividing $P(X)$ by itself is 0.

c. All multiples of $P(X)$ are in the same equivalence class as $P(X)$.

d. Elements of $\mathbb{F}_p(X)/(P(X))$ are of the form $f(X) \cdot P(X) + r(X)$ where $r(X), F(X) \in \mathbb{F}_p(X)$ and $\deg\big(r(X)\big) < \deg\big((PX)\big) = n$. Basically, one takes an element of $\mathbb{F}_p(X)$ divides by $P(X)$ and gets the remainder $r(X)$.

The elements of $\mathbb{F}_q$ are the polynomials over $\mathbb{F}_p(X)$ whose degree is strictly less than n.

- Addition and subtraction in $\mathbb{F}_q$ is defined by addition and subtraction of polynomials in $\mathbb{F}_p(X)$.
- Multiplication of two elements in $\mathbb{F}_q$ entails the product of the two elements and then taking the remainder upon long division by $P(X)$.
- The multiplicative inverse of a non-zero element may be computed with the extended Euclidean algorithm.

. . .

As an example, we construct $\mathbb{F}_4$. With respect to the above procedure, we have that $q = 2^2$ and so, $n = 2$. In $\mathbb{F}_2(X)$, there is only one irreducible polynomial of degree 2, i.e., $P(X) = X^2 + X + 1$. In $\mathbb{F}_2$, we have only two elements, i.e., $\bar{0}$ and $\bar{1}$, and so, $P(\bar{0}) = \bar{1}$ and $P(\bar{1}) = \bar{1}$ (noting that $3 \equiv 1 \pmod 2$). Thus, $P(X)$ is irreducible.

[**Note**: To prove that $P(X) = X^2 + X + 1$ is the only irreducible polynomial in $\mathbb{F}_2(X)$, one needs to consider all the polynomials of the form $aX^2 + bX + c$ where $a, b, c \in \mathbb{F}_2$ (a total of 8 cases) and then test to see which polynomial has no roots in $\mathbb{F}_2$.]

Next, we determine the elements in $\mathbb{F}_4 = \mathbb{F}_2(X)/(X^2 + X + 1)$ each of which is a remainder of a polynomial in $\mathbb{F}_2(X)$ when divided by $X^2 + X + 1$. This means the elements of $\mathbb{F}_2(X)/(X^2 + X + 1)$ are of the form $aX + b$, where $a, b \in \mathbb{F}_2$. Since $a, b \in \{\bar{0}, \bar{1}\}$, there are only four possible outcomes for $aX + b$. Working with the cosets of $\mathbb{F}_2(X)/(X^2 + X + 1)$, we have

- $\bar{0} + (X^2 + X + 1) \in \mathbb{F}_2(X)/(X^2 + X + 1)$ which is the additive identity of $\mathbb{F}_4$. We will call this element 0.
- $\bar{1} + (X^2 + X + 1) \in \mathbb{F}_2(X)/(X^2 + X + 1)$ which is the multiplicative identity of $\mathbb{F}_4$. We will call this element 1.
- $X + (X^2 + X + 1) \in \mathbb{F}_2(X)/(X^2 + X + 1)$ which we will call $\alpha \in \mathbb{F}_4$.
- $(X + 1) + (X^2 + X + 1) \in \mathbb{F}_2(X)/(X^2 + X + 1)$ which we will call $\alpha + 1 \in \mathbb{F}_4$.

The addition table for $\mathbb{F}_4$ is shown in Table 5. Keep in mind that we are working with polynomials of the form $a + b\alpha$ where $a, b \in \mathbb{F}_2$ (integers modulo 2). So, for example, $(1 + \alpha) + (1 + \alpha) = 2 + 2\alpha = 0$ since $2 \equiv 0 \pmod 2$.

Table 5. Addition table for the finite field of order 4

$+$	$\mathbf{0}$	$\mathbf{1}$	$\boldsymbol{\alpha}$	$\mathbf{1} + \boldsymbol{\alpha}$
$\mathbf{0}$	0	1	α	$1 + \alpha$
$\mathbf{1}$	1	0	$1 + \alpha$	α
$\boldsymbol{\alpha}$	α	$1 + \alpha$	0	1
$\mathbf{1} + \boldsymbol{\alpha}$	$1 + \alpha$	α	1	0

We need one other fact to compute the multiplicative group tables for $\mathbb{F}_4$. Since $X^2 + X + 1$ is the 0 elements, we have $X^2 = -X - 1 = X + 1$ since $1 = -1$ in $\mathbb{F}_2$. So, within the quotient ring $\mathbb{F}_2(X)/(X^2 + X + 1)$, $\alpha^2 = X^2 = X + 1 = \alpha + 1$.

The multiplication table for $\mathbb{F}_4$ is shown in Table 6. For example, $(1 + \alpha)(1 + \alpha) = 1 + 2\alpha + \alpha^2 = 1 + 0 + \alpha^2 = 1 + (1 + \alpha) = 2 + \alpha = \alpha$.

Table 6. Multiplication table for the finite field of order 4

$\times$	$\mathbf{0}$	$\mathbf{1}$	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha + 1}$
$\mathbf{0}$	0	0	0	0
$\mathbf{1}$	0	1	α	$\alpha + 1$
$\boldsymbol{\alpha}$	0	α	$\alpha + 1$	1
$\boldsymbol{\alpha + 1}$	0	$\alpha + 1$	1	α

. . .

As a second example, we construct $\mathbb{F}_8$. In this case, we have that $q = 2^3$ and so, $n = 2$ again. In $\mathbb{F}_2(X)$, there several irreducible polynomials of degree 3, e.g., $P(X) = X^3 - X - 1$ and $Q(X) = X^3 + X^2 + 1$. We will use $P(x)$ in the following calculations.

Now, let us determine the elements in $\mathbb{F}_8 = \mathbb{F}_2(X)/(X^3 - X - 1)$ each of which is a remainder of a polynomial in $\mathbb{F}_2(X)$ when divided by $X^3 - X - 1$. Thus, elements of $\mathbb{F}_2(X)/(X^2 + X + 1)$ are of the form $aX^2 + bX + c$, where $a, b, c \in \mathbb{F}_2$. Since $a, b, c \in \{\bar{0}, \bar{1}\}$, there are eight possible outcomes for $aX^2 + bX + c$. Letting $\alpha = X$, we have the following cosets of $\mathbb{F}_2(X)/(X^3 - X - 1)$:

$$0, 1, \alpha, \alpha + 1, \alpha^2, \alpha^2 + 1, \alpha^2 + \alpha, \alpha^2 + \alpha + 1$$

[**Author's Remark**: Letting $X = \alpha$ appears to be a convention in some textbooks and online articles. This is not necessary, but I'll go along with the de facto convention.]

The addition table for $\mathbb{F}_8$ is shown in Table 7. Note that $\mathbb{F}_4$ is embedded in the first 4 rows and columns.

Table 7. Addition table for the finite field of order 8

$+$	$\mathbf{0}$	$\mathbf{1}$	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha + 1}$	$\boldsymbol{\alpha^2}$	$\boldsymbol{\alpha^2 + 1}$	$\boldsymbol{\alpha^2 + \alpha}$	$\boldsymbol{\alpha^2 + \alpha + 1}$
$\mathbf{0}$	0	1	α	$\alpha + 1$	0	0	0	0
$\mathbf{1}$	1	0	$\alpha + 1$	α	α^2	$\alpha^2 + 1$	$\alpha^2 + \alpha$	$\alpha^2 + \alpha + 1$
$\boldsymbol{\alpha}$	α	$\alpha + 1$	0	1	$\alpha^2 + \alpha$	$\alpha^2 + \alpha + 1$	α^2	$\alpha^2 + 1$
$\boldsymbol{\alpha + 1}$	$\alpha + 1$	α	1	0	$\alpha^2 + \alpha + 1$	$\alpha^2 + \alpha$	$\alpha^2 + 1$	α^2
$\boldsymbol{\alpha^2}$	0	α^2	$\alpha^2 + \alpha$	$\alpha^2 + \alpha + 1$	0	1	α	$\alpha + 1$
$\boldsymbol{\alpha^2 + 1}$	0	$\alpha^2 + 1$	$\alpha^2 + \alpha + 1$	$\alpha^2 + \alpha$	1	0	$\alpha + 1$	α
$\boldsymbol{\alpha^2 + \alpha}$	0	$\alpha^2 + \alpha + 1$	α^2	$\alpha^2 + 1$	α	$\alpha + 1$	0	1
$\boldsymbol{\alpha^2 + \alpha + 1}$	0	$\alpha^2 + \alpha$	$\alpha^2 + 1$	α^2	$\alpha + 1$	α	1	0

Since $X^3 - X - 1$ is the 0 element, we have $X^3 = X + 1$. So, within the quotient ring $\mathbb{F}_2(X)/(X^3 - X - 1)$, $\alpha^3 = X^3 = X + 1 = \alpha + 1$. With this information, we can compute the powers of α.

- $\alpha^3 = \alpha + 1$

- $\alpha^4 = \alpha^3\alpha = (\alpha + 1)\alpha = \alpha^2 + \alpha$
- $\alpha^5 = \alpha^4\alpha = (\alpha^2 + \alpha)\alpha = \alpha^3 + \alpha^2 = \alpha^2 + \alpha + 1$
- $\alpha^6 = \alpha^3\alpha^3 = (\alpha + 1)^2 = \alpha^2 + 2\alpha + 1 = \alpha^2 + 1$, since $2 \equiv 0 \pmod 2$
- $\alpha^7 = \alpha^3\alpha^4 = (\alpha + 1)(\alpha^2 + \alpha) = \alpha^3 + 2\alpha^2 + \alpha = \alpha^3 + \alpha = (\alpha + 1) + \alpha = 1$

So, the powers of α generate $\mathbb{F}_8$, i.e., $\mathbb{F}_8 = \langle\alpha\rangle$ is a cyclic group with respect to multiplication.

The multiplication table for $\mathbb{F}_8$ is shown in Table 8.

Table 8. Multiplication table for the finite field of order 8

$\times$	$\mathbf{0}$	$\mathbf{1}$	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha+1}$	$\boldsymbol{\alpha^2}$	$\boldsymbol{\alpha^2+1}$	$\boldsymbol{\alpha^2+\alpha}$	$\boldsymbol{\alpha^2+\alpha+1}$
$\mathbf{0}$	0	0	0	0	0	0	0	0
$\mathbf{1}$	0	1	α	$\alpha+1$	α^2	α^2+1	$\alpha^2+\alpha$	$\alpha^2+\alpha+1$
$\boldsymbol{\alpha}$	0	α	α^2	$\alpha^2+\alpha$	$\alpha+1$	1	$\alpha^2+\alpha+1$	α^2+1
$\boldsymbol{\alpha+1}$	0	$\alpha+1$	$\alpha^2+\alpha$	α^2+1	$\alpha^2+\alpha+1$	α^2	1	α
$\boldsymbol{\alpha^2}$	0	α^2	$\alpha+1$	$\alpha^2+\alpha+1$	$\alpha^2+\alpha$	α	α^2+1	1
$\boldsymbol{\alpha^2+1}$	0	α^2+1	1	α^2	α	$\alpha^2+\alpha+1$	$\alpha+1$	$\alpha^2+\alpha$
$\boldsymbol{\alpha^2+\alpha}$	0	$\alpha^2+\alpha$	$\alpha^2+\alpha+1$	1	α^2+1	$\alpha+1$	α	α^2
$\boldsymbol{\alpha^2+\alpha+1}$	0	$\alpha^2+\alpha+1$	α^2+1	α	1	$\alpha^2+\alpha$	α^2	$\alpha+1$

In general, we have the following result.

Theorem 50. The multiplicative group of all nonzero elements of a finite field is cyclic.

Proof: See Theorem 22.10 and Corollary 22.11 in Section 22.1 of Judson [40]. ■

5.5 Vector Spaces

5.5.1 Definition

A **vector space** over a field F is a non-empty set V together with two binary operations defined below. The elements of V are called vectors, and the elements of F are called scalars. Vectors are usually represented either in bold (e.g., $\boldsymbol{v}$) or with an arrow above, e.g., $\vec{v}$.

- One of the binary operations is vector addition. This operation assigns to any two vectors $\boldsymbol{u}, \boldsymbol{v} \in V$ a third vector in V which is written as $\boldsymbol{u} + \boldsymbol{v}$, and called the sum of these two vectors.
- The other operation is scalar multiplication. This operation assigns to any scalar $a \in F$ and any vector $\boldsymbol{v} \in V$ another vector in V, denoted $a\boldsymbol{v}$.

In addition to the two closure axioms above, a vector space must satisfy the following axioms.

- The set V must be an abelian group under the vector addition operation. The group identity is represented by the symbol $\mathbf{0}$. [**Caution**: the field F also has an additive identity which is represented by 0 (not bold).]
- Compatibility of scalar multiplication with field multiplication, i.e., $a(b\boldsymbol{v}) = (ab)\boldsymbol{v}$ for $a, b \in F$ and $\boldsymbol{v} \in V$.
- Identity element of scalar multiplication, i.e., there exists $1 \in F$ such that $1\boldsymbol{v} = \boldsymbol{v}$ for every $\boldsymbol{v} \in V$ where 1 is the multiplicative identity in the field F.
- Distributivity of scalar multiplication with respect to vector addition, i.e., $a(\boldsymbol{u} + \boldsymbol{v}) = a\boldsymbol{u} + a\boldsymbol{v}$ for $a, b \in F$ and $\boldsymbol{u}, \boldsymbol{v} \in V$.
- Distributivity of scalar multiplication with respect to field addition, i.e., $(a + b)\boldsymbol{v} = a\boldsymbol{v} + b\boldsymbol{v}$ for $a, b \in F$ and $\boldsymbol{v} \in V$.

Some basic results for vector spaces are presented in the following theorem.

Theorem 51. The following relationships hold true:

i. $0\boldsymbol{v} = \mathbf{0}$

ii. $a\mathbf{0} = \mathbf{0}$

iii. $(-1)\boldsymbol{v} = -\boldsymbol{v}$

iv. $a\boldsymbol{v} = \mathbf{0}$ implies either $a = 0$ or $\boldsymbol{v} = \mathbf{0}$.

Proof:

i. Using the distributive rules for vector spaces, we have

$$0\boldsymbol{v} = (0 + 0)\boldsymbol{v} = 0\boldsymbol{v} + 0\boldsymbol{v}$$

Adding the additive inverse of $0\boldsymbol{v}$ to both sides of the above equation, we get

$$\mathbf{0} = 0\boldsymbol{v} - 0\boldsymbol{v} = (0\boldsymbol{v} + 0\boldsymbol{v}) - 0\boldsymbol{v} = 0\boldsymbol{v}$$

ii. Using the distributive rules for vector spaces, we have

$$a\mathbf{0} = a(\mathbf{0} + \mathbf{0}) = a\mathbf{0} + a\mathbf{0}$$

Adding the additive inverse of $a\mathbf{0}$ to both sides of the above equation, we get

$$\mathbf{0} = a\mathbf{0} - a\mathbf{0} = (a\mathbf{0} + a\mathbf{0}) - a\mathbf{0} = a\mathbf{0}$$

iii. For any $\boldsymbol{v} \in V$, we have

$$\boldsymbol{v} + (-1)\boldsymbol{v} = 1\boldsymbol{v} + (-1)\boldsymbol{v} = (1 + (-1))\boldsymbol{v} = 0\boldsymbol{v} = \mathbf{0}$$

and thus, $(-1)\boldsymbol{v} = -\boldsymbol{v}$.

iv. If $a = 0$, we are done. If $a \neq 0$ then a (being an non-zero element of a field) has an inverse, i.e., a^{-1} exists. Multiplying both sides of $a\boldsymbol{v} = \mathbf{0}$ by a^{-1} and using part ii of this theorem, we have that $aa^{-1}\boldsymbol{v} = 1\boldsymbol{v} = \boldsymbol{v} = a^{-1}\mathbf{0} = \mathbf{0}$. ■

5.5.2 Examples

The following are examples of vector spaces:

- The set of all real numbers $\mathbb{R}$ over the real numbers $\mathbb{R}$ is a vector space under the usual addition and multiplication of real numbers. In this case, $V = F = \mathbb{R}$.
- The set of all ordered pairs of real numbers is a vector space over the field of real numbers. Addition is pairwise, i.e., if $(x, y), (s, t) \in V$ then $(x, y) + (s, t) = (x + s, y + t)$. Scalar multiplication is as follows: for $(x, y) \in V$ and $a \in \mathbb{R}$, $a(x, y) = (ax, ay)$.
- The set of ordered n-tuples of real numbers is a vector space over the field of real numbers. Addition and scalar multiplication are defined in a similar manner to the previous case.
- The set of all polynomials with real coefficients is a vector space over the field of real numbers. Vector addition is accomplished by adding coefficients of like terms (i.e., terms of the same power) and scalar multiplication entails the multiplication of the given scalar times each coefficient in the given polynomial.
- The set of all matrices with real coefficients over the real numbers is a vector space. Standard matrix addition is used here, and multiplication of a matrix by a scalar results in every element of the given matrix being multiplied by the scalar.
- The set of all continuous functions from the real numbers to the real numbers is a vector space under the usual addition of functions and the pointwise multiplication of functions by real numbers. The associated field is $\mathbb{R}$.
- The set of all solutions to a system of linear equations with real coefficients is a vector space over the field of real numbers.

5.5.3 Concepts

In what follows (unless stated otherwise), we assume V is a vector space over field F.

A **linear combination of vectors** results in a new vector that is created by adding together scalar multiples of a set of vectors. For example, take vectors $\boldsymbol{u}, \boldsymbol{v} \in V$ and scalars $a, b \in F$ where V is a vector space over the field F. The vector $a\boldsymbol{u} + b\boldsymbol{v} \in V$ is a linear combination of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Linear combinations can be created for any number of vectors.

A set of vectors is **linearly independent** if no vector in the set can be written as a linear combination of the other vectors in the set. Equivalent definitions:

- A set of vectors is linearly independent if a linear combination results in the zero vector if and only if all its coefficients (i.e., all the scalars in the linear combination) are zero.
- A set of vectors is linearly independent if two linear combinations of the set define the same element of V if and only if they have the same coefficients.

For example, in 3-dimensional Euclidean space $\mathbb{R}^3$, the vectors $(1,0,0), (0,1,0), (0,0,1)$ are linearly independent since there is no way of writing any one of the vectors as a linear combination of the other two.

A set of vectors that is not linearly independent is said to be **linearly dependent**. For example, in 3-dimensional Euclidean space $\mathbb{R}^3$, the vectors $(1,1,0), (0,1,1), (2,-1,-3)$ are linearly dependent since we can write $(2,-1,-3) = 2 \cdot (1,1,0) - 3 \cdot (0,1,1)$.

A **linear subspace of a vector space** is a subset of the vector space that is also a vector space. In other words, a linear subspace is a set of vectors that is closed under vector addition and scalar multiplication. For example, 2-dimensional Euclidean space $\mathbb{R}^2$ is a linear subspace of $\mathbb{R}^3$.

Theorem 52. A subset U of a vector space V over a field F is a subspace of V if and only if $a\boldsymbol{x} + b\boldsymbol{y} \in \boldsymbol{U}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{U}$ and all $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{F}$.

Proof: If U is a subspace (and thus a vector space in its own right), then closure under vector addition and scalar multiplication implies $a\boldsymbol{x} + b\boldsymbol{y} \in U$ for all x, y $\in U$ and all $\mathbf{a}, \mathbf{b} \in$ F.

If $a\boldsymbol{x} + b\boldsymbol{y} \in$ U for all $\boldsymbol{x}, \boldsymbol{y} \in U$ and all $a, b \in F$, then

- $\boldsymbol{x} + \boldsymbol{y} = 1\boldsymbol{x} + 1\boldsymbol{y} \in U$ for every $\boldsymbol{x}, \boldsymbol{y} \in U$ (closure under vector addition)
- $0 = 0\boldsymbol{x} + 0\boldsymbol{y} \in U$
- For $\boldsymbol{x} \in U$, we have that $-\boldsymbol{x} = (-1)\boldsymbol{x} \in U$
- The additive associative and commutative laws are inherited by U from V.
- The above points prove that U is an abelian subgroup of V.
- $a\boldsymbol{x} \in U$ for every x $\in U$ and $a \in F$ (closure under scalar multiplication)
- The other properties for a vector space (i.e., compatibility of scalar multiplication and the two distributive rules) are inherited by U from V.

Thus, U is a subspace of V. ■

The **linear span** of a set of vectors is the smallest linear subspace that contains the set of vectors. In other words, the linear span of a set of vectors is the set of all vectors that can be created as linear combinations of the vectors in the set.

More formally, the **linear span** (or simply **span**) of vector $v_1, v_2, \dots, v_n \in V$ is defined as

$$span(\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_n}) = \{a_1\boldsymbol{v_1} + a_2\boldsymbol{v_2} + \cdots + a_n\boldsymbol{v_n} \mid a_1, a_2 \dots, a_n \in F\}$$

A **basis of a vector space** is a set of vectors that spans the vector space and is linearly independent. In other words, a basis for a vector space is a set of vectors that can be used to create any vector in the space, and no two vectors in the set are linearly dependent.

For example, the vectors $(1,0,0), (0,1,0), (0,0,1)$ form a basis for $\mathbb{R}^3$. There are an infinite number of bases for $\mathbb{R}^3$.

In general, there are many bases for any vector space. However, it can be shown that the number of vectors in a basis for a given vector space is the same for all bases.

The **dimension of a vector space** is the number of vectors in a basis for the space. It is possible to have infinite dimensional vector spaces. For example, the set of all polynomials with real coefficients, the set of all continuous functions on the real line, and the set of all sequences of real numbers are infinite dimensional vector spaces.

The above concepts are tightly related as evidenced by the following theorems.

Theorem 53. Let V be a vector space with $\boldsymbol{v_1, v_2, \dots, v_n} \in V$, then the following holds true

i. $\boldsymbol{v_j} \in span(\boldsymbol{v_1, v_2, \dots, v_n})$ for $j = 1,2, \dots, n$

ii. $span(\boldsymbol{v_1, v_2, \dots, v_n})$ is a subspace of V

iii. If $U \subset V$ is a subspace such that $\boldsymbol{v_1, v_2, \dots, v_n} \in U$, then $span(\boldsymbol{v_1, v_2, \dots, v_n}) \subset U$.

Proof:

i. $\boldsymbol{v_j} = 0\boldsymbol{v_1} + 0\boldsymbol{v_2} + \cdots + 1\boldsymbol{v_j} + \cdots + 0\boldsymbol{v_n}$

ii. By the definition of linear span, $span(\boldsymbol{v_1, v_2, \dots, v_n})$ is closed under vector addition and scalar multiplication and thus, is a subspace of V (by Theorem 52).

iii. Since U is a subspace of V, it is closed under vector addition and scalar multiplication which implies $span(\boldsymbol{v_1, v_2, \dots, v_n}) \subset U$. ■

Theorem 54. If $V = \boldsymbol{span}(v_1, v_2, \dots, v_n)$, then either $\boldsymbol{S} = \{v_1, v_2, \dots, v_n\}$ is a basis for V or some subset of S is a basis for V.

Proof: If some $\boldsymbol{v} \in S$ can be written as a linear combination of the other vectors in S, then $S - \{\boldsymbol{v}\}$ spans V. If some $\boldsymbol{u} \in S$ can be written as a linear combination of the other vectors in $S - \{\boldsymbol{v}\}$, then $S - \{\boldsymbol{v}, \boldsymbol{u}\}$ spans V. We continue in this manner until the remaining vectors are linearly independent while still spanning V (i.e., until we have found a basis). ■

Theorem 55. Every finite-dimensional vector space has a basis.

Proof: This follows from Theorem 55 if we let S be all the vectors in V, and then pare down the set until we eventually get a basis. ■

Theorem 56. If V is a finite-dimensional vector space, then any two bases of V have the same number of elements.

Proof: See Theorem 5.4.2 in the book "Linear Algebra" [75].

. . .

Take vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in vector space V over a field F with basis $\boldsymbol{v_1, v_2, \dots, v_n}$. We can uniquely represent $\boldsymbol{x}$ and $\boldsymbol{y}$ in terms of the basis vectors as follows:

$$\boldsymbol{x} = x_1\boldsymbol{v_1} + x_2\boldsymbol{v_2} + \cdots + x_n\boldsymbol{v_n}, \quad x_1, x_2, \dots, x_n \in F$$

$$\boldsymbol{y} = y_1\boldsymbol{v_1} + y_2\boldsymbol{v_2} + \cdots + y_n\boldsymbol{v_n}, \quad y_1, y_2, \dots, y_n \in F$$

The scalars (i.e., the x_i and y_i terms in the above equations) are called the coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$ over the given basis. They are also said to be the coefficients of the decomposition of $\boldsymbol{x}$ and $\boldsymbol{y}$ over the given of the basis. We can also write $\boldsymbol{x}$ and $\boldsymbol{y}$ in terms of their coordinates as follows:

$$\boldsymbol{x} = (x_1, x_2, \dots, x_n)$$

$$\boldsymbol{v} = (y_1, y_2, \dots, y_n)$$

A **coordinate vector** is a representation of a vector as an ordered list of numbers (a tuple) that describes the vector in terms of a particular ordered basis.

The set of n-tuples with elements from F is a vector space (denoted F^n) with addition and scalar multiplication defined component-wise, i.e.,

$$\boldsymbol{x} + \boldsymbol{y} = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)$$

$$a\boldsymbol{x} = (ax_1, ax_2, \dots, ax_n), \qquad a \in F$$

. . .

Take vectors $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$, the **inner product** of $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$\boldsymbol{x} \cdot \boldsymbol{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

For example, take vectors $\boldsymbol{x} = (1,2,0,1,1)$ and $\boldsymbol{y} = (1,1,1,0,0)$ where the elements of $\boldsymbol{x}$ and $\boldsymbol{y}$ come from $\mathbb{F}_3$, then

$$\boldsymbol{x} \cdot \boldsymbol{y} = 1 + 2 + 0 + 0 + 0 = 0$$

If $\boldsymbol{x} \cdot \boldsymbol{y} = 0$, then $\boldsymbol{x}$ and $\boldsymbol{y}$ are said to be **orthogonal vectors**.

When a vector space is defined over a finite field, it is possible for a vector to be orthogonal to itself. For example, if $\boldsymbol{x} = (1,1,1,1)$ over $\mathbb{F}_2$, then $\boldsymbol{x} \cdot \boldsymbol{x} = 0$. For vectors defined over infinite fields such as $\mathbb{R}^3$ (3-dimensional Euclidean space), a non-zero vector cannot be orthogonal to itself.

Theorem 57. For any vectors $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ in an n-dimensional vector space $\boldsymbol{V}$ over a field $\boldsymbol{F}$, and scalar $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{F}$, the following holds true

i. $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{y} \cdot \boldsymbol{x}$

ii. $(a\boldsymbol{x} + b\boldsymbol{y}) \cdot \boldsymbol{z} = a(\boldsymbol{x} \cdot \boldsymbol{z}) + b(\boldsymbol{y} \cdot \boldsymbol{z})$

Proof: The proof of part i follow directly from the definition of inner product.

For part ii, let $\boldsymbol{x} = (x_1, x_2, \dots, x_n), \boldsymbol{y} = (y_1, y_2, \dots, y_n)$ and $\boldsymbol{z} = (z_1, z_2, \dots, z_n)$. We then have

$$(a\boldsymbol{x} + b\boldsymbol{y}) \cdot \boldsymbol{z} = (ax_1 + by_1, \dots, ax_n + by_n) \cdot (z_1, z_2, \dots, z_n)$$

$$= \big((ax_1 + by_1)z_1, \dots, (ax_n + by_n)z_n\big)$$

$$= \big((ax_1 z_1 + by_1 z_1), \dots, (ax_n z_n + by_n z_n)\big)$$

$$= (ax_1 z_1, \dots, ax_n z_n) + (by_1 z_1, \dots, by_n z_n)$$

$$= a(\boldsymbol{x} \cdot \boldsymbol{z}) + b(\boldsymbol{y} \cdot \boldsymbol{z})$$

. . .

A standard basis for an n-dimensional vector space V over a field F is a linearly independent subset of the vector space that spans the vector space. The standard basis is a sequence of orthogonal unit vectors, i.e., the following set

$$B = \{(1,0,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,0,0,\ldots,1)\}$$

where 1 is the multiplicative identity in F and 0 is the additive identity in F.

Clearly, the vectors in B are linear independent, span F^n, and are orthogonal to each other.

5.6 Metric Spaces

A **metric space** is a set together with a notion of distance between its elements, usually called points. The idea is to generalize the concept of distance.

Formally, a metric space is an ordered pair (M, d) where M is a set and d is a metric (i.e., distance function) on M that satisfies the following properties:

- Non-negativity: $d(x, y) \geq 0$ for all $x, y \in M$.
- Identity of indiscernibles: $d(x, y) = 0$ if and only if $x = y$.
- Symmetry: $d(x, y) = d(y, x)$ for all $x, y \in M$.
- Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in M$.

The set of all real numbers, with the distance between $x, y \in \mathbb{R}$ is defined to be $|x - y|$ (i.e., the absolute value of the difference), is a metric space.

The set of all ordered pairs in $\mathbb{R}^2$, with the distance between $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$ defined as $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, is a metric space. This is known as the Euclidean metric.

Other metric spaces over $\mathbb{R}^2$ are possible. For example, in Taxicab geometry [76] the distance between points (x_1, y_1) and (x_2, y_2) is defined to be $|x_1 - x_2| + |y_1 - y_2|$. The idea is that one "travels" between points using a combination of horizontal and vertical moves (like a taxicab traveling over the grid of streets in a city). Figure 75 shows an example of the taxicab versus the Euclidean metric in $\mathbb{R}^2$. The Euclidean metric measures the straight line between the two points in the figure. The taxicab metric measures the distance between the two point is one is only able to travel horizontally or vertically. One such route is the dashed line, and another taxicab route is the dotted line in the figure.

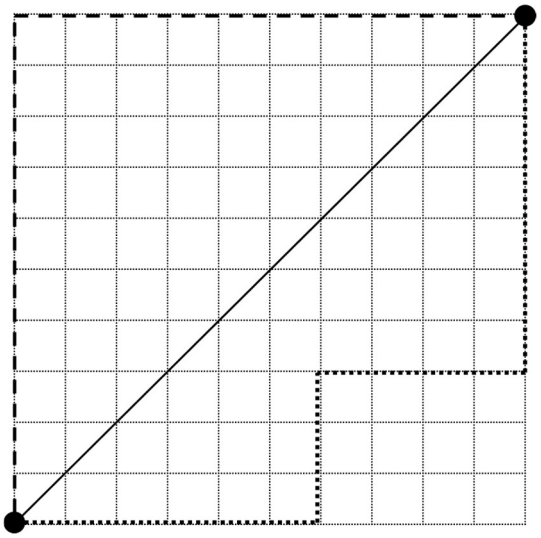


Figure 75. Taxicab versus Euclidean metric

6 Error Detecting and Correcting Codes

"Study hard what interests you the most in the most undisciplined, irreverent and original manner possible." — Richard Feynman

6.1 Overview

Error detecting and correcting codes are techniques used to ensure the reliable delivery of digital data over unreliable communication channels. These codes are used in a variety of applications, including data storage, telecommunications, and computer networking.

Error detection codes can only detect errors, while error correction codes can both detect and correct errors.

Error detection codes work by adding redundant bits to the data stream. These redundant bits are used to check the integrity of the data stream. If an error is detected, the receiver can request that the sender retransmit the data.

Error correction codes also work by adding redundant bits to the data stream, but in such a way that the receiver can correct the errors that are detected. This is done by using a mathematical algorithm to generate the redundant bits. The focus in this section is on error correcting codes.

It is a function of error correction techniques to introduce controlled redundancy with the intent of allowing corrupted messages to be identified and then corrected. Messages are comprised of **words** (typically bit strings). When employing controlled redundancy, only a subset of all possible transmitted words are valid. The subset of valid words is called a **code**, and the valid words are called **codewords** or (less commonly) code-vectors. The set of words include codewords and non-codewords (errored messages). Only codewords are intentionally transmitted. One goal in designing error correction codes is to maximize the “distance” among codewords so that the likelihood of errors corrupting one codeword into another codeword is small.

A **block code** is a type of error correcting code that encodes data in blocks. A block code consists of a fixed number of elements, called the block size. The elements can be numbers (binary numbers in many cases) or any set of symbols (e.g., letters from an alphabet). There are other types of error correcting codes, e.g., convolutional codes are a type of error correcting code that encodes data in a continuous stream. In what follows, our attention is focused exclusively on block codes.

Error detection is reduced to the task of determining whether a received word (in a message) is a codeword or not. If it is a codeword, it is assumed (with high probability) that no errors have occurred. The probability of an undetected error is the probability of sufficient errors occurring to transform the transmitted codeword into another codeword (falsely giving the appearance of a correctly transmitted codeword).

Figure 76 shows a message consisting of five words. Four of the words are codewords (intended information) and one word is corrupted, i.e., a non-codeword which needs to be mapped (via error correction) to the closest codeword.

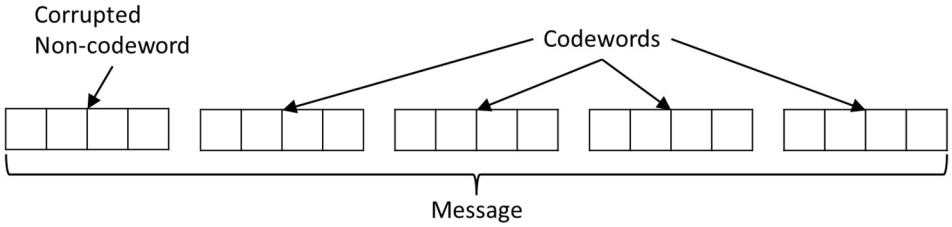


Figure 76. Message comprised of words

A message with a detected error (or errors) can be corrected in two basic ways, i.e.,

- The recipient rejects the entire message (consists of one or more errored words) as erroneous and requests retransmission of the message.
- The recipient detects and then corrects each errored word by mapping to the "nearest" codeword (i.e., least number of differences), under the assumption that the nearest codeword is the most likely correct (i.e., what was intended to be sent).

As a first example, we consider the **repetition code**. In this code, each bit in a codeword is repeated three times and we have just two codewords, i.e., 000 and 111. If the original (not encoded) binary message is "01101", it would be encoded as "000 111 111 000 111". If a single bit is changed in a codeword during transmission, the receiver can make the most likely correction. For example, if the received message is "100 111 101 000 110", the receiver can see that the majority of bits in the first group are 0, so the first bit must be 0. Similarly, the other groups (codewords) can be corrected to get “000 111 111 000 111”. This repetition code can only correct for single bit errors. As we shall see, more complex error correcting codes can correct for multiple bit errors.

If we represent the codewords in the above repetition code as vectors, we can represent the situation graphically as shown in Figure 77. One can think of a sphere of radius 1 about 000 which includes codewords that differ in one position, i.e., codewords 001, 010 and 100. There is also a sphere of radius 1 about 111 which includes codewords 011, 101 and 110. In this case (and in general for error codes represented by vectors), distance is measured by the number of differences in corresponding positions. For example, the distance between 111 and 000 is three since the two vectors have different values in all three positions.

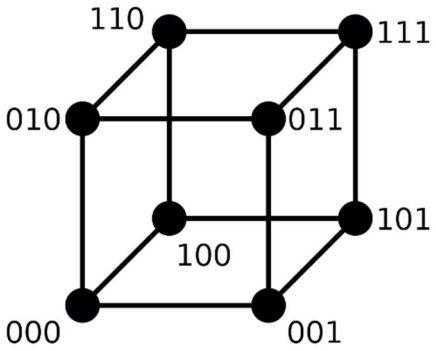


Figure 77. Triple repetition code

If we increase the repetition from say 3 to 5 (with resulting codewords 00000 and 11111), we can correct 2 errors, but the price is an increase in the size of the transmitted message.

6.2 Definitions and Basic Concepts

6.2.1 Error Correcting Codes in Terms of Vector Spaces

In terms of vector spaces, an error-correcting code can be viewed as a subspace of a vector space, where the vector space is defined over a finite field such as $\mathbb{F}_2$. The codewords are the vectors in the subspace, and the errors are the vectors that are not in the subspace. In our triple repetition example, the vector space is $\mathbb{F}_2^3$ and the codewords are $\{000,111\}$.

More formally, let $\mathbb{F}_q^n$ denote the vector space of all n-tuples over the finite field $\mathbb{F}_q$. An (n, M) code $\mathcal{C}$ over $\mathbb{F}_q$ is a subset of $\mathbb{F}_q^n$ of size M. A vector $(a_1, a_2, \ldots, a_n) \in \mathbb{F}_q^n$ may also be represented as $a_1 a_2 \ldots a_n$, which is easier to write, e.g., $(1,0,1,1,1,0)$ versus 101110. The vectors in $\mathcal{C}$ are the codewords of the code. If $\mathcal{C}$ is a subspace of $\mathbb{F}_q^n$, then $\mathcal{C}$ is known as a **linear code** over $\mathbb{F}_q$; otherwise, $\mathcal{C}$ is a non-linear code. In what follows, we focus our attention on linear codes.

Since a linear code $\mathcal{C}$ is a subspace of $\mathbb{F}_q^n$, it is also an additive subgroup of $\mathbb{F}_q^n$. By Lagrange's theorem (Theorem 36), $|\mathcal{C}|$ divides $\left|\mathbb{F}_q^n\right| = q^n$ and so, $|\mathcal{C}| = q^k$ for some positive integer $k \leq n$.

6.2.2 Distance

In the context of error correcting codes, the distance between two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{F}_q^n$, denoted by $d(\boldsymbol{u}, \boldsymbol{v})$, is the number of coordinates in which they differ. This is known as the **Hamming distance**.

For example, $d(11010,01011) = 2$ since the two vectors differ in their 1st and 5th positions.

Theorem 58. The Hamming distance over $\mathbb{F}_q^n$ meets the conditions for a metric space.

Proof: The first three properties (i.e., non-negativity, identity of indiscernibles and symmetry) are obviously true. Concerning the triangle inequality, we need to show $d(\boldsymbol{u}, \boldsymbol{v}) \leq d(\boldsymbol{u}, \boldsymbol{w}) + d(\boldsymbol{w}, \boldsymbol{v})$ for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{F}_q^n$. There are two cases:

- $u_i = v_i$ implies 0 is contributed to the distance $d(\boldsymbol{u}, \boldsymbol{v})$. For the i^{th} position on the right of the inequality, referring to the i^{th} position in $d(\boldsymbol{u}, \boldsymbol{w}) + d(\boldsymbol{w}, \boldsymbol{v})$, we at least contribute 0 but could contribute 1 or 2 depending on whether $u_i \neq w_i$ or $v_i \neq w_i$.
- $u_i \neq v_i$ implies 1 is contributed to the distance $d(\boldsymbol{u}, \boldsymbol{v})$. For the i^{th} position on the right of the inequality, the only way to get a 0 contribution is for $u_i = w_i = v_i$ (a contradiction). So, contribution on the right is at least 1.

In either case and for each position, the contribution to the distance on the left side is less than or equal to the contribution on the right. Thus, $d(\boldsymbol{u}, \boldsymbol{v}) \leq d(\boldsymbol{u}, \boldsymbol{w}) + d(\boldsymbol{w}, \boldsymbol{v})$ for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{F}_q^n$. ■

The **weight of a vector** $\boldsymbol{v} \in \mathbb{F}_q^n$, denoted $wt(v)$, is the number of nonzero coordinates in $\boldsymbol{v}$. For example, if we take the vector $\boldsymbol{u} = 1201202 \in \mathbb{F}_3^7$, the $wt(\boldsymbol{u}) = 2$ since the 3rd and 6th positions are non-zero.

The **minimum distance of a linear code** $\mathcal{C}$ is defined as the smallest Hamming distance between any two distinct codewords in $\mathcal{C}$. In our triple repeat example, the minimum distance of the code is 3 (which is easy to compute since there are only two elements in the code, i.e., 000 and 111).

A code $\mathcal{C}$ is said to be e-error correcting code if and only if the minimum of the Hamming distances between any two of its codewords is at least $2e + 1$. This is more easily understood geometrically as any closed balls of radius e centered on distinct codewords being disjoint. These balls are also called Hamming spheres in this context. In simple terms, an e-error correcting code can correct an errored word with as many as e errors.

The following theorem gives us an easier way to compute the minimum distance of a code (rather than computing all the distance between pairs of codewords).

Theorem 59. the minimum distance $d(\mathcal{C})$ of a linear code $\mathcal{C}$ equals the minimum of the weights of nonzero codewords.

Proof: Let $f = \min wt(\boldsymbol{v})$ for $\boldsymbol{v} \in \mathcal{C}$. We want to show that $f = d(\mathcal{C})$.

Let $\mathbf{0}$ be the vector with all zeros. Since $\mathcal{C}$ is a subspace, $\mathbf{0} \in \mathcal{C}$.

Let $\boldsymbol{w}$ be a codeword of length f. We have that $d(\boldsymbol{w}, \mathbf{0}) = f$ which implies $f \geq d(\mathcal{C})$.

Let $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{C}$ such that $d(\boldsymbol{u}, \boldsymbol{v}) = d(\mathcal{C})$. Since $\mathcal{C}$ is a linear code, and thus a subspace, we have closure under addition. So, $\boldsymbol{u} + (-\boldsymbol{v}) = \boldsymbol{u} - \boldsymbol{v} \in \mathcal{C}$. But $wt(\boldsymbol{u} - \boldsymbol{v}) = d(\boldsymbol{u}, \boldsymbol{v}) = d(\mathcal{C})$, and so, $f \leq d(\mathcal{C})$.

Thus, $f = d(\mathcal{C})$. ■

6.2.3 Notation for Codes

The general notation for an error correcting code is $(n, M, d)_q$ where

- n is the number of elements (positions, coordinates) in each codeword
- M is the number of codewords
- d is the minimum distance between any two codewords
- q is the size of the "alphabet" used to comprise the elements in each codeword.

A q-ary code $(n, M, d)_q$ is a collection of M vectors (codewords) of length n such that any two different codewords are at Hamming distance of at least d.

In the case of a q-ary linear code $\mathcal{C}$ a slightly different notation is sometimes used. As a subspace of $\mathbb{F}_q^n$, $\mathcal{C}$ spans a space of dimension k such that $q^k \leq q^n$. The notation for such a code is $[n, k, d]_q$ where n, d and q are as defined above, but k is the dimension of the subspace spanned by $\mathcal{C}$ rather than the number of codewords in $\mathcal{C}$ (which is q^k). So, $[n, k, d]_q$ and $\left(n, M = q^k, d\right)_q$ are two different ways of representing the same code.

6.2.4 Some Examples

A central problem in coding theory is to determine the maximum M such that an $(n, M, d)_q$ code exists.

In the following example, we demonstrate that a $(6,2^3 = 8,3)_2$ linear code does exist. Consider the code within $\mathbb{F}_2^6$ generated (i.e., spanned) be the basis vectors

$$100110$$
$$010101$$
$$001011$$

The subspace of $\mathbb{F}_2^6$ spanned by the above vectors consists of all linear combinations, i.e.,

$$a(1,0,0,1,1,0) + b(0,1,0,1,0,1) + c(0,0,1,0,1,1), \qquad a, b, c \in \mathbb{F}_2$$

Since a, b and c can only be 0 or 1, this gives us 8 combinations, leading to the following set of codewords:

$$\begin{array}{cc} 000000 & 110011 \\ 100110 & 101101 \\ 010101 & 011110 \\ 001011 & 111000 \end{array}$$

The minimum weight in the set of non-zero vectors listed above is 3. Thus, by Theorem 59, the minimum distance for the code is 3.

. . .

Next, let's consider the ternary code within $\mathbb{F}_3^4$ generated by the basis vectors

$$1002$$
$$0102$$
$$0012$$

The subspace of $\mathbb{F}_3^4$ spanned by the above vectors consists of all linear combinations, i.e.,

$$a(1,0,0,2) + b(0,1,0,2) + c(0,0,1,2), \qquad a, b, c \in \mathbb{F}_3$$

Since a, b and c can be 0, 1 or 2, this gives us $3^3 = 27$ combinations, leading to the following set of codewords:

$$\begin{array}{ccccccccc} 0000 & \mathbf{0012} & 2001 & 0021 & 1101 & 2100 & 0201 & 1200 & 2202 \\ \mathbf{1002} & 1011 & 2010 & 0111 & 1110 & 2112 & 0210 & 1212 & 2211 \\ \mathbf{0102} & 1020 & 2022 & 0120 & 1122 & 2121 & 0222 & 1221 & 2220 \end{array}$$

The minimum weight among the non-zero vectors in the above list is 2, and so, the minimum distance for the code is 2.

In terms of our notation, this is a $(4,3^3 = 27,2)_3$ linear code. Note that a $(3,27,2)_3$ code is not possible since it would necessarily consist of all the ternary triples and thus, have minimum distance 1.

. . .

A $(4,3,3)_2$ code is not possible. First off, all linear codes (being subspaces of a vector space) have the $\mathbf{0}$ vector which in this case, is $(0,0,0,0)$. Any other vector in $(4,3,3)_2$ would need to have at least three 1s to be at least a distance 3 from $(0,0,0,0)$. With loss of generality, say this second vector is $(1,1,1,0)$. A third vector is not possible since it could not differ from both $(0,0,0,0)$ and $(1,1,1,0)$ in three positions.

. . .

In some cases, we can construct a code from a known code. For example, we can construct a $(5,4,3)_2$ code from the $(6,8,3)_2$ code we developed earlier. Just take the 4 codewords in the $(6,8,3)_2$ code that end in a 0, and then omit the 0 to get the $(5,4,3)_2$ listed below.

$$00000$$

$$10011$$

$$01111$$

$$11100$$

One needs to check that this set is closed under vector addition (which it is) and that the minimum distance between any two codewords is 3 (which is also true).

6.2.5 Bounds on the Number of Codewords

The **Singleton bound**, named after Richard Collom Singleton, defines an upper bound on the size (i.e., number of codewords) of an arbitrary block code $\mathcal{A}$ with block length n and minimum distance d. [In this context, "Arbitrary code" means a code whose alphabet can be any set of characters and not necessarily from a finite field.]

Let $M_q(n,d)$ represent the maximum possible number of codewords in a block code $\mathcal{C}$ over an alphabet $\mathcal{A}$ of size q, where each codeword is a vector of length n with elements from $\mathcal{A}$. The Singleton bound theorem [77] states that

$$M_q(n,d) \leq q^{n-d+1}$$

If $\mathcal{A}$ is a linear code with block (i.e., vector) length n, dimension k and minimum distance d over the finite field $\mathbb{F}_q$, then the maximum number of codewords is q^k and the Singleton bound implies:

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

Thus, $k \leq n - d + 1$.

We can use this result to show various codes cannot exist. For example, $(5, 2^5, 3)_2$ code cannot exist because $k = 5$ is greater than $n - d + 1 = 5 - 3 + 1 = 3$.

Similarly, a $(6,2^9,3)_2$ code cannot exist because $k = 9$ is greater than $n - d + 1 = 6 - 3 + 1 = 4$.

. . .

Another type of bound for error correcting codes is known as the **Hamming bound** (or sphere packing bound). This bound applies to any block code $\mathcal{C}$ over an alphabet $\mathcal{A}_q$ of size q (including but not limited to finite fields).

We first need some definitions and a theorem before getting to the Hamming bound.

- Let $\mathcal{A}_q$ be a collection of distinct symbols of size q.
- Let $\mathcal{A}_q^n$ be the set of all vectors of length n whose elements are chosen from $\mathcal{A}_q$.
- Given a vector $\boldsymbol{x} \in \mathcal{A}_q^n$, the set of vectors whose distance from $\boldsymbol{x}$ is less than or equal to $\boldsymbol{r}$ is known as the **ball** of radius r centered at $\boldsymbol{x}$.
- The number of vectors at distance less than or equal to r from a given vector in $\mathcal{A}_q^n$ is denoted by $V_q(r,n)$. $V_q(r,n)$ is known as the volume of a ball of radius r.

Theorem 60. The volume of the ball of radius r is given by the formula

$$V_q(r,n) = \sum_{j=0}^{r} \binom{n}{j} (q-1)^j$$

Proof: Pick a distance $j \leq r$ and count the number of vectors at distance exactly j from the given vector. There are $\binom{n}{j}$ choices for the set of coordinates where the entries are different from the given vector. Once this set is fixed, there are $q-1$ possibilities for the entries in each of the j coordinates in the set. This gives us $\binom{n}{j}(q-1)^j$ vectors for each value of $j \leq r$. Summing from $j = 0$ to $j = n$, we get the desired result. ■

Theorem 61 (Hamming or sphere packing bound). Every q-ary code $\mathcal{C}$ of length n and minimum distance d satisfies

$$M_q(n,d) \leq \frac{q^n}{V_q(e,n)}$$

where $e = \left\lfloor \frac{d-1}{2} \right\rfloor$, i.e., the greatest integer less than or equal to $\frac{d-1}{2}$.

Proof: The condition $e = \left\lfloor \frac{d-1}{2} \right\rfloor$ ensures that the balls of radius e centered at each codeword must be disjoint. Since each such ball contains $V_q(e,n)$ vectors (including one codeword at its center), we must have

$$M_q(n,d) \cdot V_q(e,n) \leq q^n$$

since the number of vectors in all the balls so described cannot be greater than the number of vectors in the entire space $\mathcal{A}_q^n$, i.e., q^n. ■

Codes for which equality holds in the Hamming bound are known as **perfect codes**. Examples include codes that have only one codeword, and codes that include the entire vector space $\mathcal{A}_q^n$. Another set of examples are the repeat codes, where each symbol of the message is repeated an odd number of times to obtain a codeword where $q = 2$. All of these examples are called trivial perfect codes. In 1973,

Tietäväinen [78] proved that any non-trivial perfect code over an alphabet with p^n elements (p a prime number and n a positive integer) is either a Hamming code or Golay code (both of these types of codes will be discussed in Section 6.2.7). Thus, all perfect codes have been classified.

6.2.6 Generator Matrices

A **generator matrix** is a $k \times n$ matrix G whose rows form a basis for a $\left(n, q^k, d\right)_q$ linear code.

For example, the generator matrix for the $(6,2^3,3)_2$ linear code $\mathcal{C}$ that we saw in Section 6.2.4 is

$$G = \begin{bmatrix} 100110 \\ 010101 \\ 001011 \end{bmatrix}$$

Each codeword of $\mathcal{C}$ can be written as a linear combination of the row vectors of G. More formally, if $\boldsymbol{x} \in \mathcal{C}$, then we can write

$$\boldsymbol{x} = y_1\,[1\ 0\ 0\ 1\ 1\ 0] + y_2\,[0\ 1\ 0\ 1\ 0\ 1] + y_3\,[0\ 0\ 1\ 0\ 1\ 1]$$

or in matrix notation

$$\boldsymbol{x} = \boldsymbol{y}G$$

where $\boldsymbol{y} = [y_1\ y_2\ y_3]$.

In the linear coding scheme $(n, q^k, d)_q$, the length of each message is increased from k to n, with the first k positions of each codeword representing the intended information to be transmitted and the last $n - k$ positions containing redundant information used for error correction. In the case of binary linear codes, the redundant information is referred to as **parity bits**. For non-binary linear codes, there does not appear to be any common terminology for the redundant information used for error correction.

As an example of this concept, consider the previous example. There is a bijective mapping of the binary numbers from 000 to 111 (i.e., 0 to 7 in base 10) to the code $\mathcal{C}$, i.e.,

$$000 \rightarrow 000\ 000$$
$$001 \rightarrow 001\ 011$$
$$010 \rightarrow 010\ 101$$
$$011 \rightarrow 011\ 110$$
$$100 \rightarrow 100\ 110$$
$$101 \rightarrow 101\ 101$$
$$110 \rightarrow 110\ 011$$
$$111 \rightarrow 111\ 000$$

The first three bits in each codeword is used to represent the intended information to be transmitted and the last 3 bits ensure a distance of 3 between each pair of codewords.

. . .

The **information rate** of an $(n, q^k, d)_q$ linear code is the ration of pure information to message length, and is given by $R = \frac{k}{n}$. The goal is to make R as large for a given error correcting capability (as defined by the value of $e = \left\lfloor \frac{d-1}{2} \right\rfloor$). Further, the ratio $\frac{d}{n}$ is known as the **relative distance**.

For the $(6,2^3,3)_2$ code, the information rate is $.5$ and the relative distance is $.5$.

. . .

Consider the linear code $\mathcal{C}$ with parameters $(7,2^4,3)_2$ and the matrix

$$G_1 = \begin{bmatrix} 1\,0\,0\,1\,1\,0\,1 \\ 0\,1\,0\,1\,0\,1\,1 \\ 0\,0\,1\,0\,1\,1\,1 \\ 0\,0\,0\,1\,1\,1\,0 \end{bmatrix}$$

To verify that G_1 is a generator matrix for $\mathcal{C}$, we need to

i. check that the row vectors are of length 7 (obviously true)

ii. check that the row vectors span a space of dimension 4 (this is true since the rows are linearly independent)

iii. check that the vectors (codewords) generated by all linear combinations of the row vectors all have minimum weight 3. The row vectors each have minimum weight 3 and differ from each other in at least 3 positions. In general, if two vectors with elements from $\mathbb{F}_2$ are added, the resulting vector has a 1 in positions where the two input vectors differ and has a 0 in positions where the two input vectors are the same. So, if we add two of the row vectors in G, we will get another vector of weight at least 3.

. . .

A generator matrix of the form $G = [I_k \mid P]$ where I_k is the $k \times k$ identity matrix and P contains the parity bits for each basis vector is known as the **standard form of a generator matrix**. The generator matrix from our $(6,2^3,3)_2$ example is in standard form, but the generator matrix from our $(7,2^4,3)_2$ is not in standard form.

6.2.7 Equivalent Linear Codes

Two codes $\mathcal{C}_1$ and $\mathcal{C}_2$, with respective generator matrices G_1 and G_2, are said to be **permutation equivalent**, if a permutation of the columns of G_1 yields G_2.

For example, the code with the generator matrix G_2 (shown below) is equivalent to the $(7,2^4,3)_2$ code that we discussed previously. If we permutate the 4th and 7th columns of G_2, we get the generator matrix G_1 used in our previous example concerning the $(7,2^4,3)_2$ code. Thus, the two codes are permutation equivalent.

$$G_2 = \begin{bmatrix} 1\,0\,0\,1\,1\,0\,1 \\ 0\,1\,0\,1\,0\,1\,1 \\ 0\,0\,1\,1\,1\,1\,0 \\ 0\,0\,0\,0\,1\,1\,1 \end{bmatrix}$$

More generally, if there exists an $n \times n$ monomial matrix $M: \mathbb{F}_q^n \to \mathbb{F}_q^n$ which isomorphically maps the generator matrix for code $\mathcal{C}_1$ to the generator matrix for code $\mathcal{C}_2$, then $\mathcal{C}_1$ and $\mathcal{C}_2$ are said to be **equivalent**.

A monomial matrix over $\mathbb{F}_q$ has exactly one nonzero entry in each row and each column. For example, the following is a monomial matrix over $\mathbb{F}_3$

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

For example, consider following generator matrix for a $(6, 3^3, 3)_3$ linear code $\mathcal{C}$

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 2 & 1 & 1 \end{bmatrix}$$

and the monomial matrix

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Taking the product of M and G, we get a generator matrix for code that is equivalent to $\mathcal{C}$, i.e.,

$$GM = \begin{bmatrix} 1 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 1 & 2 & 2 \\ 0 & 1 & 0 & 2 & 2 & 1 \end{bmatrix}$$

Notice that the effect of the multiplication is the permutation of columns 2 and 3 of G, and the multiplication of column 5 in G by 2.

In general, multiplication by a monomial matrix on the right has two possible effects, i.e., transposition of columns and/or the multiplication of a column by a scalar. So, two linear codes are equivalent if the generator matrix of one code can be made equal to the other generator matrix of the other code by a combination of column permutations and scalar multiplications applied to one or more columns.

The following theorem provides even more general conditions for two codes to be equivalent.

Theorem 62. Two $k \times n$ matrices generate equivalent $(n, q^k, d)_q$ linear codes if one matrix can be obtained from the other by a sequence of operations of the following types:

- (R1) Permutation of the rows.
- (R2) Multiplication of a row by a non-zero scalar.
- (R3) Addition of a scalar multiple of one row to another.
- (C1) Permutation of the columns.
- (C2) Multiplication of any column by a non-zero scalar.

Proof: From basic linear algebra, we know that the row operations R1, R2 and R3 preserve the linear independence of the rows of a generator matrix and replace one basis by another of the same code. Thus, any combination of operations R1, R2 and R2 on a generator matrix for a code $\mathcal{C}$, produce another matrix that generates $\mathcal{C}$ exactly, and clearly, $\mathcal{C}$ is equivalent to itself.

Operations (C1) and (C2) convert a generator matrix for a code C to a generator matrix for a code that is equivalent to $\mathcal{C}$. ■

. . .

Using some basic results from matrix algebra and the concept of equivalent linear codes, we are going to show that all generator matrices can be put into standard form. However, we first need some additional background.

A matrix is in **row echelon form** if

- All rows consisting of only zeroes are at the bottom of the matrix.
- The leading entry (that is the left-most nonzero entry) of every nonzero row is to the right of the leading entry of every row above.

A matrix is in **reduced row echelon** form if

- It is in row echelon form.
- The leading entry in each nonzero row is a 1 (called a leading 1).
- Each column containing a leading 1 has zeros in all its other entries.

The following matrix is in reduced row echelon form

$$\begin{bmatrix} 1 & 0 & x & 0 & a & 0 \\ 0 & 1 & y & 0 & b & 0 \\ 0 & 0 & 0 & 1 & c & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

By means of a finite sequence of elementary row operations (i.e., R1, R2 and R3 as defined above) any matrix can be transformed into reduced row echelon form. The reduced row echelon form of a matrix is unique. This means that no matter what sequence of elementary row operations are used to reduce a matrix to reduced row echelon form, the resulting matrix will be the same.

Elementary row operations preserve the row space of the matrix [79], i.e., the space spanned by the row vectors. Thus, the row space of the reduced row echelon form of a matrix is the same as that of the original matrix. In the context of a generator matrix for a linear code, we can say that it is always possible to modify a generator matrix via elementary row operations to get another generator matrix (for the same code) which is in reduced row echelon form.

We are now in a position to state and prove the following theorem.

Theorem 63. If G is a generator matrix for an $(n, q^k, d)_q$ linear code $\mathcal{C}$, then it is always possible to transform G, using operations of the types R1, R2, R3 and C2, into the standard form of a generator matrix for a code that is equivalent to $\mathcal{C}$.

Proof: From the previous discussion, we saw that it is always possible to transform G into reduced row echelon form. However, this may not be in standard form. In order to achieve standard form, we may

need to transpose some of the columns (using C2 type operations). In this case, we will get a generator matrix in standard form for an equivalent code to $\mathcal{C}$ (this follows from Theorem 62). ■

For example, the matrix below is a generator matrix for a linear code of type $(10,2^3,5)_2$.

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

G is in reduced row echelon form but not in standard form. We can put G in standard form by transposing columns 2 and 3, and then transposing (the new) column 3 with column 5. The result is

$$G' = \begin{bmatrix} 1&0&0&0&1&0&1&1&0&1 \\ 0&1&0&1&0&0&1&0&1&1 \\ 0&0&1&0&0&1&0&1&1&1 \end{bmatrix} = [I_3 \mid P]$$

where

$$P = \begin{bmatrix} 0&1&0&1&1&0&1 \\ 1&0&0&1&0&1&1 \\ 0&0&1&0&1&1&1 \end{bmatrix}$$

G' is a generator matrix (in standard form) for a $(10,2^3,5)_2$ linear code that is equivalent to $\mathcal{C}$.

6.2.8 Duality

Given an $(n, q^k, d)_q$ linear code $\mathcal{C}$, the dual of $\mathcal{C}$ (denoted $\mathcal{C}^\perp$) is the set of all vectors in $\mathbb{F}_q^n$ that are orthogonal to every codeword in $\mathcal{C}$. In notation,

$$\mathcal{C}^\perp = \{\boldsymbol{x} \in \mathbb{F}_q^n : \boldsymbol{x} \cdot \boldsymbol{y}\ \forall \boldsymbol{y} \in \mathcal{C}\}$$

[The symbol ∀ means "for every."]

The concept of orthogonality for error correcting codes is different from that in geometry. In geometry, (for example) a line cannot be contained within a plane to which it is orthogonal (i.e., perpendicular). In error correcting codes, a codeword (vector) can be orthogonal to a space in which it is contained.

Consider the $(4,3^2,3)_3$ linear code $\mathcal{C}$ with generator matrix

$$G = \begin{bmatrix} 1&0&1&1 \\ 0&1&1&2 \end{bmatrix}$$

Represent the rows of G as $\boldsymbol{r_1} = (1,0,1,1)$ and $\boldsymbol{r_2} = (0,1,1,2)$. The vectors $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ are linearly independent, and

$$\boldsymbol{r_1} \cdot \boldsymbol{r_2} = 0 + 0 + 1 + 2 = 0$$
$$\boldsymbol{r_1} \cdot \boldsymbol{r_1} = 1 + 0 + 1 + 1 = 0$$
$$\boldsymbol{r_2} \cdot \boldsymbol{r_2} = 0 + 1 + 1 + 1 = 0$$

For any codeword $a\boldsymbol{r_1} + b\boldsymbol{r_2}$ in $\mathcal{C}$,

$$\boldsymbol{r_1} \cdot (a\boldsymbol{r_1} + b\boldsymbol{r_2}) = a\boldsymbol{r_1} \cdot \boldsymbol{r_1} + b\boldsymbol{r_1} \cdot \boldsymbol{r_2} = 0$$
$$\boldsymbol{r_2} \cdot (a\boldsymbol{r_1} + b\boldsymbol{r_2}) = a\boldsymbol{r_2} \cdot \boldsymbol{r_1} + b\boldsymbol{r_2} \cdot \boldsymbol{r_2} = 0$$

Thus, $\mathcal{C}^{\perp}$ contains all of $\mathcal{C}$.

The following theorem generalizes the result in the previous example.

Theorem 64. Given an $(n, q^k, d)_q$ linear code $\mathcal{C}$ with generator matrix G. A vector $x \in \mathbb{F}_q^n$ is in $\mathcal{C}^{\perp}$ if and only if x is orthogonal to every row of G, i.e., $x \in \mathcal{C}^{\perp}$ if and only if $Gx^T = 0$, where x^T is the transpose of x.

Proof: If $\boldsymbol{x} \in \mathcal{C}^{\perp}$ then, by definition, $\boldsymbol{r} \cdot \boldsymbol{x}^T = 0$ for every row $\boldsymbol{r}$ of G (as well as every vector in $\mathcal{C}$). Thus, $G\boldsymbol{x}^T = \boldsymbol{0}$.

Going in the other direction, assume $G\boldsymbol{x}^T = \boldsymbol{0}$ and let $\boldsymbol{r_i}$ be the i^{th} row of G.

$G\boldsymbol{x}^T = \boldsymbol{0}$ implies that $\boldsymbol{r_i} \cdot \boldsymbol{x}^T = 0$ for $i = 1,2, \ldots, k$.

If $\boldsymbol{y} \in \mathcal{C}$, then we can write $\boldsymbol{y}$ as a linear combination of the rows of *G*, i.e.,

$$\boldsymbol{y} = \sum_{i=1}^{k} a_i \boldsymbol{r_i}$$

Using Theorem 57 (Part ii), we have

$$\boldsymbol{y} \cdot \boldsymbol{x}^T = \left(\sum_{i=1}^{k} a_i \boldsymbol{r_i} \right) \cdot \boldsymbol{x}^T = \sum_{i=1}^{k} a_i (\boldsymbol{r_i} \cdot \boldsymbol{x}^T) = \sum_{i=1}^{k} a_i 0 = 0$$

So, $\boldsymbol{x}^T$ (or equivalently, $\boldsymbol{x}$) is orthogonal to every codeword in $\mathcal{C}$, and thus, $\boldsymbol{x} \in \mathcal{C}^{\perp}$. ■

The next theorem tells us that the dimension of an $(n, q^k, d)_q$ linear code plus the dimension of its dual equals n.

Theorem 65. If $\mathcal{C}$ is a linear code of dimension k in $\mathbb{F}_q^n$, then its dual $\mathcal{C}^{\perp}$ is linear code of dimension $n - k$ in $\mathbb{F}_q^n$.

Proof: First, we show that $\mathcal{C}^{\perp}$ is a subspace of $\mathbb{F}_q^n$. Take any linear combination of vectors in $\mathcal{C}^{\perp}$, e.g., $a\boldsymbol{x} + b\boldsymbol{y}$. For all $\boldsymbol{z} \in \mathcal{C}$, we have (by Theorem 57, Part ii)

$$(a\boldsymbol{x} + b\boldsymbol{y}) \cdot \boldsymbol{z} = a(\boldsymbol{x} \cdot \boldsymbol{z}) + b(\boldsymbol{y} \cdot \boldsymbol{z}) = a0 + b0 = 0 + 0 = 0$$

Thus, $a\boldsymbol{x} + b\boldsymbol{y} \in \mathcal{C}^{\perp}$ and by Theorem 52, $\mathcal{C}^{\perp}$ is a subspace of $\mathbb{F}_q^n$.

Next, we show that $\mathcal{C}^{\perp}$ has dimension $n - k$. Let G be the generator matrix for $\mathcal{C}$. By Theorem 64, $G\boldsymbol{x}^T = 0$ for all $\boldsymbol{x} \in \mathcal{C}^{\perp}$. This is a system of k independent homogeneous equations in n unknowns and by a standard result from linear algebra (see Theorem 1.5.1 in Kuttler [80]) its solution space (i.e., $\mathcal{C}^{\perp}$) is of dimension $n - k$. ■

Going back to our example concerning the dual of the $(4,3^2,3)_3$ linear code. We can now apply Theorem 65, and conclude that $\mathcal{C}^{\perp}$ is of dimension $n - k = 4 - 2 = 2$. Since we already determined that $\mathcal{C}^{\perp}$ contains all of $\mathcal{C}$, and $\mathcal{C}$ is of dimension 2, we can say that $\mathcal{C}^{\perp} = \mathcal{C}$.

Theorem 66. For any $(n, q^k, d)_q$ linear code $\mathcal{C}$, $(\mathcal{C}^\perp)^\perp = \mathcal{C}$.

Proof: Since every vector in $\mathcal{C}$ is orthogonal to every vector in $\mathcal{C}^\perp$, $\mathcal{C} \subseteq (\mathcal{C}^\perp)^\perp$. By Theorem 65, the dimension of $(\mathcal{C}^\perp)^\perp$ is $n - (n-k) = k$ which is the dimension of $\mathcal{C}$. Thus, $\mathcal{C} = (\mathcal{C}^\perp)^\perp$. ■

. . .

A **parity-check matrix** is a matrix that describes the linear relationships between the elements of a codeword. It can be used to determine whether a vector is a valid codeword, and it can also be used to decode received messages that have been corrupted by errors. Formally, a parity-check matrix H of a linear code $\mathcal{C}$ is a generator matrix of the dual code $\mathcal{C}^\perp$. So, vector $\boldsymbol{x}$ is a codeword in $\mathcal{C}$ if and only if the $H\boldsymbol{x}^T = \mathbf{0}$.

A linear code is completely specified by its parity-check matrix. For example, assume the following is a parity-check matrix for code $\mathcal{C}$, which is defined over $\mathbb{F}_2^5$.

$$H = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The vector $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5) \in \mathbb{F}_2^5$ is a codeword of $\mathcal{C}$ only if $H\boldsymbol{x}^T = \mathbf{0}$, which implies

$$x_1 + x_2 + x_3 = 0$$

$$x_2 + x_4 = 0$$

$$x_1 + x_5 = 0$$

Letting $x_2 = s$ and $x_1 = t$, we have

$$x_3 = -s - t = s + t$$

$$x_4 = -s = s$$

$$x_5 = -t = t$$

So, $x = (t, s, s+t, s, t) = t(1,0,1,0,1) + s(0,1,1,1,0)$ which implies that $(1,0,1,0,1)$ and $(0,1,1,1,0)$ are a basis for the subspace spanned by code $\mathcal{C}$. Thus, a generator matrix for $\mathcal{C}$ is

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

The following theorem provides a mechanism for constructing a parity-check matrix given the generator matrix for a linear code, and vice versa.

Theorem 67. Given a generator matrix in standard form $G = [I_k \mid P]$ for an $(n, q^k, d)_q$ linear code $\mathcal{C}$, the associated parity-check matrix for $\mathcal{C}$ is $H = [-P^t \mid I_{n-k}]$.

Proof: The rows of H are linear independent, given the placement of the identity matrix. Further, H is of dimension $(n-k) \times n$ which is the correct size (as implied by Theorem 65). We are left to show that every row of H is orthogonal to every row of G (which implies that the rows H span the code $\mathcal{C}^\perp$. Expanding G and H, we have

$$G = \begin{bmatrix} 1\,0\,\ldots\,0 & p_{11}\,\ldots\,p_{1,n-k} \\ 0\,1\,\ldots\,0 & p_{21}\,\ldots\,p_{2,n-k} \\ & \vdots \\ 0\,0\,\ldots\,1 & p_{k1}\,\ldots\,p_{k,n-k} \end{bmatrix}$$

and

$$H = \begin{bmatrix} -p_{11} & -p_{21} & \ldots & -p_{k1} & 1\,0\,\ldots\,0 \\ -p_{12} & -p_{22} & \ldots & -p_{k2} & 0\,1\,\ldots\,0 \\ & & \vdots & & \\ -p_{1,n-k} & -p_{2,n-k} & \ldots & -p_{k,n-k} & 0\,0\,\ldots\,1 \end{bmatrix}$$

With the advantage of viewing the above expansions, it is clear that the inner product of the i^{th} row of G with the j^{th} row of H is

$$-p_{ij} + p_{ij} = 0$$

So, every row of H is orthogonal to every row of G. ■

For the example at the end of Section 6.2.7, we found the following to be a generator matrix for the linear code of type $(10,2^3,5)_2$

$$G = \begin{bmatrix} 1\,0\,0 & 0\,1\,0\,1\,1\,0\,1 \\ 0\,1\,0 & 1\,0\,0\,1\,0\,1\,1 \\ 0\,0\,1 & 0\,0\,1\,0\,1\,1\,1 \end{bmatrix} = [I_3 \mid P]$$

By Theorem 67, the parity-check matrix is

$$H = [-P^t \mid I_7] = \begin{bmatrix} 0\,1\,0 & 1\,0\,0\,0\,0\,0\,0 \\ 1\,0\,0 & 0\,1\,0\,0\,0\,0\,0 \\ 0\,0\,1 & 0\,0\,1\,0\,0\,0\,0 \\ 1\,1\,0 & 0\,0\,0\,1\,0\,0\,0 \\ 1\,0\,1 & 0\,0\,0\,0\,1\,0\,0 \\ 0\,1\,1 & 0\,0\,0\,0\,0\,1\,0 \\ 1\,1\,1 & 0\,0\,0\,0\,0\,0\,1 \end{bmatrix}$$

Since we are working in $\mathbb{F}_2$, $-1 \equiv 1 \pmod 2$.

6.3 Linear Codes

6.3.1 Simplex and Hamming Codes

For positive integer $k \geq 2$, form a matrix whose columns are all the non-zero bitstrings of length k. This results in a total of $2^k - 1$ columns and a matrix of size $k \times (2^k - 1)$ which is denoted by S_k. If we order the bitstrings such that the first k columns are I_k, then clearly, the rows are linearly independent and span a subspace of $\mathbb{F}_2^{(2^k-1)}$. Thus, S_k is a generator for a $\left(2^k - 1, 2^k, d\right)_2$ linear code. Such codes are known as the binary **Simplex codes**. Rather than using the generic notation, a Simplex code formed with bitstrings of length k is represented by $\mathcal{S}(2,k)$.

We still need to determine the minimum distance d for $\mathcal{S}(2,k)$. To that end, we construct the generator matrices for the binary Simple codes as follows. We start with the following

$$S_3 = \begin{bmatrix} 1\,0\,0\,1\,1\,0\,1 \\ 0\,1\,0\,1\,0\,1\,1 \\ 0\,0\,1\,0\,1\,1\,1 \end{bmatrix}$$

Next, we define S_4 in terms of S_3 as follows:

$$S_4 = \begin{bmatrix} \mathbf{0} & 1 & \mathbf{1} \\ S_3 & \mathbf{0} & S_3 \end{bmatrix}$$

Basically, we just add either a 0 or a 1 to each of the rows in S_3. Note that the $\mathbf{1}$ in the upper right is a vector with $2^3 - 1 = 7$ ones, and the 0 in the upper left is a vector with $2^3 - 1 = 7$ zeros. One of the possible bitstring of length 4 is missing since the all-zero column is excluded in S_3. We address this issue by adding the column $\begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix}$ in the middle of S_4.

In general,

$$S_k = \begin{bmatrix} \mathbf{0} & 1 & \mathbf{1} \\ S_{k-1} & \mathbf{0} & S_{k-1} \end{bmatrix}$$

where $\mathbf{1}$ is a vector with $2^{k-1} - 1$ ones, the $\mathbf{0}$ in the upper right is a row vector with $2^{k-1} - 1$ zeros, and the $\mathbf{0}$ in the middle of the second row is a column vector with $k - 1$ zeros.

We are now in a position to prove the following theorem.

Theorem 68. The minimum distance for the $\mathcal{S}(2,k)$ Simplex code is 2^{k-1}.

Proof: The proof is by induction.

For $k = 3$, note that each row of S_3 is of weight $2^{3-1} = 4$. By inspection, we see that adding any combination of rows in S_3 also yields a codeword of weight 4. Thus, the minimum distance for $\mathcal{S}(2,3)$ is 4. So, the theorem is true for $k = 3$.

Next, we assume that the theorem is true for the case $k - 1$ and show that the theorem holds for the case k. Using our construction methodology, we have

$$S_k = \begin{bmatrix} \mathbf{0} & 1 & \mathbf{1} \\ S_{k-1} & \mathbf{0} & S_{k-1} \end{bmatrix}$$

By the induction hypothesis, each linear combination of rows (not involving the first row) has weight

$$2^{k-2} + 2^{k-2} = 2^{k-1}$$

A linear combination involving the first row and any of the other rows has weight

$$2^{k-2} + 1 + \left(\left(2^{k-1} - 1\right) - 2^{k-2}\right) = 2^{k-1}$$

So, we have shown the theorem is true for the case k, and the completes the induction proof. ■

In the above proof, we showed that all codewords in $\mathcal{S}(2,k)$ are of exactly the same weight, i.e., 2^{k-1}. Codes with the property that every codeword is of the same weight are known as **constant-weight codes**.

In our generic notation for linear codes, an $\mathcal{S}(2,k)$ Simplex code is a $\left(2^k-1, 2^k, 2^{k-1}\right)_2$ code.

Theorem 69. Every binary Simplex code $\mathcal{S}(2,k), k \geq 3$, is self-orthogonal (i.e., every codeword is orthogonal to every other codeword).

Proof: Consider the generator matrix S_k for $\mathcal{S}(2,k)$. Since each codeword has even weight (2^{k-1} in this case), the dot produce of a codeword with itself is the sum of an even number of ones which in binary adds to zero. So, each codeword is orthogonal to itself.

For the case of different codewords, we proceed by induction. For $k=3$, we can verify the property by inspection.

Assume that for the case $k-1$, any two differ codewords in $\mathcal{S}(2,k)$ are orthogonal, i.e., have a dot product of zero. As we did in the previous theorem, we write

$$S_k = \begin{bmatrix} \mathbf{0} & 1 & \mathbf{1} \\ S_{k-1} & 0 & S_{k-1} \end{bmatrix}$$

By the induction hypothesis, the dot product of any two different rows (excluding the first row) is zero. The dot product of the first row with any of the other rows is just the sum bits in a basic vector from $\mathcal{S}(2,k-1)$ which we know to be zero since each codework (including the basis vectors) have even weight.

We are not quite done. To complete the induction proof, we need to show that any linear combination of the basis vectors (itself a codeword) is orthogonal to any other linear combination of the basis vectors. Let the following be linear combinations of the basis vectors (i.e., rows of S_k):

$$u = x_1 + x_2 + \cdots + x_r$$

$$v = y_1 + y_2 + \cdots + y_s$$

Taking the dot product, we have

$$u \cdot v = \sum_{i=1}^{r}\sum_{j=1}^{s} x_i \cdot y_j = \sum_{i=1}^{r}\sum_{j=1}^{s} 0 = 0$$

This completes the proof. ■

. . .

A **Hamming code** is typically defined by specifying its parity-check matrix as follows:

For a positive integer k, let H_k be a $k \times (2^k-1)$ matrix whose columns are the distinct non-zero vectors of $\mathbb{F}_2^k$. The code having H_k as its parity-check matrix is known as the binary Hamming code and is denoted by $\mathcal{H}(2,k)$. [**Warning**: In the literature, different notations are used for the Hamming codes.]

Since $H_k = S_k$ (as defined for the Simplex code), we conclude that the Hamming code $\mathcal{H}(2,k)$ and the Simplex code $\mathcal{S}(2,k)$ are dual codes, i.e., $\mathcal{H}(2,k) = \mathcal{S}(2,k)^{\perp}$. By Theorem 65, $\mathcal{H}(2,k)$ is of dimension $\left(2^k-1\right)-k = 2^k-k-1$.

Another notation for $\mathcal{H}(2,k)$, perhaps more common in the industry, is $\mathcal{H}(2^k-1, 2^k-k-1)$ where the first component is the length of a codeword, and the second component is the dimension of the code (i.e., number of basis vectors).

Theorem 70. The Hamming $\mathcal{H}(2,k)$ code has minimum distance 3.

Proof: Suppose that $\mathcal{H}(2,k)$ has a codeword $\boldsymbol{v}$ of weight 1 (with a 1 in position i). Since (by definition) $\boldsymbol{v}$ is orthogonal to every row of H_k, the i^{th} position of every row of H_k must be zero. This implies that the i^{th} column of H_k is an all-zero vector, which contradicts the definition of H_k, Thus, there are no codewords of weight 1.

Suppose that $\mathcal{H}(2,k)$ has a codeword $\boldsymbol{v}$ of weight 2 (with a 1 in the i^{th} and j^{th} positions). Let row r of H_k be represented as $[h_{r1}, h_{r2}, \dots, h_{rn}]$. Since v is orthogonal to each row of H_k, we have

$$h_{ri} + h_{rj} = 0, \;\; r = 1,2, \dots, k$$

which implies (noting that we are working in $\mathbb{F}_2$)

$$h_{ri} = h_{rj}, \;\; r = 1,2, \dots, k$$

Thus, columns i and j of H_k are equal, which is a contradiction. Thus, there are no codewords of length 2.

So far, we have shown that the minimum distance for $\mathcal{H}(2,k)$ is greater than 2. To show that the minimum distance is equal to 3, we only need to exhibit a codeword of weight 3. Since, by definition, the columns of H_k cover all possible distinct non-zero vectors in $\mathbb{F}_2^k$, H_k has the initial three columns as shown below (noting that we can rearrange the columns of H_k and still have an equivalent code by Theorem 62).

$$H_k = \begin{bmatrix} 0\,0\,0\,\dots \\ \vdots \\ 0\,0\,0\,\dots \\ 0\,1\,1\,\dots \\ 1\,0\,1\,\dots \end{bmatrix}$$

The codeword $1110 \dots 0$ is orthogonal to every row in H_k, and so, we have found a codeword of weight 3. ■

Using our generic notation, we can say that $\mathcal{H}(2,k)$ is a $\left(2^k - 1, 2^{(2^k-1)-k}, 3\right)_2$ which can correct for a single error in a word.

. . .

The **rate of a code** is the ratio of data bits to total bits in a codeword. For example, in an $\mathcal{H}(2,k)$ code, each codeword has a total of $2^k - 1$ bits of which $2^k - k - 1$ are data bits and k are parity bits. So, the rate for $\mathcal{H}(2,k)$ is

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

Table 9 shows the rates for the first several Hamming codes. While the rate does increase and quickly approach 1 as k increases, keep in mind that only one error can be corrected even as the word length gets larger. For example, $\mathcal{H}(2,9)$ has 502 data bits per each 511 bit message but can only correct one error in each word. So, there is a trade-off between increased rate, and the ratio of "errors than can be corrected" to "word size", which is $\frac{1}{2^9-1}$ for $\mathcal{H}(2,9)$.

Table 9. Rates for various Hamming codes

Parity bits (k)	Total bits ($2^k - 1$)	Data bits	Code	Rate
2	3	1	$\mathcal{H}(2,2)$	$\frac{1}{3} \cong 0.333$
3	7	4	$\mathcal{H}(2,3)$	$\frac{4}{7} \cong 0.571$
4	15	11	$\mathcal{H}(2,4)$	$\frac{11}{15} \cong 0.733$
5	31	26	$\mathcal{H}(2,5)$	$\frac{26}{31} \cong 0.839$
6	63	57	$\mathcal{H}(2,6)$	$\frac{57}{63} \cong 0.905$
7	127	120	$\mathcal{H}(2,7)$	$\frac{120}{127} \cong 0.945$
8	255	247	$\mathcal{H}(2,8)$	$\frac{247}{255} \cong 0.969$
9	511	502	$\mathcal{H}(2,9)$	$\frac{502}{511} \cong 0.982$

In the $\mathcal{S}(2,k)$ Simplex code each codeword has a total of $2^k - 1$ bits of which $2^k - k - 1$ are parity bits and k are data bits. So, the rate for $\mathcal{S}(2,k)$ is

$$\frac{k}{2^k - 1}$$

As k increases, the rate of $\mathcal{S}(2,k)$ decreases, but the number of errors that can be corrected increases. So, the Simplex codes are better suited for noisy communication environments where many errors are likely to occur.

. . .

Hamming codes can be defined over any finite field $\mathbb{F}_q$ (recall q is a prime number squared).

The following theorem is required for our development of the q-ary Hamming codes. However, it should be emphasized that that theorem is applicable to any linear code over a finite field $\mathbb{F}_q$.

Theorem 71. Let $\mathcal{C}$ be a linear code over $\mathbb{F}_q^n$ of dimension k and with parity-check matrix H. The minimum distance for the code is d if and only if any collection of $d-1$ columns from H are linearly independent and there exists d columns that are linearly dependent.

Proof: If the minimum distance for the code is d, then there is a codeword (vector) $\boldsymbol{x} = (x_1, x_2, \dots, x_n) \in \mathbb{F}_q^n$ of weight d. By definition of a parity-check matrix, $H\boldsymbol{x}^{\boldsymbol{T}} = 0$ which is equivalent to

$$x_1\boldsymbol{h_1} + x_2\boldsymbol{h_2} + \cdots + x_n\boldsymbol{h_n} = \boldsymbol{0}$$

where $\boldsymbol{h_i}$ is the i^{th} column of H, and $\boldsymbol{0}$ is a column vector of zeros. Since $\boldsymbol{x}$ is of weight d, we have that exactly d of the x_i elements are non-zero. This implies that the d columns of H corresponding to the d non-zero elements of $\boldsymbol{x}$ are linearly dependent (by the definition of linear dependence). If

there were $d-1$ linearly dependent columns of H, then there would exist $y=(y_1,y_2,\ldots,y_n)\in\mathbb{F}_q^n$ of weight $d-1$ such that

$$y_1\boldsymbol{h_1}+y_2\boldsymbol{h_2}+\cdots+y_n\boldsymbol{h_n}=\mathbf{0}$$

but this would contradict our assumption that d is the minimum weight of the code. So, any collection of $d-1$ columns of H must be linearly independent.

The above arguments can be reversed to prove the theorem in the other direction. ■

For a given minimum distance d, one can construct a set of vectors from $\mathbb{F}_q^k$, any $d-1$ of which are linearly independent, by adding vectors from $\mathbb{F}_q^k$, one at a time while at each step, while making sure that the additional vector is not a linear combination of any $d-2$ of the previous vectors in the set. This process leads to a q-ary linear code known as the Hamming code and denoted $\mathcal{H}(q,k)$. [Again, we warn the reader that the notation varies in the literature.]

For example, a parity check matrix for $\mathcal{H}(7,2)$ with $d=3$ is

$$\begin{bmatrix}1&0&1&1&1&1&1&1\\0&1&1&2&3&4&5&6\end{bmatrix}$$

In the above matrix, any $d-1=2$ columns are linearly independent (i.e., in this case, one is not a multiple of the other). However, there are sets consisting of 3 columns that are linearly dependent, e.g., column #3 is equal to column #1 plus column #2.

Taking $d=3$ again, a parity check matrix for $\mathcal{H}(3,3)$ is

$$\begin{bmatrix}1&0&0&0&0&1&1&1&1&1&1&1&1\\0&1&0&1&1&0&0&1&1&1&2&2&2\\0&0&1&1&2&1&2&0&1&2&0&1&2\end{bmatrix}$$

The difficulty of the general problem is summarized in the book by Hill [81], and paraphrased below.

> For $d=3$, it is reasonable to construct a set of vectors from $\mathbb{F}_q^k$, any $d-1=2$ of which are linearly independent, by writing down vectors from $\mathbb{F}_q^k$, one at a time, each time making sure that the new vector is not a multiple of any of the earlier ones. However, this approach is not practical for $d\geq 4$, since we are likely to run out of choices for the new vector at a relatively early stage. In fact, the problem of finding the maximum possible number of vectors in $\mathbb{F}_q^k$ such that any $d-1$ are linearly independent is extremely difficult for $d\geq 4$ and very little is known except for cases where $k\leq 4$. The problem is of much interest in other branches of mathematics, namely in finite geometries and in the theory of factorial designs in statistics.

6.3.2 Golay Codes

From the Wikipedia article on the **binary Golay code** [82]:

> In mathematics and electronics engineering, a binary Golay code is a type of linear error-correcting code used in digital communications. The binary Golay code, along with the ternary Golay code, has a particularly deep and interesting connection to the theory of finite sporadic groups in mathematics. These codes are named in honor of Marcel J. E. Golay whose

1949 paper [83] introducing them has been called, by E. R. Berlekamp, the "best single published page" in coding theory.

The binary Golay code was used in the Voyager 1 and 2 spacecrafts to encode and decode the general science and engineering data for the missions.

The binary version of the Golay code, denoted $\mathcal{G}(23)$, is a $(23, 2^{12}, 7)_2$ binary linear code consisting of $2^{12} = 4096$ codewords of length 23, and minimum distance $d = 7$. This code can correct $\frac{d-1}{2} = 3$ errors.

To construct the binary Golay code, we start with the following vector from $\mathbb{F}_2^{23}$

$$(1,1,0,0,0,1,1,1,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0)$$

We shift the above vector one position to the right and wrap-around (i.e., move the entry in position 23 to position 1) to get the following vector

$$(0,1,1,0,0,0,1,1,1,0,1,0,1,0,0,0,0,0,0,0,0,0,0)$$

We do the one-position shift a total of 11 times. The first vector along with the 11 other vectors forms a basis for a 12 dimensional subspace of $\mathbb{F}_2^{23}$. This gives us the following generator matrix for $\mathcal{G}(23)$

$$\begin{matrix}
1&1&0&0&0&1&1&1&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&1&1&0&0&0&1&1&1&0&1&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&1&1&0&0&0&1&1&1&0&1&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&1&1&0&0&0&1&1&1&0&1&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&1&1&0&0&0&1&1&1&0&1&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&1&1&0&0&0&1&1&1&0&1&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&1&1&0&0&0&1&1&1&0&1&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&1&1&0&0&0&1&1&1&0&1&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&1&1&0&0&0&1&1&1&0&1&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&1&1&0&0&0&1&1&1&0&1&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&1&1&1&0&1&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&1&1&1&0&1&0&1
\end{matrix}$$

If we add a parity check bit to each row such that each row adds to zero, then we get a generator matrix for something called the **extended binary Golay code**, denoted $\mathcal{G}(24)$. This is a $(24,2^{12},8)_2$ linear code. $\mathcal{G}(24)$ is a self-dual code, i.e., $\mathcal{G}(24) = \mathcal{G}(24)^{\perp}$.

6.4 Cyclic Codes

Cyclic codes are error-correcting codes that have algebraic properties that are convenient for efficient error detection and correction. They are a type of block code, where the circular shifts of each codeword gives another codeword. We already saw an example in the previous section concerning the Golay codes.

A cyclic code $\mathcal{C}$ has two basic properties:

- Linearity property: $\mathcal{C}$ is a subspace of a vector space V over a field F which is equivalent to $\mathcal{C}$ be closed under linear combinations, i.e., if $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and $a, b \in F$ then $a\boldsymbol{x} + b\boldsymbol{y} \in \mathcal{C}$.
- Cyclic property: Any cyclic shift of a code word in the code is also a code word.

In this section, we develop some of the theory underlying cyclic codes.

. . .

The mathematics behind cyclic codes relies on rings, ideals, quotient rings and fields.

Recall from Section 5.3.2, we discussed a polynomial ring over the real numbers. For the task at hand, we are going focus on the set of polynomials with coefficients from a finite field $\mathbb{F}_q$, denoted $\mathbb{F}_q(X)$. For example, the following is an element of $\mathbb{F}_q(X)$

$$f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_kX^k$$

where each coefficients (the a_i terms) are elements of $\mathbb{F}_q$.

Addition and multiplication is done in the same way as in elementary algebra, while taking into account that the coefficients obey the laws of modular arithmetic for $\mathbb{F}_q$. For example, consider the following two polynomials in $\mathbb{F}_2(X)$

$$f(X) = X^2 + 1$$

$$g(X) = X + 1$$

We have that

$$f(X) + g(X) = X^2 + X, \text{ noting that } 1 + 1 = 0 \text{ in } \mathbb{F}_2$$

$$f(X) * g(X) = X^3 + X^2 + X + 1$$

[In what follows, we will sometimes use juxtaposition to indicate polynomial multiplication, e.g., $f(X)g(X)$ instead of $f(X) * g(X)$.]

If we restrict the polynomials in $\mathbb{F}_q(X)$ to powers of X less than or equal to n, then we have an n-dimensional vector space, denoted by $\mathbb{F}_q^n(X)$. A vector $\boldsymbol{v}$ in $\mathbb{F}_q^n(X)$ can be presented either as an ordered list or as an n^{th} degree polynomial, i.e.,

$$\boldsymbol{v} = (a_0, a_1, a_2, \dots, a_n) \sim a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$$

Addition of vectors in this space is per component. Multiplication is based on polynomial multiplication as described in the previous example.

In this context, we can view the generator vector from the Golay code as a member of the vector space $\mathbb{F}_2^{23}(X)$. However, we still need to show mathematically how the other codewords are generated. So, let's continue to develop the necessary mathematics.

Given a pair of polynomials $a(X), b(X) \in \mathbb{F}_q(X)$, the **division algorithm** states that there exists a unique pair of polynomials ($q(X)$, the quotient, and $r(X)$, the remainder) such that

$$a(X) = q(X)b(X) + r(X), \qquad \deg(r(X)) < \deg(b(X))$$

For example, take $a(X) = X^3 + X^2 + X + 1$ and $b(X) = X^3 + 1$ in $\mathbb{F}_2(X)$. Using long division of polynomials, we have

$$X^3 + X^2 + X + 1 = 1 \cdot (X^3 + 1) + (X^2 + X)$$

So, $q(X) = 1$ and $r(X) = X^2 + X$.

For a fixed polynomial $f(X) \in \mathbb{F}_q$, we can form the quotient ring $\mathbb{F}_q(X)/(f)$ which is read as "$\mathbb{F}_q(X)$ modulo the principal ideal (f)" or just "$\mathbb{F}_q(X)$ modulo $f(X)$".

Polynomials $a(X), b(X) \in \mathbb{F}_q(X)$ are said to be congruent modulo $f(X)$, written as $a(X) \equiv b(X) (\mathrm{mod}\ f(X))$, if $a(X) - b(X)$ is exactly divisible by $f(X)$. For example, consider the quotient ring $\mathbb{F}_2(X)/(X^3 + 1)$, and polynomials $a(X) = X^4 + X$ and $b(X) = X^2 + X + 1$. $a(X)$ and $b(X)$ are congruent modulo $X^3 + 1$ since

$$a(X) - b(X) = X^4 + X = (X^3 + 1)X$$

Let $f(X)$ be a polynomial of degree n in $\mathbb{F}_q(X)$. By the division algorithm, the elements of the quotient ring $\mathbb{F}_q(X)/(f)$ are all the polynomials in $\mathbb{F}_q(X)$ of degree less than n. Thus, the number of elements in $\mathbb{F}_q(X)/(f)$ is q^n.

For $a(X), b(X) \in \mathbb{F}_q(X)/(f)$, the sum $a(X) + b(X)$ in $\mathbb{F}_q(X)/f(X)$ is the same as the sum in $\mathbb{F}_q(X)$, because $\deg\big(a(X) + b(X)\big) \leq \deg(f(X))$. The product $a(X)b(X)$ in $\mathbb{F}_q(X)/(f)$ is the unique polynomial of degree less than $\deg(f(X))$ to which $a(X)b(X)$ is congruent modulo $f(X)$.

For example, take $a(X) = X^2 + 1$ and $b(X) = X^3 + X^2 + 1$, and let $f(X) = x^5 + 1$. In $\mathbb{F}_2(X)/(f)$, we have

$$a(X) + b(X) = X^3 + 2X^2 + 2 \equiv X^3 (\mathrm{mod}\ f(X))$$

$$a(X) * b(X) = (X^2 + 1)(X^3 + X^2 + 1) = X^5 + X^4 + X^3 + 2X^2 + 1$$

$$= X^5 + X^4 + X^3 + 1 \equiv (X^4 + X^3) (\mathrm{mod}\ f(X))$$

where the last equality on the right is gotten from taken the remainder when $X^5 + 1$ is divided into $X^5 + X^4 + X^3 + 1$, or we could just notice that $X^5 + 1 \equiv 0 (\mathrm{mod}\ f(X))$

. . .

The following theorem defines an equivalence for cyclic codes.

Theorem 72. A code $\mathcal{C}$ in $\mathbb{F}_q^n(X)$ is a cyclic code if and only if $\mathcal{C}$ satisfies the following conditions:

- $a(X), b(X) \in \mathcal{C}$ implies $a(X) + b(X) \in \mathcal{C}$
- $a(X) \in \mathcal{C}$ and $r(X) \in \mathbb{F}_q^n(X)$ implies $r(X)a(X) \in \mathcal{C}$

Proof: See Theorem 12.6 in Hill [81]. ■

While we don't need the following theorem for our analysis of cyclic error correcting codes, it is an important result in abstract algebra, and so, we state it here for the reader's reference.

Theorem 73. The quotient ring $\mathbb{F}_q(X)/(f)$ is a field if and only if $f(X)$ is irreducible over $\mathbb{F}_q(X)$ (i.e., $f(X)$ cannot be factored into polynomials of lesser degree).

Proof: A brief proof is given in the Wikipedia article on polynomial rings [84]. ■

. . .

For the task at hand (i.e., modeling cyclic codes via abstract algebra), we focus on quotient rings of the form $\mathbb{F}_q(X)/(X^n-1)$.

The fact that $X^n \equiv 1 (\mathrm{mod}\ X^n-1)$ gives us a shortcut to reduce polynomials modulo X^n-1, i.e., we can replace X^n by 1, X^{n+1} by X, …, X^{2n-1} by X^{n-1}, X^{2n} by 1, and so on.

For example, consider $\mathbb{F}_2(X)/(X^3-1)$ and the polynomial $a(X)=X^7+X^5+X^4+1$. [Note that $X^3-1=X^3+1$ when working in $\mathbb{F}_2$.] Using our shortcut, we can reduce $a(X)$ to $X+X^2+X+1=X^2+1$. As a check, we can use long division of polynomials to get

$$X^7+X^5+X^4+1=(X^3+1)(X^4+X^2)+(X^2+1)$$

which implies

$$X^7+X^5+X^4+1 \equiv (X^2+1)(\mathrm{mod}\ X^3+1)$$

. . .

We can now state the criterion for a code to be cyclic using the terminology of abstract algebra.

Theorem 74. A code $\mathcal{C}$ in $\mathbb{F}_q(X)/(X^n-1)$ is a cyclic code if and only if $\mathcal{C}$ is an ideal.

Proof: First, note that $\mathbb{F}_q(X)/(X^n-1)$ is a commutative ring, and so, we don't need to deal with left and right ideals.

If $\mathcal{C}$ is a cyclic code in $\mathbb{F}_q(X)/(X^n-1)$, then $\mathcal{C}$ has the linearity and cyclic properties stated in the definition of cyclic codes at the beginning of this section. As defined in Section 5.3.3, an ideal must be closed and have the absorption property.

The linearity property implies that $\mathcal{C}$ is closed under addition and thus, a subgroup of the associated ring by Theorem 33.

To show the absorption property, take $a(X) \in \mathcal{C}$ and any $b(X)=b_{n-1}X^{n-1}+\cdots+b_1X+b_0 \in \mathbb{F}_q(X)/(X^n-1)$. Multiplying $b(X)$ by X yields

$$b_{n-1}X^n+b_{n-2}X^{n-1}+\cdots+b_1X^2+b_0X=b_{n-2}X^{n-1}+\cdots+b_1X^2+b_0X+b_{n-1} \in \mathcal{C}$$

where we use the fact that $X^n \equiv 1 (\mathrm{mod}\ X^n-1)$. Multiplication by X produces a cyclic shift, which implies multiplication by X^m cyclically shifts a code m places and produces another code in $\mathcal{C}$. So, $X^i * b(X) \in \mathcal{C}$ for $i=1,2,\ldots,n-1$. Further, by the linearity property, $\mathcal{C}$ is closed under scalar multiplication which implies $b_iX^{i-1} * a(X) \in \mathcal{C}$. Since $\mathcal{C}$ is closed under addition, we have

$$b(X) * a(X)=b_{n-1}X^{n-1} * a(X)+\cdots+b_1X * a(X)+b_0 * a(X) \in \mathcal{C}$$

Going in the other direction, assume $\mathcal{C}$ is an ideal in $\mathbb{F}_q(X)/(X^n-1)$, i.e., $\mathcal{C}$ has the closure and absorption properties. Consider $au(X)+bv(X)$ where $u(X), v(X) \in \mathcal{C}$ and $a, b \in \mathbb{F}_q$. By the absorption property, $au(X), bv(X) \in \mathcal{C}$ and by closure, $au(X)+bv(X) \in \mathcal{C}$. Thus, $\mathcal{C}$ has the linearity property.

For any $u(X) \in \mathcal{C}$, $X * u(X) \in \mathcal{C}$ by the absorption property and so, $\mathcal{C}$ has the cyclic property. ■

In what follows, our focus will be on principle ideal within quotient rings of the form $\mathbb{F}_q(X)/(X^n-1)$. Recall that a principal ideal is generated by one element of the associated ring. For $f(X) \in \mathbb{F}_q(X)/(X^n-1)$, the principal ideal generated by $f(X)$ is the set of all multiples of $f(X)$ by elements in $\mathbb{F}_q(X)/(X^n-1)$, i.e.,

$$\langle f(X)\rangle = \{a(X)f(X): a(X) \in \mathbb{F}_q(X)/(X^n-1)\}$$

Such principal ideals generate cyclic codes.

For example, take $f(X)=X+1$ in $\mathbb{F}_2(X)/(X^4-1)$. To determine the cycle code generated by $f(X)$, we need to multiple $f(X)$ by each element in $\mathbb{F}_2(X)/(X^4-1)$, i.e., all possible polynomials of the form $g(X)=aX^3+bX^2+cX+d$ where $a, b, c, d \in \mathbb{F}_2$. Keep in mind that $X^4=1$ and we are working in $\mathbb{F}_2$. The generated code is shown in Table 10. The codewords are generated twice. A shift in any of the codewords produces another codeword, and thus, we have a cyclic code.

Table 10. Cyclic code generated by X+1

Polynomial $g(X)$ from $\mathbb{F}_2(X)/(X^4-1)$	$g(X) * f(X)$	Codeword
0	0	(0,0,0,0)
1	$X+1$	(0,0,1,1)
X	X^2+X	(0,1,1,0)
$X+1$	X^2+1	(0,1,0,1)
X^2	X^3+X^2	(1,1,0,0)
X^2+1	X^3+X^2+X+1	(1,1,1,1)
X^2+X	X^3+X	(1,0,1,0)
X^2+X+1	X^3+1	(1,0,0,1)
X^3	X^3+1	(1,0,0,1)
X^3+1	X^3+X	(1,0,1,0)
X^3+X	X^3+X^2+X+1	(1,1,1,1)
X^3+X+1	X^3+X^2	(1,1,0,0)
X^3+X^2	X^2+1	(0,1,0,1)
X^3+X^2+1	X^2+X	(0,1,1,0)
X^3+X^2+X	$X+1$	(0,0,1,1)
X^3+X^2+X+1	0	(0,0,0,0)

The next theorem says that any cyclic code can be generated by a unique monic polynomial of smallest degree. [A monic polynomial is a polynomial whose leading coefficient is 1.]

Theorem 75. If $\mathcal{C}$ is a non-zero cyclic code in $\mathbb{F}_q(X)/(X^n-1)$, then

i. there exist a unique non-zero, monic polynomial of smallest degree in $\mathcal{C}$ such that $\mathcal{C} = \langle g(X) \rangle$

ii. $g(X)$ is a factor of $X^n - 1$

Proof: Assume that $\mathcal{C}$ has two monic polynomials of smallest degree, i.e., $f(X)$ and $g(X)$. By the linearity property, $h(X) = f(X) - g(X) \in \mathcal{C}$. However, the highest degree terms of $f(X)$ and $g(X)$ cancel, and so, $h(X)$ is of smallest degree (a contradiction). Thus, there can be only one non-zero, monic polynomial in $\mathcal{C}$ of least degree which we designate as $g(X)$.

Next, we show that $a(X) \in \langle g(X) \rangle$ for every codeword $a(X) \in \mathcal{C}$. By the division algorithm, we can write $a(X) = q(X)g(X) + r(X)$ where the degree of $r(X)$ is less than the degree of $g(X)$. By the properties noted in Theorem 72, $r(X) = a(X) - q(X)g(X) \in \mathcal{C}$. But $g(X)$ is of minimal degree out of all the non-zero elements in $\mathcal{C}$, and so, $r(X) = 0$ which implies $a(X) = q(X)g(X)$, i.e., $a(X) \in \mathcal{C}$. So, $\mathcal{C} = \langle g(X) \rangle$.

By the division algorithm, $X^n - 1 = q_1(X)g(X) + r_1(X)$ where the degree of $r_1(X)$ is less than the degree of $g(X)$. However, $r_1(X) \equiv -q_1(X)g(X) (\text{mod } X^n - 1)$ which implies $r_1(X) \in \langle g(X) \rangle = \mathcal{C}$. Since we have already proven that $g(X)$ is the non-zero polynomial of minimum degree in $\mathcal{C}$, it must that that $r_1(X) = 0$. Thus, $g(X)$ is a factor of $X^n - 1$. ■

The monic polynomial of least degree for a non-zero cyclic code $\mathcal{C}$ is known as the **generator polynomial** for $\mathcal{C}$.

Using the result of Theorem 75, we can determine all cyclic codes of length n by factoring $X^n - 1$ into irreducible monic polynomials.

For example, consider $\mathbb{F}_2(X)/(X^6 - 1)$. We can factor $X^6 - 1$ as $(X^3 + 1)(X^3 + 1)$, noting that $-1 \equiv 1 \pmod 2$. In $\mathbb{F}_2$, $X^3 + 1$ has one root, i.e., 1. Factoring the $(X + 1)$ term out of $X^3 + 1$, we get $X^3 + 1 = (X + 1)(X^2 + X + 1)$. In $\mathbb{F}_2$, $X^2 + X + 1$ has no roots, i.e., it is irreducible. Thus, $X^6 - 1$ written in terms of irreducible factors is equal to $(X + 1)^2(X^2 + X + 1)^2$.

This gives us a total of nine factors of $X^6 - 1$ (see the table below), each one of which generates a cyclic code.

1	$(X + 1)$	$(X + 1)^2$
$(X^2 + X + 1)$	$(X^2 + X + 1)(X + 1)$	$(X^2 + X + 1)^2$
$(X^2 + X + 1)(X + 1)^2$	$(X^2 + X + 1)^2(X + 1)^2$	$X^6 - 1$

We can generate the associated cyclic codes by brute force, as we did for the example associated with Table 10, but there is an easier way which makes use of Theorem 77. We first need the following preliminary result.

Theorem 76. If $g(X) = g_0 + g_1X + \cdots + g_kX^k$ is the generator polynomial for a cyclic code $\mathcal{C}$ in $\mathbb{F}_q(X)/(X^n - 1)$, then $g_0 \neq 0$.

Proof: Assume $g_0 = 0$, and consider $X^{n-1}g(X) \in \mathcal{C}$. Noting that $X^n = 1$, we have

$$X^{n-1}g(X) = g_1X^n + g_2X^{n+1} + \cdots + g_kX^{n+k-1}$$
$$= g_1 + g_2X + \cdots + g_kX^{k-1}$$

Thus, we have shown there is a polynomial of degree less than $g(X)$ in $\mathcal{C}$, which is a contradiction. So, our initial assumption is false and it must be that $g_0 \neq 0$. ■

Theorem 77. If $\mathcal{C}$ is a cyclic code in $\mathbb{F}_q(X)/(X^n - 1)$ with generator $g(X) = g_0 + g_1X + \cdots + g_kX^k$, then the dimension of $\mathcal{C}$ is $n - k$. Further, $\mathcal{C}$ has an $(n - k) \times n$ generator matrix G given by

g_0	g_1	g_2	g_3	…	g_k	0	0	0	…	0
0	g_0	g_1	g_2	…	g_{k-1}	g_k	0	0	…	0
0	0	g_0	g_1	…	g_{k-2}	g_{k-1}	g_k	0	…	0
…	…	…	…	…	…	…	…	…	…	…
0	0	0	…	0	g_0	g_1	…	g_{k-2}	g_{k-1}	g_k

[Author's Remark: The above matrix is not drawn exactly. The idea is that you start with the first row as shown, with the coefficients of $g(X)$ followed by $n - k$ zeros. Each row is shifted one position to the right relative to the previous row. This process is repeated until you have $n - k$ rows.**]**

Proof: From Theorem 76, we know that $g_0 \neq 0$ and so, matrix G is in upper echelon form which, in turn, implies that the rows of G are linearly independent.

The $n - k$ rows of G are all codewords in $\mathcal{C}$ as each row is a cyclic shift of $g(X)$, i.e.,

$$g(X), Xg(X), X^2g(X), \ldots, X^{n-k-1}g(X)$$

To complete the proof, we need to show that every codeword in $\mathcal{C}$ can be expressed as a linear combination of the above polynomials (vectors).

By Theorem 75, every codeword $a(X) \in \mathcal{C}$ can be represented in the form $a(X) = b(X)g(X)$ where $b(X) \in \mathbb{F}_q(X)/(X^n - 1)$. Since the degree of $a(X)$ is less than or equal to $n - 1$, then $\deg b(x) \leq n - k - 1$. So, we have

$$b(X)g(X) = \left(b_0 + b_1X + \cdots + b_{n-k-1}X^{n-k-1}\right)g(X)$$

$$= b_0g(X) + b_1Xg(X) + \cdots + b_{n-k-1}X^{n-k-1}g(X)$$

Thus, any codeword in $\mathcal{C}$ can be written as a linear combination of the rows of G (represented as polynomials). ■

We now apply Theorem 77, to our example concerning $\mathbb{F}_2(X)/(X^6 - 1)$. If we take $g(X) = X^2 + X + 1$ as a generator polynomial, then have the following generator matrix for a cyclic code

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

The code generated by the above matrix consists of all possible linear combinations of the rows. In addition to the 4 codewords in the generator matrix, we have the following codewords

$$\begin{array}{ll} 1\,0\,0\,1\,0\,0 & 1\,0\,1\,0\,1\,0 \\ 1\,1\,0\,1\,1\,0 & 1\,1\,0\,0\,0\,1 \\ 1\,1\,1\,1\,1\,1 & 1\,0\,0\,0\,1\,1 \\ 0\,1\,0\,0\,1\,0 & 0\,1\,0\,1\,0\,1 \\ 0\,1\,1\,0\,1\,1 & 1\,0\,1\,1\,0\,1 \\ 0\,0\,1\,0\,0\,1 & 0\,0\,0\,0\,0\,0 \end{array}$$

While a cyclic shift of one position for any of the above codes does lead to another codeword in the list, there is more than one cycle. The codewords $(1\ 1\ 1\ 1\ 1\ 1)$ and $(0\ 0\ 0\ 0\ 0\ 0)$ each generate a cycle of length 1. The are 4 other cycles are

$$\{(1\ 0\ 1\ 0\ 1\ 0), (\ 0\ 1\ 0\ 1\ 0\ 1)\}$$

$$\{(1\ 1\ 0\ 1\ 1\ 0), (0\ 1\ 1\ 0\ 1\ 1), (1\ 0\ 1\ 1\ 0\ 1)\}$$

$$\{(1\ 0\ 0\ 1\ 0\ 0), (0\ 1\ 0\ 0\ 1\ 0), (0\ 0\ 1\ 0\ 0\ 1)\}$$

$$\{(1\ 1\ 1\ 0\ 0\ 0), (0\ 1\ 1\ 1\ 0\ 0), (0\ 0\ 1\ 1\ 1\ 0), (0\ 0\ 0\ 1\ 1\ 1), (1\ 0\ 0\ 0\ 1\ 1), (1\ 1\ 0\ 0\ 0\ 1)\}$$

Let's try a different generator, i.e.,

$$g(X) = (X^2 + X + 1)^2 = X^4 + X^3 + X^2 + X^3 + X^2 + X + X^2 + X + 1$$

$$= X^4 + X^2 + 1$$

This leads to the generator matrix

$$\begin{bmatrix} 1\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,1 \end{bmatrix}$$

which generates the additional two codewords:

$$0\ 0\ 0\ 0\ 0\ 0$$

$$1\ 1\ 1\ 1\ 1\ 1$$

So, we have two cycles of length 1, and one cycle of length 2.

. . .

The **check polynomial of a cyclic code** $\mathcal{C}$ generated by $g(X)$in $\mathbb{F}_q(X)/(X^n - 1)$ is (by definition) the unique non-zero polynomial $h(X)$ satisfying $g(X)h(X) = X^n - 1$.

From the previous example, the check polynomial for the code generate by $g(X) = (X^2 + X + 1)^2$ is $h(X) = (X + 1)^2 = X^2 + 1$. The generator matrix associated with $h(X)$ is

$$\begin{bmatrix} 1\,0\,1\,0\,0\,0 \\ 0\,1\,0\,1\,0\,0 \\ 0\,0\,1\,0\,1\,0 \\ 0\,0\,0\,1\,0\,1 \end{bmatrix}$$

The following theorem describes the general relationship between the codewords associated with a generator polynomial $g(X)$ and the associated check polynomial $h(X)$.

Theorem 78. Let $\mathcal{C}$ be a cyclic code in $\mathbb{F}_q(\boldsymbol{X})/(\boldsymbol{X^n-1})$ with generator polynomial $\boldsymbol{g(X)}$ and the associated check polynomial $\boldsymbol{h(X)}$. A polynomial $\boldsymbol{c(X)} \in \mathbb{F}_q(\boldsymbol{X})/(\boldsymbol{X^n-1})$ is a codeword of $\mathcal{C}$ if and only if $\boldsymbol{c(X)h(x)=0}$.

Proof: By definition of $h(X)$, if $g(X)$ is of degree $n-k$ then $h(X)$ is of degree k. Further, in $\mathbb{F}_q(X)/(X^n-1)$, $g(X)h(X)=X^n-1=0$.

Going in one direction, assume $c(X) \in \mathcal{C}$ which implies $c(X)=a(X)g(X)$ for $a(X) \in \mathbb{F}_q(X)/(X^n-1)$. Multiplying to $h(X)$, gives us

$$c(X)h(X)=a(X)g(X)h(X)=a(X)*0=0$$

Going in the other direction, assume $c(X)h(X)=0$. Applying the division algorithm, we have that

$$c(X)=q(X)g(X)+r(X)$$

where $\deg r(X) < \deg g(X) = n-k$. Multiplying both sides of the above equation by $h(X)$, and noting that $c(X)h(X)=0$ and $g(X)h(X)=0$, we have that $r(X)h(X)=0$.

Since $\deg\big(r(X)h(X)\big) < (n-k)+k=n$, $r(X)h(X)=0$ in $\mathbb{F}_q(X)$. To finish the proof, we need a result from abstract algebra, i.e., a polynomial ring over a field, e.g., $\mathbb{F}_q(X)$, is an integral domain and as such, does not have zero divisors [85]. Since $h(X) \neq 0$, it must be that $r(X)=0$ since $\mathbb{F}_q(X)$, being an integral domain, does not have zero divisors. Thus, $c(X)=q(X)g(X)$ and $c(X) \in \mathcal{C}$. ■

The following theorem provides a way to compute the parity check matrix for a cyclic code, and also describes how to determine the generator polynomial for the dual of a cyclic code.

Theorem 79. If $\mathcal{C}$ is a cyclic code in $\mathbb{F}_q(\boldsymbol{X})/(\boldsymbol{X^n-1})$ which check polynomial $\boldsymbol{h(X)=h_0+h_1X+\cdots+h_kX^k}$ then a parity-check matrix for $\mathcal{C}$ is given by

h_k	h_{k-1}	h_{k-2}	h_{k-3}	…	h_0	0	0	0	…	0
0	h_k	h_{k-1}	h_{k-2}	…	h_1	h_0	0	0	…	0
0	0	h_k	h_{k-1}	…	h_2	h_1	h_0	0	…	0
…	…	…	…	…	…	…	…	…	…	…
0	0	0	…	0	h_k	h_{k-1}	…	h_2	h_1	h_0

and $\mathcal{C}^{\perp}$ is generated by $\boldsymbol{h'(X)=h_0^{-1}\big(h_k+h_{k-1}X+\cdots+h_1X^{k-1}+h_0X^k\big)=h_0^{-1}h_k+h_0^{-1}h_{k-1}X+\cdots+h_0^{-1}h_1X^{k-1}+X^k}$.

Proof: See Theorem 12.15 in the book by Hill [81]. ■

From our example concerning $\mathbb{F}_2(X)/(X^6-1)$ with code $\mathcal{C}$ generate by $g(X)=(X^2+X+1)^2$ and check polynomial $h(X)=(X+1)^2=X^2+1$, the parity-check matrix for $\mathcal{C}$ is

$$H=\begin{bmatrix}1&0&1&0&0&0\\0&1&0&1&0&0\\0&0&1&0&1&0\\0&0&0&1&0&1\end{bmatrix}$$

The reader is invited to check that each row of the generator matrix for $\mathcal{C}$ is orthogonal to every row of H.

The generator polynomial for the dual of $\mathcal{C}$ is $h'(X)=X^2+1$.

7 Geometric Packing Problems

The study of mathematics is, if unprofitable, a perfectly innocent and harmless occupation.
G.H. Hardy

7.1 Overview

Mathematical packing problems are a class of optimization problems that involve the packing (placement) of objects into containers. The goal is to either pack a single container as densely as possible with a given set of objects, or to pack (place) all objects in a given set using as few containers as possible. "Density" is defined as the area (or volume) of the items being placed in a given contained divided by the area (or volume) of the container.

- In some situations, all the objects are identical in terms of geometry (e.g., circles of the same radius), and in other cases, the objects are of different shapes and/or sizes.
- Usually, but not always, it is assumed that the objects do not overlap. In theoretical problems, the objects are sometimes allowed to overlap.
- In yet other situations, the task is to fill an infinite space as efficiently as possible with a given type of object (or set of object types). When the space is to be filled exactly (i.e., with no gaps or overlaps), we call the covering a tiling or tessellation [86].

Many of these problems can be related to real-life packaging, storage and transportation issues, e.g.,

- Bin Packing Problem: In this type of problem, a set of objects of different sizes are to be packed into a limited number of containers (bins) of fixed capacity. The goal is to minimize the number of bins used or to maximize the space utilization. As noted in the Wikipedia article on bin packing [87], the problem has many applications, such as filling up containers, loading trucks with weight capacity constraints, creating file backups in media, and technology mapping in Field-Programmable Gate Array (FPGA) semiconductor chip design.
- Knapsack Problem (see [88] and Section 11.2.1 of [1]): Given a set of items, each with a weight and a value, determine which items to include in the collection so that the total weight is less than or equal to a given limit and the total value is as large as possible.
- The cutting-stock problem [89] is the problem of cutting standard-sized pieces of stock material, such as paper rolls or sheet metal, into pieces of specified sizes while minimizing wasted material. There is also the related strip packing problem [90], i.e., given a set of axis-aligned rectangles and a strip of bounded width and indetermined height, determine an overlapping-free packing of the rectangles into the strip minimizing its height.

Industrial packing problems, such as those listed above, are typically solved using the following algorithmic techniques:

- Integer programming (see Section 11 of [1]): This is a mathematical technique for solving optimization problems with integer variables. Integer programming can be used to solve a variety of packing problems, including bin packing and strip packing.
- Heuristics: Heuristics are algorithms that do not guarantee to find the optimal solution, but can often find good solutions quickly. There are a number of heuristics that have been developed for solving packing problems.

- Metaheuristics: Metaheuristics are a class of algorithms that use heuristics to search for solutions to optimization problems. Metaheuristics can be used to solve a variety of packing problems, including bin packing and strip packing.

The choice of technique depends on the specific packing problem and the desired level of accuracy. For small problems, an exact algorithm or closed analytic solution may be feasible. For large industrial problems, a heuristic or metaheuristic algorithm may be the best option.

In addition to the techniques listed above, there are a number of other techniques that can be used to solve packing problems. These techniques include:

- Geometric techniques: These techniques use the geometry of the objects to solve the packing problem. For example, the no-fit polygon (NFP) approach [91] is a method for solving two-dimensional packing problems. It is based on the concept of the no-fit polygon, which is the set of all possible positions at which one polygon can be placed without overlapping with another polygon.
- Probabilistic techniques: These techniques use probability to solve the packing problem. For example, Monte Carlo simulation [92] can be used to generate a large number of random packings and then select the best one.

The choice of technique depends on the specific packing problem and the desired level of accuracy. For small problems, a geometric technique may be feasible. For larger problems, a probabilistic technique may be the best option.

In what follows, the focus is on theoretical packing problems.

7.2 Packing in 2-dimensional containers

7.2.1 Circle Packing

7.2.1.1 Packing circles in a circle

Of the theoretical packing problems, those that involve the packing of congruent circles into a larger circle are among the most studied. Although dated, the paper by Graham, et al [93] provides a good summary of the early work on this topic (from its beginnings in the 1960s until the publication of said paper in 1998).

As our first example, consider the problem of packing three congruent circles into a unit circle. The goal is for the three congruent circles to be as large as possible. Construction of the optimal solution is shown in Figure 78. On the right, we divided the unit circle into 6 equal sectors. In the middle diagram, we place three circles as shown. Each circle is bisected by one of the three lines. Uniformly expand the circles until they touch, as shown on the right of the figure.

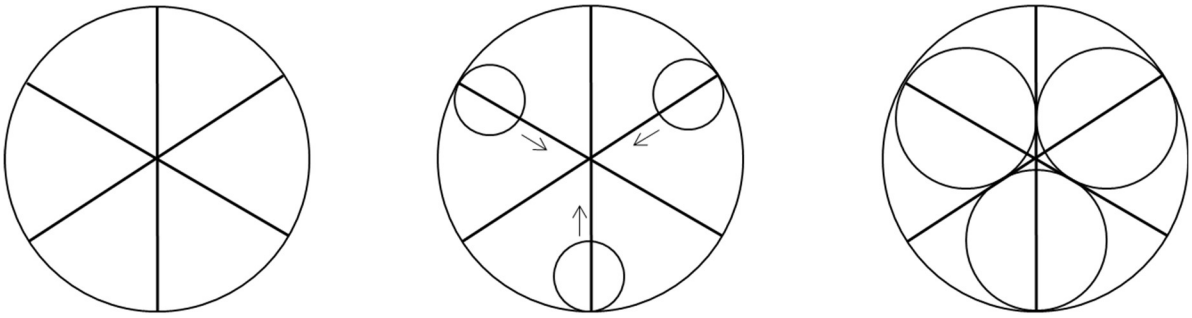


Figure 78. Optimal packing of 3 congruent circles into a unit circle

Let r be the radius of each of the interior circles. We want to compute the value of r. To that end, create an equilateral triangle whose vertices are the centers of the interior circles, as shown in Figure 79. Each side of the triangle is of length $2r$. The dashed lines within the triangle are the perpendicular bisectors of each side (going from the midpoint of one side to the opposite side). By the symmetry of the construction, the intersection of the three perpendicular bisectors is the center of the larger circle.

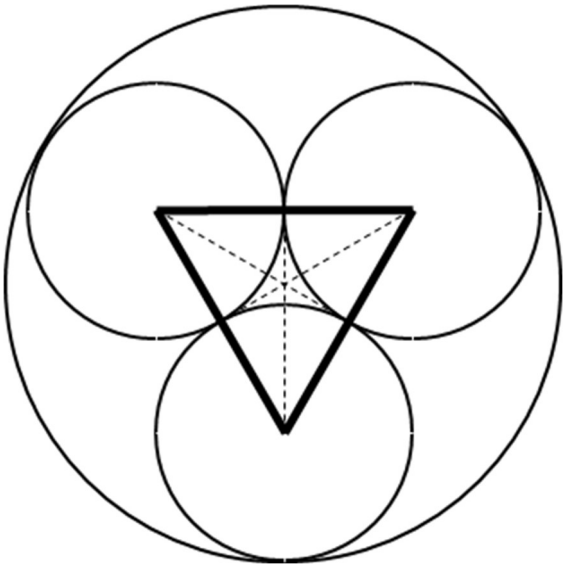
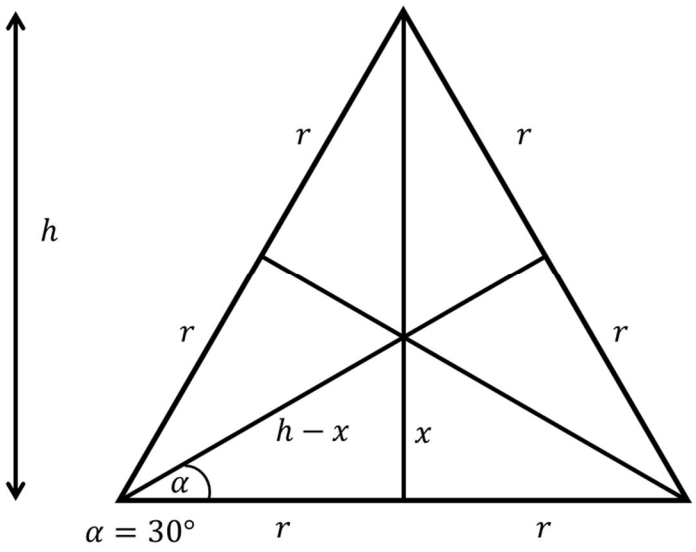


Figure 79. Equilateral triangle whose vertices are the centers of the interior circles

Next, we compute the value of the radius r of the smaller circles. By assumption, the radius of the outer circle is 1. Using the calculations shown in Figure 80, and keeping in mind the position of the triangle in Figure 79, we have

$$1 = r + (h - x) = r + \frac{2\sqrt{3}}{3}r = r\left(1 + \frac{2\sqrt{3}}{3}\right)$$

$$r \cong 0.4641$$



$$h^2 + r^2 = (2r)^2$$
$$h = r\sqrt{3}$$

$$x = r\tan(30°) = \frac{\sqrt{3}}{3}r$$

$$h - x = \frac{2\sqrt{3}}{3}r$$

Figure 80. Calculation of r

. . .

Figure 81 shows the optimal packing of 4 congruent circles into a unit circle. If we let the radius of the smaller circles be r, then we have $1 = r + r\sqrt{2}$ which implies

$$r = \frac{1}{1+\sqrt{2}} \cong .4142$$

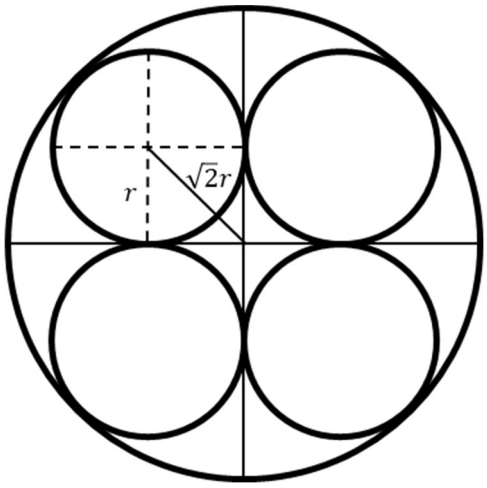


Figure 81. Optimal packing of 4 congruent circles into a unit circle

. . .

Figure 82 depicts the optimal packing of 5 congruent circles into a unit circle. The interior pentagon has vertices corresponding to the 5 interior circles. If we let x be the radius of each interior circle, then each side of the pentagon is of length $2x$. We will use the properties of pentagons to find x.

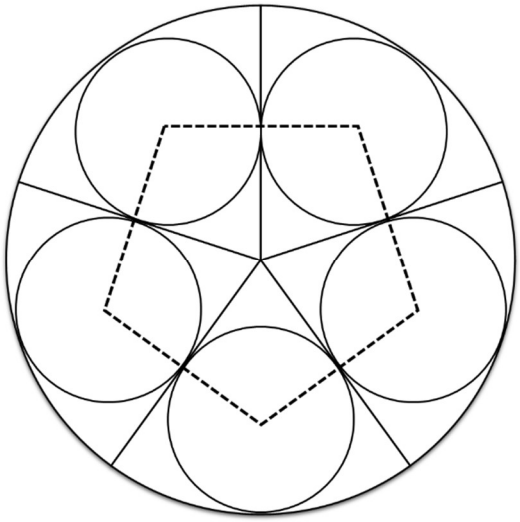


Figure 82. Optimal packing of 5 congruent circles into a unit circle

The perpendicular bisectors of each edge of a regular pentagon, and the bisectors of each angle meet at one interior point. Figure 83 shows the regular pentagon from Figure 82 in isolation. The formula for R is as shown in the figure. (Formulas for the various components of a regular pentagon can be found in the Wikipedia article on pentagons [94].) For the problem at hand, we see that the radius of the unit circle is equal to R plus the radius x of an interior circle. So, we have

$$1 = R + x = 2x\sqrt{\frac{5+\sqrt{5}}{10}} + x = x\left(2\sqrt{\frac{5+\sqrt{5}}{10}} + 1\right)$$

$$x = \frac{1}{\left(2\sqrt{\frac{5+\sqrt{5}}{10}} + 1\right)} \cong .370192$$

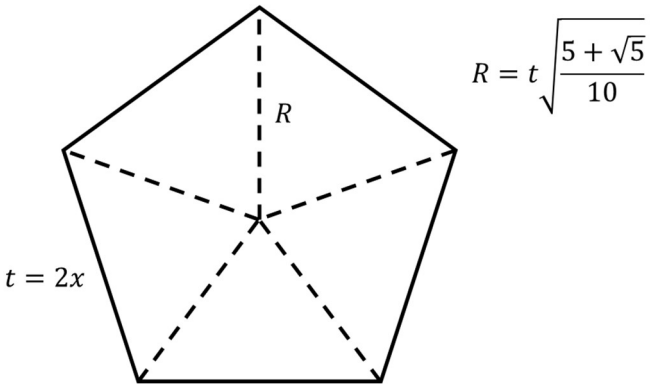


Figure 83. Circumradius of a pentagon

The previous examples may give the impression that the packing of smaller circles into a unit circle follows some regular pattern but that is far from true. In fact, optimal solutions are only known for cases 1-13 and 19. For all other cases that have been studied, only "best known" solutions are available. The website www.packomania.com has a subpage (http://hydra.nat.uni-magdeburg.de/packing/cci/cci.html) that shows the best known packings of equal circles into a circle up to 2600.

Figure 84 shows circle packing for cases 37, 38 and 39. The dark circles in cases 38 and 39 are free standing, i.e., not touching any other circle. The marking within each circle indicates the number tangencies with surrounding circles (including the outer circle).

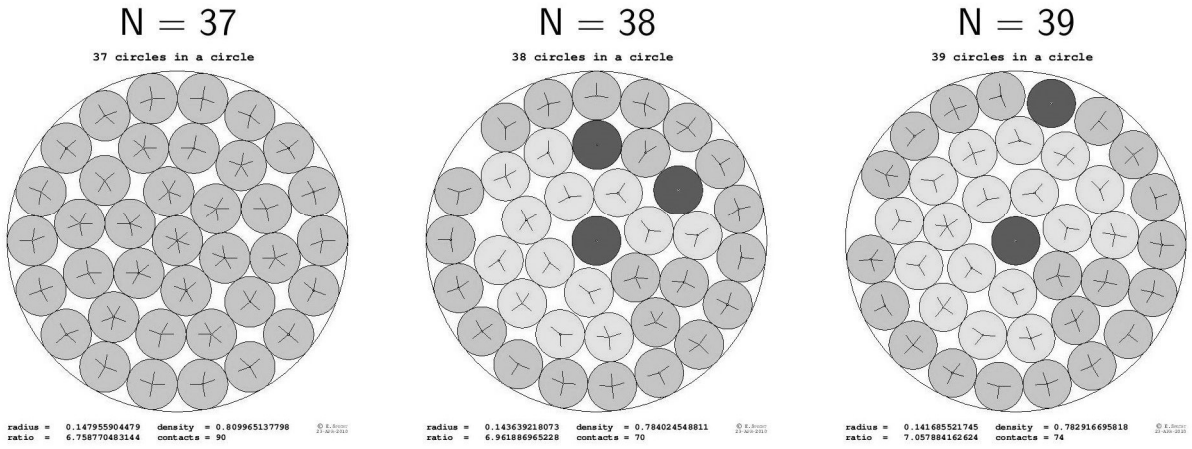


Figure 84. Circle packing for cases 37, 38 and 39

Only 26 optimal packings are thought to be rigid (with no circles free to "rattle" about in the configuration) [95].

- Proven for $n = 1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 19$
- Conjectured for $n = 14, 15, 16, 17, 18, 22, 23, 27, 30, 31, 33, 37, 61, 91$

7.2.1.2 Packing circles in a squares and other shapes

Consider the problem of packing n of congruent circles into a square of a given size. The goal is to make the diameter of the circles as large as possible and still fit n of them into the square. Further, consider the problem of arranging n points in a square so that the minimum distance between any two points is maximized. The two problems are equivalent, since if a collection of points in a unit square are at a distance of at least d from each other, the points can serve as the centers of a collection of circles of diameter d that will pack into a square with side length $1 + d$.

[The two equivalent problems also pertain to circles with circles, and other similar packing problems.]

The optimal solutions for the two variations of the problem for the case $n = 5$ are depicted in Figure 85. The distance between the centers of adjacent circles is d.

Consider the shaded triangle in the figure. By the Pythagorean theorem, we have $\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = d^2$ and so, $d = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}$.

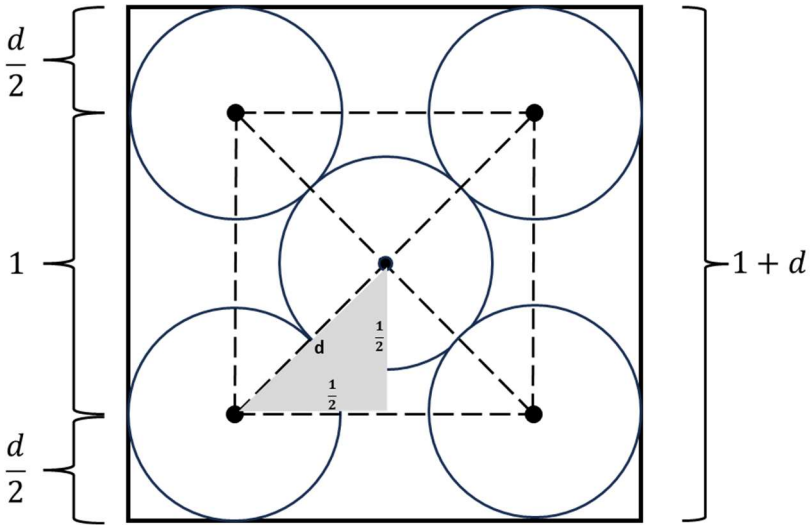
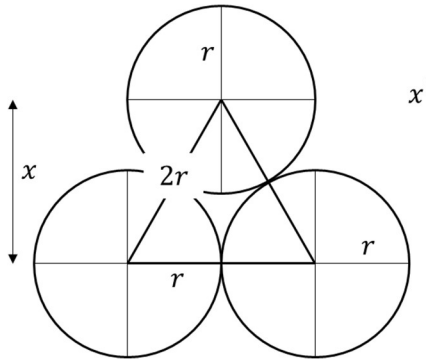


Figure 85. Packing 5 congruent circles into a square

. . .

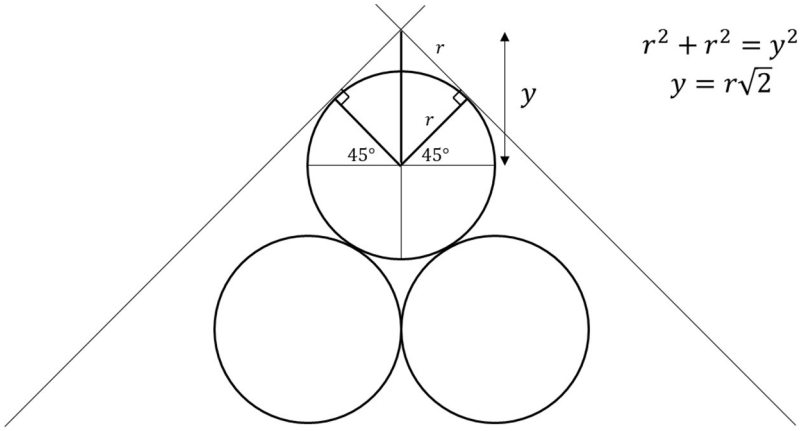
The geometry for three congruent circles in a unit square is a bit more involved that the previous example. The optimal configuration (shown in Figure 86) is two tangent congruent circles at the bottom with one congruent circle at the top (tangent to other two circles). The figure also shows an equilateral triangle whose vertices coincide with the centers of the circles. The calculation for the height of the triangle is shown in the figure.



$$x^2 + r^2 = (2r)^2$$
$$x = r\sqrt{3}$$

Figure 86. Three congruent circles in a unit square – Part 1

The two lines shown in Figure 87 are part of the unit square surrounding the circles. The length y is part of the diagonal of the surrounding unit square. The plan is to compute the length of the diagonal of the surrounding unit square in terms of r and then solve for r.



$$r^2 + r^2 = y^2$$
$$y = r\sqrt{2}$$

Figure 87. Three congruent circles in a unit square – Part 2

In Figure 88, we complete the unit square surrounding the circles. The two additional lines are perpendicular to each other, and each is tangent to one of the bottom circles. We have already computed x and y in terms of r. So, we only need to compute z in terms of r to determine the length of the diagonal of the unit square in terms of r. Further, we already know that the length of the diagonal of the unit square is $\sqrt{2}$.

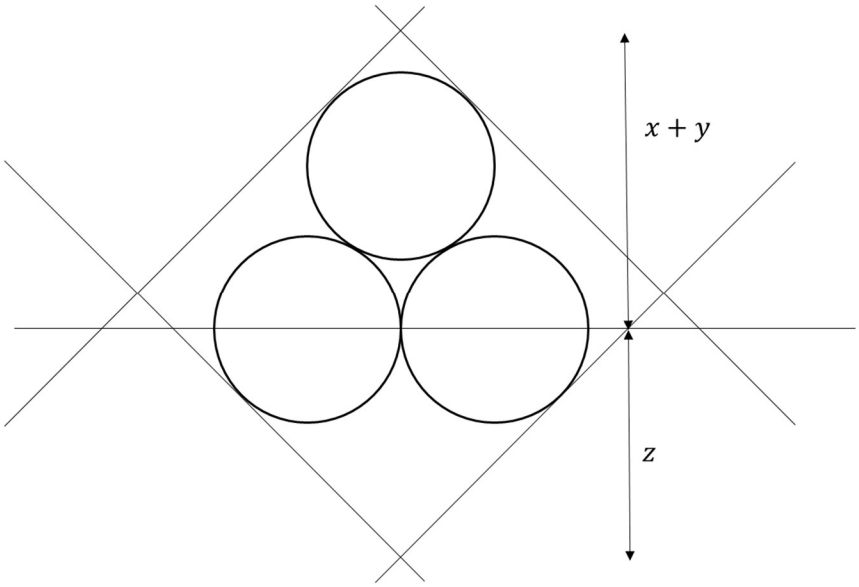


Figure 88. Three congruent circles in a unit square – Part 3

The computations for the length z are shown in Figure 89. The half-angle formula for tangents was used to compute $\tan 67.5$. [If you are interested in the details, input "tan (67.5)" at https://www.mathway.com/).

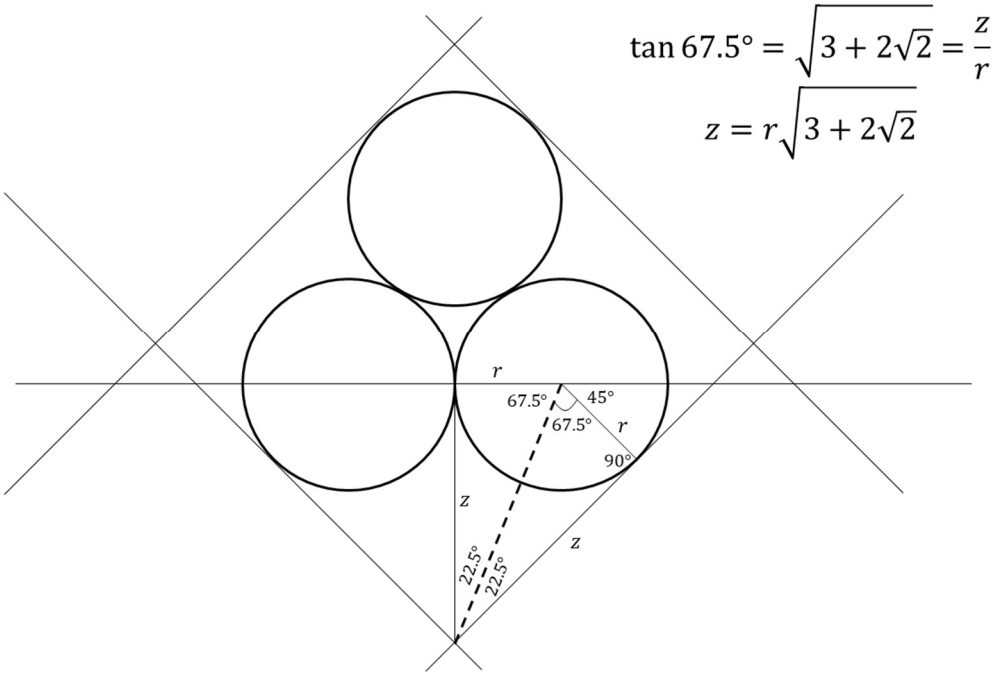


Figure 89. Three congruent circles in a unit square – Part 4

Putting the pieces together, we have

$$x + y + z = \sqrt{2}$$

$$r\sqrt{2} + r\sqrt{3} + r\left(\sqrt{3 + 2\sqrt{2}}\right)$$

$$r = \frac{\sqrt{2}}{\sqrt{2} + \sqrt{3} + \sqrt{3 + 2\sqrt{2}}} \cong 0.254333$$

$$\ldots$$

At the time of this writing, optimal solutions for packing n circles into a unit square are known for $n = 1 - 30$ and $n = 36$ (see the list at http://hydra.nat.uni-magdeburg.de/packing/csq/csq.html).

For the cases $n = 1,4,9,16,25,36$, the optimal packing is a lattice. The packings for $n = 16,25$ and 36 are shown in Figure 90.

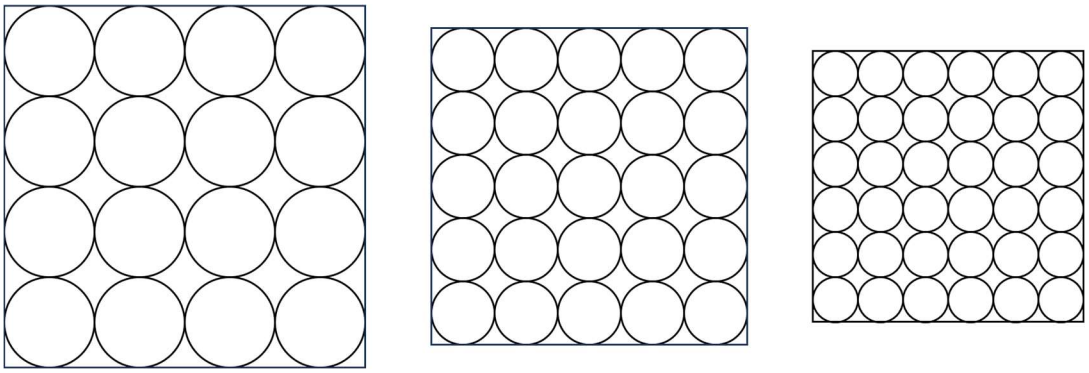


Figure 90. Optimal packing of circles into a unit square for n=16,25,36

One might think that the pattern continues for all squares of integers, but this is not the case. The pattern stops at 6^2. For example, a more efficient pattern than a lattice (not yet proven optimal) for $n = 49$ has been found. Currently, the best known solution for packing 49 circles into a unit square is shown in Figure 91 (see the source of the figure at http://hydra.nat.uni-magdeburg.de/packing/csq/d5.html). The dark circle is free standing (i.e., not tangent to any other circle).

[**Author's Remark**: Equally strange results have been found for 64, 81 and other square numbers. To me, this is both perplexing and interesting. These solutions are all more efficient than the regular lattice configuration. I wonder if there is not some hidden pattern.]

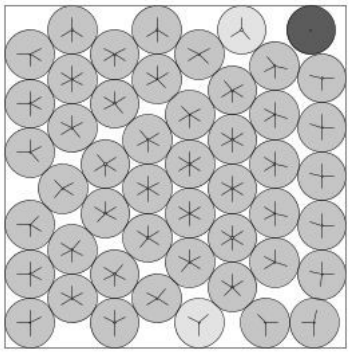


Figure 91. Packing 49 circles into a unit square

. . .

The packing of circles into various shapes (other than circles) is also well studied. The sites listed below track the most current results of circle packings into squares, rectangles, triangles, hexagons, and several other shapes:

- Erich's Packing Center: https://erich-friedman.github.io/packing/
- Packomania: www.packomania.com

7.2.1.3 Circle packing theorem

On the left of Figure 92, we have a collection of connected circles with disjoint interiors. In the middle of the figure, we have overlayed a graph whose vertices are the centers of each circle, and whose edges are between tangent circles. This is known as a tangency graph or contact graph. On the right, we see the tangency graph in isolation.

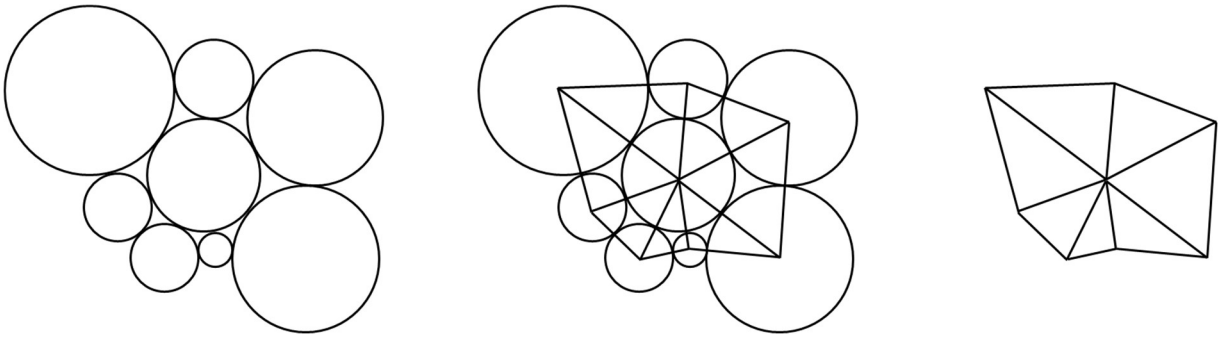


Figure 92. Mapping of circle packing to tangency graph

In general, it is always possible to map between a circle packing and a simple connected planar graph. This result is known as the **circle packing theorem**.

Theorem 80. (Circle packing theorem also known as the Koebe–Andreev–Thurston theorem) For every connected simple planar graph G there is a circle packing in the plane whose tangency graph is isomorphic to G.

Proof: The Wikipedia article on this topic [96] gives references to several proofs. The theorem can be proven for Euclidean, spherical and hyperbolic geometries, see Stephenson [97]. Stephenson devotes about 100 pages to the preliminaries and the proof (including some examples).

7.2.1.4 Descartes' theorem

Descartes’ theorem (also known as the kissing circles theorem), named after René Descartes who proposed it in 1643, is a concept in geometry that deals with the relationship between the radii of four circles that are tangent to each other. According to the theorem, the radii of these circles satisfy a specific quadratic equation. By finding the solution to this equation, it is possible to construct a fourth circle that is tangent to three given circles that are also tangent to each other.

Descartes' theorem is typically stated in terms of the circles' curvatures. In the context of Descartes’ theorem, if the radius of a circle is r then its curvature is $k = \pm\frac{1}{r}$. The sign is positive for a circle that is externally tangent to the other three circles. The sign is negative if a circle circumscribes the other circles.

For four circles (with curvatures k_1, k_2, k_3, k_4) that are tangent to each other at six distinct points, Descartes’ theorem tells us that

$$(k_1 + k_2 + k_3 + k_3)^2 = 2(k_1^2 + k_2^2 + k_3^2 + k_4^2)$$

If we know the curvatures of three of the circles, we can compute the curvature of the fourth by the following formula

$$k_4 = k_1 + k_2 + k_3 \pm 2\sqrt{k_1k_2 + k_2k_3 + k_1k_3}$$

As one can see, the above equation has two possible solutions. This arises from the fact that any triple of tangent circles has two tangent circles (or degenerate circles, i.e., tangent straight lines).

In Figure 93, we are given three circles with curvatures $k_1 = 2, k_2 = 2$ and $k_3 = 3$, respectively. Using Descartes’ theorem, we can find the curvatures of circles A and B as follows:

$$k_4 = k_1 + k_2 + k_3 \pm 2\sqrt{k_1k_2 + k_2k_3 + k_1k_3} = 7 \pm 2\sqrt{4 + 6 + 6} = 7 \pm 8 = -1,15$$

In Figure 93, the curvature of circle A is 15, and the curvature of circle B is -1.

The converse of Descartes’ theorem is also true, i.e., every four integers that satisfy the equation in Descartes' theorem form the curvatures of four tangent circles [98].

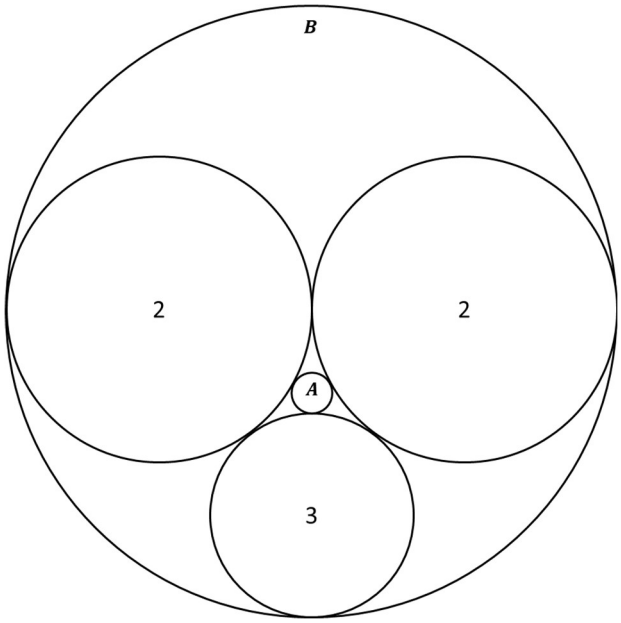


Figure 93. Example of two solutions to Descartes' theorem

7.2.1.5 Some Specialized Circle Packings

The **Apollonian gasket** is a fractal that is created by starting with three tangent circles and then adding more circles that are tangent to three existing circles. This process is repeated indefinitely, resulting in a complex and beautiful pattern. The Apollonian gasket is named after the Greek mathematician Apollonius of Perga, who studied the properties of circles.

Figure 94 shows the first several iterations of the Apollonian gasket. As a fractal, the pattern continues indefinitely and thus, cannot be completely drawn.

The figure is taken from the Wikipedia article entitled "Apollonian gasket" [99].

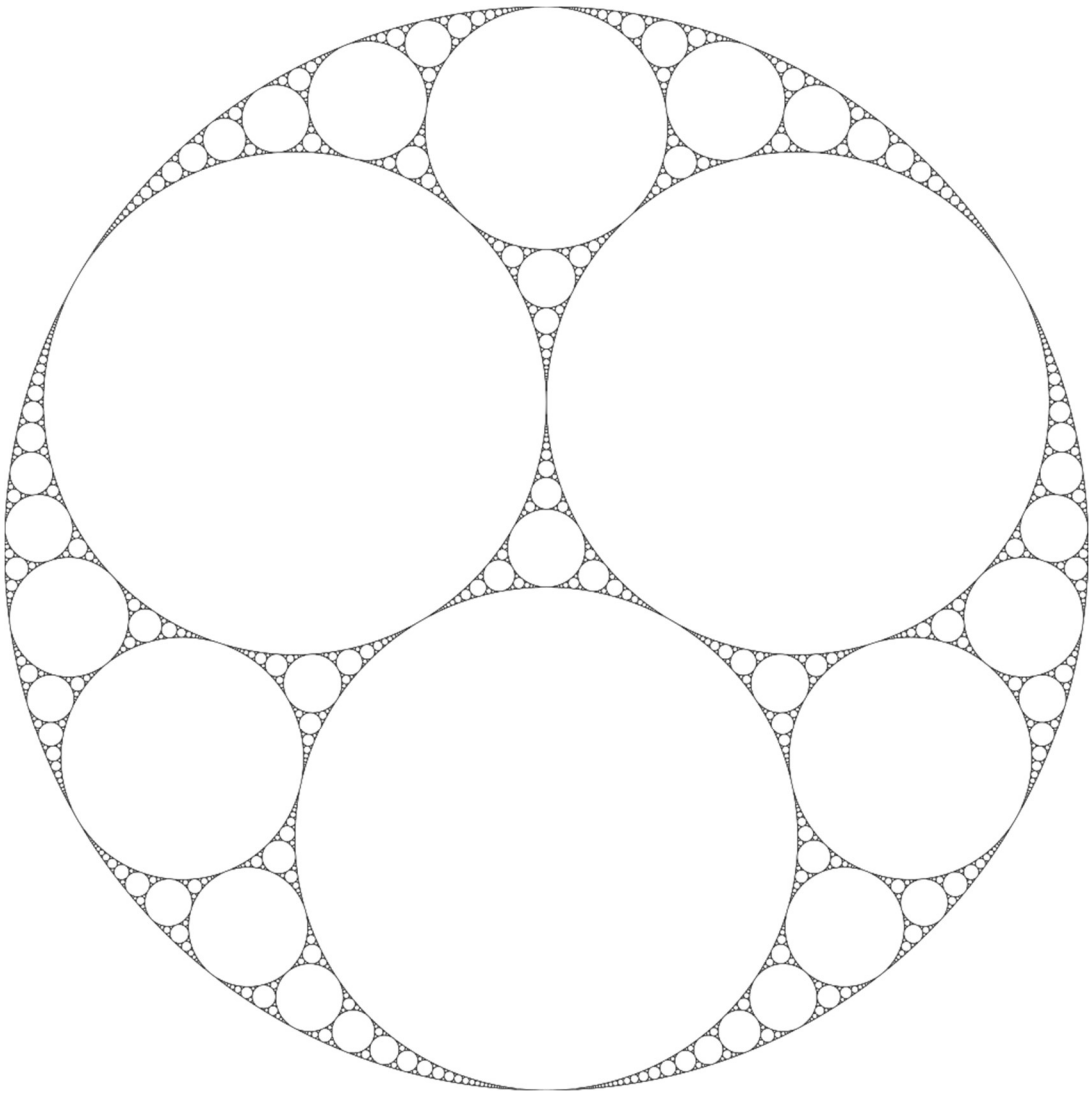


Figure 94. Apollonian gasket

It is also possible to have Apollonian gaskets where the circles are of different curvatures. Further, there are cases were all the circles have integer curvatures. Figure 95 depicts an Apollonian gasket with integer curvatures, generated by four mutually tangent circles with curvatures −10 (the outer circle), 18, 23, and 27. The figure is taken from the Wikipedia article entitled “Descartes’ theorem” [100].

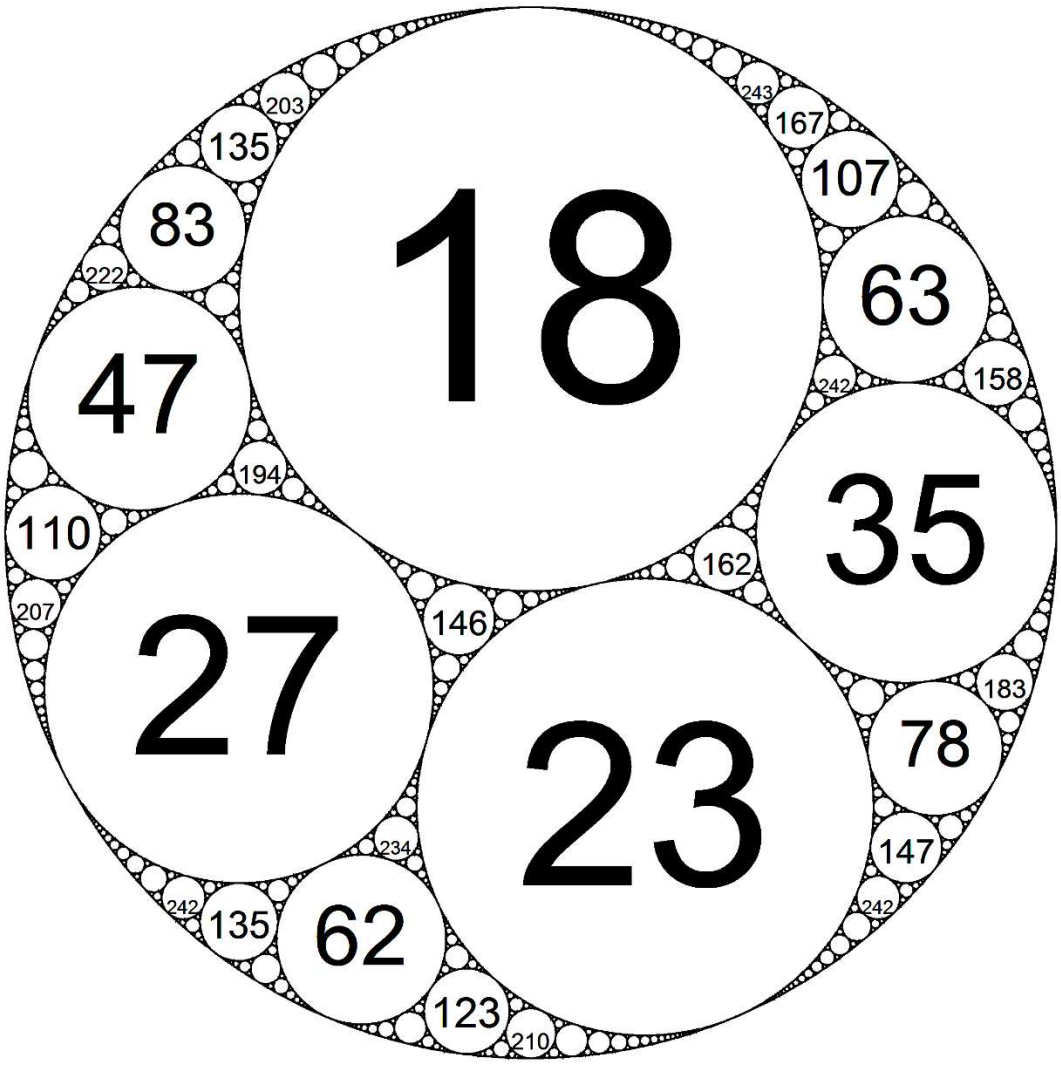


Figure 95. Apollonian gasket with circles having integer curvatures

If any four mutually tangent circles in an Apollonian gasket all have integer curvature then all circles in the gasket will have integer curvature [98].

The following is a list of some beginning curvatures that lead to Apollonian gaskets that only have circles with integer curvatures.

$$-1, 2, 2, 3, 3$$
$$-2, 3, 6, 7, 7$$
$$-3, 4, 12, 13, 13$$
$$-3, 5, 8, 8, 12$$
$$-4, 5, 20, 21, 21$$
$$-4, 8, 9, 9, 17$$
$$-5, 6, 30, 31, 31$$
$$-5, 7, 18, 18, 22$$
$$\ldots$$

A **Doyle spiral** is a type of circle packing that consists of infinitely many circles in the plane, with no two circles having overlapping interiors. Doyle spirals are special logarithmic spirals of touching circles, where each circle is surrounded by a set of six touching circles. The circles extend indefinitely outwards across the plane with ever increasing radii, and indefinitely inward with ever decreasing radii towards the spiral center.

From the Wikipedia article on this topic [101]:

> Doyle spirals are named after mathematician Peter G. Doyle, who made an important contribution to their mathematical construction in the late 1980s or early 1990s. However, their study in phyllotaxis (the mathematics of plant growth) dates back to the early 1900s.

The center circle and six surrounding circles (used to initiate a Doyle spiral) are collectively known as a flower. If the flower has the relative dimensions as indicated in Figure 96, the pattern can be extended to an infinite hexagonal circle packing of the plane.

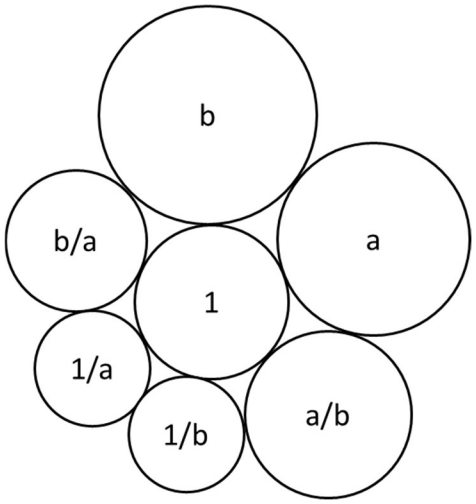
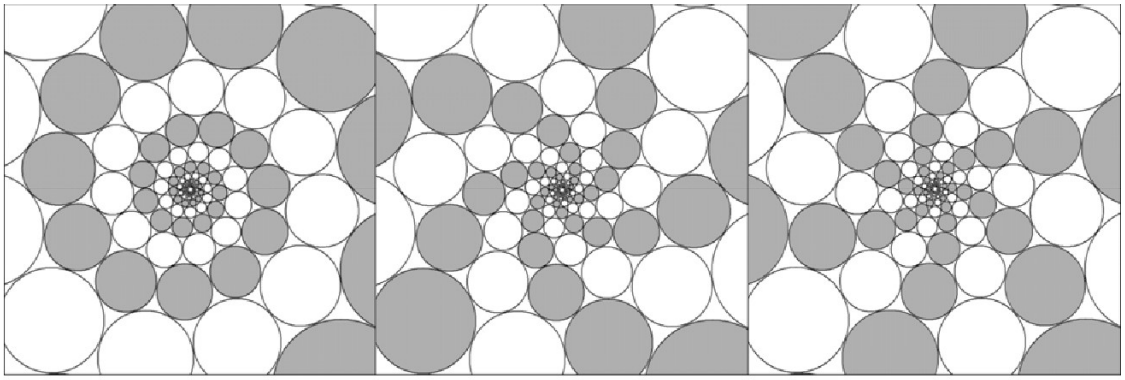


Figure 96. Relative radii dimensions for a Doyle spiral

Each Doyle spiral contains multiple spiral arms formed by circles linked through opposite points of tangency, with their centers on logarithmic spirals of three different shapes. The diagram below is Figure 2 taken from the paper “Doyle Spiral Circle Packings Animated” [102]. It shows the same Doyle spiral from three different perspectives.



2 Spiral Arms *10 Spiral Arms* *12 Spiral Arms*

Figure 2: *Three Views of the Packing $p_1 = 2$, $p_2 = 10$ and $Q = 12$.*

. . .

A Steiner chain [103] is defined as a set of circles that are tangent to two given non-intersecting circles, referred to as the "base circles" or "fixed circles". In a Steiner chain, each circle in the chain is tangent to both the preceding and succeeding circles in the chain. Figure 97 depicts a Steiner chain with one base circle inside the other. In this figure (and those that follow) the base circles have white interiors, and the other circles have gray interiors.

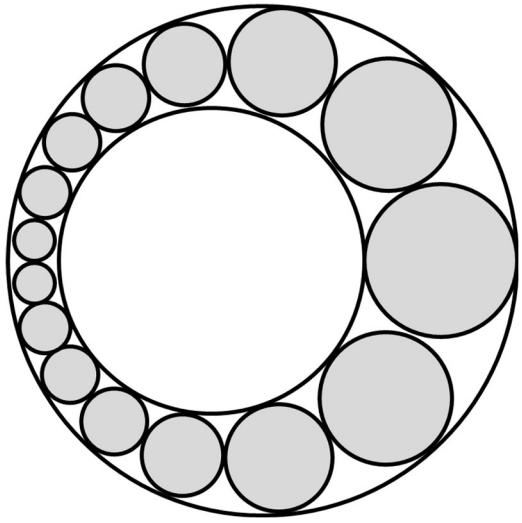


Figure 97. Steiner chain with one base circle inside the other

In a closed Steiner chain, the first and last circles are also tangent to each other, forming a closed loop, e.g., the Steiner chain in Figure 97. In an open Steiner chain, tangency between the first and last circles is not necessary, and the chain remains open-ended. An example of an open Steiner chain is shown in Figure 98.

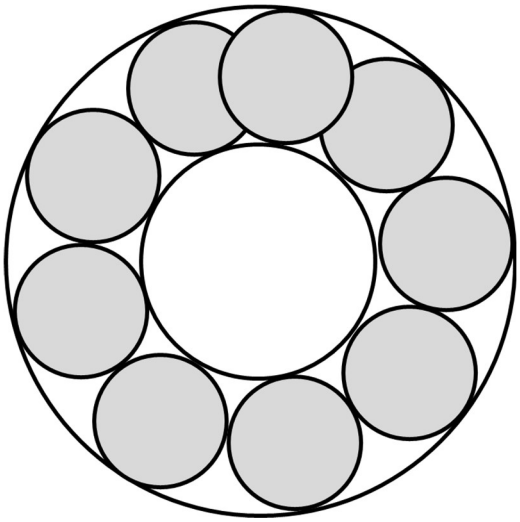


Figure 98. Open Steiner chain

While the base circles, around which the chain is constructed, are not allowed to intersect each other, they can otherwise be positioned freely. The base circles can be disjoint from each other, e.g., see Figure 99.

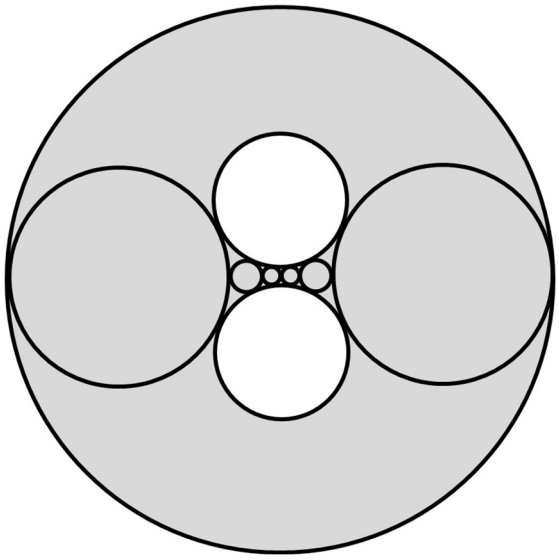


Figure 99. Steiner chain with disjoint base circles

Depending on the configuration of the fixed circles, the centers of the circles in the Steiner chain may follow different geometric curves. If the smaller circle lies completely inside the larger circle, the centers of the Steiner-chain circles will lie on an ellipse. Conversely, if the smaller circle lies completely outside the larger circle, the centers will lie on a hyperbola.

7.2.2 Square Packing

Next, we consider the packing of n congruent squares into a unit square, or equivalently, the packing of n unit squares into the smallest containing square.

The cases where the number of smaller squares is a square have natural solutions, e.g., see the optimal solutions for $n = 4,9,16$ in Figure 100. For these cases, the optimal solution leaves no empty space at all.

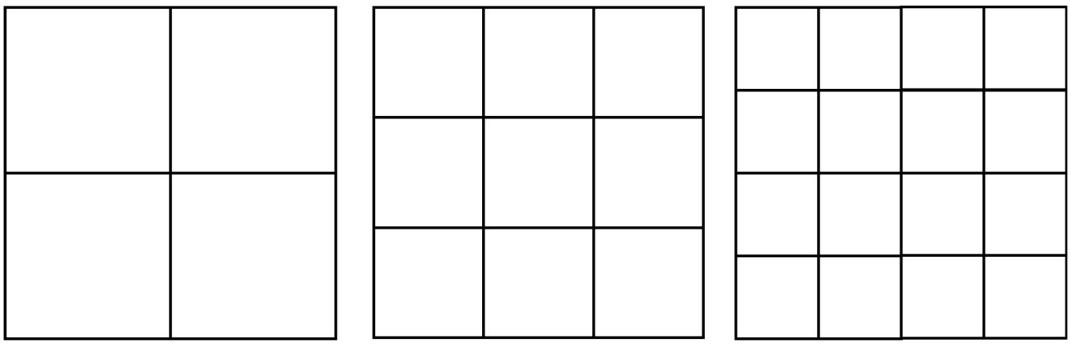


Figure 100. Optimal packing of congruent squares into a unit square for n=4,9,16

In some cases, when n is close to a square number, the optimal packing is a matter of removing one or more squares from the optimal packing of the nearby square. The optimal packings for $n = 13,14,15$ are shown in Figure 101. The packings are not rigid in the sense that one can move the squares around and still get an optimal packing.

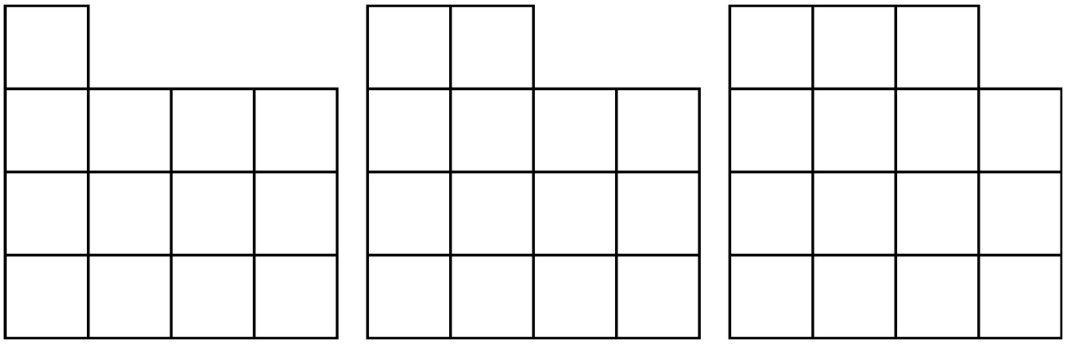


Figure 101. Optimal packing of congruent squares into a unit square for n=13,14,15

The optimal square packing for $n = 5$ (proved by Frits Göbel in 1979) is shown in Figure 102. All the squares are rigidly placed in the larger square.

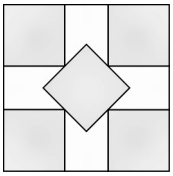


Figure 102. Optimal square packing for n=5

There are four optimal packings in the case of $n = 10$ (see the paper by Stromquist [104]). The packing on the left is rigid, but the other three are not.

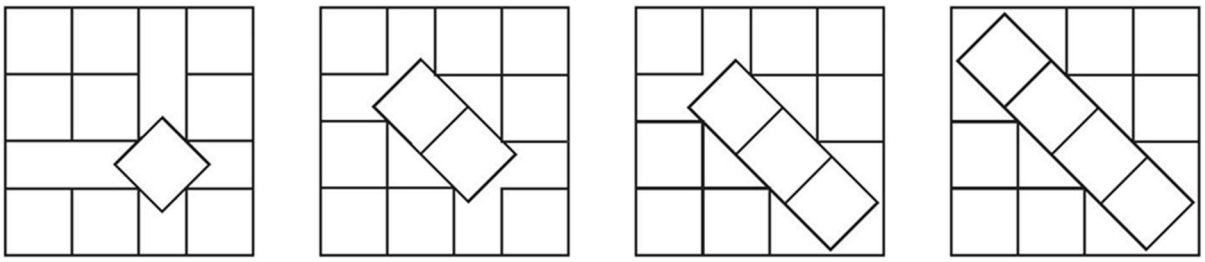


Figure 103. Optimal square packing solutions for n=10

At the time of this writing and according to the webpage "Squares in Squares" [105], optimal solutions are known when n is a square, and for the cases n = 2, 3, 5, 6, 7, 8, 10, 13, 14, 15, 24, 34, 35, 46, 47, 48, 62, 63, 79 and 80. The paper by Friedman [106] indicates there are also known optimal solutions for $n = 98$ and 99.

Best known solutions (not proven to be optimal) are available for many other values of n, see the survey paper by Friedman [106]). The best known solution for $n = 272$ is particularly unusual, see Figure 104.

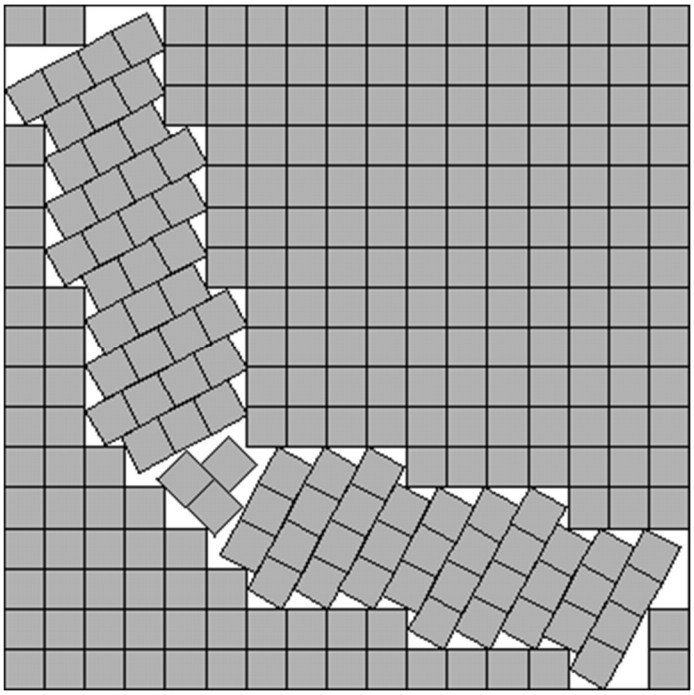


Figure 104. Best known square packing for n=272

7.3 Sphere Packing

Three-dimensional sphere packing is a problem in geometry that asks how many congruent spheres can be packed inside a unit sphere. For a given number of smaller spheres (say N), the goal is to

determine the largest radius of the smaller spheres such that they can all be packed into a unit sphere.

The optimal sphere packings for cases $N = 7,8$ and 9 are shown in Figure 105. The figure is taken from the Packomania webpage on this topic:

http://hydra.nat.uni-magdeburg.de/packing/ssp/ssp.html

This same site list hundreds of best known solutions for all values of N up to 600, and for select values of N up to 2,982,989.

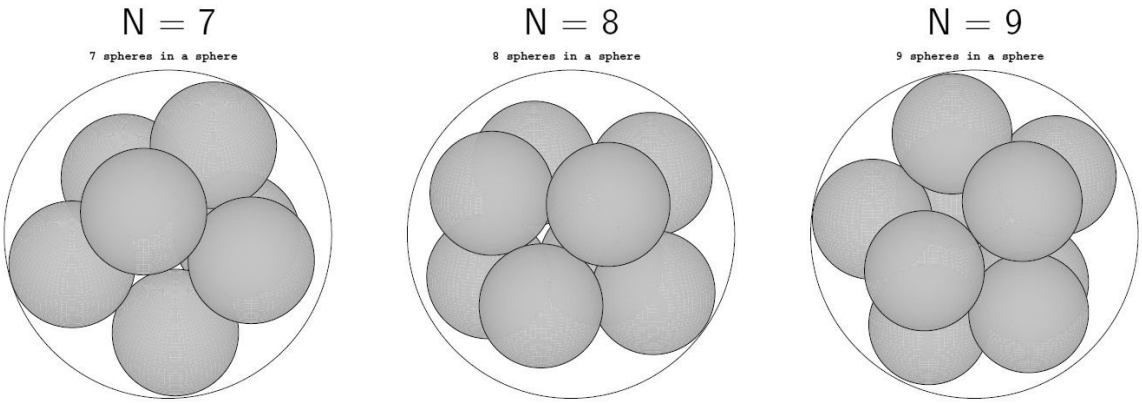


Figure 105. Optimal sphere packing for N=7,8,9

For dimension 3, the **Kepler conjecture** is about the densest possible sphere packing. From the Wikipedia article on the topic [107]:

> The Kepler conjecture, named after the 17th-century mathematician and astronomer Johannes Kepler, is a mathematical theorem about sphere packing in three-dimensional Euclidean space. It states that no arrangement of equally sized spheres filling space has a greater average density than that of the cubic close packing (face-centered cubic) and hexagonal close packing arrangements. The density of these arrangements is around 74.05% *[or* $\frac{\pi}{3\sqrt{2}}$ *to be exact]*.
>
> In 1998, Thomas Hales, following an approach suggested by Fejes Tóth (1953), announced that he had a proof of the Kepler conjecture. Hales' proof is a proof by exhaustion involving the checking of many individual cases using complex computer calculations. Referees said that they were "99% certain" of the correctness of Hales' proof, and the Kepler conjecture was accepted as a theorem. In 2014, the Flyspeck project team, headed by Hales, announced the completion of a formal proof of the Kepler conjecture using a combination of the Isabelle and HOL Light proof assistants *[both are automated theorem proving software]*. In 2017, the formal proof was accepted by the journal Forum of Mathematics, Pi.

In the above quote, the cubic close packing (also called face-centered cubic (FCC) packing) and hexagonal close-packed (HCP) arrangements refer to the two ways of stacking layers of spheres. In the following figures, we explain the difference between the two packings (noting that any combination leads to the most efficient packing in space).

Figure 106 shows a layer of hexagonally packed spheres (looking down from above). The packing extends indefinitely (left and right, and up and down).

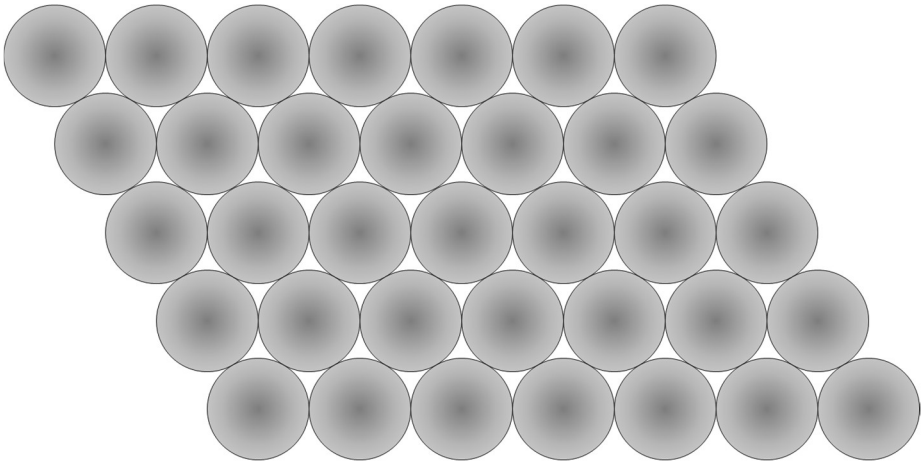


Figure 106. Hexagonal packing of spheres

Figure 107 shows a second layer of spheres stacked on top of the first layer (again, looking down from above). The key thing to notice here is that half of the gaps between the spheres in layer 1 are directly covered by spheres in layer 2, and the other gaps in layer 1 coincide with gaps in layer 2.

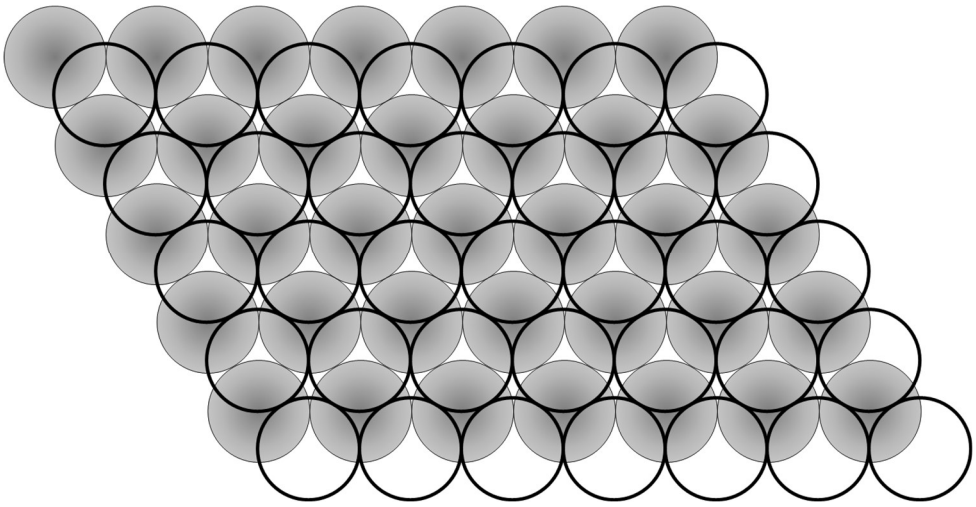


Figure 107. Two layers of hexagonally packed spheres

When adding a third layer, there are two options. In the option shown in Figure 108, the centers of the spheres in the third layer are directly over the common gaps in layers 1 and 2.

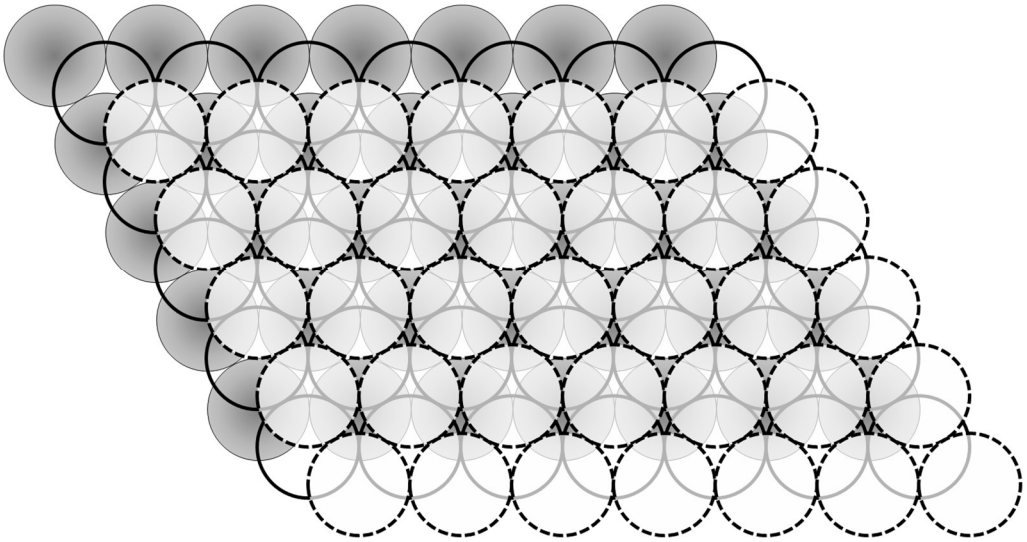


Figure 108. Hexagonal close-packed alternative

The other option for adding the third layer is to put each sphere in the third layer directly over a sphere in the first layer, as shown in Figure 109.

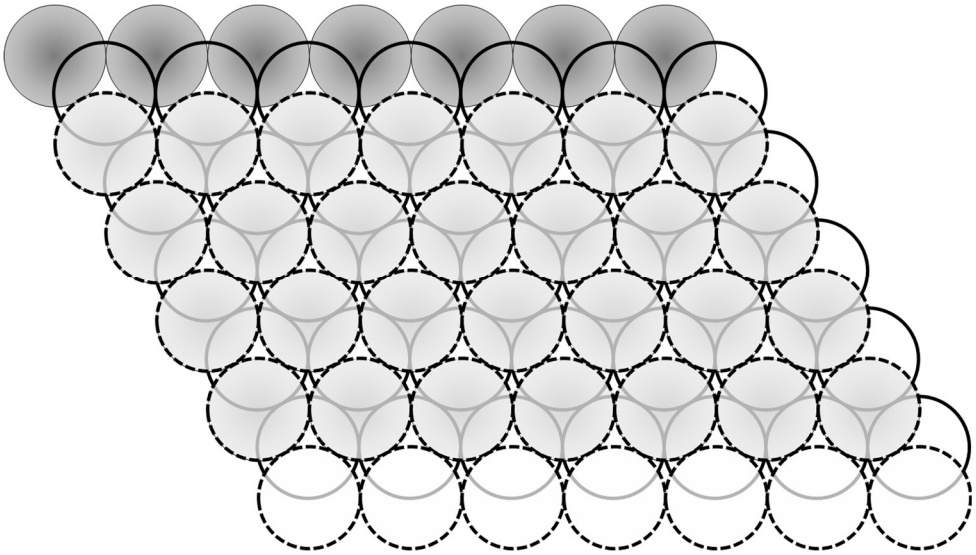


Figure 109. Face-centered cubic packing

For the fourth layer, we can place the spheres in either HCP or FCC pattern in relation to the second layer. For each subsequent layer, we have a binary decision in terms of placement relative to the second layer below.

The Wikipedia article entitled “Close-packing of equal spheres” [108] provides some additional explanation and diagrams concerning the HCP and FCC patterns. Section 7.8: Cubic Lattices and Close Packing of the online chemistry book by Lower [109] describes HCP and FCC patterns in terms of crystal arrangements.

. . .

It is also possible to pack generalized spheres in higher dimensions. The formula for the equivalent of a sphere in n-dimensions is

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2 \leq r^2$$

where the center of the entity is point $(y_1, y_2, \ldots, y_n)$ and the radius is r. Such an entity is known as an n-dimensional ball. The equation defines all points $(x_1, x_2, \ldots, x_n)$ on the surface and the interior of the n-dimensional ball. The equation for the surface of the n-dimensional ball is

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2 = r^2$$

For $n = 2$, the above formula reduces to that for the boundary of a circle, and for $n = 3$, it is the formula for the surface of a sphere.

For $n \geq 3$, we cannot visualize ball packings, but nevertheless, mathematicians have proved some results on this topic. Maryna Viazovska has proved that the best packing density for eight-dimensional balls into 8-dimensional space is bounded by $\frac{\pi^4}{385}$ which is about 25%. Viazovska's result [110] is the first time that an exact number has been put on the maximum ball packing density for a space of dimension larger than three. For her work in this area, Viazovska was awarded the prestigious Fields Medal in 2022. Within a week, Viazovska and four other mathematicians, successfully extended her method to 24-dimensional balls [111]. In particular, they showed that something called the Leech lattice [112] achieves the optimal ball packing density for 24-dimensional Euclidean space. The packing density of the Leech lattice is $\frac{\pi^{12}}{12!}$ or about 0.193%.

In higher dimensions, the sphere packing density decreases as compared to lower dimensions, which is a counterintuitive result. One of the key reasons for this behavior is related to the "empty space" (voids) that arises in higher dimensions.

In higher dimensions, the volume of these voids increases significantly. The higher the dimension, the more space is wasted, and it becomes increasingly difficult to pack spheres more densely. This behavior is a consequence of the nature of higher-dimensional space, where there is more room for spheres to move away from each other, resulting in more empty space.

This phenomenon can be explained by considering the volume-to-surface-area ratio of spheres which is r/n where r is the radius of the ball and n is the dimension of the space (for the details of this formula, see “Surface-area-to-volume ratio” [113]). So, as we move to higher dimensions, this ratio decreases, leading to a less efficient packing of spheres.

8 Knot Theory

A new scientific truth triumphs, not because it convinces its opponents and makes them see the light, but because the opponents eventually die, and a new generation that is familiar with it grows up. — Max Planck

8.1 Overview

Mathematical knot theory is the study of closed curves in three-dimensional space (which are known as knots). From the Wikipedia article on knot theory [114]:

> In topology, knot theory is the study of mathematical knots. While inspired by knots which appear in daily life, such as those in shoelaces and rope, a mathematical knot differs in that the ends are joined so it cannot be undone, the simplest knot being a ring (or "unknot"). In mathematical language, a knot is an embedding of a circle in 3-dimensional Euclidean space, $\mathbb{R}^3$. Two mathematical knots are equivalent if one can be transformed into the other via a deformation of $\mathbb{R}^3$ upon itself (known as an **ambient isotopy**[1]); these transformations correspond to manipulations of a knotted string that do not involve cutting it or passing it through itself.

From Wolfram MathWorld:

> A knot is defined as a closed, non-self-intersecting curve that is embedded in three dimensions and cannot be untangled to produce a simple loop (i.e., the unknot). *[Some definitions allow unknots to be consider as knots.]*

Knots can be classified by their topology, which is the study of the properties of shapes that remain the same when they are continuously deformed.

The basic question in knot theory is whether two knots are equivalent. This can be difficult to answer, as knots can be very complex. One way to classify knots is to count the number of times they cross themselves. Another way is to use invariants, which are mathematical objects that are associated with a knot and do not change when the knot is deformed.

Some of the practical applications of knot theory include [115]:

- Chemistry and Molecular Biology: Knot theory is used to analyze and classify the complexity of DNA and protein structures. In DNA research, it helps in understanding the topological properties of DNA molecules and their role in replication, recombination, and other biological processes.
- Physics: Knot theory has found applications in theoretical physics, particularly in the study of quantum mechanics and gauge theories. Knot invariants have been used to classify and understand the quantum states of particles and gauge fields.
- Robotics and Computer Graphics: In robotics, knot theory is applied to study the movements of robot arms and how to avoid entanglement in complex environments. In computer graphics and animation, it is used to simulate realistic and natural-looking deformations of virtual objects.

[1] The rearrangement of a knot in three-dimensional space without letting it pass through itself is called an ambient isotopy. The word "isotopy" refers to equivalence of the knot before and after the deformation. The word "ambient" refers to the fact that the knot is being rearranged through the three-dimensional space in which it sits.

- Material Science and Nanotechnology: Knot theory plays a role in understanding the behavior of polymers and other complex materials. It helps analyze the topological aspects of entangled structures in materials, which has implications for their mechanical properties.
- Fluid Dynamics: In fluid mechanics, knots and links arise in the study of vortex lines and fluid turbulence. Knot theory provides tools to analyze and classify these structures, leading to a better understanding of fluid dynamics.
- Statistical Mechanics: Knot theory is used to study the behavior of random walks and polymers, which have applications in understanding phase transitions and critical phenomena in statistical mechanics.
- Coding and Cryptography: Knot theory has been used in the development of error-correcting codes and cryptographic algorithms.
- Art and Design: Knot theory has inspired artists and designers to create intricate and aesthetically pleasing patterns and sculptures.
- Spatial Mapping and Geographic Information Systems (GIS): Knot theory can be applied to analyze and represent complex spatial networks, such as transportation networks, river systems, and power grids.
- Image and Signal Processing: Knot theory tools have been applied to process and analyze images, particularly in areas such as computer vision and medical imaging.

For an introductory course on knot theory, see “Open Algebra and Knots” [116].

8.2 Examples and Basic Concepts

The **trefoil knot** is the simplest example of a nontrivial knot (see Figure 110). The trefoil can be obtained by joining together the two loose ends of a common overhand knot, resulting in a knotted loop.

Figure 110. Trefoil knot

[**Warning**: It should be emphasized that knots are one-dimensional entities existing in 3-dimensional space. Some of the knot diagrams in this book (for example the one above) are drawn in a style that might give the false impression that a knot has a surface area and associated volume. This is typical of drawings in the literature and can help with viewing a knot when drawn in 2-dimensions, but please keep in mind that knots are not surfaces or solids.]

A **figure-eight knot** (also called Listing's knot) is the only knot with a crossing number of four (see Figure 111). This makes it the knot with the third-smallest possible crossing number, after the unknot and the trefoil knot.

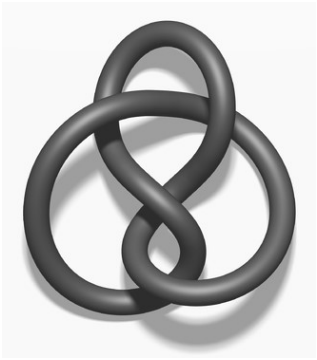


Figure 111. Figure-eight knot

The **cinquefoil knot** (also known as Solomon's seal knot or the pentafoil knot) is one of two knots with crossing number five, the other being the **three-twist knot**. The cinquefoil knots appears on the left of Figure 112, and the three-twist knot is shown on the right.

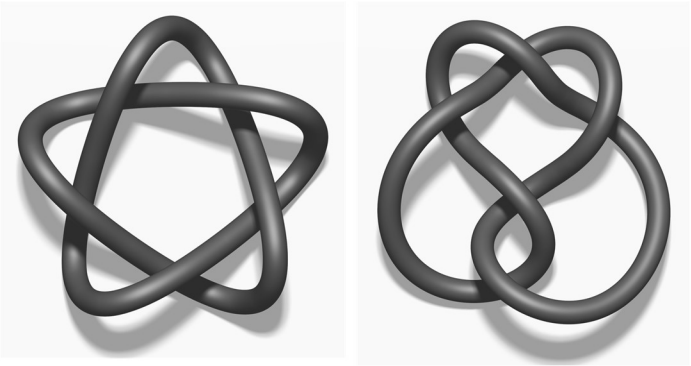


Figure 112. Cinquefoil and three-twist knots

Knot theory comes under an area of mathematics known as topology (the study of the properties of geometric objects that are preserved under deformations). The unknot is depicted in the center of Figure 113, with deformations of the knot on either side. All three knots are topologically the same. The little bit of white space at each crossing is meant to show one part of the knot going underneath and the other part.

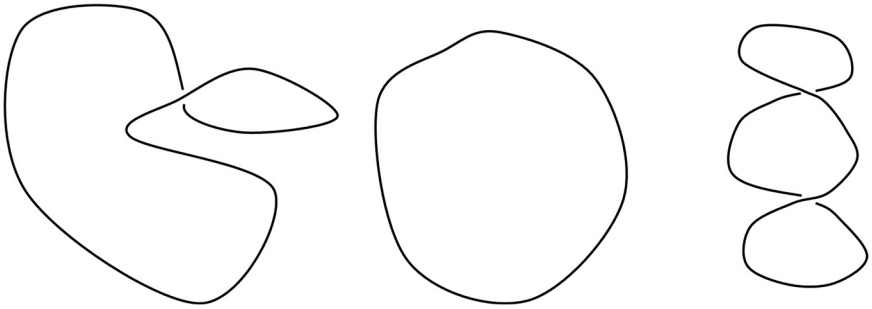


Figure 113. Deformations of the unknot

Knots are essentially 1-dimensional objects existing in 3-dimensional space. A **knot projection** is a drawing of a knot in 2 dimensions, e.g., there are three knot projections in Figure 113. The **crossing number of a knot** is the minimum number of crossings in a projection of the knot.

As the number of crossings increases, it becomes extremely difficult to determine equivalency among knots. In fact, the central problem of knot theory is determining whether two knots can be rearranged (without cutting) to be exactly alike.

A knot whose crossings alternate between under and over is known as an alternating knot. The trefoil, figure-eight, cinquefoil and three-twist knots are all alternating knots. The so-called 8_{19} knot (shown in Figure 114) is the simplest example of a non-alternating knot. [The figure is taken from the Wikipedia article “List of prime knots” [122]. This shows yet another style for rendering knots in 2-dimensions.]

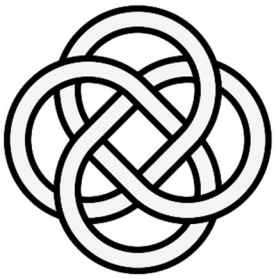


Figure 114. Example of an non-alternating knot

C. N. Little, a professor at the University of Nebraska, was the first to attempt an enumeration of non-alternating knots. In 1899, he published a table of 43 non-alternating knots with 10 crossings. His table was believed to be correct for 75 years. However, in 1974, an amateur mathematician (attorney Kenneth A. Perko) discovered that two of the knots in Little's table were the same knot (see Figure 115). The two equivalent knots are now known as the Perko pair [118]. This example gives an indication of how difficult it can be to distinguish equivalent knots.

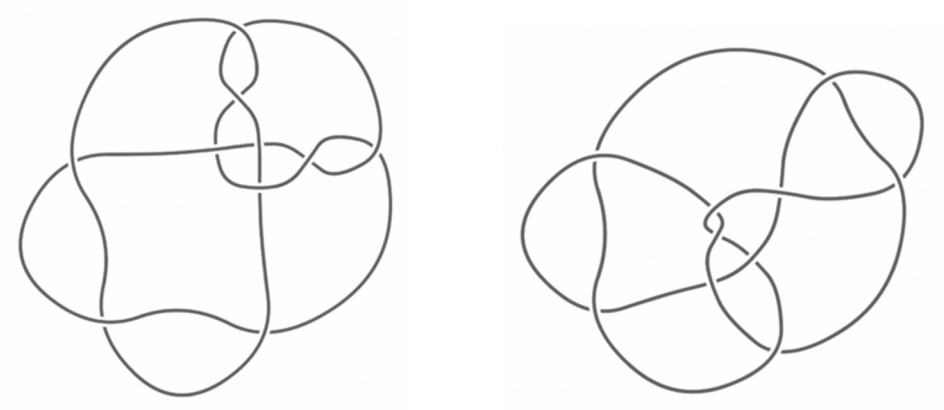


Figure 115. Perko pair

The **unknotting number of a knot** is the minimum number of crossing changes needed to transform the knot into an unknotted (i.e., trivial) loop. In other words, it represents the minimal number of changes you need to make to a knot's minimal projection to untangle it completely. For example, the trefoil, fight-eight and three-twist knots each have an unknotting number of 1. The

cinquefoil has an unknotting number of 2. The unknotting number of a knot is always less than half of its crossing number [117].

. . .

The mirror image of a knot is gotten by changing all the crossings of the knot, i.e., change all the under-crossings to over-crossings, and all the over-crossings to under-crossings. While not obvious, it turns out the mirror image of a knot (as defined above) is equivalent to the knot one gets one reflecting the knot in a mirror.

Knots that are equivalent to their mirror images are called **amphicheiral knots**, and those that are not equivalent are called **chiral knots** [127]. For example, the trefoil knot is chiral, and the figure-eight knot is amphicheiral.

. . .

A knot is said to be oriented if one defines a direction along the entire path of the knot. Figure 116 shows the two orientations of the trefoil knot. The two orientations are not equivalent as there is no way to deform one knot into the other while preserving the orientation.

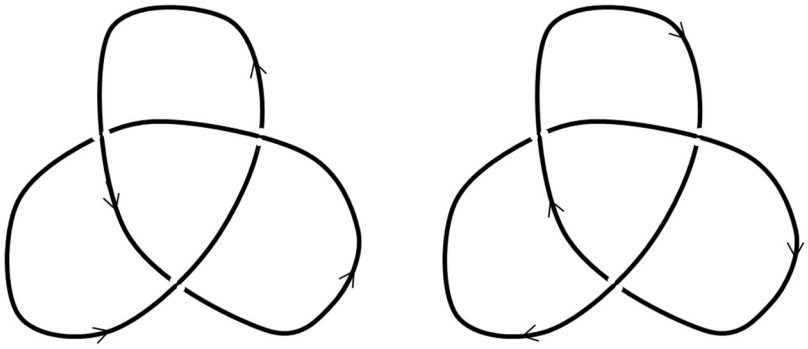


Figure 116. The two orientations of the trefoil knot

. . .

So far, we have discussed single closed curves that do not self-intersect (i.e., knots). It is also possible to define related collections of knots known as **links**. Several of the simplest links are shown in Figure 117. The Hopf link is comprised of two interlocked unknots. Borromean rings consist of three interlocked unknots but no two of the knots are interlocked if considered minus the other ring. The Whitehead link consists of two interlocked unknots, but differs from the Hopf link in that the twist in one unknot cannot be unraveled due to the positioning of the other unknot.

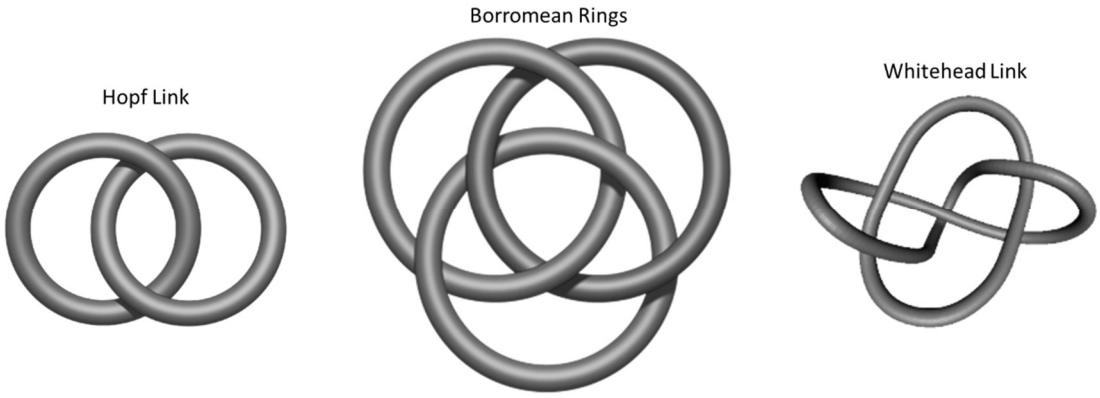


Figure 117. Several example links

From the Wikipedia article on links (relative to knot theory) [119]:

> A link is a collection of knots which do not intersect, but which **may** be linked (or knotted) together. A knot can be described as a link with one component. Links and knots are studied in a branch of mathematics called knot theory. Implicit in this definition is that there is a trivial reference link, usually called the unlink, but the word is also sometimes used in context where there is no notion of a trivial link.

The term "link" is also used in graph theory but the meaning is much different.

Notice that the above definition says "may". This means that several knots (which are not interlocked) could be considered a link.

The knots comprising a link are referred to as the components of the link. A knot is a link with one component.

The concepts of crossing number, unknotting number, chiral, amphicheiral, alternating, and orientation also apply to links.

Two links are said to be **equivalent** if they have the same number of components and one link (doesn't matter which) can be deformed in space (e.g., rotated, twisted, stretched) without cutting such that it is identical to the other link.

A **trivial link** is a type of link that consists of one or more knots that are not entangled or interlocked with each other.

A **Brunnian link** [120] is a nontrivial link that becomes a set of unknots if any one component is removed. In other words, cutting any loop frees all the other loops (so that no two loops can be directly linked). Borromean rings are the simplest example of a Brunnian link. Three Brunnian links are depicted in Figure 118. The link on the left is comprised of four components (each an unknot). The middle link is formed by three interlocked unknots, with a total of 12 crossings. The link on the right is also formed by three interlocked unknots, but with a total of 18 crossings.

The journal article by Bai and Wang [121] provides some advanced techniques for constructing complex Brunnian links.

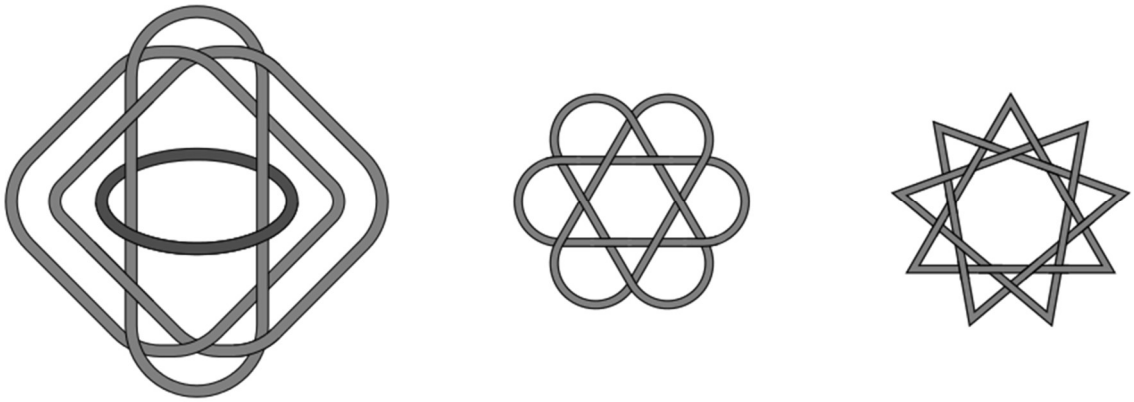


Figure 118. Examples of Brunnian links

8.3 Composition and Decomposition of Knots

One can join (compose) two knots (call them A and B) together to form another knot which may or may not be different from the original two knots. The operation is known as the knot sum, or the connected sum or composition of two knots, and is written as $A\#B$.

At the top of Figure 119 are shown a trefoil knot on the left and a figure-eight knot on the right. To sum the two knots, we remove a segment from each knot and then connect the two knots as shown on the bottom of the figure.

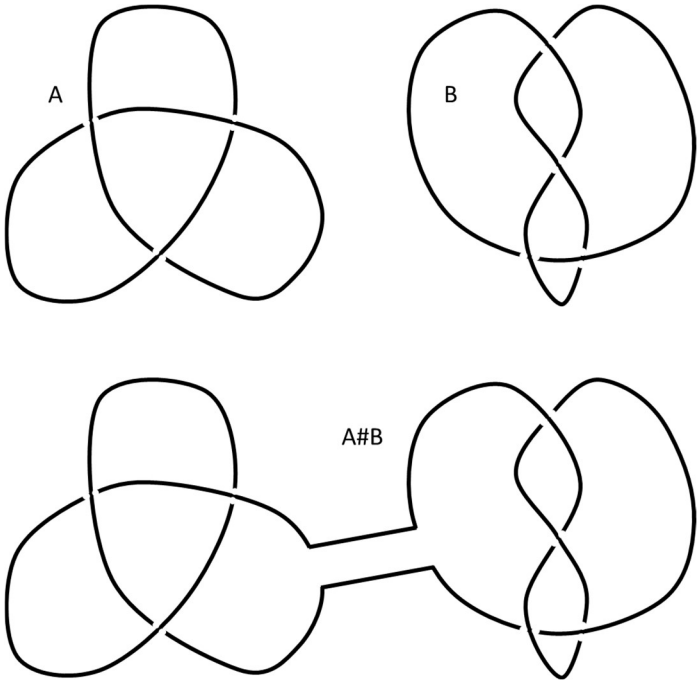


Figure 119. Sum of trefoil and figure-eight knots

The composition of any knot with the unknot results in the same knot, assuming the joining does not span over a crossover point. So, the unknot serves as an identity element under knot summation. Figure 120 depicts the sum of a figure-eight knot and an unknot, resulting in a figure-eight knot.

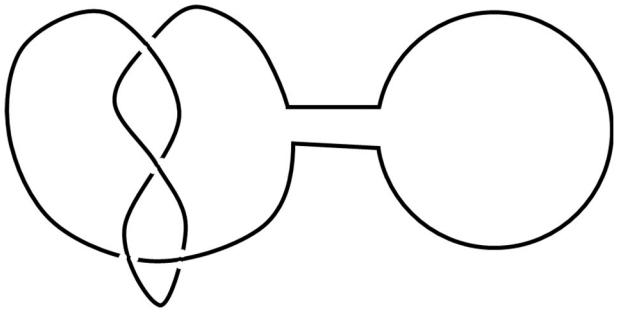


Figure 120. Sum of figure-eight know with the unknot

Knots that cannot be expressed as the sum of two other knots are known as **prime knots**. The Wikipedia article "List of prime knots" [122] shows all the prime knots with 10 crossings or less. The number of different prime knots with a given number of crossings is shown in Table 11. The source of the table is The Online Encyclopedia of Integer Sequences [123]. The table does not count both a knot and its mirror image. When a prime knot is equivalent to its mirror image (i.e., the knot is **amphicheiral** [124] in the technical language of knot theory), no information is lost. However, when the prime knot is not equivalent to its mirror image, it is only counted once and information is lost. [**Author's Remark**: This is not my idea but it is how the counting is done. In any event, a list of the number of amphicheiral knots (for prime knots with 16 or less crossings) is available [125] which allows one to reconstruct the lost information. Also, see the tables provided in the Wolfram MathWorld article on knots [126]. It is interesting that for $n = 13,14,15,16$ there are more non-alternating than alternating knots, where n is the crossing number.]

Table 11. Number of prime knots with a given number of crossings

Number of Crossings	Number of Prime Knots
1	0
2	0
3	1
4	1
5	2
6	3
7	7
8	21
9	49
10	165
11	552
12	2176
13	9988
14	46972

Number of Crossings	Number of Prime Knots
15	253293
16	1388705
17	8053393
18	48266466
19	294130458

. . .

The composition of two knots is independent of the location of the joining (i.e., will result in the same knot) with the exception of when the joining spans a crossover point in one of the knots. As an example of this exception, consider the composition of the unknot and the trefoil knot at the location shown in Figure 121.

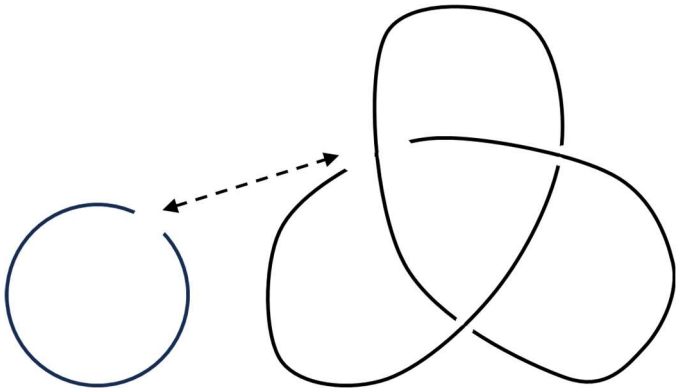


Figure 121. Composition of unknot with trefoil knot at crossover point

There are two options in this case, i.e., the unknot can be joined to the trefoil knot from below (left diagram in Figure 122) or from above (right diagram in Figure 122). When joining from below, the result is the same trefoil knot. When joining from above, we get the unknot (after some unraveling).

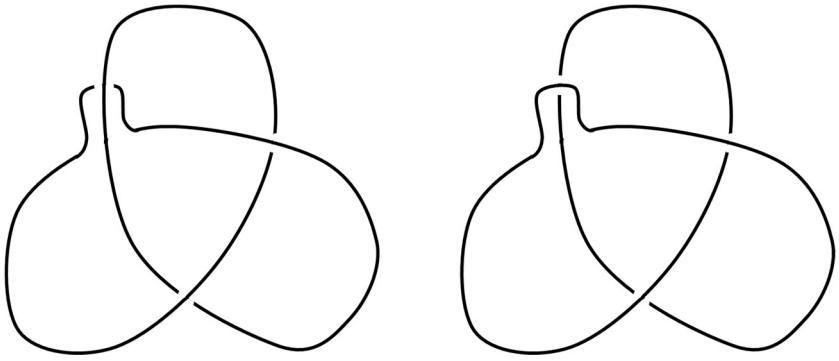


Figure 122. Options for joining an unknot and trefoil knot at a crossover point

The composition of two non-oriented knots will have the same result regardless of the joining location (with the exception of the case noted above, i.e., when the joining spans a crossing point).

In order to avoid such ambiguities, knot theorists often focus on oriented knots. For example, we have the following fundamental theorem due to Schubert (see Theorem 2.1 in the survey paper by Sakuma [128]).

Theorem 81 (Unique prime decomposition of knots). Every nontrivial oriented knot K can be decomposed as the sum of finitely many nontrivial prime oriented knots. Further, if $K \cong K_1 \# K_2 \# \ldots \# K_n$ and $K \cong J_1 \# J_2 \# \ldots \# J_m$ with each K_i and J_i being nontrivial knots, then $m = n$, and after reordering, $K_i \cong J_i$ as oriented knots.

8.4 Knot Notations

8.4.1 Alexander–Briggs

The Alexander-Briggs numbering scheme for knots is based on the numbering of crossings. If there is more than one type of knot with a given number of crossing, then a subscript is used to distinguish the knots. However, the subscript provides no information about the structure of the knot. For example, the notation for the Cinquefoil knot is 5_1 and the notation for the Three-twist knot is 5_2.

From the Wikipedia article on knot theory [114]:

> This *[Alexander-Briggs]* is the most traditional notation, due to the 1927 paper of James W. Alexander and Garland B. Briggs and later extended by Dale Rolfsen in his knot table ... The notation simply organizes knots by their crossing number. One writes the crossing number with a subscript to denote its order amongst all knots with that crossing number. This order is **arbitrary** and so has no special significance (though in each number of crossings the twist knot comes after the torus knot).

8.4.2 Dowker

The Dowker notation describes the structure of a knot although there are some cases where there is ambiguity. As an example, consider the figure-eight knot. We first orient the knot (i.e., pick a direction to traverse the knot) and then pick a starting point from which to number the crossings, see Figure 123. In general, each cross is labeled twice (once on the way out and once on the way back). This results in an odd number and an even number being assigned to each crossing.

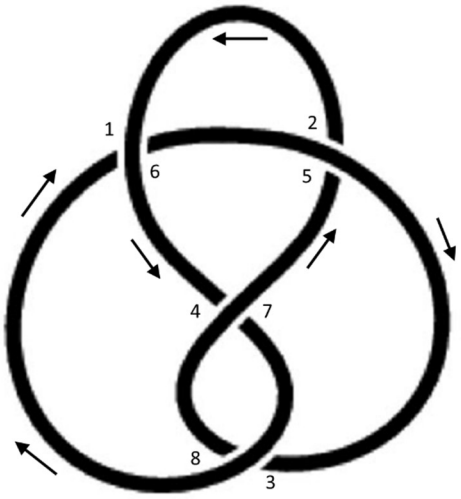


Figure 123. Dowker notation applied to figure-eight knot

For the example at hand, we have the following labels for the crossings

1	3	5	7
6	8	2	4

As shorthand, the Dowker notation only lists the even numbers. So, for our example, the Dowker representation for the figure-eight knot is 6 8 2 4. This is not unique since the result depends on the starting position and orientation. For example, the Wikipedia article "List of prime knots" [122] lists the Dowker notation for the figure-eight knot as 4 6 8 2. Further, this procedure assumes that the knot alternates between under and over crossings. The procedure can be modified for non-alternating knots.

One can also construct the knot associated with a given Dowker number. For example, take the notation 4 8 10 2 6. We first reconstruct the labels for the crossings, i.e.,

1	3	5	7	9
4	8	10	2	6

Figure 124 shows the steps in reconstructing the knot associated with the above Dowker notation.

- At the top-left of the figure, we start with the crossing labeled 1,4 (by convention, we make this an under-crossing).
- In the second diagram down on the left, we alternate between under and over crossings while adding the crossings 2,7 and 3,8.
- Since 4 is already on the diagram, we need to circle back from 3,8 to 1,4 in the next step (bottom-left of the diagram).
- Next, we add crossings 5,10 and 6,9 (as shown in the top-right diagram).
- At this point, all the crossings are on our diagram, but we still need to continue making connections until we circle back to 1. In the bottom-right diagram, we go from 6,9 to 2,7 to 3,8 to 6,9 to 5,10 and finally back to 1,4. This is the three-twist knot.

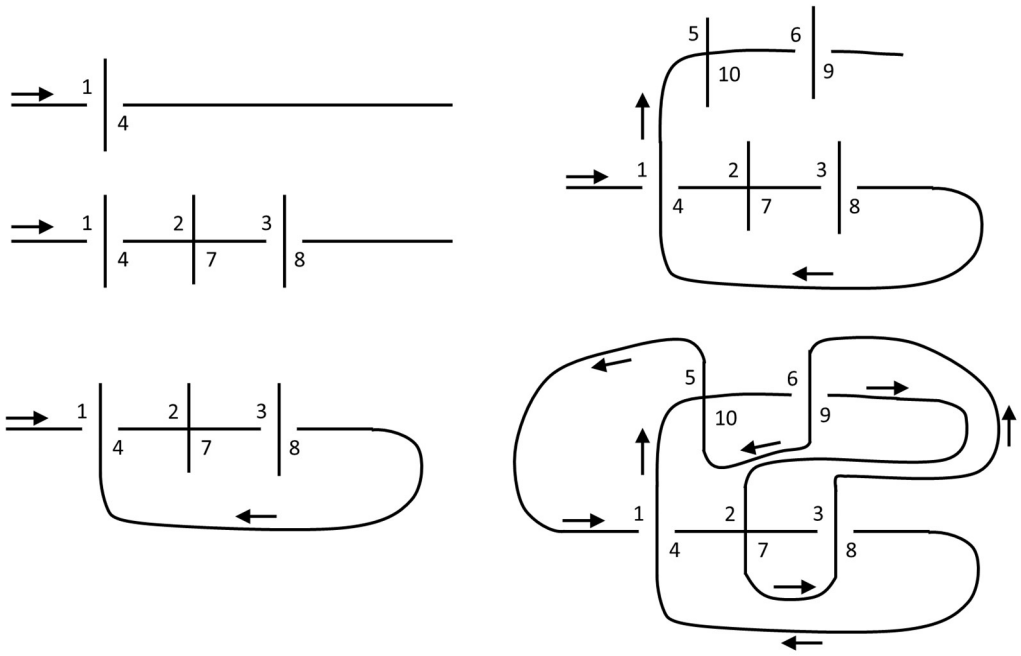


Figure 124. Reconstruction of a knot from its Dowker notation

In the case of composite knots, the Dowker notation leads to multiple alternatives (i.e., composite knots are not uniquely determined by their Dowker representation). Figure 2.8 in the book by Adams [129] provides a detailed example of this ambiguity. For prime knots, a given Dowker representation leads to either a particular knot or its mirror image (see Figure 2.9 in Adams [129]). If the knot is amphicheiral, then the Dowker representation leads to a unique knot.

For non-alternating knots, the Dowker representation needs to be modified as follows:

> If an even integer is assigned while traversing over a crossing, we leave the even integer positive.
>
> However, if the even integer is assigned while traversing under a crossing, we change the sign of the even number to negative.

The smallest non-alternating prime knots have 8 crossings. They are $8_{19}, 8_{20}$ and 8_{21}. If you look at the associated Dowker representation for these three knots in the list of prime knots [122], you will see this is the first place in the list where negative numbers are used.

. . .

Wolfram Alpha (at https://www.wolframalpha.com/) will return a drawing of a knot (and a lot of other information) if you supply the associated Dowker representation. For example, if you enter "knot 4 8 -12 2 14 -6 16 10", Wolfram Alpha returns the knot shown in Figure 125.

Figure 125. 8_21 knot

8.4.3 Conway

The Conway notation for knots (named after mathematician John Horton Conway) is based on the concepts of tangles. In general, a **tangle** is a collection of n disjoint lines enclosed in a sphere. The $2n$ endpoints of the lines are visible at the boundary of the sphere. As applied to knot theory, tangles are restricted to 2 lines bounded by a circle.

The diagram on the left of Figure 126 is an example of a tangle. The two strands extend between the points labeled as NW and SW, and between NE and SE. In general, the four points extending outside of a tangle are labeled as directions on a compass (NW, NE, SW and SE). If we connect NW to NE, and SW to SE, the tangle becomes the figure-eight knot.

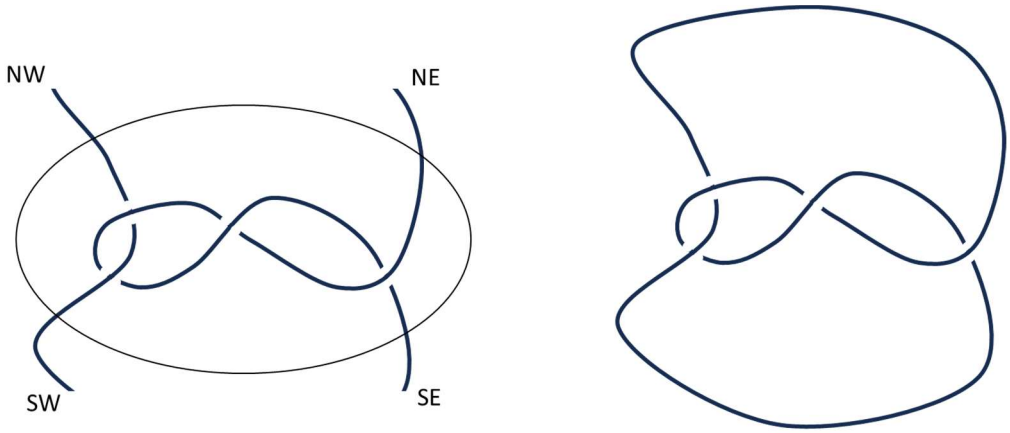


Figure 126. Example of a tangle and associated knot

The insight from Conway was to construct more complex tangles from a set of simple tangles, and to define operations for combining the simple tangles into more complex ones. Conway used tangles to classify all the prime knots up to and including 11 crossings and all prime links up to and including 10 crossings in 1970 [130].

Several of the most basic tangles are shown in Figure 127. In this figure and the following figures, we omit the compass directions (NW, NE, SW and SE) and the surrounding circle. With the exception of the infinity tangle, the number associated with each tangle indicates the number of

crossings. The plus sign means the over-strand has a positive slope at each crossing, and the negative sign means the over-strand has a negative slope at each crossing. The 2 tangle is formed by adding (joining) two instance of the 1 tangle. The -2 tangle is formed in a similar manner by adding two instances of the -1 tangle. In general, two tangles are added by connecting the NE endpoint of the first tangle to the NW endpoint of the second tangle, and the SE endpoint of the first tangle to the SW endpoint of the second tangle.

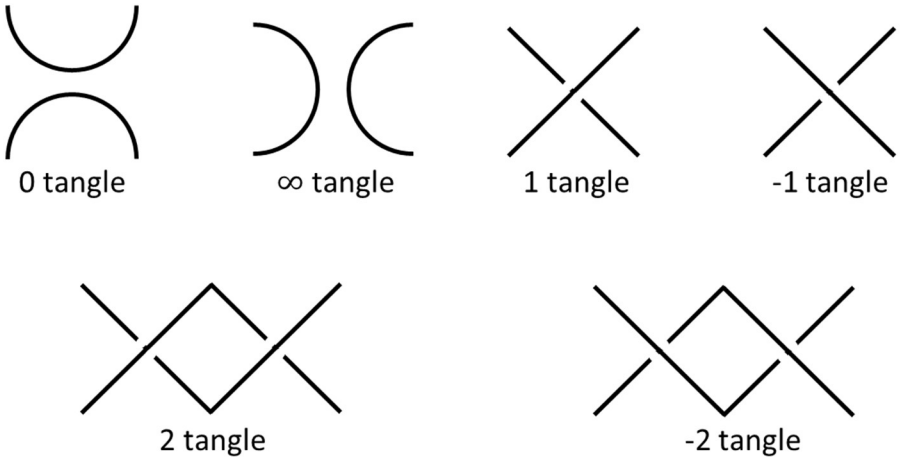


Figure 127. Some basic tangles

Two tangles are equivalent if we can get from one to the other by a sequence of Reidemeister moves [131] while the four endpoints of the strings in the tangle remain fixed and the strings of the tangle never travel outside the circle defining the tangle. The three types of Reidemeister moves are listed below and illustrated in Figure 128.

- Type I: Twist and untwist in either direction
- Type II: Move one loop completely over another
- Type III: Move a string completely over or under a crossing.

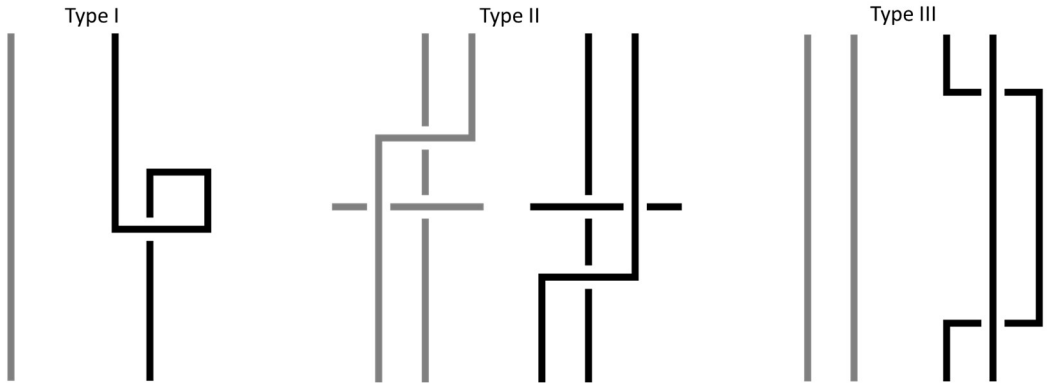


Figure 128. Reidemeister moves

Figure 129 depicts the four tangle operations for creating new (more complex) tangles. (Source: Wikipedia article on tangles [132].)

- The tangle operation on the left of the figure represents a reflection of a tangle about a diagonal line. The reflection of tangle a is denoted $-a$.
- At the top right is the tangle addition operation of which we've already seen two examples, i.e., the 2 and -2 tangles in Figure 127.
- At the center right is the tangle product. The product of tangles a and b is denoted by ab; it is equivalent to $-a + b$.
- At the bottom right is the tangle ramification operation. The ramification of tangle a and tangle b is denoted by a, b; it is equivalent to $-a + -b$.

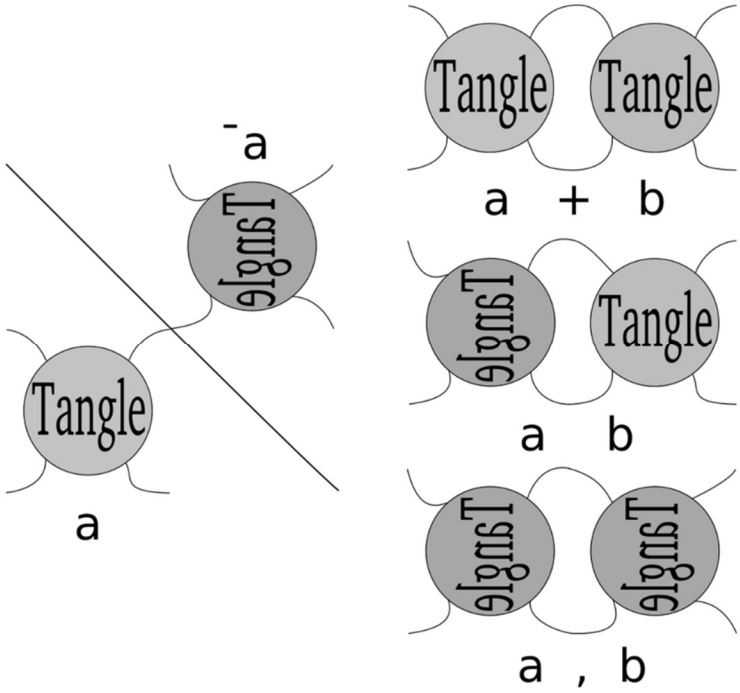


Figure 129. Tangle operations

Figure 130 illustrates the addition of two tangles to get a figure-eight knot. After the addition, the exposed endpoints are connected as shown in the figure.

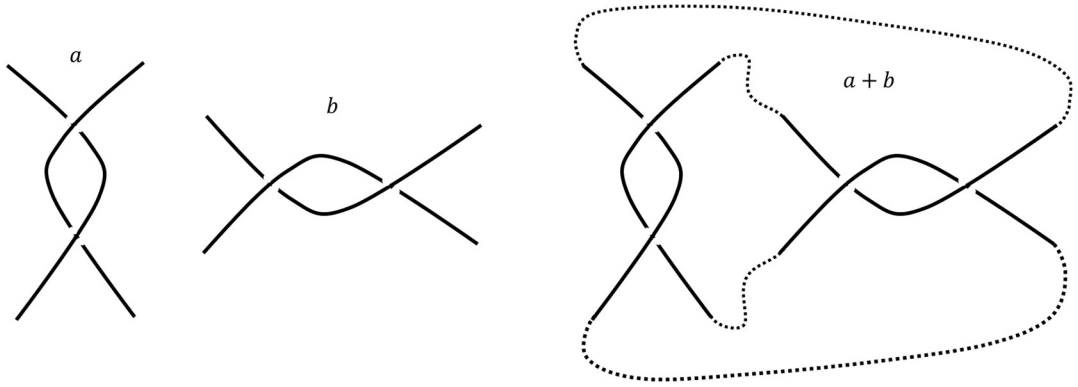


Figure 130. Adding two tangles to get a figure-eight knot

The various online tables of knots (e.g., "List of prime knots" [122] and "the Rolfsen knot table" [133]) provide diagrams of knots that are not drawn from the perspective of the Conway notation. Further, it is usually difficult to associate the Conway notation with the drawings provided in these tables. For example, the reader is invited to reconcile the drawing of 6_2 knot at "List of prime knots" [122] with its Conway notation 3 1 2.

. . .

A **rational tangle** is a special kind of 2-tangle (i.e., a tangle with two strands) that may be unwound into one of the two elementary 2-tangles (i.e., the 0 or ∞ tangle) by twisting the endpoints (two at a time). Not all 2-tangles are rational tangles. The 2-tangle in Figure 131 cannot be untangled by twisting 2 endpoints at a time, and so, it is not a rational tangle.

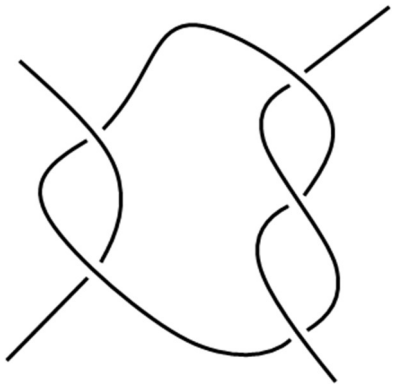


Figure 131. Example of a 2-tangle that is not rational

[**Warning**: There is some potential for confusion here. The term "2-tangle" refers to a tangle with 2 strands. In general, an n-tangle has n strands but in the contexts of knots, we only consider 2-tangles. On the other hand, the term "2 tangle" (as used here) refers to a tangle with two positive twists.]

As an example, consider the −2 3 7 sequence of tangles known as the pretzel knot (see Figure 132). It has two right-handed twists in its first tangle, three left-handed twists in its second, and seven left-handed twists in its third.

The figure is from the Wikipedia article entitled “Tangle (mathematics)” [132].

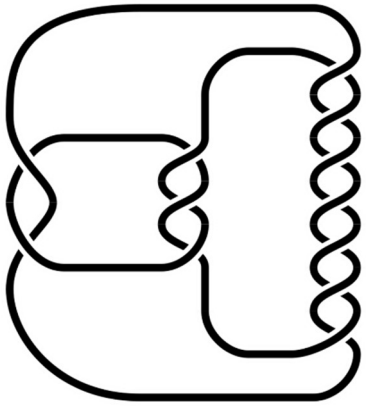


Figure 132. Pretzel knot

A rational tangle can be represented by its continued fraction. For example, the continued fraction for the tangle represented by -2 3 7 in the Conway notation is

$$-2+\frac{1}{3+\left(\frac{1}{7}\right)}=-\frac{37}{22}$$

The convention here is to work from right to left. So, the tangle $u\ v\ w\ x\ y\ z$ would be represented by the continued fraction

$$u+\cfrac{1}{v+\cfrac{1}{w+\cfrac{1}{x+\cfrac{1}{y+\left(\frac{1}{z}\right)}}}}$$

The amazing punching line to all of this is that two rational tangles are equivalent if and only if the continued fraction representations are equal. The theorem was proven by Conway [130]. The journal article “Rational Tangles” [134] gives an elementary and self-contained proof of Conway's theorem on rational tangles. For example, the continued fractions for the tangles $-2\ 3\ 2$ and $3\ -2\ 3$ are both equal to the same number, i.e., $\frac{12}{5}$. Thus, by Conway’s theorem, the two tangles are equivalent.

Further reading: Section 4.3 of the book “An Interactive Introduction to Knot Theory” [135] provides comprehensive details concerning Conway’s notation for knots.

9 Voting Theory

"I care not who casts the votes of a nation, provided I can count them." – Napoleon

"Indeed, you won the elections, but I won the count."
Nicaraguan dictator Anastasio Somoza (1896-1956)

9.1 Overview and Concepts

Voting theory is a branch of mathematics and political science that deals with the study of different methods and systems for aggregating individual preferences into collective decisions, typically in the context of elections or decision-making processes. The goal of voting theory is to design fair, efficient, and meaningful voting systems that accurately represent the preferences of a group while avoiding issues like manipulation and paradoxes [136].

More generally, voting theory falls within the area of social choice theory [137]:

> **Social choice theory** or social choice is a theoretical framework for analysis of combining individual opinions, preferences, interests, or welfares to reach a collective decision or social welfare in some sense. Whereas choice theory is concerned with individuals making choices based on their preferences, social choice theory is concerned with how to translate the preferences of individuals into the preferences of a group. A non-theoretical example of a collective decision is enacting a law or set of laws under a constitution. Another example is voting, where individual preferences over candidates are collected to elect a person that best represents the group's preferences.

The following are some key concepts and methods in voting theory:

- Preference Aggregation or **Rank Choice Voting** [138]: This type of voting involves aggregating the preferences of individual voters to determine a group choice. Each voter ranks a set of alternatives according to their preferences. The individual rankings are then combined to produce a collective ranking.
- Voting Systems: A voting system is a method used to determine the winner of an election based on the preferences of voters. There are many different voting systems, each with its own rules and properties. Some common voting systems include:
    - Plurality Voting: The candidate with the most first-place votes wins, regardless of whether they have a majority.
    - Majority Runoff Voting: If no candidate has a majority of first-place votes, a second round between the top two candidates is held.
    - Instant Runoff Voting (IRV) [139]: Voters in IRV elections rank the candidates in order of preference. Ballots are initially counted to establish the number of votes for each candidate. If a candidate has more than half of the first-choice votes, that candidate wins. If not, the rankings are “instantly” recalculated (i.e., without another election) after removing the candidate with the fewest votes. The process continues until one candidate has more than half of the votes, and that person is declared the winner. IRV is not a proportional voting system but a "winner-takes-all" method, because it results in only one winner. This is a form of rank choice voting.

- Borda Count [140]: This is a family of positional voting rules which gives each candidate, for each ballot, a number of points corresponding to the number of candidates ranked lower. In the original variant, the lowest-ranked candidate gets 0 points, the next-lowest gets 1 point, etc., and the highest-ranked candidate gets $n - 1$ points, where n is the number of candidates. Once all votes have been counted, the option or candidate with the most points is the winner. The Borda count is intended to elect broadly acceptable options or candidates, rather than those preferred by a majority, and so, it is often described as a consensus-based voting system rather than a majoritarian one.
- Condorcet Method [141]: an election method that elects the candidate who wins a majority of the vote in every head-to-head election against each of the other candidates, that is, a candidate preferred by more voters than any others, whenever there is such a candidate. A candidate with this property, the pairwise champion or beats-all winner, is formally called the Condorcet winner. The head-to-head elections need not be done separately; a voter's choice within any given pair can be determined from the ranking.
- Approval Voting: Voters can vote for as many candidates as they like, and the candidate with the most approvals wins.

- Arrow's Impossibility Theorem [142]: According to Arrow's impossibility theorem, in all cases where preferences are ranked, it is impossible to formulate a social ordering (i.e., group preference) while meeting some very reasonable criteria (described in Section 9.4).
- Gibbard-Satterthwaite Theorem [143]: In simple terms, this theorem states that a dictatorship is the only voting system for three or more candidates that cannot be manipulated.

. . .

The following are some definition that we will use in the remainder of this section.

- A **majority**, also called a simple majority or absolute majority to distinguish it from related terms, is more than half of the total. It is a subset of a set consisting of more than half of the set's elements.[144]
- A majority can be compared to a **plurality** (sometimes called relative majority), which is a subset larger than any other subset but not necessarily larger than all other subsets combined, and not necessarily greater than half of the set.[144] For example, consider a group with 40 members which is divided into subgroups with 18, 12, and 10 members, then the 18-member group would be the plurality.
- An **election** is a process where voters decide among several choices. In this section and in the context of voting theory, this term is meant to be used generically in any situation where a vote is taken to decide an issue, and is not restricted to political elections.
- A **candidate** (or option) is one of several choices to be selected in an election. A candidate could be a person or an option to be selected (e.g., choice of color scheme to be used for website).
- **Instant runoff voting** entails several voting procedures where (1) voters lists there choices in rank order, (2) some scheme is used to eliminate candidates until a candidate with the

majority of votes is left. This is done “instantly” (perhaps “automatically” is a better word) based on an agreed elimination scheme and without any further elections.

9.2 Simple Elections

For the purposes of this document, a simple election is one in which there are multiple candidates (choices) and only one is elected (chosen). Further, there is a well-defined set of voters (known as the electorate) and each voter casts one vote. [In contrast, a complex election entails the selection of two or more winners in an election.]

For the simplest case, i.e., only two candidates, the situation is straightforward. The only complication is when there are a small number of voters (e.g., a city council voting on a law), and there is a tie. In this case, a tie resolution procedure is necessary. If there are a large number of voters, then the possibility of a tie is almost zero.

For three or more candidates, complications arise (even for simple elections). For example, consider an election for mayor in a city where candidate X gets 40% of the vote, candidate Y gets 35% of the vote and candidate Z gets 25% of the vote. There is no majority. Also, there is no preferencing ranking by the voters, i.e., just one vote per voter for the candidate of their choice.

- One option is for the city to only require a plurality to win an election. In this case, candidate X would win. However, it may be that voters for candidate Y and Z hate candidate Z. Moreover, the voters for candidate Y favor candidate Z over X, and the voters for candidate Z favor Y over X. With this approach, a candidate unfavored by 60% of the voters wins the election.
- Another option would be to eliminate the candidate with the least number of votes (candidate Z) and then have a run-off election between the remaining candidates (candidates X and Y). Using this scheme, candidate Y would win 60% to 40% (under the assumption that the voters who previously chose X, prefer Y over Z).

. . .

Another approach is to allow the voters to rank their choices. Table 12 shows the results of a rank choice election. There are a total of 49 voters, each of whom ranks their preferences from among options A, B, C and D. Each column in the table represents the number of voters that have selected a given combination. For example, column #1 indicates that 10 voters favor A, then C, then D and finally B. There are a total of $4! = 24$ possible combinations but for the example, only 6 of the possible combinations have been chosen by any of the voters.

Table 12. Rank Choice Voting example

10	**15**	**8**	**6**	**6**	**4**	**0**
A	B	A	C	D	D	All other combinations
C	D	D	D	C	C	
D	A	C	B	A	B	
B	C	B	A	B	A	

If majority voting is used, no one wins. A gets the most votes first-place votes (18) but that is less than 50%.

If only a plurality is required, then A wins with 18 votes, and B, C and D get 15, 6 and 10 votes, respectively.

If the runoff method is used, we eliminate all but the two candidates with the most votes. In this case, we eliminate C and D, and hold another election between A and B. Assuming voter preferences don't change, we have the following table:

10	**15**	**8**	**6**	**6**	**4**
A	B	A	B	A	B
B	A	B	A	B	A

This would result in B winning over A by a vote of 25 to 24.

If we use the instant runoff method, we eliminate candidates one at a time until one candidate has a majority of the votes. So, we first eliminate C (who got the least number of votes), and get the following table:

10	**14**	**8**	**6**	**6**	**4**
A	B	A	D	D	D
D	D	D	B	A	B
B	A	B	A	B	A

The new results are (instantly) computed, and we see that A gets 18, B gets 14 and D gets 16. Since no candidate has a majority (25 in this case), we do another round and eliminate B. This gives us

10	**14**	**8**	**6**	**6**	**4**
A	D	A	D	D	D
D	A	D	A	A	A

In the final instant runoff, D wins over A by a vote of 30 to 18. This may seem odd since D came in third place in the initial vote.

Based on three different voting methods (plurality, run-off and instant run-off), we got 3 different winners. This example illustrates the power of being able to select the voting method, especially if one has some indication of voting patterns (e.g., via polling before an election).

. . .

Let's try another example using the instant runoff method. This time using percentages and five candidates. The initial vote and associated preferences are shown below. Again, only some of the possible voting patterns occurred in the election.

$$30\%: A > B > C > D > E$$
$$25\%: B > C > D > A > E$$
$$20\%: C > D > E > B > A$$
$$15\%: D > E > C > B > A$$
$$10\%: E > D > C > B > A$$

Since no candidate has a majority, we eliminate the candidate with the smallest percentage of votes (i.e., E) and do a second round. The updated vote is

$$30\%: A > B > C > D$$
$$25\%: B > C > D > A$$
$$20\%: C > D > B > A$$
$$15\%: D > C > B > A$$
$$10\%: D > C > B > A$$

There is still no candidate with a majority. At this point, we have A at 30%, B at 25%, C at 20% and D at 25%. We eliminate the candidate with the smallest percentage of votes (i.e., C) and do a third round.

$$30\%: A > B > D$$
$$25\%: B > D > A$$
$$20\%: D > B > A$$
$$15\%: D > B > A$$
$$10\%: D > B > A$$

Now, D is in the lead with 45% of the votes, A is in second with 30% and B trailing with 25%. Next, we eliminate B and have the final round.

$$30\%: A > D$$
$$25\%: D > A$$
$$20\%: D > A$$
$$15\%: D > A$$
$$10\%: D > A$$

In the final round, D wins the majority with 70% of the votes. Again, the conclusion seems at least a little odd since D was in fourth place in the initial vote.

. . .

The elimination method that we used in the previous instant runoff elections is called the **Hare method** (invented by the English lawyer Sir Thomas Hare in 1859). As we saw, the candidate with

the least number (or percentage) of first place votes gets eliminated in each round when using the Hare method. There is also the **Coombs rule** (named after American psychologist Clyde Coombs) where the candidate with the most last place votes is eliminated in each round.

If we apply Coombs rule to the previous example (with 5 candidates), then E is eliminated after the first round since E has the most last place votes. Going into round #2, we have

$$30\%: A > B > C > D$$
$$25\%: B > C > D > A$$
$$20\%: C > D > B > A$$
$$15\%: D > C > B > A$$
$$10\%: D > C > B > A$$

This time we eliminate A since A has 70% of the last place votes. Going into round #3, we have

$$30\%: B > C > D$$
$$25\%: B > C > D$$
$$20\%: C > D > B$$
$$15\%: D > C > B$$
$$10\%: D > C > B$$

In this round B wins with a 55% majority.

. . .

In a variation of rank choice voting, the voters are not required to rank all the candidates. For example, if there are five candidates (A, B, C, D and E), a voter may only rank two, e.g., E first and C second.

In yet another variation, the preferences are weighted, e.g., 3 for first place vote, 2 for second place vote, 1 for third place vote, and 0 for 4th place and beyond.

9.3 Condorcet's Method

A **Condorcet method** (named after the Marquis de Condorcet) is an election method that selects the candidate who wins a majority of the vote in every head-to-head election against each of the other candidates. A candidate with this property, the pairwise champion or beats-all winner, is known as the Condorcet winner. The head-to-head elections need not be done separately; a voter's choice relative to any given pair can be determined using rank choice voting followed by instant runoffs.

For example, take the ranked voting results shown in Table 13.

Table 13. Example of the Condorcet method with 3-way tie

7	**5**	**11**	**3**	**13**
D	A	C	C	A
C	B	D	A	B
B	C	A	D	D
A	D	B	B	C

In the initial vote, A gets 18 votes, C gets 14 votes and D gets 7 votes. So, no candidate get a majority of the 39 votes. If we do an instant runoff using the Condorcet method, the head-to-head results are as follows:

- A wins over B by 32-7
- C wins over A by 21-18
- A wins over D by 21-18
- C wins over B by 21-18
- D wins over B by 21-18
- D wins over C by 20-19

A, C and D each win 2 of the head-to-head competitions, and so, we have a three-way tie.

Removing B does not help in this case. As one can see from the table below and associated results, there is still a three-way tie if we remove B.

7	**5**	**11**	**3**	**13**
D	A	C	C	A
C	C	D	A	D
A	D	A	D	C

- C wins over A by 21-18
- A wins over D by 21-18
- D wins over C by 20-19

As a second example, we slightly change the vote from our previous example (see the gray cells in Table 14 as compared to the same cells in Table 13).

Table 14. Example of the Condorcet method with one winner

7	**5**	**11**	**3**	**13**
D	A	C	C	A
C	B	D	A	B
B	C	A	D	C
A	D	B	B	D

The results of the instant runoff using the Condorcet method (shown below) give us a single winner, i.e., C who wins every head-to-head competition (in which C is involved).

- A wins over B by 32-7
- C wins over A by 21-18
- A wins over D by 21-18
- C wins over B by 21-18
- D wins over B by 21-18
- C wins over D by 32-7

. . .

The Condorcet method can sometimes lead to paradoxical conclusions. For example, consider the vote shown in Table 15. In head-to-head competition, A beats B (8-4), B beats C (9-3) and C beats A (7-5). So, transitivity does not necessarily hold when employing the Condorcet method, and there is no winner in such cases (just a multi-candidate tie). This anomaly is known as the **Condorcet paradox**.

Table 15. Example of the Condorcet paradox

5	**4**	**3**
A	B	C
B	C	A
C	A	B

. . .

Other strange things can happen with the Condorcet method. In the election shown in Table 16, candidate B has (by far) the least number of votes. However, in head-to-head competition, B beats A by 59-41 and B beats C by 60-40. So, B is the Condorcet winner.

Table 16. Candidates with least number of votes wins by Condorcet method

41	**19**	**40**
A	B	C
B	C	B
C	A	A

Further, if we use the Hare method, B is eliminated first and we then have the election shown in the table below. In this case, C is the winner.

41	**19**	**40**
A	C	C
C	A	A

. . .

In general, an electoral system satisfies the **Condorcet winner criterion** if it always chooses the Condorcet winner when one exists. The Wikipedia article on this topic [145] provides an extensive list of electoral systems with an indication of which systems satisfy the Condorcet winner criterion.

An electoral system that never allows a Condorcet loser to win is said to satisfy the **Condorcet loser criterion** [146]. A Condorcet loser is defeated in every head-to-head competition against each other candidate.

. . .

In addition to the cyclic paradox noted previously, there is another way for the Condorcet method not to result in a winner, i.e., when no candidate wins a majority of the head-to-head elections. In cases where there is no Condorcet winner, the **Condorcet's extended method** can be used to select a winner. In this approach, we list each head-to-head election in order of the largest margin of victory, and then compute an ordered lists.

Consider the election results in Table 17. In head-to-head elections, A beats B by 13-6, B beats C by 15-4 and C beats A by 10-9. So, no candidate wins all their head-to-head elections, and thus, there is no Condorcet winner.

Table 17. Election with no Condorcet winner

9	**6**	**4**
A	B	C
B	C	A
C	A	B

Using Condorcet's extended method, we order the results by the largest margin of victory and proceed as follows:

- B over C by 15-4 implies $B > C$.
- A over B by 13-6 implies $A > B$. The two results thus far imply $A > B > C$.
- C beats A by 10-9 which implies $C > A$ but we already have $A > C$. In such cases where a latter result contradicts a former result in the computation, Condorcet's extended method tells us to ignore the latter result.

So, the final result using Condorcet's extended method is A first, B second and C third.

9.4 Arrow's Impossibility Theorem and the Gibbard-Satterthwaite Theorem

In a **cardinal voting scheme**, each voter rates the candidates independently. This allows for the possibility of ties in a voter's rating of the candidates. For example, consider an election using a cardinal voting scheme that has four candidates (A, B, C and D). Possible votes are 1 for "approve", 0 for "neutral" and -1 for "disapprove". A given voter could give a vote of 1 to candidates A and B, and a vote of 0 for C and D. The Wikipedia article on cardinal voting [147] provides additional details and some examples.

Ordinal voting, also known as ranked-choice voting or preferential voting, is a voting system in which voters express their preferences by ranking candidates or options in order of preference. In ordinal voting, a given voter cannot indicate a tie in their preferences. In this section, the focus is on ordinal voting.

A voting system that allows for the creation of a rank ordered list of the candidates is called a **rank order voting system**. One example is Condorcet's extended method. In the absence of ties (which are very unlikely with a large number of voters), plurality voting also allows one to produce a rank ordered list of candidates.

Some ordinal voting schemes are not rank order voting systems, e.g., the instant runoff approach finds a winner and then stops (possibly before ranking all the candidates). However, any ordinal voting approach can be modified so that it produces a ranking of all the candidates. For example, use a given scheme to compute a winner, say it is A. Next, modify the preference profile by deleting A and then recomputing the winner, say B. Continue by removing A and B from the preference list, and recomputing the winner. This process will produce an ordered list of all the candidates. The question is whether this procedure or some other procedure can convert the set of ranked preferences (by individual voters) in an ordinal election into a preference ranking for the entire electorate (i.e., all voters as a group) while meeting the following criteria:

- Non-dictatorship: No single voter should be able to determine the outcome of an election.
- Independence of Irrelevant Alternatives [148]: If A is selected over B out of the choice set {A,B} by a voting rule for given voter preferences of A, B, and an unavailable third alternative X, then if only preferences for X change, the voting rule must not lead to B's being selected over A.
- Pareto Efficiency or Unanimity: If every voter prefers A to B, then the voting scheme cannot conclude that the electorate as a whole prefers B to A.

In 1950, Kenneth J. Arrow answered this question in the negative in what is now known as Arrow's Impossibility Theorem [149]. Arrow received the 1972 Nobel Prize in Economics for his work on this topic. Arrow's impossibility theorem states that whenever an election has more than 2 candidates, then the following three conditions become incompatible: non-dictatorship, independence of irrelevant alternatives and Pareto efficiency. The Wikipedia article entitled "Arrow's impossibility theorem" [150] provides a proof of the theorem.

As a follow-up to Arrow's work, Gibbard (in 1973) and Satterthwaite (in 1975) independently proved a stronger theorem. The so called Gibbard-Satterthwaite Theorem deals with deterministic ordinal electoral systems that choose a single winner [143]. It states that for every voting system, one of the following three things must hold:

- The system is dictatorial, i.e., there exists a distinguished voter who can choose the winner.
- The system limits the possible outcomes to two alternatives.
- The system is susceptible to tactical voting [151]. Tactical voting (also known as strategic voting, sophisticated voting or insincere voting) occurs when a voter chooses a candidate, other than their sincere preference, to prevent an undesirable outcome. For example, a voter (or block of voters) may choose an option they perceive as having a greater chance of winning versus an option they prefer.

10 Application of Probability to Genetics

"The apple does not fall far from the tree."
Old proverb

10.1 Background

10.1.1 Probability

Probability is the branch of mathematics that concerns numerical descriptions of how likely an event is to occur (or not occur). The probability of an event is a number between 0 and 1, where a probability 0 indicates impossibility of the event and a probability 1 indicates an event will occur, e.g., the probability of tossing a coin and getting either heads or tails is 1 since one of the other must occur. The higher the probability of an event, the more likely it is that the event will occur. A simple example is rolling a fair (unbiased) die. Since the die is fair, there are six equally probable outcomes each with probability $\frac{1}{6}$.

In some cases, the probability of an event can be computed using combinatorics. In such cases, one calculates the total number of possible outcomes for some situation (e.g., rolling two dice) and the number of outcomes that yield the desired event (e.g., a roll of 7 when summing the outcome of each die). For the two dice example, the total number of outcomes is 36 and the number of ways to roll a 7 is 6, i.e., (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), and (6, 1). So, the probability of rolling a 7 with two dice is $\frac{6}{36} = \frac{1}{6}$ (i.e., number of outcomes resulting in the desired event divided by the total number of possible outcomes for the given situation).

Two events A and B are **mutually exclusive** if both cannot happen. Two events A and B are **independent** if the occurrence of one does not affect the probability of the other.

Some basic probability rules:

- Complement Rule: The probability of the complement of an event A (not A, written $\neg A$) is 1 minus the probability of A, i.e., $P(\neg A) = 1 - P(A)$ where $P(A)$ is read "the probability of event A."
- Union Rule: The probability of A or B (written $A \cup B$ and read "A union B") is the sum of their individual probabilities minus the probability of their intersection, i.e., $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. In words, the probability of A or B happening equals the probability of A plus the probability of B minus the probability of both A and B happening.
- If two events A and B are mutually exclusive, then $P(A \cap B) = 0$, and we have $P(A \cup B) = P(A) + P(B)$.
- Conditional Probability: The probability of event A occurring given that event B has occurred is denoted by $P(A \mid B)$ and is calculated as $\frac{P(A \cap B)}{P(B)}$, where $P(B)$ is not equal to 0.
- Independence: If events A and B are independent, then $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$. In this case, $P(A \cap B) = P(A)P(B)$, i.e., probability of A and B happening for independent events is just the product of the probability of each event. For example, the probability of rolling a 3 with a die and getting tails with the flip of a coin is $\frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$.

This is about all we need for the task at hand. For a more detailed introduction to probability, see "Chapter 3: Probability" from the online book "College Mathematics for Everyday Life" [157].

10.1.2 Genetics

Genetics is a branch of biology that focuses on the study of heredity (the process by which traits are passed from one generation to the next). It encompasses the study of genes, DNA, and how these molecules contribute to the variation and inheritance of traits in living organisms. [158]

10.1.2.1 Definitions and Concepts

DNA and Genes:

- A **molecule** is the smallest particle of a substance that retains all the properties of the substance and is composed of one or more atoms, i.e., a molecule of water is comprised to two hydrogen and one oxygen atoms.
- Deoxyribonucleic acid (DNA) is a molecule that carries the genetic information in all living organisms. It is a double-stranded helical structure made up of four nucleotide bases: adenine (A), thymine (T), cytosine (C), and guanine (G).
- Ribonucleic acid (RNA) is a molecule used to code, decode, regulate, and express genes. Forms of RNA include messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). RNA codes for amino acid sequences, which may be combined to form proteins. Where DNA is used, RNA acts as an intermediary, transcribing the DNA code so that it can be translated into proteins.
- **Genes** are specific segments of DNA that contain instructions for the synthesis of proteins and other essential molecules for sustaining life. Genes are the basic unit of heredity. There are two types of molecular genes: protein-coding genes and non-coding genes. Protein-coding genes contain instructions (genetic code) for the synthesis of specific proteins. Protein-coding genes play a crucial role in determining an organism's traits, functions, and overall biology. Non-coding genes comprise a diverse category that includes DNA coding for non-translated ribonucleic acid (RNA), such as that for ribosomal RNA, transfer RNA, ribozymes, small nuclear RNAs, and several types of regulatory RNAs.
- An **allele** is a term used to describe one of two or more alternative forms of a gene that can occupy a specific position, or locus, on a chromosome. Alleles are responsible for the variations in traits and characteristics that can be inherited from one generation to the next. One can view an allele as a variant of a gene. Each gene can have one or more alleles.

Chromosomes:

- The focus in this section is on the eukaryotes, i.e., organisms whose cells have a membrane-bound nucleus. All animals, plants, fungi, and many unicellular organisms are eukaryotes. They constitute a major group of life forms alongside the two groups of prokaryotes, i.e., bacteria and archaea.
- In eukaryotic cells, DNA is organized into structures called chromosomes. Humans, for example, have 46 chromosomes (23 pairs) in each cell.
- Each chromosome contains numerous genes, and the combination of genes on chromosomes determines an individual's traits. The human genome has about 19,000 -

20,000 protein-coding genes and an even larger number of non-coding genes (see Table 1 in the journal article “Open questions: How many genes do we have?” [160]).

- **Homologous chromosomes**, or homologs, are a set of one maternal and one paternal chromosome that pair up with each other inside a cell during fertilization. Homologs have the same genes (but possibly different alleles) in the same location in the containing chromosome.
- **Diploid** cells have two homologous copies of each chromosome, usually one from the mother and one from the father. All or nearly all mammals are diploid organisms. Human diploid cells have 46 chromosomes and human gametes (egg and sperm) have 23 chromosomes. Gametes (having only one set of chromosomes) are said to be **haploid**.
- For the sake of completeness (but not to be discussed further in this document), we note that the cells of some organisms have more than two homologous chromosomes. Such cells are referred to as having the **polyploidy** characteristic. Most species whose cells have nuclei (eukaryotes) are diploid, meaning they have two complete sets of chromosomes, one from each of two parents; each set contains the same number of chromosomes, and the chromosomes are joined in pairs of homologous chromosomes. However, some organisms are polyploid. Polyploidy is especially common in plants.
- A **sex chromosome** (also referred to as an **allosome**) is a type of chromosome involved in sex determination. Humans and most other mammals have two sex chromosomes, X and Y, that in combination determine the sex of an individual. Females have two X chromosomes in their cells, while males have one X and one Y chromosome. An **autosome** is one of the numbered chromosomes in an organism, as opposed to the sex chromosomes. For example, humans have 22 pairs of autosomes and one pair of sex chromosomes (XX or XY). Autosomes are numbered roughly in relation to their sizes. The largest human autosome (chromosome 1) has approximately 2,800 genes; the smallest autosome (chromosome 22) has approximately 750 genes. [161]

Genotype and **Phenotype** [162]:

- The genotype of an organism is its complete set of genetic material. Genotype can also be used to refer to the alleles or variants an individual carries in a particular gene or genetic location. The number of alleles an individual can have in a specific gene depends on the number of copies of each chromosome found in that species, also referred to as ploidy. In diploid species like humans, two full sets of chromosomes are present, meaning each individual has two alleles for any given gene. If both alleles are the same, the genotype is referred to as **homozygous**. If the alleles are different, the genotype is referred to as **heterozygous**.
- Genotype contributes to phenotype, i.e., the observable traits and characteristics of an individual organism. The degree to which genotype affects phenotype depends on the trait. For example, the blossom color in a pea plant is exclusively determined by genotype. The blossoms can be purple or white depending on the alleles present in the pea plant. However, other traits are only partially influenced by genotype. These traits are often called complex traits because they are influenced by additional factors, such as environmental or epigenetic factors. Not all individuals with the same genotype look or act the same way because appearance and behavior are modified by environmental and growing conditions. Likewise, not all organisms that look alike necessarily have the same genotype.

On the left of Figure 133, we have a cell with a nucleus containing multiple chromosome pairs. Moving to the right, we see a magnification of a chromosome pair and two genes (represented as white and checkered rectangles). The two genes code for two different variants (alleles) of the same type of molecule (e.g., a protein that controls eye color). The two alleles constitute a genotype for the organism containing the given cell. There are various ways that different alleles can interact, e.g., the checkered allele (for a protein that results in brown eyes) might dominate the white allele (for a protein that results in blue eyes), resulting in a protein for brown eyes being produced which, in turn, leads to brown eyes (the observable trait, i.e., a phenotype).

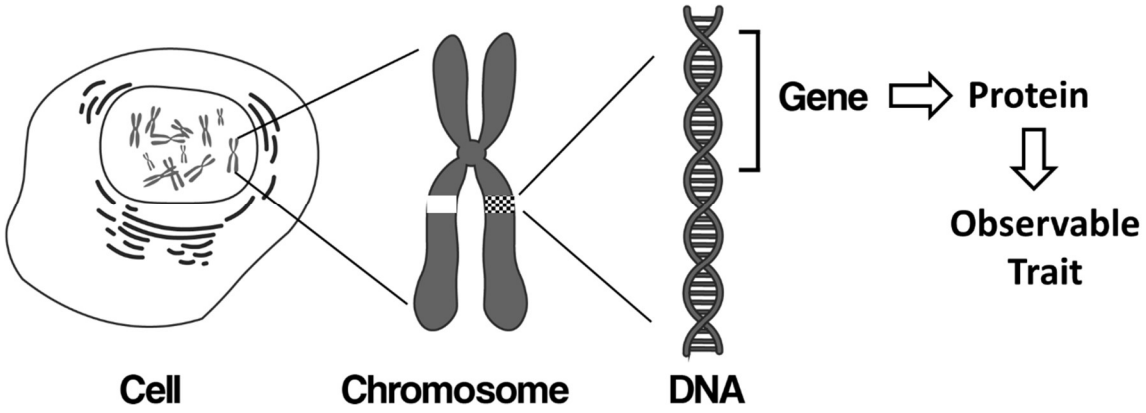


Figure 133. Cell, chromosome, gene and observable trait

Reproduction: Reproduction is the process by which organisms produce offspring. There are two main types of reproduction: asexual and sexual.

- Asexual reproduction involves the production of genetically identical offspring (except in the case of mutations) from a single parent. It does not involve the formation of gametes (sex cells) and is common in simpler organisms.
- Sexual reproduction involves the fusion of **gametes**, typically a sperm and an egg, resulting in genetic diversity in the offspring. A **somatic cell** is any biological cell forming the body of a multicellular organism other than a gamete. Somatic cells form the body of an organism and divide through the process of mitosis or binary fission (for some single celled organisms).

There are two types of cell division: **mitosis** and **meiosis**. Our focus here is on the latter of the two.

- Mitosis is a type of cell division that occurs in somatic (non-reproductive) cells of multicellular organisms. It is a process through which a single eukaryotic cell [159] divides into two genetically identical daughter cells, each with the same number of chromosomes as the original cell. Mitosis is essential for growth, tissue repair, and the maintenance of the body's cell population. It plays a key role in ensuring that each daughter cell inherits a complete and identical set of genetic material from the parent cell.
- Meiosis is a specialized type of cell division that occurs in sexually reproducing organisms. Its primary purpose is to reduce the chromosome number in half, ensuring that the resulting gametes (e.g., sperm and egg cells) have half the genetic material of the parent cells. Meiosis is essential for maintaining genetic diversity in populations and for the inheritance of traits from one generation to the next.

For readers not familiar with mitosis and meiosis, it is recommended to view the following videos before proceeding:

- Mitosis: The Amazing Cell Process that Uses Division to Multiply! [153]
- Meiosis [154]
- Meiosis – Plants and Animals [155]
- Mitosis vs. Meiosis: Side by Side Comparison [156].

Figure 134 provides a high-level summary of the two phases of meiosis. The figure only shows the nuclei of each cell. Starting on the left, we see the nucleus of a diploid somatic cell (non-gamete) with two pairs of chromosomes. The chromosomes labeled A form one pair (coming from one parent) and the chromosomes labeled B form another pair (coming from the other parent). Next, the chromosomes replicate, form homologous chromosomes and exchange some genes. The cell and associated nucleus then splits into two daughter nuclei. Finally, the two nuclei split into four gametes which, in turn, may be paired with the gamete of another organism in sexual reproduction. A key point is that each of the four gametes has a different combination of alleles from the chromosomes in the original somatic cell.

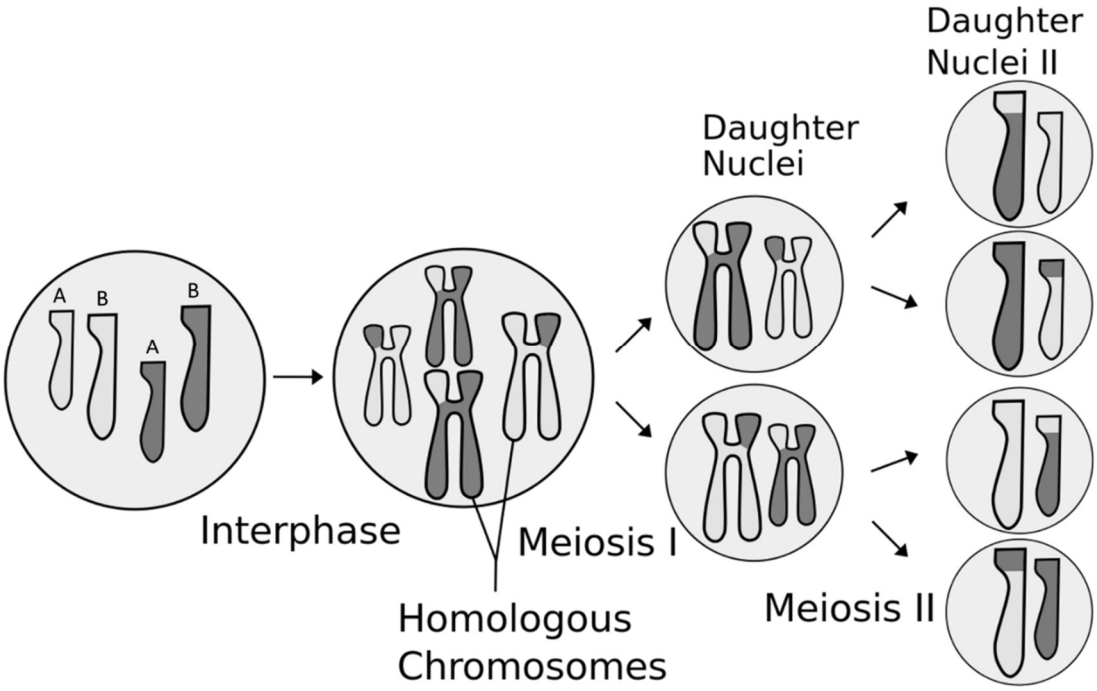


Figure 134. Two stages of meiosis

10.1.2.2 Lifecycle and Heredity

Figure 135 (Source: Wikipedia article entitled “Sexual reproduction” [163]) is a high-level diagram of the lifecycle process for sexually reproducing organisms. This is typical in animals, though the number of chromosome sets and how that number changes in sexual reproduction varies, especially among plants, fungi, and other eukaryotes. Section 11.2 of the online text *Biology* [164] shows the details of several variations for the lifecycles of animals, fungi and plants.

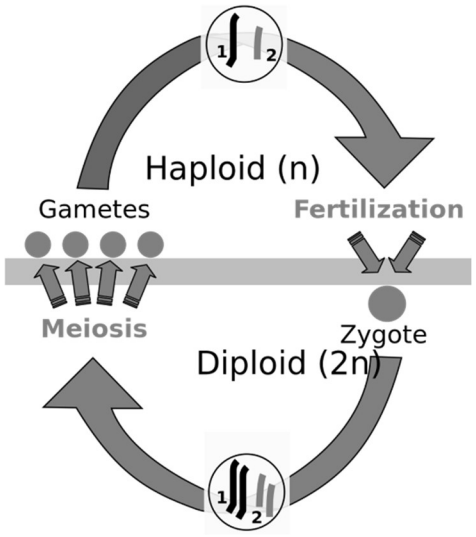


Figure 135. Generic lifecycle diagram

The general lifecycle is as follows:

- Starting with a single cell zygote, a fully formed new organism arises after many cell divisions and cell specialization. [Yes, we could have started anywhere in the process – the old chicken and egg quandary.]
- Each somatic cell in the adult organism has several pairs of chromosomes (one from each parent). Each pair of chromosomes has the same genes in the same locations. However, the allele from the male parent and the associated allele from the female parent may provide different variations of the same gene. See the example in Figure 136.
    - Laws for how different alleles interact have been studied since the time of their discovery by Gregor Mendel. Mendel published his results in 1866 but the profound significance of Mendel's work was not recognized until the turn of the 20th century with the rediscovery of his laws. This will be discussed in more detail in the next section.
- Four gametes are created from a single somatic cell through the process of meiosis. Each chromosome in a gamete has a mixture of the alleles from the two associated chromosomes in the given somatic cell. This process is repeated many times. Each gamete cell has half the number of chromosomes than the somatic cells for a given species.
- A gamete from a male and a gamete from a female are united during sexual reproduction. The female gamete is said to be fertilized by the male gamete, and a zygote is formed – thus, completing the cycle.

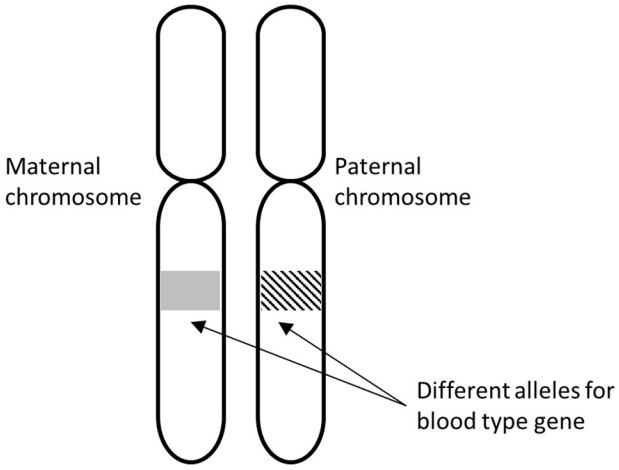


Figure 136. Homologous chromosomes in a somatic cell

10.2 Laws of Heredity

10.2.1 Mendelian inheritance

Mendelian inheritance is a biological pattern that follows the principles of segregation and independent assortment, which were first proposed by Gregor Mendel (1822-1884). Mendel was a German-Czech biologist, meteorologist, mathematician, Augustinian friar and abbot of St. Thomas' Abbey in Brno. Mendel's principles were derived from his experiments in crossing pea plants with different observable traits such as blossom color. He found that the traits were passed down to the offspring in predictable patterns.

The main principles of Mendelian inheritance are the following [182]:

- Principle of Segregation (The First Law): Mendel's first law states that an individual organism has two alleles (gene variants) for each gene, one inherited from each parent. These alleles segregate (separate) during gamete formation (the process that produces eggs and sperm), so each gamete carries only one allele for each gene. When fertilization occurs, the offspring inherits one allele from each parent, restoring the diploid number of alleles.
- Principle of Independent Assortment (The Second Law): Mendel's second law states that the alleles of different genes assort (distribute) independently during gamete formation. In other words, the inheritance of one gene does not influence the inheritance of another gene located on a different chromosome. This principle assumes that genes are located on different chromosomes or are far apart on the same chromosome.
- Principle of Dominance: Mendel's experiments showed that some alleles are dominant, while others are recessive. Dominant alleles mask the expression of recessive alleles in heterozygous individuals (those with two different alleles for a particular gene). Only when an individual carries two recessive alleles (recessive homozygous) will the recessive trait be expressed.
- Principle of Unit Characters: Mendel proposed that genes are responsible for the inheritance of specific traits, and each gene corresponds to a specific characteristic (unit character). These unit characters are inherited independently, and the combination of unit characters determines an individual's phenotype (observable traits).

Mendelian inheritance is a fundamental concept in genetics and is used to explain the inheritance of many different traits, including eye color, hair color, and blood type. However, there are extensions and exceptions to Mendel's laws of inheritance, as we shall see in the next section.

Some additional details about Mendel's experiments:

- Mendel chose to study pea plants because they are easy to grow and have a relatively short lifespan.
- He crossed pea plants that were purebred for different traits, such as seed color (yellow vs. green) and seed shape (round vs. wrinkled).
- He carefully counted the number of offspring with each trait in each generation.
- From his experiments, Mendel concluded that the traits he was studying were inherited in a particulate manner, meaning that they were passed down from parents to offspring in discrete units. He also concluded that the two alleles for each trait segregated into the gametes and assorted independently of each other.

Amazingly, Mendel theories came well before the discovery of the internal structures within cells such as chromosomes and DNA.

. . .

For example, consider a pea gene that codes for flower color. There are two alleles for this gene, i.e., the purple blossom allele (represented by P) and the white blossom allele (represented by w). The purple blossom allele is dominant, and the white blossom allele is recessive. So, if a pea plant inherits the P allele from one parent and the w allele from the other, the plant will have the Pw genotype and the purple blossom phenotype since P is dominant and w is recessive.

Table 18 show the cross of two different pure breed pea plants with respect to blossom color, i.e., one plant is of genotype ww and the other is PP. The only possible result is a purple plant with genotype Pw, because P (purple blossoms) is a dominant trait.

This type of table is known as a Punnett square [165]. Analyses using such squares were devised by Reginald C. Punnett in 1905.

Table 18. Punnett square for pea blossom color – pure breed parents

		Pollen (male) from purple plant Genotype PP	
		P	P
Pistil (female) of white plant Genotype ww	w	Pw (purple)	Pw (purple)
	w	Pw (purple)	Pw (purple)

Table 19 shows the cross of two purple pea plants both of which have genotype Pw. Each gamete produced by a parent plant has only one of the two alleles for blossom color (i.e., either P or w) and the distribution of alleles to gametes is random. This follows from the law of segregation. Further, assuming the inheritance of alleles from each parent is equally likely, each genotype combination shown in the table has a .25 chance of occurrence. In terms of phenotype, the child plant has a .75 chance of exhibiting purple blossoms, and a .25 chance of exhibiting white blossoms. The resulting phenotypes are governed by the law dominance (with trait P dominating the recessive trait w).

Mendel performed experiments using pure breed pea plants that he developed over several generations. He then crossed the different pure breeds in experiments (similar in concept to the situation shown in Table 19). Mendel's experimental results were similar (not exact) to the ratios suggested in the Punnett square analysis [166].

Table 19. Punnett square for pea blossom color – parents of genotype Pw

		Pollen (male) from purple plant Genotype Pw	
		P	w
Pistil (female) of purple plant Genotype Pw	P	PP (purple)	Pw (purple)
	w	Pw (purple)	ww (white)

If we change the genotypes of the parents in our example to Pw and ww, then the child plant has a .5 chance of having purple blossoms and a .5 chance of having white blossoms.

Table 20. Punnett square for pea blossom color – parents of genotype Pw and ww

		Pollen (male) from purple plant Genotype ww	
		w	w
Pistil (female) of purple plant Genotype Pw	P	Pw (purple)	Pw (purple)
	w	ww (white)	ww (white)

. . .

The law of segregation allows us to determine how an allele associated with a single gene is inherited. However, there are instances where we want to predict the inheritance of alleles associated with two different genes. In such cases, we need to know whether the two genes are inherited independently or not. Within the scope of his experiments on pea plants, Mendel discovered that different genes were inherited independently of one another, and followed what is called the law of independent assortment. An example will help illustrate the point. We start with two types of pure breed pea plants. One breed has purple blossoms P and yellow seeds Y (both dominant alleles) and is of genotype $PPYY$. The other breed has white blossoms w and green seeds g (both recessive alleles) and is of genotype $wwgg$. In the first generation of crossing the pure breed pea plants, we only get genotype $PwYg$, and the phenotype consisting of purple blossoms and yellow seeds, as shown in Table 21.

Table 21. Crossing of pure breed $PPYY$ and $wwgg$ pea plants

		Pollen (male) from purple plant Genotype $PPYY$	
		PY	PY
Pistil (female) of white plant Genotype $wwgg$	wg	$PwYg$ (purple, yellow)	$PwYg$ (purple, yellow)
	wg	$PwYg$ (purple, yellow)	$PwYg$ (purple, yellow)

In the second generation, we breed instances of the plants with genotype $PwYg$. By the law of segregation, the gametes from the parent plants have one of each allele of the blossom color and seed color genes. Further, the alleles from the different genes are inherited independently (this is the law of independent assortment). The Punnett square below (Table 22) shows the predicted outcome of the breeding experiment. By counting cells with a given phenotype, we can compute the probability of each possible phenotype. For example, there are a total of 16 possibilities of which 9 result in phenotype purple blossoms and yellow seeds (see the cells in gray). Thus, the probability of getting this phenotype is $\frac{9}{16}$. The probabilities for the other phenotypes are as follows:

- Prob (purple, green) $= \frac{3}{16}$
- Prob (white, yellow) $= \frac{3}{16}$
- Prob (white, green) $= \frac{1}{16}$

Table 22. Breeding of pea plants with genotype $PwYg$

		Pollen (male) from purple plant Genotype $PwYg$			
		PY	Pg	wY	wg
Pistil (female) of white plant Genotype $PwYg$	PY	$PPYY$ (purple, yellow)	$PPYg$ (purple, yellow)	$PwYY$ (purple, yellow)	$PwYg$ (purple, yellow)
	Pg	$PPYg$ (purple, yellow)	$PPgg$ (purple, green)	$PwYg$ (purple, yellow)	$Pwgg$ (purple, green)
	wY	$PwYY$ (purple, yellow)	$PwYg$ (purple, yellow)	$wwYY$ (white, yellow)	$wwYg$ (white, yellow)
	wg	$PwYg$ (purple, yellow)	$Pwgg$ (purple, green)	$wwYg$ (white, yellow)	$wwgg$ (white, green)

. . .

To compute the various phenotype probabilities involving more than two genes, we need to divide the problem into parts. For example, consider the case of an organism with 4 genes of interest, where each gene has two alleles (one dominant and the other recessive). Label the alleles A, a, B, b, C, c, and D, d where the uppercase letters represent the dominant alleles. If two organisms, both with genotype $AaBbCcDd$ are bred together, what is the probability of getting the phenotype that results from all

dominant forms of each of the four gene (or equivalent genotype $AxByCzDw$ where x, y, z and w can be either of the alleles for the associated gene)? Assume the experiment is run multiple times.

To solve the problem, we consider one gene at a time (which we can do when the law of independent assortment holds). For the desired phenotype, we need to determine the probability of an child with the AA or Aa genotype. The probably is $\frac{3}{4}$, using an analysis identical to that used in Table 19. Similarly, the probability of the genotypes BB or Bb is $\frac{3}{4}$, and the same for the other two genes. So, to determine the probability of the desired phenotype (resulting from a dominant allele for each gene) we multiple the probability of the 4 independent events to get

$$\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}=\frac{81}{256}\cong .3164$$

For the same setup, what is the probability of children having genotype $AxByccdd$ where x can be A or a, and y can be B or b? Again, we break the problem into parts. The probability of $AxBy$ is $\frac{3}{4}\cdot\frac{3}{4}=\frac{9}{16}$. The probability of cc is $\frac{1}{4}$, and the same for dd. So, the probability for the genotype in question is

$$\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}=\frac{9}{256}\cong .0352$$

. . .

As an example of when independent assortment does not hold, assume we have an type of animal (a Jabberwock) with genes for hair color and eye color. The hair color gene has alleles red R (dominant) and white w (recessive). The eye color gene has alleles green G (dominant) and hazel h (recessive). Further, and this is where independent assortment breaks down, the hair and eye color genes are always inherited as a pair, e.g., the allele pair (R, G) is always inherited together, and (w, h) are always inherited together.

If we start with pure breed Jabberwocks having genotypes (R, G) (R, G) or $(w, h)(w, h)$, then after the first generation of cross breeding, all the children have genotype (R, G) (w, h). In the second generation, we breed males and females of genotype $(R, G)(w, h)$. A Punnett square can be used to predict the genotypes and phenotypes for the second generation (see Table 23). As we can see from the table, there is a .75 chance of having a 2nd generation child with genotype $(R, G)(w, h)$ or $(R, G)(R, G)$ and phenotype (red hair, green eyes), and a .25 chance of children with genotype $(w, h)(w, h)$ and phenotype (white hair, hazel eyes).

Table 23. Breading of Jabberwocks with genotype (R, G) (w, h)

		Males with genotype $(R, G)(w, h)$	
		(R, G)	(w, h)
Females with genotype $(R, G)(w, h)$	(R, G)	$(R, G)(R, G)$ (red hair, green eyes)	$(R, G)(w, h)$ (red hair, green eyes)
	(w, h)	$(R, G)(w, h)$ (red hair, green eyes)	$(w, h)(w, h)$ (white hair, hazel eyes)

10.2.2 Pedigree Diagrams

A **pedigree diagram** is used to study the expression of phenotypes (related to a particular gene) over several generations. An example pedigree diagram is shown in Figure 137 (adapted from a figure in the Wikipedia article "Pedigree chart" [167]). The following conventions are used:

- Each row represents a generation. The generations are indicated by Roman numerals.
- Circles represent females and squares represent males.
- An organism that exhibits the phenotype under study is represented by a black (filled-in) symbol.
- Parents are connected by a horizontal line, and a vertical line leads to their offspring. The offspring are connected by a horizontal sibship line and listed in birth order from left to right, e.g., I.1 and I.2 are the parents of II.2, II.3 and II.4.
- Notice that some of the organisms "marry into" the tree, e.g., II.1 and II.5.

Figure 137 represents a Y-linked disorder, i.e., a malady that only affects males. If a male parent is affected then all male offspring will be affected. The particular disorder in the figure does not skip a generation.

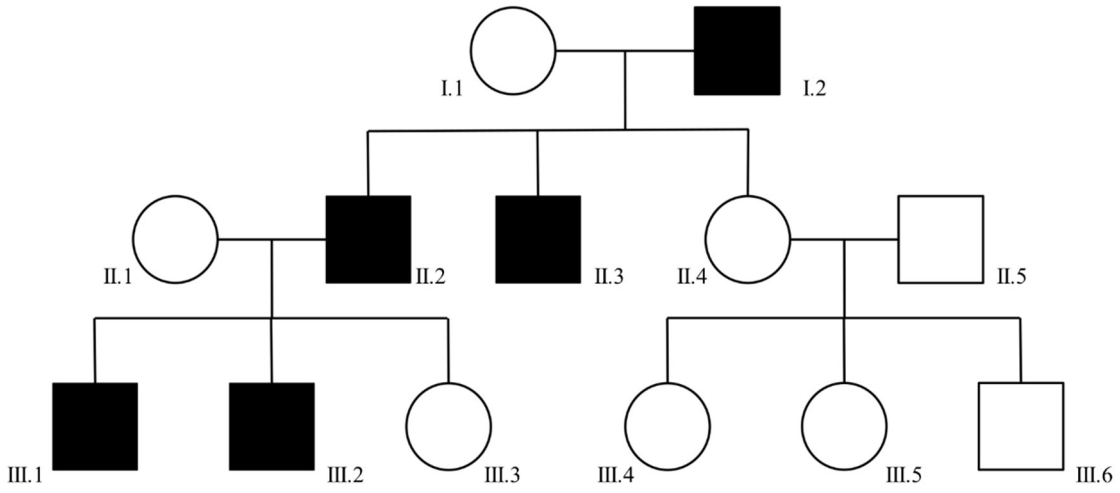


Figure 137. Pedigree diagram for Y-linked (male) disorder

. . .

"Autosomal recessive" is a pattern of inheritance characteristic of some genetic disorders such as sickle cell anemia and cystic fibrosis. Figure 138 shows an example pedigree diagram for sickle cell anemia (on the left) and the associated chromosome pattern (on the right). (Figure source: NIH article entitled Autosomal Recessive Disorder [168].)

In the second generation, there is a marriage line between a male and female who are carriers of sickle cell anemia but are unaffected. So, their genotype is Dd. Their four children are shown in the third generation. On average, only 1 of the 4 children will have two copies of the recessive d allele and thus, exhibit sickle cell anemia. This comes from a Punnett square analysis identical in structure to that in Table 19.

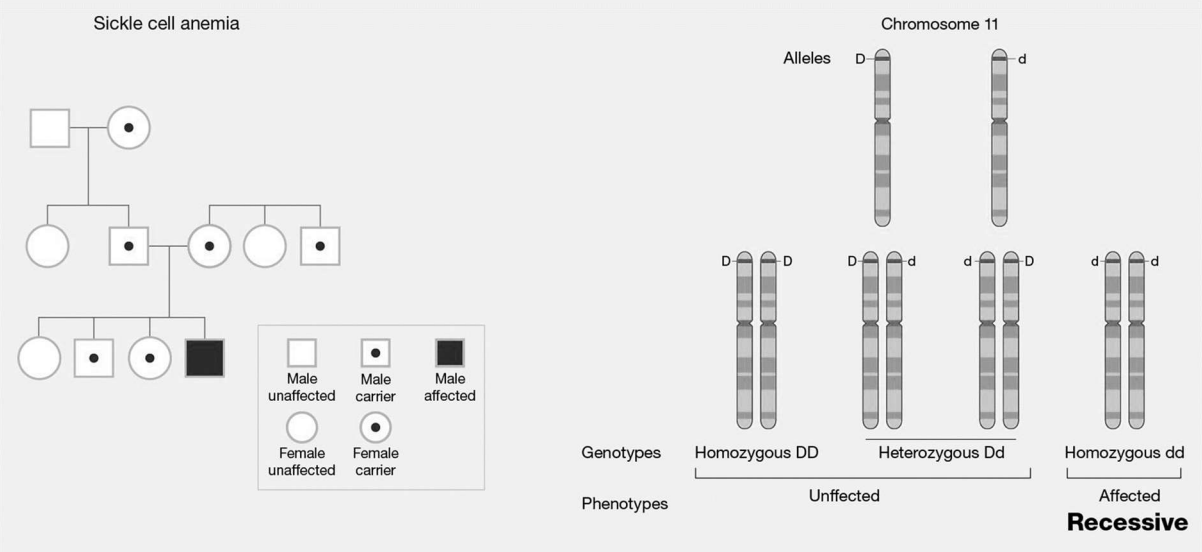


Figure 138. Pedigree diagram for sickle cell anemia

. . .

"Autosomal dominant" is a pattern of inheritance characteristic of some genetic disorders such as Huntington's disease, achondroplasia, and Marfan syndrome. In these disorders, a single copy of a mutated gene from one parent is enough to cause the disorder. Figure 139 illustrates an example pedigree diagram for Huntington's disease (on the left) and the associated chromosome pattern (on the right). (Figure source: NIH article entitled Autosomal Dominant Disorder [169].)

The male parent in the first generation is affected with Huntington's, which means his genotype (with respect to Huntington's) is either DD or Dd. Since the female in the first generation is unaffected, she must have genotype dd. If the first generation male was DD, then all his children would necessarily have one D allele and thus, be affected. Since that is not the case, he must be Dd. Using a Punnett squares analysis (similar to that in Table 20), one parent with genotype Dd and the other with genotype dd would (on average) have 50% of their children affected by the disease (being of genotype Dd) and 50% of their children unaffected (being of genotype dd).

On the extreme left and extreme right of the 2nd generation in the diagram, we have parents of genotype Dd and dd. Using a Punnett squares analysis, one would expect (on average) half their children to be affected and the other half not to be affected. This pattern is reflected in the 3rd generation of the pedigree diagram (for children whose parents have genotype Dd and dd).

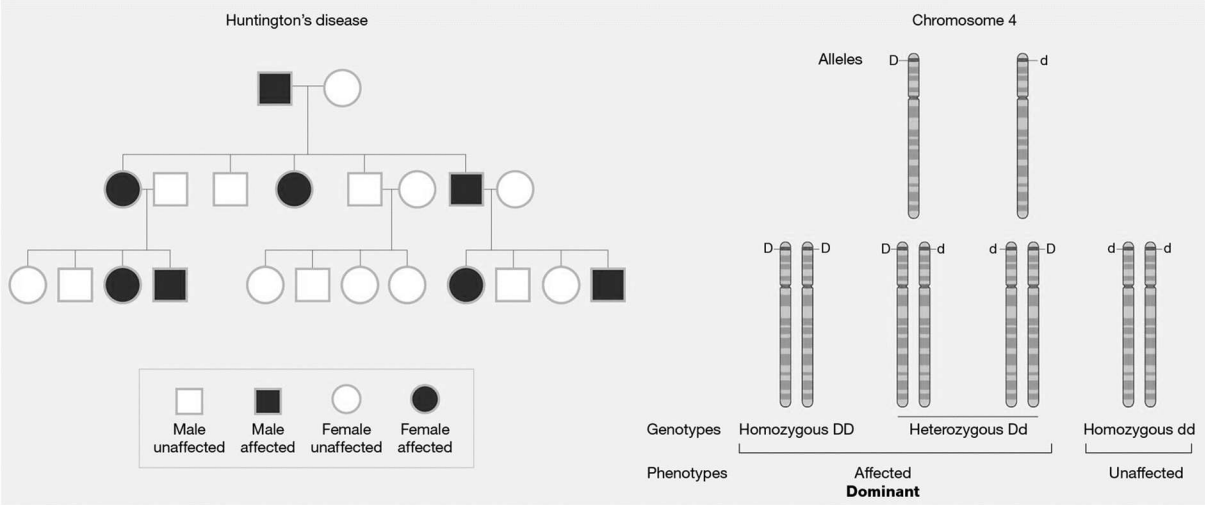


Figure 139. Pedigree diagram for Huntington's disease

. . .

Figure 140 shows an example of what is sometimes called "skip generation of a trait." Assume the gene associated with the trait has two alleles (A which is dominant, and a which is recessive). The trait is recessive, and so, for the trait to be expressed, an organism needs to have genotype aa.

- In the first generation, we have (thus far) two unrelated couples. Their genotypes and phenotypes are as indicated in the figure.
- In the second generation, all the children (of both couples) inherit one A allele and one a allele, and thus none has the phenotype associated with the trait. This is the skip generation.
- In the second generation, a child from each couple has a child of their own (shown in the third generation). The third generation child inherits the a allele from each parent, and the given trait again appears.

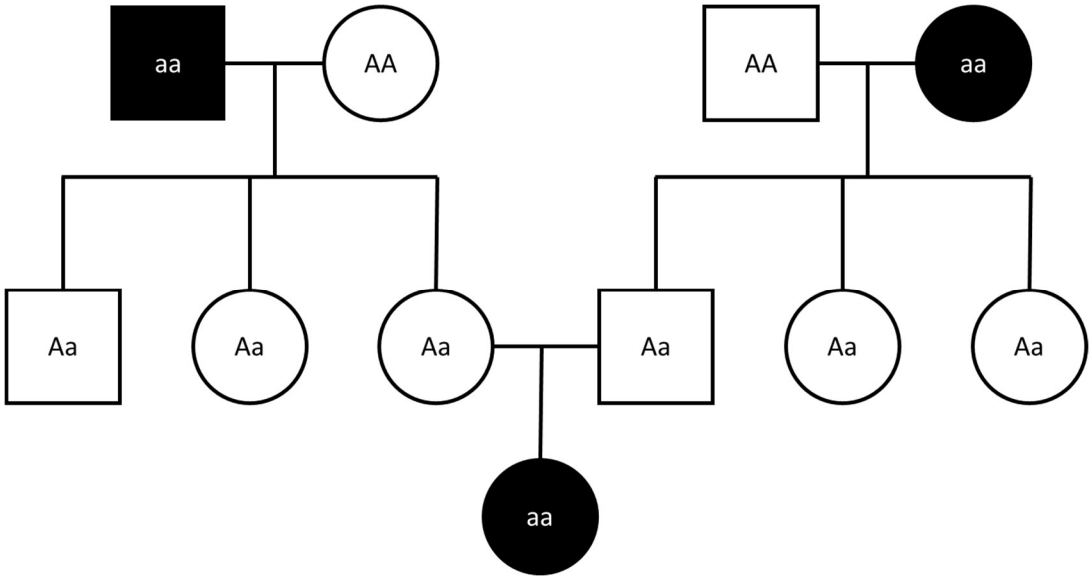


Figure 140. Trait that skips a generation

10.2.3 Non-Mendelian inheritance

10.2.3.1 Overview

Non-Mendelian inheritance, as the name suggest, entails patterns of genetic inheritance that do not follow the principles outlined by Gregor Mendel. In Mendelian genetics, traits are typically controlled by a single gene with two alleles (variants), and the inheritance of these traits can be explained by simple rules such as dominant and recessive alleles, independent assortment, and segregation.

On the other hand, non-Mendelian inheritance involves more complex genetic mechanisms and patterns that do not adhere to the classical Mendelian rules. Here are some key characteristics of non-Mendelian inheritance:

- Multiple Alleles: Instead of just two alleles for a gene, non-Mendelian inheritance can involve multiple alleles for a single gene. For example, blood type inheritance (A, B, AB, O) is determined by multiple alleles.
- Codominance: In some cases, two alleles for a gene are expressed equally in the phenotype, leading to codominance. This means that both traits (i.e., phenotypes) associated with the alleles are visibly present in the individual.
- Incomplete Dominance: In incomplete dominance, neither allele is completely dominant over the other. Instead, the heterozygous individual displays an intermediate phenotype that is a blend of the two alleles.
- Pleiotropy: Non-Mendelian traits may exhibit pleiotropy, where a single gene can affect multiple seemingly unrelated traits or characteristics.
    - An example of pleiotropy is **phenylketonuria**, an inherited disorder that affects the level of phenylalanine, an amino acid that can be obtained from food, in the human body. Phenylketonuria causes this amino acid to increase in amount in the body, which can be very dangerous. The disease is caused by a defect in a single gene on

chromosome 12 that codes for enzyme phenylalanine hydroxylase, that affects multiple systems, such as the nervous and integumentary system. [171]

    - An example comes from a recent discovery that several alleles of a gene involved in an immune response to pathogens are protected against both Alzheimer's and Parkinson's disease. [172] [173]

- Polygenic Inheritance: Some traits are controlled by multiple genes (polygenic traits), and their inheritance patterns are often more complex than those governed by a single gene.
    - An example of a polygenic trait is human skin color variation. Several genes factor into determining a person's skin color, so modifying only one of those genes can change skin color slightly or in some cases, such as for SLC24A5, moderately. Many disorders with genetic components are polygenic such as autism, cancer and diabetes. Most phenotypic characteristics are the result of the interaction of multiple genes. [174]
- Epistasis [176]: In epistasis, the expression of one gene masks or adds to the expression of another gene, leading to complex interactions between different genes.
    - The main difference between epistasis and polygenic inheritance is that epistasis refers to the interaction between two or more genes, while polygenic inheritance refers to the combined effects of multiple genes. In epistasis, the effect of one gene can be masked or altered by the allele of another gene. In polygenic inheritance, the phenotype of a trait is determined by the combined effects of many different genes. [177]
- Environmental Factors: Non-Mendelian traits may also be influenced by environmental factors that can modify gene expression and phenotype. This is known as gene-environment interaction.
- Sex-Linked Traits [175]: Traits located on sex chromosomes (X and Y) follow non-Mendelian patterns of inheritance, as these chromosomes have unique inheritance patterns. In humans, there are three types of sex-lined traits, i.e., X-linked recessive, X-linked dominant and Y-linked. The inheritance and presentation of all three differ depending on the sex of both the parent and the child. This makes them characteristically different from autosomal dominance and recessiveness.

10.2.3.2 Codominance

Human blood types have two non-Mendelian characteristics, i.e., multiple alleles and codominance. The gene related to human red blood cells has three alleles known as A (red blood cells with the A antigen), B (red blood cells with the B antigen) and O (red blood cells with neither the A nor B antigen). The O comes from the German word “ohne” which means “without”. The so-called ABO blood group system was discovered by Austrian scientist Karl Landsteiner who was awarded the Nobel prize in physiology / medicine (1930) for his work in this area.

Alleles A and B both dominate the O allele. However, when a child inherits allele A from one parent and allele B from the other parent, the child has both the A and B antigens presents on their red blood cells. In this case, the A and B alleles are said to be codominant since they are both expressed in the resulting child. Table 24 summarizes the resulting phenotypes from various combinations of alleles.

Table 24. Punnett square for human red blood cells

		Alleles inherited from father		
		A	B	O
Alleles inherited from mother	A	A	AB	A
	B	AB	B	B
	O	A	B	O

Depending on the blood group, human red blood cells and blood serum have various antigens and antibodies, respectively. Table 26 summarizes the relationships among human blood groups, antigens present (or not) on red blood cells, antibodies in the blood serum and the associated genotypes.

Table 25. Blood antigens, antibodies and genotypes

Blood Group (Phenotype)	**Antigen(s) present on red blood cells**	**Antibodies present in serum**	**Genotype**
A	A antigen	Anti-B	AA or AO
B	B antigen	Anti-A	BB or BO
AB	A and B antigen	None	AB
O	None	Anti-A and Anti-B	OO

Table 24 and Table 25 are adapted from the article “Blood Groups and Red Cell Antigens” [170].

10.2.3.3 Incomplete Dominance

Incomplete dominance (also called partial dominance, or intermediate inheritance) occurs when the phenotype of the heterozygous genotype (organism has two different alleles for a given gene) is distinct from and often intermediate to the phenotypes of the homozygous genotypes (organism has two of the same alleles for a given gene). The phenotypic result often appears as a blended form of characteristics in the heterozygous case. For example, Mirabilis Jalapa (also known as “the marvel of Peru” or “four o'clock flower”) has red or white flowers when the genotype is homozygous. However, when the red homozygous flower is paired with the white homozygous flower, the result is a plant with pink flowers.

Table 26 shows the results of breeding two heterozygous Mirabilis Jalapa plants. Each plant (or possibly the male and female parts of one plant, i.e., self-fertilization) has genotype rw with respect to flower color (with r being the red genotype and w being the white genotype). The resulting phenotypes are shown in parenthesis. The red and white alleles are said to blend since they produce a pink flower phenotype.

Table 26. Punnett square for Mirabilis Jalapa flower color

		Genotype rw (male)	
		r	w
Genotype rw (female)	r	rr (red)	rw (pink)
	w	rw (pink)	ww (white)

10.3 Epigenetics

Gene expression is the process by which information from a gene (encoded in a DNA segment) is used in the synthesis of a functional gene product (e.g., a protein). The gene products, in turn, affect one or more phenotypes (e.g., eye color, muscle elasticity). The process of gene expression is used by all known life, i.e., eukaryotes (including multicellular organisms) and prokaryotes (bacteria and archaea), and viruses (considered non-life since they cannot reproduce on their own) to generate the macromolecular machinery for life.

In both Mendelian and non-Mendelian genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype. The genetic information stored in DNA represents the genotype, whereas the phenotype results from the "interpretation" of that information. Such phenotypes are often displayed by the synthesis of proteins that control the organism's structure and development, or that act as enzymes catalyzing specific metabolic pathways.

In contrast to genetics governed by DNA, there are processes that affect gene expression without modifying the DNA. Such processes are known as **epigenetics**. Some definitions of epigenetics:

- In biology, epigenetics is the study of stable changes in cell function (known as marks) that do not involve alterations in the DNA sequence. [178]
- Epigenetics is the study of how your behaviors and environment can cause changes that affect the way your genes work. Unlike genetic changes, epigenetic changes are reversible and do not change your DNA sequence, but they can change how your body reads a DNA sequence. [179]
- Epigenetics is the study of heritable changes in gene expression or cellular phenotype that occur without alterations to the underlying DNA sequence. In other words, it deals with changes in the way genes are turned on or off, or how they are expressed, without changes to the actual DNA code itself. These changes can be influenced by various factors, including environmental factors, lifestyle choices, and developmental processes.

  Common epigenetic mechanisms include DNA methylation (the addition of methyl groups to specific DNA bases), histone modification (chemical alterations to histone proteins), and non-coding RNA molecules that can influence gene expression. Epigenetic changes can be stable and heritable through cell divisions, and they play a crucial role in regulating various

biological processes, including development, differentiation, disease susceptibility, and even responses to environmental stimuli. [180]

Types of epigenetic changes include [179]:

- DNA Methylation: DNA methylation works by adding a chemical group to DNA. Typically, this group is added to specific places on the DNA, where it blocks the proteins that attach to DNA to “read” the gene. This chemical group can be removed through a process called demethylation. Typically, methylation turns genes “off” and demethylation turns genes “on.”
- Histone modification: DNA wraps around proteins called histones. When histones are tightly packed together, proteins that ‘read’ the gene cannot access the DNA as easily, so the gene is turned “off.” When histones are loosely packed, more DNA is exposed or not wrapped around a histone and can be accessed by proteins that ‘read’ the gene, so the gene is turned “on.” Chemical groups can be added or removed from histones to make the histones more tightly or loosely packed, turning genes “off” or “on.” Also, see the Wikipedia article “Histone acetylation and deacetylation” [183].
- Non-coding RNA: Your DNA is used as instructions for making coding and non-coding RNA. Coding RNA is used to make proteins. Non-coding RNA helps control gene expression by attaching to coding RNA, along with certain proteins, to break down the coding RNA so that it cannot be used to make proteins. Non-coding RNA may also recruit proteins to modify histones to turn genes “on” or “off.”

Additional educational resources concerning epigenetics can be found at Genetic Scient Learning Center [181].

“A book is a writer's will: ideas are the inheritance.”
Matshona Dhliwayo

“What really interests me is whether god had any choice in the creation of the world."
Albert Einstein to Ernst Strauss

Acronyms

BIBD – Balanced Incomplete Block Design

CRC – Cyclic Redundancy Check

DNA – Deoxyribonucleic acid

EAA – Essential Amino Acid

FCC – Face-Centered Cubic (packing)

FPGA – Field-Programmable Gate Array

GIS – Geographic Information System

HCP – Hexagonal Close-Packed

IRV – Instant Runoff Voting

MOLS – Mutually Orthogonal Latin Square

mRNA – messenger RNA

NFP – No-Fit Polygon

OEIS – On-line Encyclopedia of Integer Sequences®

RCV – Ranked-Choice Voting

rRNA – ribosomal RNA

RNA – Ribonucleic Acid

SBIBD – Symmetric Balanced Incomplete Block Design

SLC24A5 – Solute Carrier Family 24 Member 5 (a gene that has a major influence on natural skin color variation)

tRNA – transfer RNA

References

[1] Fratini, S., *Mathematical Vignettes: Number theory, stochastic processes, game theory, cryptography, linear programming and more*, self-published on Amazon, August 2022. https://github.com/sfratini33/art-of-managing-things-external/blob/master/free_books/MathVignettes-I.pdf

[2] *Combinatorial design*, Wikipedia, https://en.wikipedia.org/wiki/Combinatorial_design, accessed on 1 February 2023.

[3] Stinson, D.R., *Combinatorial Design: Constructions and Analysis*,* Springer, 2004. Available from the Internet Archive at https://archive.org/details/springer_10.1007-b97564.

[4] Colbourn, C.J., Dinitz, J.H., Handbook of Combinatorial Designs, 2nd Edition, Chapman and Hall / CRC, 2006.

[5] *Equivalence relation*, Wikipedia, https://en.wikipedia.org/wiki/Equivalence_relation, accessed on 9 January 2023.

[6] *Modular arithmetic,* https://en.wikipedia.org/wiki/Modular_arithmetic, Wikipedia, accessed on 27 January 2023.

[7] A002860, *Number of Latin squares of order n*, The On-line Encyclopedia of Integer Sequences®, https://oeis.org/A002860, accessed on 12 January 2023.

[8] A040082, *Number of inequivalent Latin squares (or isotopy classes of Latin squares) of order n*, The On-line Encyclopedia of Integer Sequences®, https://oeis.org/A040082, accessed on 12 January 2023.

[9] *Mutually orthogonal Latin squares - Thirty-six officers problem*, Wikipedia, https://en.wikipedia.org/wiki/Mutually_orthogonal_Latin_squares#Thirty-six_officers_problem, accessed on 14 January 2023.

[10] *Mutually orthogonal Latin squares*, Wikipedia, https://en.wikipedia.org/wiki/Mutually_orthogonal_Latin_squares, accessed on 14 January 2023.

[11] N.P. Uto, R.A. Bailey, *Constructions for regular-graph semi-Latin rectangles with block size two*, Journal of Statistical Planning and Inference, Volume 221, 2022, Pages 81-89, ISSN 0378-3758, https://doi.org/10.1016/j.jspi.2022.02.007.

[12] *Sudoku*, Wikipedia, https://en.wikipedia.org/wiki/Sudoku, accessed on 30 January 2023.

[13] A107739, *Number of (completed) sudokus (or Sudokus) of size* $n^2 \times n^2$, On-Line Encyclopedia of Integer Sequences (OEIS), https://oeis.org/A107739, accessed on 30 January 2023.

[14] A109741, *Number of inequivalent (completed)* $n^2 \times n^2$ *sudokus (or Sudokus)*, Encyclopedia of Integer Sequences (OEIS), https://oeis.org/A109741, 30 January 2023.

[15] Rosenhouse, J., Taalman, L., *Taking Sudoku Seriously: The Math Behind the World's Most Popular Pencil Puzzle*, Oxford University Press, 2012.

[16] *KenKen®*, Wikipedia, https://en.wikipedia.org/wiki/KenKen, accessed on 31 January 2023.

[17] *Fisher's inequality*, Wikipedia, https://en.wikipedia.org/wiki/Fisher%27s_inequality, accessed on 4 February 2023.

[18] Hughes, D. R., Piper, F., *Design Theory**, Cambridge University Press, 1985.

[19] *Steiner Triple System*, Wolfram MathWorld, https://mathworld.wolfram.com/SteinerTripleSystem.html, accessed on 17 September 2023.

[20] *Kirkman's schoolgirl problem*, Wikipedia, https://en.wikipedia.org/wiki/Kirkman%27s_schoolgirl_problem, 8 February 2023.

[21] *Hadamard matrix: Hadamard conjecture*, Wikipedia, https://en.wikipedia.org/wiki/Hadamard_matrix#Hadamard_conjecture, 9 February 2023.

[22] *Factorial experiment*, Wikipedia, https://en.wikipedia.org/wiki/Factorial_experiment, 11 February 2023.

[23] Peter Woolf et al., *Chemical Process Dynamics and Controls*, LibreTexts™, https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Book%3A_Chemical_Process_Dynamics_and_Controls_(Woolf), accessed 11 February 2023.

[24] A006052, *Magic Squares*, Online Encyclopedia of Integer Sequences (OEIS), https://oeis.org/A006052, accessed on 19 February 2023.

[25] Pickover, C.A., *The Zen of Magic Squares, Circles, and Stars An Exhibition of Surprising Structures across Dimensions*, Princeton University Press, 2002.

[26] Swetz, Frank. *Legacy of the Luoshu: The 4,000 Year Search for the Meaning of the Magic Square of Order Three**. United States: CRC Press, 2008.

[27] *Melencolia I*, Wikipedia, https://en.wikipedia.org/wiki/Melencolia_I, accessed on 22 February 2023.

[28] *Magic square*, Wikipedia, https://en.wikipedia.org/wiki/Magic_square, accessed on 27 February 2023.

[29] *Benjamin Franklin*, Wikipedia, https://en.wikipedia.org/wiki/Benjamin_Franklin, accessed on 19 September 2023.

[30] *Broken diagonal*, Wikipedia, https://en.wikipedia.org/wiki/Broken_diagonal, accessed on 27 February 2023.

[31] *Franklin's Magic Squares*, MathPages, https://mathpages.com/home/kmath155.htm, accessed on 28 February 2023.

[32] Napolitano, V., Olanda, D. *A simple new proof of the fundamental theorem for finite linear spaces*. Ricerche mat. 63, 41–45 (2014). https://doi.org/10.1007/s11587-013-0160-x.

[33] Batten, L.M., *Combinatorics of finite geometries**, Cambridge University Press, 1986.

[34] Kiss, G., Szonyi, T., *Finite Geometries*, CRC Press, Taylor and Francis Group, 2020.

[35] *Bijection*, Wikipedia, https://en.wikipedia.org/wiki/Bijection, accessed on 28 March 2023.

[36] Lam, Clement W. H., *The Search for a Finite Projective Plane of order 10*, The American Mathematical Monthly, 98 (4): 305–318, 1991, doi:10.1080/00029890.1991.12000759, https://www.maa.org/sites/default/files/pdf/upload_library/22/Ford/Lam305-318.pdf.

[37] *Projective space*, Wikipedia, https://en.wikipedia.org/wiki/Projective_space, accessed on 29 March 2023.

[38] *Equivalence class*, Wikipedia, https://en.wikipedia.org/wiki/Equivalence_class, accessed on 31 March 2023.

[39] *Projective plane: Construction of projective planes from affine planes*, Wikipedia, https://en.wikipedia.org/wiki/Projective_plane#Construction_of_projective_planes_from_affine_planes, accessed on 1 April 2023.

[40] Judson, T.W., *Abstract Algebra: Theory and Applications*, Publisher: University of Puget Sound, available from the Open Textbook Library at https://open.umn.edu/opentextbooks/textbooks/217.

[41] Neumann, P.M., *A breakthrough in Algebra: Classification of the Finite Simple Groups*, YouTube video, https://youtu.be/s88bfJzyA78, accessed on 7 April 2023.

[42] *Modular arithmetic*, Wikipedia, https://en.wikipedia.org/wiki/Modular_arithmetic, accessed on 2 April 2023.

[43] *Multiplicative group of integers modulo n*, Wikipedia, https://en.wikipedia.org/wiki/Multiplicative_group_of_integers_modulo_n, accessed on 2 April 2023.

[44] *Symmetric group*, Wikipedia, https://en.wikipedia.org/wiki/Symmetric_group, accessed on 3 April 2023.

[45] *Alternating group*, Wikipedia, https://en.wikipedia.org/wiki/Alternating_group, accessed on 6 April 2023.

[46] *Dihedral group*, Wikipedia, https://en.wikipedia.org/wiki/Dihedral_group, accessed on 12 April 2023.

[47] *General linear group*, Wikipedia, https://en.wikipedia.org/wiki/General_linear_group, accessed on 12 April 2023.

[48] *Determinant*, Wikipedia, https://en.wikipedia.org/wiki/Determinant, 12 April 2023.

[49] *Special linear group*, Wikipedia, https://en.wikipedia.org/wiki/Special_linear_group, 12 April 2023.

[50] Rotman, J., *A First Course in Abstract Algebra (Third Edition)**, Prentice Hall, 2005.

[51] *Lagrange's theorem (group theory)*, Wikipedia, https://en.wikipedia.org/wiki/Lagrange%27s_theorem_(group_theory), accessed on 8 April 2023.

[52] *Determinant: Multiplicativity and matrix groups*, Wikipedia, https://en.wikipedia.org/wiki/Determinant#Multiplicativity_and_matrix_groups, accessed on 27 September 2023.

[53] Baumslag, Benjamin (2006), "A simple way of proving the Jordan-Hölder-Schreier theorem", American Mathematical Monthly, 113 (10): 933–935, doi:10.2307/27642092.

[54] O. Hölder, Die einfachen Gruppen in ersten und zweiten Hundert der Ordnungszahlen, Math. Annalen 40 (1892), 55–88.

[55] W. Burnside, On a class of groups of finite order, Trans. Cambridge Phil. Soc. 18 (1899), 269–276.

[56] Gorenstein, D., Finite Simple Groups; *An Introduction to Their Classification*, Plenum, New York, 1982.

[57] Gorenstein, D., *The Classification of the Finite Simple Groups*, Volume I, Plenum, New York, 1983.

[58] Aschbacher, Michael (2004). "The Status of the Classification of the Finite Simple Groups". Notices of the American Mathematical Society. Vol. 51, no. 7. pp. 736–740. https://www.ams.org/notices/200407/fea-aschbacher.pdf

[59] *Classification of finite simple groups*, Wikipedia, https://en.wikipedia.org/wiki/Classification_of_finite_simple_groups, accessed on 14 April 2023.

[60] *Abelian Group is Simple iff Prime, Proof Wiki*, https://proofwiki.org/wiki/Abelian_Group_is_Simple_iff_Prime, accessed on 28 September 2023.

[61] *Fundamental Theorem of Finite Abelian Groups*, Proof Wiki, https://proofwiki.org/wiki/Fundamental_Theorem_of_Finite_Abelian_Groups, accessed on 28 September 2023.

[62] *List of finite simple groups*, Wikipedia, https://en.wikipedia.org/wiki/List_of_finite_simple_groups, accessed on 15 April 2023.

[63] *Polynomial ring*, Wikipedia, https://en.wikipedia.org/wiki/Polynomial_ring, accessed on 16 April 2023.

[64] *Ring homomorphism*, Wikipedia, https://en.wikipedia.org/wiki/Ring_homomorphism, accessed on 28 September 28, 2023.

[65] *Intersection Distributes over Symmetric Difference*, Proof Wiki, https://proofwiki.org/wiki/Intersection_Distributes_over_Symmetric_Difference, 16 April 2023.

[66] *Set Intersection Not Cancellable*, Proof Wiki, https://proofwiki.org/wiki/Set_Intersection_Not_Cancellable, accessed on 16 April 2023.

[67] *Quadratic integer*, HandWiki, https://handwiki.org/wiki/Quadratic_integer, accessed on 17 April 2023.

[68] *Why is quadratic integer ring defined in that way?*, StackExchange: Mathematics, https://math.stackexchange.com/questions/1198188/why-is-quadratic-integer-ring-defined-in-that-way, accessed on 18 April 2023.

[69] *Pointwise product*, Wikipedia, https://en.wikipedia.org/wiki/Pointwise_product, accessed on 19 April 2023.

[70] *Ideal (ring theory)*, Wikipedia, https://en.wikipedia.org/wiki/Ideal_(ring_theory), accessed on 29 September 2023.

[71] *Polynomial long division*, Wikipedia, https://en.wikipedia.org/wiki/Polynomial_long_division, accessed on 15 June 2023.

[72] *Field (mathematics)*, Wikipedia, https://en.wikipedia.org/wiki/Field_(mathematics), accessed on 20 April 2023.

[73] *Principle ideal*, Wikipedia, https://en.wikipedia.org/wiki/Principal_ideal, accessed on 20 April 2023.

[74] *Finite field*, Wikipedia, https://en.wikipedia.org/wiki/Finite_field, accessed on 20 April 2023.

[75] Schilling, A., Nachtergaele, B., and Lankham, I., *Linear Algebra*, LibreTexts™, https://math.libretexts.org/Bookshelves/Linear_Algebra/Book%3A_Linear_Algebra_(Schilling_Nachtergaele_and_Lankham), accessed on 30 April 2023.

[76] *Taxicab geometry*, Wikipedia, https://en.wikipedia.org/wiki/Taxicab_geometry, accessed on 2 October 2023.

[77] *Singleton bound*, Wikipedia, https://en.wikipedia.org/wiki/Singleton_bound, accessed on 26 May 2023.

[78] Tietäväinen, A. (1973). "On the nonexistence of perfect codes over finite fields". SIAM J. Appl. Math. 24: 88–96. doi:10.1137/0124010.

[79] *Gaussian elimination*, Wikipedia, https://en.wikipedia.org/wiki/Gaussian_elimination, accessed on 31 May 2023.

[80] Kuttler, K., *A First Course in Linear Algebra,* LibreTexts™, https://math.libretexts.org/Bookshelves/Linear_Algebra/A_First_Course_in_Linear_Algebra_(Kuttler), accessed on 4 June 2023.

[81] Hill, R., *A First Course in Coding Theory**, Oxford University Press, 1986.

[82] *Binary Golay code*, Wikipedia, https://en.wikipedia.org/wiki/Binary_Golay_code, accessed on 12 June 2023.

[83] Golay, M.J.E., "Notes on digital coding," Proc. IRE, 37, 657. 1949. https://web.archive.org/web/20161007122006/http://www.maths.manchester.ac.uk/~ybazlov/code/golay_paper.pdf

[84] *Polynomial ring: Quotient ring*, Wikipedia, https://en.wikipedia.org/wiki/Polynomial_ring#Quotient_ring, accessed on 14 June 2023.

[85] *Integral Domain*, Wikipedia, https://en.wikipedia.org/wiki/Integral_domain, accessed on 22 June 2023.

[86] *Tessellation*, Wikipedia, https://en.wikipedia.org/wiki/Tessellation, accessed on 1 July 2023.

[87] *Bin packing problem*, Wikipedia, https://en.wikipedia.org/wiki/Bin_packing_problem, accessed on 1 July 2023.

[88] *Knapsack problem*, Wikipedia, https://en.wikipedia.org/wiki/Knapsack_problem, accessed on 1 July 2023.

[89] *Cutting stock problem*, Wikipedia, https://en.wikipedia.org/wiki/Cutting_stock_problem, accessed on 1 July 2023.

[90] *Strip packing problem*, Wikipedia, https://en.wikipedia.org/wiki/Strip_packing_problem, accessed on 1 July 2023.

[91] Burke, E. K., R. S. R. Hellier, G. Kendall, and G. Whitwell. "Irregular Packing Using the Line and Arc No-Fit Polygon." Operations Research 58, no. 4 (2010): 948–70. http://www.jstor.org/stable/40792736.

[92] *Monte Carlo method*, Wikipedia, https://en.wikipedia.org/wiki/Monte_Carlo_method, accessed on 3 July 2023.

[93] R.L. Graham, B.D. Lubachevsky, K.J. Nurmela, P.R.J. Östergård, *Dense packings of congruent circles in a circle*, Discrete Mathematics, Volume 181, Issues 1–3, 1998, Pages 139-154, ISSN 0012-365X, https://doi.org/10.1016/S0012-365X(97)00050-2. (https://www.sciencedirect.com/science/article/pii/S0012365X97000502)

[94] *Pentagon*, Wikipedia, https://en.wikipedia.org/wiki/Pentagon, accessed on 6 July 2023.

[95] *Circle packing in a circle*, Wikipedia, https://en.wikipedia.org/wiki/Circle_packing_in_a_circle, accessed on 6 July 2023.

[96] *Circle packing theorem*, Wikipedia, https://en.wikipedia.org/wiki/Circle_packing_theorem, accessed on 12 July 2023.

[97] Stephenson, K., *Introduction to Circle Packing: The Theory of Discrete Analytic Functions*, Cambridge University Press, 2005.

[98] Graham, Ronald L.; Lagarias, Jeffrey C.; Mallows, Colin L.; Wilks, Allan R.; Yan, Catherine H. (2003), "Apollonian circle packings: number theory", Journal of Number Theory, 100 (1): 1–45, https://arxiv.org/abs/math/0009113.

[99] *Apollonian gasket*, Wikipedia, https://en.wikipedia.org/wiki/Apollonian_gasket, accessed on 12 July 2023.

[100] *Descartes' theorem*, Wikipedia, https://en.wikipedia.org/wiki/Descartes%27_theorem, accessed on 13 July 2023.

[101] *Doyle Spiral*, Wikipedia, https://en.wikipedia.org/wiki/Doyle_spiral, accessed on 13 July 2023.

[102] Sutcliffe, A., *Doyle Spiral Circle Packings Animated*, 2008, https://archive.bridgesmathart.org/2008/bridges2008-131.pdf.

[103] *Steiner chain*, Wikipedia, https://en.wikipedia.org/wiki/Steiner_chain, accessed on 14 July 2023.

[104] Stromquist, Walter R.. "Packing 10 or 11 Unit Squares in a Square." Electron. J. Comb. 10 (2003): n. pag. https://www.combinatorics.org/ojs/index.php/eljc/article/view/v10i1r8

[105] Ellsworth, D., *Squares in Squares*, https://kingbird.myphotos.cc/packing/squares_in_squares.html, accessed on 17 July 2023.

[106] Friedman, E., *Packing Unit Squares in Squares: A Survey and New Results*, https://erich-friedman.github.io/papers/squares/squares.html, accessed on 17 July 2023.

[107] Kepler conjecture, Wikipedia, https://en.wikipedia.org/wiki/Kepler_conjecture, accessed on 18 July 2023.

[108] *Close-packing of equal spheres*, Wikipedia, https://en.wikipedia.org/wiki/Close-packing_of_equal_spheres, accessed on 19 July 2023.

[109] Lower, S., *Chem1 Virtual Textbook*, LibreTexts™, https://chem.libretexts.org/Bookshelves/General_Chemistry/Chem1_(Lower), accessed on 19 July 2023.

[110] Viazovska, M., *The sphere packing problem in dimension 8*, arXiv:1603.04246v2 [math.NT], https://arxiv.org/abs/1603.04246.

[111] Cohn, H., Kumar, A., Miller, S., Radchenko, D., Viazovska, M., *The sphere packing problem in dimension 24*, arXiv:1603.06518v3 [math.NT], https://arxiv.org/abs/1603.06518.

[112] *Leech lattice*, Wikipedia, https://en.wikipedia.org/wiki/Leech_lattice, accessed on 19 July 2023.

[113] *Surface-area-to-volume ratio*, Wikipedia, https://en.wikipedia.org/wiki/Surface-area-to-volume_ratio, accessed on 19 July 2023.

[114] *Knot theory*, Wikipedia, https://en.wikipedia.org/wiki/Knot_theory, accessed on 20 July 2023.

[115] *Conversation with ChatGPT*, 20 July 2023.

[116] Salomone, M., *Open Algebra and Knots*, http://matthematics.com/oak/section-tanglearithmetic.html. The videos are available on YouTube at https://www.youtube.com/playlist?list=PLL0ATV5XYF8BfT8CmmzKnfTlf3V9hQgj9

[117] Taniyama, K., *Unknotting numbers of diagrams of a given nontrivial knot are unbounded*, Journal of Knot Theory and its Ramifications, 18 (8): 1049–1063, 2009. https://arxiv.org/abs/0805.3174

[118] *Perko pair*, Wikipedia, https://en.wikipedia.org/wiki/Perko_pair, accessed on 1 August 2023.

[119] *Link (knot theory)*, Wikipedia, https://en.wikipedia.org/wiki/Link_(knot_theory), accessed on 24 July 2023.

[120] *Brunnian link*, Wikipedia, https://en.wikipedia.org/wiki/Brunnian_link, accessed on 25 July 2023.

[121] Bai, Sheng; Wang, Weibiao, *New criteria and constructions of Brunnian links*, Journal of Knot Theory and Its Ramifications, 29 (13): 2043008, November 2020.

[122] *List of prime knots*, Wikipedia, https://en.wikipedia.org/wiki/List_of_prime_knots, accessed on 28 July 2023.

[123] *Number of prime knots with n crossings*, A002863, The Online Encyclopedia of Integer Sequences®, https://oeis.org/A002863, accessed on 1 August 2023.

[124] *Amphichiral Knot*, Wolfram MathWorld, https://mathworld.wolfram.com/AmphichiralKnot.html, accessed on 2 August 2023.

[125] *Number of amphichiral prime knots with n crossings*, A052401, The Online Encyclopedia of Integer Sequences®, https://oeis.org/A052401, accessed on 1 August 2023.

[126] *Knot*, Wolfram MathWorld, https://mathworld.wolfram.com/Knot.html, accessed on 4 August 2023.

[127] *Chiral knot*, Wikipedia, https://en.wikipedia.org/wiki/Chiral_knot, accessed on 30 July 2023.

[128] Sakuma, M., *A survey of the impact of Thurston's work on Knot Theory*, Preprint available on ResearchGate, February 2020. https://www.researchgate.net/publication/339015556_A_survey_of_the_impact_of_Thurston%27s_work_on_Knot_Theory

[129] Adams, C.C., *The knot book : an elementary introduction to the mathematical theory of knots**, W.H. Freeman and Company, 1994. https://archive.org/details/knotbookelementa0000adam

[130] Conway, J. H., *An Enumeration of Knots and Links, and Some of Their Algebraic Properties*, In Leech, J. (ed.). Computational Problems in Abstract Algebra, Oxford, England: Pergamon Press, pp. 329–358, 1970. https://www.maths.ed.ac.uk/~v1ranick/papers/conway.pdf

[131] *Reidemeister move*, Wikipedia, https://en.wikipedia.org/wiki/Reidemeister_move, accessed on 4 October 2023.

[132] *Tangle (mathematics)*, Wikipedia, https://en.wikipedia.org/wiki/Tangle_(mathematics), accessed on 8 August 2023.

[133] *The Rolfsen Knot Table*, http://katlas.math.toronto.edu/wiki/The_Rolfsen_Knot_Table, accessed on 8 August 2023.

[134] Jay R. Goldman, Louis H. Kauffman, *Rational Tangles*, Advances in Applied Mathematics, Volume 18, Issue 3, 1997, Pages 300-332, ISSN 0196-8858, https://doi.org/10.1006/aama.1996.0511. (https://www.sciencedirect.com/science/article/pii/S0196885896905114)

[135] Johnson, I., Henrich, A., *An Interactive Introduction to Knot Theory*, Dover Publications, 2017,

[136] *Conversation with ChatGPT*, 10 August 2023.

[137] *Social choice theory*, Wikipedia, https://en.wikipedia.org/wiki/Social_choice_theory, 10 August 2023.

[138] *Ranked-choice voting (RCV)*, Ballotpedia, https://ballotpedia.org/Ranked-choice_voting_(RCV), accessed on 10 August 2023.

[139] *Instant runoff voting*, Wikipedia, https://en.wikipedia.org/wiki/Instant-runoff_voting, accessed on 10 August 2023.

[140] *Borda count*, Wikipedia, https://en.wikipedia.org/wiki/Borda_count, accessed on 10 August 2023.

[141] *Condorcet method*, Wikipedia, https://en.wikipedia.org/wiki/Condorcet_method, accessed on 10 August 2023.

[142] Liberto, D., Kelly, R.C., Arrow's Impossibility Theorem Definition, Investopedia, https://www.investopedia.com/terms/a/arrows-impossibility-theorem.asp, accessed on 10 August 2023.

[143] Gibbard-Satterthwaite theorem, Wikipedia, https://en.wikipedia.org/wiki/Gibbard%E2%80%93Satterthwaite_theorem, accessed on 10 August 2023.

[144] *Majority*, Wikipedia, https://en.wikipedia.org/wiki/Majority, accessed on 12 August 2023.

[145] *Condorcet winner criterion*, Wikipedia, https://en.wikipedia.org/wiki/Condorcet_winner_criterion, accessed on 16 August 2023.

[146] *Condorcet loser criterion*, Wikipedia, https://en.wikipedia.org/wiki/Condorcet_loser_criterion, accessed on 16 August 2023.

[147] *Cardinal voting*, Wikipedia, https://en.wikipedia.org/wiki/Cardinal_voting, 17 August 2023.

[148] *Independence of irrelevant alternatives*, Wikipedia, https://en.wikipedia.org/wiki/Independence_of_irrelevant_alternatives, 2 October 2023.

[149] Arrow, K. J. (1950). A Difficulty in the Concept of Social Welfare. Journal of Political Economy, 58(4), 328–346. http://www.jstor.org/stable/1828886.

[150] *Arrow's impossibility theorem*, Wikipedia, https://en.wikipedia.org/wiki/Arrow's_impossibility_theorem, accessed on 18 August 2023

[151] *Strategic voting*, Wikipedia, https://en.wikipedia.org/wiki/Strategic_voting, accessed on 18 August 2023.

[152] *College Biology: Unit 5 – Heredity*, Kahn Academy Video, https://www.khanacademy.org/science/ap-biology/heredity, accessed on 26 August 2023.

[153] *Mitosis: The Amazing Cell Process that Uses Division to Multiply!*, Amoeba Sisters video on YouTube, https://youtu.be/f-ldPgEfAHI?si=CHlGp3w_Czt-JpMC, accessed on 26 August 2023.

[154] *Meiosis*, Amoeba Sisters video on YouTube, https://youtu.be/VzDMG7ke69g?si=mjmGk8cu1QKmuiy-, accessed on 26 August 2023.

[155] Raghavendra Rao, *Meiosis – Plants and Animals*, YouTube video, https://youtu.be/jjEcHra3484?si=XxJa1RiPtv38f0hP, accessed on 6 September 2023.

[156] *Mitosis vs. Meiosis: Side by Side Comparison*, Amoeba Sisters video on YouTube, https://youtu.be/zrKdz93WlVk?si=KaeD7YQNheBpBXUm, accessed on 26 August 2023.

[157] Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, Kim Sonier, *College Mathematics for Everyday Life*, LibreTexts™, https://math.libretexts.org/Bookshelves/Applied_Mathematics/Book%3A_College_Mathematics_for_Everyday_Life_(Inigo_et_al).

[158] Conversation with ChatGPT, 1 September 2023

[159] *Eukaryote*, Wikipedia, https://en.wikipedia.org/wiki/Eukaryote, accessed on 1 September 2023.

[160] Salzberg, S.L. Open questions: How many genes do we have?. BMC Biol 16, 94 (2018). https://doi.org/10.1186/s12915-018-0564-x.

[161] *Talking Glossary of Genomic and Genetic Terms*, National Institutes of Health (NIH), https://www.genome.gov/genetics-glossary, accessed on 8 September 2023.

[162] *Genotype*, Wikipedia, https://en.wikipedia.org/wiki/Genotype, accessed on 2 September 2023.

[163] *Sexual reproduction*, Wikipedia, https://en.wikipedia.org/wiki/Sexual_reproduction, accessed on 4 September 2023.

[164] Mary Ann Clark, Matthew Douglas, Jung Choi, *Biology (2e)*, OpenStax, https://openstax.org/details/books/biology-2e, accessed on 4 September 2023.

[165] *Punnett square*, Wikipedia, https://en.wikipedia.org/wiki/Punnett_square, accessed on 5 September 2023.

[166] *Mendel's experiments*, video from the Science Learning Hub, https://www.sciencelearn.org.nz/resources/1999-mendel-s-experiments, accessed on 6 September 2023.

[167] *Pedigree chart*, Wikipedia, https://en.wikipedia.org/wiki/Pedigree_chart, accessed on 9 September 2023.

[168] Hanchard, N., *Autosomal Recessive Disorder*, National Institutes of Health (NIH), https://www.genome.gov/genetics-glossary/Autosomal-Recessive-Disorder, accessed on 9 September 2023.

[169] Hanchard, N., *Autosomal Dominant Disorder*, National Institutes of Health (NIH), https://www.genome.gov/genetics-glossary/Autosomal-Dominant-Disorder, accessed on 9 September 2023.

[170] Dean L. *Blood Groups and Red Cell Antigens* [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005. Chapter 5, The ABO blood group. Available from: https://www.ncbi.nlm.nih.gov/books/NBK2267/.

[171] *Pleiotropy*, Wikipedia, https://en.wikipedia.org/wiki/Pleiotropy, accessed on 10 September 2023.

[172] Brown, J., *A Single Gene Variant Protects From Both Alzheimer's And Parkinson's*, Science Alert, https://www.sciencealert.com/a-single-gene-variant-protects-from-both-alzheimers-and-parkinsons, 12 September 2023.

[173] Le Guen, Yann et al. "Multiancestry analysis of the HLA locus in Alzheimer's and Parkinson's diseases uncovers a shared adaptive immune response mediated by HLA-DRB1*04 subtypes." Proceedings of the National Academy of Sciences of the United States of America vol. 120,36 (2023): e2302720120. doi:10.1073/pnas.2302720120.

[174] *Quantitative trait locus*, Wikipedia, https://en.wikipedia.org/wiki/Quantitative_trait_locus, accessed on 10 September 2023.

[175] *Sex linkage*, Wikipedia, https://en.wikipedia.org/wiki/Sex_linkage, accessed on 10 September 2023.

[176] *Epistasis*, Wikipedia, https://en.wikipedia.org/wiki/Epistasis, accessed on 10 September 2023.

[177] *Conversation with Google Bard*, 10 September 2023.

[178] *Epigenetics*, Wikipedia, https://en.wikipedia.org/wiki/Epigenetics, accessed on 11 September 2023.

[179] *What is Epigenetics?*, Centers for Disease Control and Prevention (CDC), https://www.cdc.gov/genomics/disease/epigenetics.htm, accessed on 11 September 2023.

[180] *Conversation with ChatGPT, 11 September 2023.*

[181] *Epigenetics*, Learn Genetics: Genetic Science Learning Center (University of Utah), https://learn.genetics.utah.edu/content/epigenetics/, accessed on 11 September 2023.

[182] *Conversation with ChatGPT, 12 September 2023.*

[183] *Histone acetylation and deacetylation*, Wikipedia, https://en.wikipedia.org/wiki/Histone_acetylation_and_deacetylation, accessed on 28 October 2023.

* Indicates the book or article is available for borrowing from the Internet Achieve at https://archive.org/. In some cases, only an earlier edition of a book will be available.

Index of Terms