Combinatorial geometry notes

David S. Gunderson

30 December 2021

To Karen

Contents

Preface

0.1 An apology

The collection of topics included in this book is a demonstration of aspects of geometry and combinatorial geometry (CG) that have attracted my attention over the last few years. Some topics given here are, in a sense, more recreational than others, and might not be appropriate for a course in CG, but since they were so much fun, I include them in case an instructor might want to spend a few minutes on some of the highlights. Over the years, I have made many wooden models for geometry and puzzles, and although some are not so good or relevant for a course in CG, I include discussions and pictures for general interest.

For reference, I have put in a brief review of vector space geometry (Chapter 23), most of which is contained in any first year linear algebra course. Chapter 1 is a review of Euclidean geometry in the plane. Many of the problems mentioned do not use combinatorics, so the combinatorial geometry student need not digest that entire chapter before proceeding on to topics in combinatorial geometry. Chapter 1 and later chapters arose from notes made for mathletics training in geometry, so there is far more there than is needed for a course in combinatorial geometry. However, early chapters, especially Chapter 1 do supply a great deal of geometric vocabulary and an introduction to many topics later in the book. Some of the standard concepts in geometry are not included here; for example, the study of conics (with foci, directrix, intersecting planes with cones) has not been included for two reasons: I don't like drawing the necessary pictures, and their use in combinatorial geometry takes back seat to other concepts discussed in these notes.

As I said above, many of the problems and solutions here come from mathematical contests, and hence these notes might also be useful for mathletics training (regarding geometric problems). Quite a few examples in Euclidean geometry came from math contests, but there a few other contest problems found in other genres given here, as well (e.g., Chapter 10 on grid problems). Three main contests cited here are The William Lowell Putnam Mathematical Competition, the North Central Section MAA (NCS/MAA) team contests, and the International Mathematical Olympiad (IMO). For recent contest questions and solutions for the Putnam, see Kiran S. Kedlaya's archive page `http://kskedlaya.org/putnam-archive/`. For contests previous, there are three books that are useful: for the years 1938–1964 [391], 1965–1984 [19], and 1985–2000 [536]. For NCS/MAA team competition problems and solutions, see the archive page `http://sections.maa.org/northcen/teamcomp.html` page by Jerry Heuer. For IMO problems, see [262]. Two other books that might be helpful for mathletics training are [27] and [294].

Some courses on combinatorial geometry are for graduate students; these notes are suitable for either undergraduate or graduate work. Some topics for a graduate course are not yet fully developed here (e.g., epsilon nets, Voronoi diagrams, Delauney triangulations), but there is enough to keep any grad student busy for a year (or decade?). I would have also liked to demonstrate more connections between extremal graph theory and finite geometries (e.g., norm graphs, or strongly regular graphs), but I guess this will have to wait. (However, I have included a few applications of graph theory in geometry.) Some topics given here may not normally be the content of a CG course (like blocking sets or discrepancy), but some of these topics might be appropriate research for the right student. I have included far more on finite projective planes than might ordinarily be used in a first (or one semester) course in combinatorial geometry.

Background material on graph theory and number theory is included here, topics that a graduate student in combinatorics can simply have as references.

0.2 Development of these notes

These notes arose from both mathletics training and teaching four courses in combinatorial geometry, two grad courses and two undergrad.

A very early version of the notes for Euclidean geometry (a few pages) was written by Andriy Prymak for mathletics training at U of M, somewhere around 2005. Andriy gave a lecture in our mathletics training seminar on geometry, and the accompanying handout contained his notes, a collection of actual competition problems (and their solutions), together with diagrams. Andriy and I then produced a handout of a dozen pages or so, adding some general principles to the problems. That geometry handout for mathletics was used in our training program for over a decade.

In 2012 I gave a graduate reading course in combinatorial geometry based on Matoušek's textbook [645] and notes I had taken while in grad school. In W2015, I agreed to give an extra semester of mathletics training devoted to geometry; during that course, I extended the previous geometry notes to 45 pages. Andrii Arman was an immense help in preparing those lectures. These notes were then expanded for another grad course I taught in combinatorial geometry in Winter 2016, however I only managed to give my students a fairly decent first draft (just over 200 pages, if I remember correctly) only after the course was over. In 2018, an undergraduate course in combinatorial geometry ran for the first time, and the notes grew to nearly 700 pages for the second time I taught the course in 2020.

Since then, I have added more material, including chapters on Euclidean Ramsey theory, distance graphs, and compass-straightedge constructions, but there is still much I want to write. Maybe there is an end to everything (except mathematics, as Paul Erdős would say). There are now 943 references in the bibliography, and each entry is given the page number(s) where it is used. The .pdf version of these notes are cross-referenced. There are 371 exercises, most of which have solutions written up (but appear after the index, so that one can print out the book without solutions).

0.3 Possible course topics

I have taught a version of this course four times (twice as a graduate course, and twice at the fourth year undergraduate level), and each time I changed the topic selection a bit. However, I feel that a delivery of a course in combinatorial geometry needs to contain some basics. Among those topics I feel are basic to an undergraduate course include (the number of days listed for each is based on 50 minute lecture days):

- Basic geometry review. (2–4 days)
- Art gallery problem. ($<$ .5 day)

- Pick's theorem. ($< .5$ day)
- Graph theory and polyhedra. (2 days)
- Polytopes. (3 days)
- Convex n-gons, Erdős–Szekeres theorems, including basic Ramsey theory. (2 days)
- Convex combinations, affine dependence. (1 day)
- Convexity theorems by Caratheodory, Helly, Radon, and Tverberg. (2 days)
- Regions formed by lines, circles, or planes. (1 day)
- Chromatic number of the plane. (2 days)
- Extremal graph theory, a few major results for later use (1 day)
- Crossing numbers. (2 days)
- Unit distances, repeated distances, geometric graphs. (2–3 days)
- Configurations: Sylvester–Gallai theorem, de Bruijn–Erdős theorem, Szemerédi–Trotter theorem, (n_a, ℓ_b) configurations. (2–4 days)
- Finite geometries, including projective and affine geometries. (2–3 days)
- Euclidean Ramsey theory. (2 days)
- Sperner's lemma and fixed point theorem. (1 day)
- Minkowski's lattice theorems. (1.5–2 days)
- Volumes in higher dimension. (1 day)
- Shattered sets and VC dimension (1 day)
- The Kakeya problem. (1 day)
- Heilbronn's problem. (1 day)

- Dissections and dissection paradoxes. (1–3 days)
- Grid problems and chessboard problems. (optional, 2 days)
- Modular lattices and geometry, dimension theory (optional, 1 or 2 days)
- Voronoi diagrams and Delauney triangulations (optional, 1 day)
- Isometries, symmetries, groups, and crystallography. (optional, 2 days)

Since an average semester runs 36–39 lectures (50 minutes each), the minimum times in the above schedule do not leave much room (some topics, such as grid problems, can be left for the student's private entertainment). Various topics from classical geometry can be given a combinatorial treatment, like Descarte's deficiency result, or Steinitz's theorem.

For most of the topics listed above, I have provided far more information here than can be delivered in lectures, so course planning (and sticking to a schedule) are imperative. For example, it would be easy to go on about Euclidean Ramsey theory for weeks, but then other major topics will be missed. Although I have chapters on blocking sets and colouring finite projective planes, these topics might ordinarily not be covered in combinatorial geometry—or only covered in more advance treatments for graduate students, but they are surely eligible topics for such a course. I think that the connection between lattice theory and certain geometries is also a topic that fits into some courses, but in the above schedule, I have not dedicated much time to this topic. In my opinion, any course called "combinatorial geometry" would probably at least introduce almost all of the topics listed above, but even an introduction to each of the topics in this book might be too ambitious. (However, since I gave notes before lectures, students did not have to write notes, and so I was able to introduce many more theorems and proofs than what might be considered normal.)

0.4 Acknowledgements

Many thanks are owed to students who helped with these notes, including Andrii Arman, Sergei Tsaturian, Kyle Monkman, Goldwyn Millar, Garett Klus, Danylo Radchenko (who recently co-authored a breakthrough paper regarding sphere packing in 24 dimensions; see [200] and [201] for background), and the three brave students in the 2016 offering of a course in combinatorial geometry: Colin Desmarais, Suraj Srinivasan, and Sina Zabanfahm. Jasmin Bissonnette, Elaine Herrera, NIchola Millman, and Ian Thompson, were also helpful with many comments on a 2018 draft. Students Tessa Reimer, Max Gutkin, Dana Kapoostinsky, Mingyang (Mia) Diao, and Wan Zhang also found typos in the 2020 draft.

Departmental colleagues also helped with this text. Rob Craigen helped with many comments and by introducing me to some theorems given here. Andriy Prymak has also been helpful with topic selection, and Ranganathan Padmanabhan (Pad) helped with references and many enlightening discussions.

I thank Professor H. K. Farahat for my first (wonderful) university course in geometry (and all of his care in helping me to improve my presentation skills) at the University of Calgary. Ted Bisztriczky attracted me to research in finite geometry and polytopes. Also in Calgary, Richard Guy also motivated my interest with his many lectures, books, articles and discussions that, at first, seemed to arise from recreational math, but contained many insights into much deeper mathematics. (Richard Guy passed away 9 March 2020 at 103 years old.) My doctoral supervisor Vojtech Rödl taught me a vast array of combinatorial geometry concepts and problems as well. I also thank Professors Stephane Durocher and Konrad Swanepoel for helpful comments on drafts of this book. Thanks are also due to Bill Kocay for many wonderful discussions about finite geometries, and for teaching me (and guest lecturing) more about configurations.

The deepest of thanks and praise are owed to my wife, Karen. First, she was generous in sharing her own notes and ideas on various topics here. It was also wonderful to be able to talk to her at any time about ideas I was trying to learn or write about. She also regularly helped me with proofreading, my struggles in drawing with Tikz, and translating articles from French. Karen was also extremely kind in making sure that my writing environment was always *most* pleasant. Thanks, my dear!

Foreword

The word "geometry" brings to mind the type of mathematics that concentrates on concepts that can be visualized, like lines, angles, and shapes, usually in two or three dimensions. In modern mathematics, the word "geometry" is used in a much broader sense.

Geometry learned in school is usually restricted to what is now called "Euclidean geometry", and many theorems are proved keeping in mind some diagram. Euclid defined geometry by listing a set of facts ("axioms" or "postulates") that are assumed to be true for the kind of geometry he was thinking of, and then, from these facts, derived many of the "truths" that are now considered standard. Many of the results in standard Euclidean geometry now have simple proofs if points are given coordinates, where linear algebra and vector space notation is helpful (see Chapter 23).

Centuries later, mathematicians observed that if one starts with different sets of axioms, different kinds of "geometry" may arise. For example, what kind of world would it be if there were no such things as "parallel lines"? For certain applications, some "geometries" served purposes better than others. Projective geometries and hyperbolic geometries are just two of the new kinds of geometry that have evolved, and these find applications in various fields, from the physics of relativity to error-correcting codes.

Spherical geometry studies the patterns available by putting points on a sphere and connecting these points by arcs (usually geodesics—shortest paths). Such geometry can be used to analyze the patterns in the heavens or navigation systems. The surface of a sphere is just one example of a structure where geometry is studied; one might also look at other surfaces, like a torus or more complicated higher dimensional structures. Many of these structures, including hyperbolic geometry or spherical geometry, are not developed in these notes. In these notes, geometries are usually Euclidean, affine, or projective (finite or not).

It seems that "combinatorial geometry" is a recent term to describe a field of mathematical questions that share some common theme. As one might expect, this field considers combinatorics as applied in various geometric worlds. Combinatorics can be used to count patterns in standard Euclidean geometry, but it also plays a vital role in other forms of geometry, particularly in the world of "finite geometries". Some questions regarding certain "patterns" can be rephrased in terms of finite geometries.

Much of what is now known as "combinatorial geometry" is restricted to "discrete" structures, those that can be described by giving lists or sequences, as opposed to "continuous" structures, like the curve of a graph of a function on the reals. The stress in this field is to count points and shapes, hence polyhedra and their higher dimensional analogues form core notions. The terms "discrete geometry" and "combinatorial geometry" seem to describe very similar, if not identical, themes. Combinatorial geometry also includes the study of how convex sets intersect, including the theory of polyhedra and polytopes.

In these notes, Euclidean geometry is first reviewed since it forms the basis for many of the geometric ideas later. It may seem obvious that geometry and simple properties of numbers are two central themes in the development of modern mathematics; however, geometry no longer seems to be studied in any great depth in North American schools, and so a large portion of these notes is devoted to Euclidean geometry. Chapter 1 alone contains enough material on Euclidean geometry in the plane for a course on its own.

While trying to understand and solve geometrical problems, certain "principles" and basic tools can be helpful. A few of these principles or tools when dealing with Euclidean geometry might include the following:

- Proper drawing helps. However, sometimes a drawing can misdirect the reasoning in a proof, so it can help if multiple drawings are considered, some with certain proportions different. (An example of a deceptive drawing is given in Section 1.7.4.)
- Vector geometry and dot product (first and second year linear algebra) in $\mathbb{R}^2$ and $\mathbb{R}^3$ (basics are reviewed in Chapter 23);
- The unit circle;
- Similar triangles, basic properties of triangles (including trigonometry);
- Mathematical induction;

- Complex numbers for planar problems;
- Calculus—applications of integration and slope;
- Symmetry;
- Coordinate systems: cartesian, polar, spherical, and cylindrical.

For other kinds of geometry, additional tools may be required. Some kinds of finite geometries are based on finite fields, a topic either introduced in number theory, discrete mathematics, or a first combinatorics course.

Throughout this text, the set of integers is denoted by $\mathbb{Z}$ and the set of positive integers by $\mathbb{Z}^+$. Depending upon the author or culture, the expression"natural numbers" has been used to denote both $\{0, 1, 2, 3, \ldots\}$ and $\{1, 2, 3, \ldots\}$, so to avoid confusion, the phrase "natural numbers" is avoided in this document. If the set of non-negative integers is required, the notation $\mathbb{Z}^+ \cup \{0\}$ can be used; the non-negative integers are often called "whole" numbers. Other standard notation in combinatorics is also used. For example, for a set S and a positive integer k, the notation $[S]^k = \{T \subseteq S : |T| = k\}$ is used. Also, for integers a, b, the notation $[a, b] = \{z \in \mathbb{Z} : a \leq z \leq b\}$ is common. The notation $[n] = \{1, 2, \ldots, n\}$ is an abbreviation for $[1, n]$. When $S = [n]$, the notation $[S]^k$ is another way to write $[[n]]^k$, however for simplicity, just use $[n]^k$ to indicate the family of k-sets chosen from $[n]$.

Chapter 1

Euclidean geometry in the plane

1.1 Introduction

Many results are given here without proof. The exercises are from various sources, ranging from standard high school material to competition problems, and so vary greatly in difficulty. Not all aspects of "basic geometry" are considered here, but there is enough to give a student a fairly thorough introduction to (or review of) what might be thought of as "classical geometry" in the plane, for either the combinatorial geometer or for the serious mathletics student.

In addition to the basics, some other interesting aspects of plane geometry are given here. For example, a long list of trig identities is given in Section 1.4.2 as exercises (most of which have solutions included).

1.2 Notation for points and lines

Most material in this section is can be found in the vector space review (Section 23.1), but is repeated here in only a slightly different context.

The cartesian plane is often denoted by

$$\mathbb{R}^2 = \{(x, y) : x, y \in \mathbb{R}\}.$$

The Euclidean plane is the cartesian plane along with a distance defined by

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

The Euclidean plane is denoted by $\mathbb{E}^2$; however, notations $\mathbb{R}^2$ and $\mathbb{E}^2$ are often used interchangeably when no confusion arises.

In general, for $k \in \mathbb{Z}^+$, the k-dimensional *Euclidean space* is the set of points

$$\mathbb{R}^k = \{(x_1, \ldots, x_k) : x_1, \ldots, x_k \in \mathbb{R}\}$$

with the distance metric

$$d((x_1, \ldots, x_k), (y_1, \ldots, y_k)) = \sqrt{\sum_{i=1}^{k} (y_i - x_i)^2}.$$

In k-dimensional Euclidean space, points can be denoted by capital letters, or if information about coordinates is required, the notation $\mathbf{x} = (x_1, \ldots, x_k)$ is often used. A point in $\mathbb{R}^k$ can be interpreted simply as a k-tuple, but the same k-tuple can also be viewed as a directed line from the origin to the point. The interplay between points and vectors is discussed in a first year course in linear algebra (and reviewed here in Chapter 23).

The set of points $\mathbb{R}^k$ is a k-dimensional vector space with addition defined by

$$(x_1, \ldots, x_k) + (y_1, \ldots, y_k) = (x_1 + y_1, \ldots, x_k + y_k),$$

and for any $c \in \mathbb{R}$, scalar multiplication is defined by

$$c \cdot (x_1, \ldots, x_k) = (cx_1, \ldots, cx_k).$$

The vector space $(\mathbb{R}^k, +, \cdot)$ is also endowed with a dot product, defined by

$$(x_1, \cdots, x_k) \bullet (y_1, \ldots, y_k) = x_1 y_1 + \cdots + x_k y_k.$$

There are two main properties of the dot product: for any two vectors $\mathbf{x} = (x_1, \ldots, x_k)$ and $\mathbf{y} = (y_1, \ldots, y_k)$, their dot product $\mathbf{x} \bullet \mathbf{y}$ is 0 if and only if the two vectors are orthogonal, and the norm (size) of a vector $\mathbf{x}$ is $\|\mathbf{x}\| = \sqrt{\mathbf{x} \bullet \mathbf{x}}$. This dot product in $\mathbb{R}^k$ is also called the Euclidean inner product.

In much of mathematics, there is a "standard" notation that uses lower case (italic) letters for elements, capital letters for sets, and capital script letters for families of sets. For example, one might write $x \in X \in \mathcal{X}$ to denote the fact that an element x is in a set X which is chosen from some family of sets $\mathcal{X}$. A vector (or n-tuple) in some n-dimensional vector space is often denoted by a boldface letter $\mathbf{x}$. (Note: in handwritten mathematics, various notations are used for a vector, including $\overline{x}$, $\vec{x}$, or $\underset{\sim}{x}$, the latter of which is often preferred because the typesetter's notation used for boldface is to put a "squiggle" under the text to be made boldfaced. In many vector spaces, a vector need not have a starting and end point, so arrows above a letter may be meaningless. I think that the arrow notation for vectors stems from the often quoted definition "a vector is something with magnitude and direction", whereas the actual definition of a vector is simply "an element of a vector space".) In mathematics, most characters are typeset in "math italics"; however, in these notes there is one exception when it comes to boldfaced characters: the notation $\mathbf{x}$ is used instead of $\boldsymbol{x}$ only because $\mathbf{x}$ is easier to typeset in LaTeX.

In geometry, some of the above conventions are not always followed. Points, which are often considered as vectors or "basic" elements in some geometric setting, are often denoted by upper case letters, and lines are often denoted by lower case. So if a point P is on a line ℓ, and if a line is considered as a set of points, then $P \in \ell$ can be used (contrary to the practise of having lower case representing an element and upper case denoting a set). In these notes, a line is usually denoted by ℓ or m. However, a line can be given a label L and a point p is on L if and only if $p \in L$. Notation is not universal, not even in these notes; often, the context dictates the notation.

If points A and B are in Euclidean space (e.g., the plane $\mathbb{E}^2$ or 3-space $\mathbb{E}^3$) or in some other "geometry", the expression AB can denote the straight line segment from A to B. The distance from A to B is denoted by $|AB|$, the length of the segment (one may also denote the distance by $\|AB\|$ if one wants to emphasize that the length comes from a norm). The line containing A and B is often denoted by $\overleftrightarrow{AB}$, and the ray containing AB starting at A is $\overrightarrow{AB}$. However, it is also sometimes convenient to let $\overrightarrow{AB}$ denote the vector from A to B. The triangle determined by points A, B, and C is denoted $\triangle ABC$.

If A is a point and O is the origin (in any $\mathbb{R}^d$), the vector $\overrightarrow{OA}$ is sometimes represented simply as $\vec{A}$ or $\mathbf{A}$. If A and B are points, the vector from A to B is sometimes denoted by $\mathbf{AB}$, with boldface so as to differentiate from the segment AB, and without the arrow so as to not be confused with the ray $\overrightarrow{AB}$ emanating from A.

The intersection of lines ℓ and m may be denoted by $\ell \wedge m$. If two lines ℓ and m are parallel, write $\ell \| m$. If A and B are points and ℓ is the unique line containing them, one can write $\ell = A \vee B$. In a lattice or other partial order $(X, \leq)$, the symbols $\wedge$ and $\vee$ are called *meet* and *join* defined by (provided they exist)

$$a \wedge b = \sup\{x \in X : x \leq a, x \leq b\} = \text{glb}\{a, b\},$$

and

$$a \vee b = \inf\{y \in X : a \leq y, b \leq y\} = \text{lub}\{a, b\}.$$

In many geometries, certain elements (like points, lines, planes, hyperplanes) are arranged by the partial order "containment", forming a (modular) lattice, so the above notation makes sense. For more on such partial orders, see Section 21.1. For example, in most geometries, distinct points A and B are contained in a unique line, so $A \vee B$ is defined (as the line containing both points).

1.3 Notation for angles

The angle (at B) determined by points A, B, and C is denoted $\angle ABC$, or on occasion, simply by $\angle B$. The measure of an angle $\angle ABC$ can be denoted $m\angle ABC$; however, it is quite common to refer to both the angle and the measure of the angle by simply $\angle ABC$ (or just $\angle B$).

Angles can be given in degrees or in radian measure; the angle of one revolution is divided into 360 degrees The distance around the unit circle is 2π, and so the angle of one revolution is also called 2π *radians*. Often, the word "radian" is dropped; for example one says that a right angle is $\pi/2$ or has measure $\pi/2$.

Exercise 1. *How many degrees are there in 1 radian?*

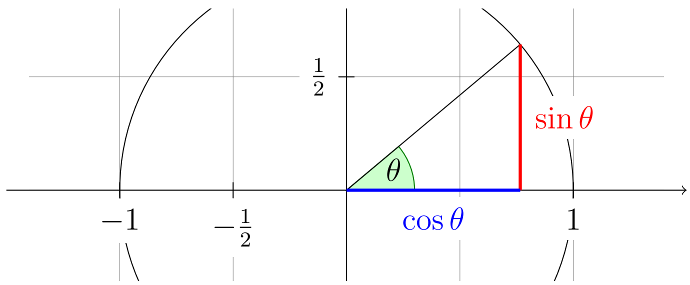
An angle is called a *right angle* if and only if its measure is $\pi/2$ (or 90 degrees). An angle is called *acute* if and only if its measure is less than $\pi/2$ and *obtuse* if and only if its measure is between $\pi/2$ and π.

Two (positive) angles are called *supplementary* if the sum of their measures is π, and *complementary* if the sum of their measures is $\pi/2$.

1.4 Trigonometry

1.4.1 The basics

The two main functions in trigonometry are *sine* and *cosine* (abbreviated sin and cos). The functions $\sin : \mathbb{R} \to [-1, 1]$ and $\cos : \mathbb{R} \to [-1, 1]$ are defined as follows: If a point (x, y) on the unit circle $x^2 + y^2 = 1$ is determined by an angle θ (measured in a counter-clockwise manner about the origin starting from the positive x-axis), then define $\cos(\theta) = x$ and $\sin(\theta) = y$. When it causes no confusion, the parentheses may be dropped, writing $\cos\theta$ and $\sin\theta$.

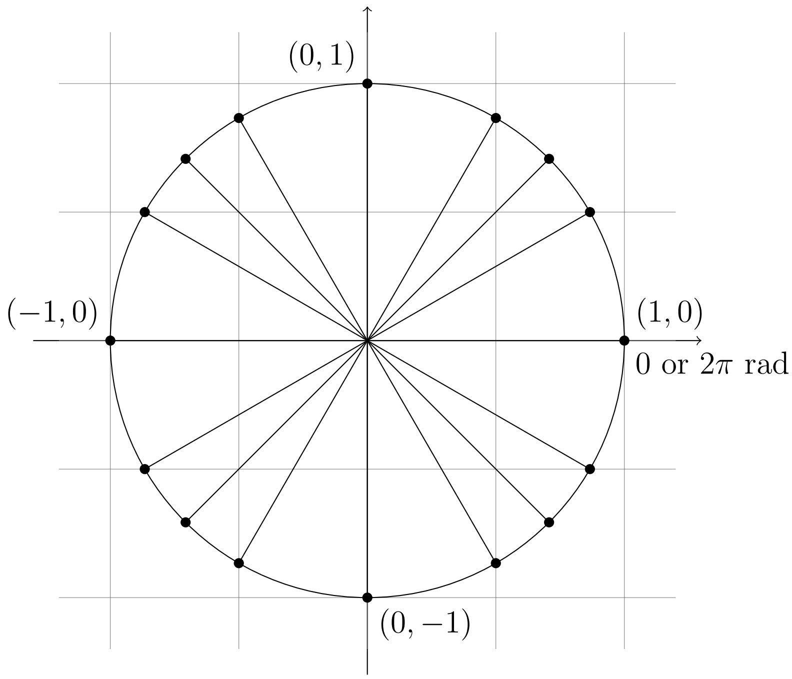


Since any multiple of 2π added to an angle gives the same point on the unit circle, for any $k \in \mathbb{Z}$, $\sin(\theta + k \cdot 2\pi) = \sin(\theta)$ and $\cos(\theta + k \cdot 2\pi) = \cos(\theta)$. For any real number (or angle) θ, the equation of the *unit circle* (the circle of radius 1 centered at the origin) becomes

$$\cos^2(\theta) + \sin^2(\theta) = 1. \tag{1.1}$$

[Note: $\cos^2(\theta)$ means $(\cos(\theta))^2$.]

Exercise 2. *In the following diagram of the unit circle, fill in the missing values. Give angles both in degrees and in radian measure. The coordinates of the point on the unit circle that corresponds to an angle* θ *are* $(x, y) = (\cos\theta, \sin\theta)$.



If $\alpha \neq 0$, the angle $-\alpha$ is measured from the x-axis in the opposite direction that α is; for example, the angle $-\pi/6$ is measured clockwise from the x-axis. By looking at the unit circle,

$$\sin(-\alpha) = -\sin(\alpha); \tag{1.2}$$

$$\cos(-\alpha) = \cos(\alpha). \tag{1.3}$$

The other four main trigonometric functions are defined by

$$\tan\theta = \frac{\sin\theta}{\cos\theta}.$$

$$\sec\theta = \frac{1}{\cos\theta}.$$

$$\csc\theta = \frac{1}{\sin\theta}.$$

$$\cot\theta = \frac{\cos\theta}{\sin\theta} = \frac{1}{\tan\theta}.$$

Standard identities for sums of angles are:

$$\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta. \tag{1.4}$$

$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta. \tag{1.5}$$

Exercise 3. *Prove formulae (1.4) and (1.5) for angle sums. You may need to separate proofs into cases depending on how large α and β are.*

With $\alpha = \beta$, equations (1.4) and (1.5) become what are called "double-angle identities":

$$\sin(2\alpha) = 2\sin\alpha\cos\alpha. \tag{1.6}$$

$$\cos(2\alpha) = \cos^2\alpha - \sin^2\alpha. \tag{1.7}$$

Equation (1.7) has two forms, depending upon which replacement $\cos^2(\alpha) = 1 - \sin^2(\alpha)$ or $\sin^2(\alpha) = 1 - \cos^2(\alpha)$ is used:

$$\cos(2\alpha) = 1 - 2\sin^2(\alpha). \tag{1.8}$$

$$\cos(2\alpha) = 2\cos^2(\alpha) - 1. \tag{1.9}$$

Replacing 2α with θ in (1.8) and (1.9) gives the standard "half-angle" identities.

$$\sin\left(\frac{\theta}{2}\right) = \sqrt{\frac{1-\cos\theta}{2}}. \tag{1.10}$$

$$\cos\left(\frac{\theta}{2}\right) = \sqrt{\frac{1+\cos\theta}{2}}. \tag{1.11}$$

$$\tag{1.12}$$

There is also a formula for the tangent of a sum of angles:

$$\tan(\alpha+\beta) = \frac{\tan\alpha + \tan\beta}{1 - \tan\alpha\tan\beta}, \tag{1.13}$$

which follows directly from (1.4) and (1.5) and the definitions:

$$\begin{aligned}
\tan(\alpha+\beta) &= \frac{\sin(\alpha+\beta)}{\cos(\alpha+\beta)} \\
&= \frac{\sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)} \\
&= \frac{\frac{\sin(\alpha)\cos(\beta)}{\cos(\alpha)\cos(\beta)} + \frac{\cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta)}}{1 - \frac{\sin(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta)}}
\end{aligned}$$

$$= \frac{\frac{\sin(\alpha)}{\cos(\alpha)} + \frac{\sin(\beta)}{\cos(\beta)}}{1 - \frac{\sin(\alpha)}{\cos(\alpha)}\frac{\sin(\beta)}{\cos(\beta)}}$$

$$= \frac{\tan(\alpha) + \tan(\beta)}{1 - \tan(\alpha)\tan(\beta)}.$$

The addition formula for tangent also applies to differences; using the fact that $\tan(-\beta) = -\tan(\beta)$,

$$\tan(\alpha - \beta) = \frac{\tan(\alpha) - \tan(\beta)}{1 + \tan(\alpha)\tan(\beta)}. \tag{1.14}$$

□

1.4.2 Exercises with trigonometric identities

The exercises here are not necessarily needed in any course in combinatorial geometry, but some may prove useful in areas of Euclidean geometry. (Many of these exercises and solutions also appear in [433].)

Exercise 4. *Prove the half-angle identity for tangent:*

$$\tan\left(\frac{\theta}{2}\right) = \sqrt{\frac{1-\cos\theta}{1+\cos\theta}}. \tag{1.15}$$

The identity in the following exercise is often used in a first year (integral) calculus course.

Exercise 5. *Show that*

$$\cos(A+B) + \cos(A-B) = 2\cos(A)\cos(B). \tag{1.16}$$

Hint: Use identity (1.5) to expand the two terms on the left.

Many of the following exercises have proofs by induction (appearing in, *e.g.*, [433]).

Exercise 6. *Show that for each $n \in \mathbb{Z}^+$,*

$$\cos(n\pi) = (-1)^n.$$

Exercise 7. *Let $x \in \mathbb{R}$. Show that for each $n \in \mathbb{Z}^+$,*

$$|\sin(nx)| \leq n|\sin(x)|.$$

The result in the next exercise is named for Abraham De Moivre (1667–1754). Recall that i is a (complex) number satisfying $i^2 = -1$.

Exercise 8 (De Moivre's Theorem)**.** *Let θ be an angle. Prove that for each positive integer n,*

$$[\cos(\theta) + i\sin(\theta)]^n = \cos(n\theta) + i\sin(n\theta).$$

Exercise 9. *Prove that for any positive integer n and any angle θ,*

$$\sin(\theta + n\pi) = (-1)^n \sin(\theta).$$

Exercise 10. *Prove that for any $n \in \mathbb{Z}^+$ and any angle θ,*

$$\cos(\theta + n\pi) = (-1)^n \cos(\theta).$$

Exercise 11. *Prove that for each $n \in \mathbb{Z}^+$, and an angle θ that is not a multiple of 2π,*

$$\sin\theta + \sin(2\theta) + \cdots + \sin(n\theta) = \frac{\sin(\frac{n+1}{2}\theta)\sin(\frac{n\theta}{2})}{\sin(\theta/2)}.$$

Exercise 12. *Prove that for $n \in \mathbb{Z}^+$ and any angle θ that is not a multiple of 2π,*

$$\cos\theta + \cos(2\theta) + \cdots + \cos(n\theta) = \frac{\cos(\frac{n+1}{2}\theta)\sin(\frac{n\theta}{2})}{\sin(\theta/2)}.$$

Exercise 13. *Prove that for each $n \in \mathbb{Z}^+$, and any angle θ that is not a multiple of π,*

$$\sin(\theta) + \sin(3\theta) + \cdots + \sin((2n-1)\theta) = \frac{\sin^2(n\theta)}{\sin(\theta)}.$$

Exercise 14. *Prove that for $n \in \mathbb{Z}^+$, and any angle θ that is not a multiple of π,*

$$\cos\theta + \cos(3\theta) + \cdots + \cos((2n-1)\theta) = \frac{\sin(2n\theta)}{2\sin(\theta)}.$$

Exercise 15. *Let θ be not a multiple of π. Let $s_0 = 0$, $s_1 = 1$, and for $n \geq 2$, recursively define*

$$s_n = 2\cos(\theta)s_{n-1} - s_{n-2}.$$

If θ is not a multiple of π, prove that for each integer $n \geq 0$,

$$s_n = \frac{\sin(n\theta)}{\sin(\theta)},$$

and for each $n \geq 1$,

$$\cos(n\theta) = \cos(\theta)s_n - s_{n-1}.$$

The next exercise uses the same recursion as in Exercise 15, however with different initial values.

Exercise 16. *Let θ be any angle. Define a sequence of real numbers recursively by $s_1 = \cos(\theta)$, $s_2 = \cos(2\theta)$ and for $n > 2$, define*

$$s_n = 2\cos(\theta)s_{n-1} - s_{n-2}.$$

Prove that for every $n \in \mathbb{Z}^+$, $s_n = \cos(n\theta)$.

Exercise 17. *Let x be any real number. Prove that for any positive integer n,*

$$\cos^{2n}(x) + \sin^{2n}(x) \geq \frac{1}{2^{n-1}}.$$

Exercise 18. *Prove that for any non-negative integer n,*

$$\cos(\alpha)\cos(2\alpha)\cos(4\alpha)\cdots\cos(2^n\alpha) = \frac{\sin(2^{n+1}\alpha)}{2^{n+1}\sin(\alpha)}.$$

Exercise 19. *Prove that for any positive integer n and any angle t that is not a multiple of 2π,*

$$\frac{1}{2} + \cos(t) + \cos(2t) + \cdots + \cos(nt) = \frac{\sin((2n+1)t/2)}{2\sin(t/2)}.$$

The expression

$$D_n(t) = \frac{1}{2} + \sum_{j=1}^{n} \cos(jt)$$

is called the *Dirichlet kernel*, arising in the theory of convex functions and Fourier series (see, *e.g.*, [280, p. 64]).

The next exercise regards the average of Dirichlet kernels, and was developed by Lipot Fejér. It might also be interesting to note that among Fejér's students was Paul Erdős, and that Fejér's advisor was Schwarz (as in "Cauchy–Schwarz inequality").

Exercise 20. *Using the notation from Exercise 19 put*

$$K_N(t) = \frac{1}{N+1}\sum_{n=0}^{N} D_n(t).$$

$K_N(t)$ is called the Fejér kernel (see, e.g., [280, p.64]). Note that when $n=0$, the sum in $D_n(t)$ is empty, so $D_0(t) = 1/2$. Prove that for $N \geq 0$,

$$K_N(t) = \frac{\sin^2((N+1)t/2)}{2(N+1)\sin^2(t/2)}.$$

Exercise 21. *Prove that for each positive integer n and any real number x that is not a multiple of 2π,*

$$\begin{aligned}\sin(x) + 2\sin(2x) + 3\sin(3x) + \cdots + n\sin(nx) \\ = \frac{(n+1)\sin(nx) - n\sin((n+1)x)}{4\sin^2(x/2)}.\end{aligned}$$

Exercise 22. *Prove that for each positive integer n and any real number x that is not a multiple of 2π,*

$$\begin{aligned}\cos(x) + 2\cos(2x) + 3\cos(3x) + \cdots + n\cos(nx) \\ = \frac{(n+1)\cos(nx) - n\cos((n+1)x) - 1}{4\sin^2(x/2)}.\end{aligned}$$

Exercise 23. *For any real number x that is not a multiple of π, prove that for each positive integer n,*

$$\frac{1}{2}\tan\left(\frac{x}{2}\right) + \frac{1}{2^2}\tan\left(\frac{x}{2^2}\right) + \cdots + \frac{1}{2^n}\tan\left(\frac{x}{2^n}\right) = \frac{1}{2^n}\cot(\frac{x}{2^n}) - \cot(x).$$

Exercise 24. *Prove that for each positive integer n,*

$$\begin{aligned}\cot^{-1}(3) + \cot^{-1}(5) + \cdots + \cot^{-1}(2n+1) \\ = \tan^{-1}(2) + \tan^{-1}\left(\frac{3}{2}\right) + \cdots + \tan^{-1}\left(\frac{n+1}{n}\right) - n\tan^{-1}(1).\end{aligned}$$

Exercise 25. *Prove that*

$$\prod_{n=1}^{\infty} \frac{1}{1-\tan^2(\frac{1}{2^n})} = \tan(1).$$

Exercise 26. *Show that for each integer $n \geq 2$, there exist (real) constants $a_0, a_1, \ldots, a_n$, and $b_0, b_1, \ldots, b_n$ so that*

$$\sin^n(x) = \sum_{r=0}^{n} \left(a_r \cos(rx) + b_r \sin(rx)\right).$$

Exercise 27. *Let x and α be real numbers so that $x + \frac{1}{x} = 2\cos(\alpha)$. Prove that for every $n \geq 1$,*

$$x^n + \frac{1}{x^n} = 2cos(n\alpha).$$

1.5 Collinear points

Three points are said to be *collinear* if and only if they all lie on a common line, and three lines are said to be *concurrent* if and only if all three lines pass through (or contain) the same point.

Lemma 1.5.1. *Let A and B be distinct points in the plane $\mathbb{E}^2$. To each point C on the line $\overleftrightarrow{AB}$, there exists a unique $t \in \mathbb{R}$ so that*

$$\vec{C} = (1-t)\vec{A} + t\vec{B}$$

(and for each such t, the corresponding point is on the line).

Proof: Let $\mathbf{v} = \vec{AB}$. Then the line ℓ through A with direction vector $\mathbf{v}$ is given by $\ell = \{\vec{A} + t\mathbf{v} : t \in \mathbb{R}\}$. Since $\vec{AB} = \vec{B} - \vec{A}$, any point on ℓ is of the form $\vec{A} + t(\vec{B} - \vec{A}) = (1-t)\vec{A} + t\vec{B}$. □

Lemma 1.5.2. *Points A, B, and C (not necessarily distinct) in $\mathbb{R}^2$ are collinear if and only if there exist α, β, γ, not all zero, so that $\alpha + \beta + \gamma = 0$ and (switching to boldface notation for vectors)*

$$\alpha\mathbf{A} + \beta\mathbf{B} + \gamma\mathbf{C} = \mathbf{0}.$$

Proof: For the first direction, assume that A, B, C are collinear. If all three points are the same, then the result is trivial (pick, for example, $\alpha = 1, \beta = -1, \gamma = 0$), so without loss of generality, suppose that $A \neq B$. By Lemma 1.5.1, there exists $t \in \mathbb{R}$ so that $\mathbf{C} = (1-t)\mathbf{A} + t\mathbf{B}$. Put $\alpha = 1-t$, $\beta = t$ and $\gamma = -1$.

Now assume that α, β, γ are not all zero, $\alpha + \beta + \gamma = 0$, and $\alpha\mathbf{A} + \beta\mathbf{B} + \gamma\mathbf{C} = \mathbf{0}$. It remains to show that A, B, C are collinear. Without loss of generality, suppose that $\alpha \neq 0$. Then

$$\mathbf{A} = \frac{-\beta}{\alpha}\mathbf{B} + \frac{-\gamma}{\alpha}\mathbf{C}.$$

Put $t = \frac{-\gamma}{\alpha}$. Then

$$\frac{-\beta}{\alpha} = \frac{\alpha + \gamma}{\alpha} = 1 + \frac{\gamma}{\alpha} = 1 - t.$$

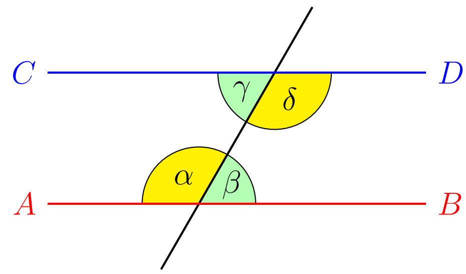
So $\mathbf{A} = (1-t)\mathbf{B} + t\mathbf{C}$. If $B = C$, then the conditions imply that $A = B$, in which case the points are trivially collinear. So suppose that $B \neq C$; then (by Lemma 1.5.1) A lies on the line $\overleftrightarrow{BC}$. $\square$

Corollary 1.5.3. *Three points (x_1, y_1), (x_2, y_2), and (x_3, y_3) in $\mathbb{R}^2$ are collinear if and only if*

$$\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} = 0.$$

1.6 Parallel lines

In some geometries, two lines in a geometry are said to be *parallel* if they do not intersect. In Euclidean geometry, two non-intersecting lines are parallel if they have the same direction vector. (There are geometries in which no two lines are parallel—see Chapter 11.) The following diagram was made by adapting the code given in the Tikz manual [859]:



If $AB \parallel CD$, then $\alpha = \delta$ and $\beta = \gamma$.

A line that cuts a pair of parallel lines is called a *transversal.*

Theorem 1.6.1 (Transversal ratios)**.** *Let ℓ_1, ℓ_2, and ℓ_3 be distinct parallel lines, and let m and n be two transversals. Let m intersect ℓ_1, ℓ_2, and ℓ_3 respectively in points A_1, A_2, A_3, and n intersect ℓ_1, ℓ_2, and ℓ_3 respectively in points B_1, B_2, B_3. Then*

$$\frac{|A_1A_2|}{|A_2A_3|} = \frac{|B_1B_2|}{|B_2B_3|}.$$

Proof: Let $\alpha = \frac{|A_1A_2|}{|A_2A_3|}$. First suppose that α is a rational number, say $\alpha = \frac{a}{b}$, where $a, b \in \mathbb{Z}^+$. Then $\frac{|A_1A_2|}{a} = \frac{|A_2A_3|}{b}$; call this common value λ. Then A_1A_2 can be partitioned into a segments of length λ by $a-1$ new points, and A_2A_3 can be partitioned into b segments of length λ by $b-1$ new points. From each of these new points, extend lines parallel to the ℓ_is, that intersect n in new points. The new points then partition B_1B_2 into a segments of some length λ' and similarly, new points partition $B2_B3$ also into b segments of length λ'. Thus $\frac{|A_1A_2|}{|A_2A_3|} = \frac{a\lambda}{b\lambda} = \frac{a}{b}$, and similarly, $\frac{|B_1B_2|}{|B_2B_3|} = \frac{a\lambda'}{b\lambda'} = \frac{a}{b}$, finishing the proof when α is rational.

To complete the proof when α is irrational, use the fact that any real number is the limit of a sequence of rational numbers. $\square$

1.7 Triangles

1.7.1 Types of triangles

- A triangle with all sides having the same length is called *equilateral.*
- A triangle with two sides the same length is called *isosceles.*
- A triangle with all side lengths different is called *scalene.*

- A triangle with one right angle is called a *right triangle*.
- A triangle with all angles less than $\pi/2$ is called *acute*.
- A triangle with one angle larger than $\pi/2$ is called *obtuse*.

1.7.2 Medians, altitudes, and angle bisectors of triangles

In this section are only three common definitions; the corresponding topics are investigated in later sections. For any triangle, there are three main types of line segments from a vertex.

Definition 1.7.1. In $\triangle ABC$ if X is a point on the line $\overleftrightarrow{BC}$, then AX is called an *altitude* of $\triangle ABC$, if and only if AX is perpendicular to $\overleftrightarrow{BC}$. Similarly, there exists an altitude from the other two vertices B and C.

Note that if a triangle is acute, then all altitudes lie inside of the triangle; if the triangle is obtuse, one altitude lies outside of the triangle. It is known that in any triangle, all altitudes intersect (see Section 1.7.9). Note that if one edge of a triangle is called the base, the length of the altitude from the opposite vertex is called "the altitude" or "height" of the triangle with a given base.

Definition 1.7.2. A *median* of a triangle is a line segment from one of the three vertices to a point bisecting the opposite side.

It is known that the three medians of a triangle have a common intersection (see Section 1.7.10.)

Definition 1.7.3. If $\triangle ABC$ is a triangle and D is a point on BC, then AD is called an angle bisector of $\triangle ABC$ if and only if AD bisects $\angle BAC$.

It is known that in any triangle, all three angle bisectors intersect in a common point (see Section 1.7.11).

1.7.3 Similar and congruent triangles

Two triangles $\triangle ABC$ and $\triangle DEF$ are said to be *similar* if and only if they have the same angles, and are called *congruent* if they are similar and all corresponding side lengths agree. Since by Exercise 82 the sum of angles in a triangle is constant ($= \pi$), if two triangles share two common angles, then all three angles agree and the triangles are similar.

If $\triangle ABC$ and $\triangle DEF$ are similar with $m\angle A = m\angle D$, $m\angle B = m\angle E$ and $m\angle C = m\angle F$, then corresponding ratios agree:

$$\frac{|AB|}{|DE|} = \frac{|BC|}{|EF|} = \frac{|AC|}{|DF|}.$$

Two triangles are congruent if they share any of the following:

- all three side lengths (SSS);
- two angles and the corresponding side length between them (ASA);
- two sides and the corresponding angle between them (SAS).

Two triangles are similar if they have the same angles. (So congruent triangles are similar, but similar triangles are not necessarily congruent.)

If two triangles share any side and two respective angles, they are congruent since two angles determine the third, in which case ASA (above) applies. So, for example, SAA also guarantees congruence.

In the above discussion, "corresponding" and "respective" are two adjectives not to be ignored. Surprisingly, it is possible for two triangles to share five of the six measures (three side lengths and three angles) and yet not be congruent. The following exercise shows the smallest such example among triangles with integer lengths.

Exercise 28. *Show that if a triangle has side lengths 12, 18, 27, and a second triangle has side lengths 8, 12, 18, then they are similar but not congruent. Can you find another such pair of triangles?*

For another exercise similar to Exercise 28, see Exercise 85.

A short discussion (and seven references, going back to 1954) of such triangles in Exercise 28 is found in Martin Gardner's *Mathematical circus* [374, Ch. 5]. Richard G. Pawley [714] calls triangles that share 5 of the 6 elements (side lengths and angles) but that are not congruent, "5-con" triangles. There are an infinite number of such triangle pairs (see also [402, Prob. 59], [509], [773], [865]). The reader might be interested to know that the classification of all such 5-con pairs involves Fibonacci numbers (see [481, Ch. 4] and [482]).

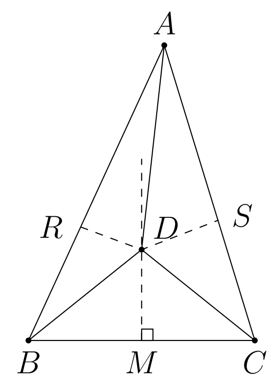
1.7.4 All triangles are isosceles?

There is an old favourite "proof" that might teach a geometer an important lesson. Of course the following is not really a theorem; can you find the error in the "proof"? Does this "proof" clearly fail when the triangle is not acute?

Theorem? All triangles are isosceles.

Proof?: Let $\triangle ABC$ be acute; the other cases are similar. Assume that the perpendicular bisector of BC misses A (otherwise the triangle is easily seen to be isosceles). Let D be the intersection of the angle bisector at A with the perpendicular bisector of BC, where M is the midpoint of BC (see diagram below).

By SAS, $\triangle BMD \cong \triangle CMD$, which shows $|BC| = |CD|$. (This also follows directly from Pythagoras and the fact that $|BM| = |MC|$.) Let R be the point on AB so that DR is perpendicular to AB, and let S be the point on AC so that DS is perpendicular to AC.



By SAA, $\triangle ARD \cong \triangle ASD$, which implies that $|DR| = |DS|$ and $|RA| = |SA|$. Also,

$$|BR| = \sqrt{|DB|^2 - |DR|^2} = \sqrt{|DC|^2 - |DS|^2} = |CS|.$$

Then

$$|BA| = |BR| + |RA| = |CS| + |SA| = |CA|,$$

and so $\triangle ABC$ is isosceles. $\square$

Exercise 29. *What is wrong with the above "proof"?*

So the lesson is "don't trust your diagrams!".

1.7.5 Circles and centers for triangles

There are two common circles associated with a triangle. The *inscribed circle* (also called an *incircle*) for a triangle is the unique circle drawn inside the triangle that is tangent to each of the sides.

If three points are not collinear, there is a unique circle containing these three points (this can be seen in any number of ways—for example, algebraically, or by looking at the family of all circles through two of the points). The *circumscribed circle* (or *circumcircle*) is the unique circle that passes through the three corners of the triangle (and so the triangle is inscribed in the circle).

How does one find the centers of these two circles (called *incenter* and *circumcenter* respectively)? It does not take too much effort to see that the circumcenter lies on each of the perpendicular bisectors of the sides of the triangle. To find the incenter, intersect the three angle bisectors (see Section 1.7.11).

See Exercise 103 for a question about these two centers of a triangle. Also see Exercises 102, 105 , and 115 for questions with circumcircles.

There are other "centers" of triangles. Another "center" of a triangle is called the *orthocenter*, the point where lines through all altitudes intersect (see Lemma 1.7.19).

In Section 1.7.10, the "center of mass" for a triangle is considered. This center (also called the *barycenter* or centroid) exists and can be computed using integral calculus, but can also be found easily by intersecting the three medians (from each vertex to the midpoint of the opposite side). In Lemma 1.7.20 it is shown that indeed all three medians of a triangle have a common intersection and that the centroid divides each median in a 1:2 ratio.

1.7.6 Lengths of sides of a triangle

For this section, a "triangle" is meant to be non-trivial, that is if A, B, C are vertices of a triangle, then A, B, C are distinct and are not all on one line. The following simple observation can be quite useful:

Lemma 1.7.4 (Triangle inequality)**.** *If a triangle has positive side lengths a, b, and c, then $a + b > c$.*

Exercise 30. *Prove Lemma 1.7.4.*

An application of Lemma 1.7.4 is found in Exercise 112.

Theorem 1.7.5 (Pythagoras)**.** *Let $\triangle ABC$ have non-zero side lengths a, b, and c, where $a = |BC|$, $b = |AC|$, and $c = |AB|$. Then the angle at C is a right angle if and only if $a^2 + b^2 = c^2$.*

As of 1968, over 400 proofs of Pythagoras's theorem were known [620], one of which was discovered by James Garfield (before he was President of the United States).

A *Pythagorean triple* is a collection of three positive integers a, b, c so that $a^2 + b^2 = c^2$. For example, $(3, 4, 5)$, $(6, 8, 10)$, and $(5, 12, 13)$ are Pythagorean triples.

Lemma 1.7.6. *In any Pythagorean triple a, b, c with $a^2 + b^2 = c^2$, at least one of a and b is even.*

Proof: Suppose, in hope of a contradiction, that both a and b are odd. Then a^2 and b^2 are odd, and so their sum, c^2 is even. Since c^2 is even, it follows that c is even, and so $c^2 \equiv 0 \pmod 4$. However, $a^2 \equiv 1 \pmod 4$ and $b^2 \equiv 1 \pmod 4$, and so $a^2 + b^2 \equiv 2 \pmod 4$, contradicting that $c^2 \equiv 0 \pmod 4$. $\square$

A standard result in number theory texts is that $\sqrt{2}$ is not rational. It follows that no two elements of a Pythagorean triple are the same.

Exercise 31. *Let (a, b, c) be a Pythagorean triple, with $a < b < c$. Show that the smallest of a or b is at least 3.*

A Pythagorean triple (a, b, c) is called *primitive* if a, b, c have no common factor. For example, $(6, 8, 10)$ is a Pythagorean triple that is not primitive since the values all have a common factor of 2. Note that if a, b, c is a Pythagorean triple, if any two share a common factor, then so do all three, so in a primitive Pythagorean triple, each pair is relatively prime.

Lemma 1.7.7. *Let a, b, c be a primitive Pythagorean triple with $a^2 + b^2 = c^2$. Then exactly one of a or b is odd and c is odd.*

Proof: If both a and b are even, then so is $a^2 + b^2 = c^2$, and so c is even, but in this case, then a, b, c all share a common factor of 2, contrary to the triple being primitive. If both a and b are odd, say $a = 2j + 1$, $b = 2k + 1$, then $a^2 + b^2 \equiv 1 + 1 = 2 \pmod 4$, but being a perfect square, c^2 is congruent to either 0 or 1 modulo 4, so a and b can not be both odd. Hence exactly one of a or b is odd. Since "odd+even=odd", c is then also odd. $\square$

The following characterization of primitive Pythagorean triples can be found in Euclid's *Elements*, book X, Proposition XXIX (see, e.g., [462]).

Theorem 1.7.8. *Let a, b, c be positive integers satisfying $a^2 + b^2 = c^2$, and without loss of generality, suppose that a is even. Then a, b, c is a primitive Pythagorean triple if and only if there exist relatively prime positive integers $m < n$, one odd, one even, so that $a = 2mn$, $b = n^2 - m^2$ and $c = m^2 + n^2$.*

Proof: Since $a^2 + b^2 = c^2$,

$$a^2 = c^2 - b^2 = (c - b)(c + b). \tag{1.17}$$

Since a is even, write $a = 2k$; since b and c are odd, both $c + b$ and $c - b$ are also even, so write $c - b = 2s$ and $c + b = 2t$.

Claim: s and t are relatively prime. Proof of claim: Solving the system $c - b = 2s$ and $c + b = 2t$ gives $b = t - s$ and $c = s + t$. If s and t have a common divisor, then so do b and c, contrary to (a, b, c) being a primitive triple, finishing the proof of the claim.

Rewriting equation (1.17), $4k^2 = (2s)(2t)$, and so $k^2 = st$. Since s and t are relatively prime, both s and t are perfect squares, say $s = m^2$ and $t = n^2$ (where $n > m$) and m and n are relatively prime. Since $b = t - s$ and $c = s + t$ are both odd, precisely one of s or t is odd, and so one of m or n is odd, the other even.

So $n^2 + m^2 = t + s = c$, $n^2 - m^2 = t - s = b$ and $a^2 = (c - b)(c + b) = 2s2t = 4(m^2)(n^2)$ implies that (since all numbers are positive) $a = 2mn$, concluding the first part of the proof.

To see the other direction, let $a = 2mn$, $b = n^2 - m^2$, and $c = m^2 + n^2$, where m and n are relatively prime and exactly one of m, n is even. Simple algebra confirms that $a^2 + b^2 = c^2$. It remains only to show that no two of a, b, c share a common factor. For example, if a and b share a common factor, then so do $2mn$ and $n^2 - m^2 = (n - m)(n + m)$. Since m is relatively prime to n, m is also relatively prime to $n + m$ and $n - m$, and so m is relatively prime to $b = n^2 - m^2$; a similar calculation shows that n is also relatively prime to b, and since b is odd, b is then relatively prime to $2mn = a$. The remaining calculations are nearly identical. $\square$

So any Pythagorean triple (a, b, c) is either a primitive Pythagorean triple or there exists a positive integer ℓ and a primitive triple (a', b', c') so that $(a, b, c) = (\ell a', \ell b', \ell c')$. Any right triangle with integer side lengths is called a Pythagorean triangle. The two Pythagorean triples $(20, 21, 29)$ and $(12, 35, 37)$ give triangles with the same area.

Exercise 32. *Find three Pythagorean triangles with the same area. Hint: The smallest such common area of three different Pythagorean triangles is 420.*

Exercise 33. *Prove or disprove: every Pythagorean triangle contains a side whose length is a multiple of 5.*

Exercise 34. *Show that in any Pythagorean triangle with the two legs having length a and b, then 12 divides ab.*

Exercise 35. *Show that the only Pythagorean triples (a, b, c) that are in arithmetic progression are the multiples of $(3, 4, 5)$.*

For more elementary properties of Pythagorean triples (including some of those given above), I recommend a small inexpensive book *Pythagorean triples* [794] by Wacław Sierpiński (1882–1969) first appearing in 1962 as a set of notes published by Yeshiva University, NY.

Theorem 1.7.9 (Cosine law)**.** *For a triangle with side lengths a, b, c if θ is the (measure of) the angle across from the side with length c,*

$$c^2 = a^2 + b^2 - 2ab \cdot \cos\theta.$$

Note that Pythagoras's theorem is a special case when $\theta = \pi/2$.

Theorem 1.7.10 (Sine law)**.** *Let $\triangle ABC$ be a triangle, and let α, β, γ be the measures of the angles at A, B, C respectively. If a, b, c are the respective lengths of the sides opposite A, B, C, then*

$$\frac{\sin(\alpha)}{a} = \frac{\sin(\beta)}{b} = \frac{\sin(\gamma)}{c}.$$

In some textbooks, the sine law is written as

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c},$$

where A, B, C are meant to indicate the angles at the respective points.

1.7.7 Area of a triangle

There are many ways to compute the area of a triangle. If a triangle $\triangle XYZ$ is drawn so that XY is horizontal and Z is above the line containing XY, then XY can be referred to as a *base* of $\triangle XYZ$, and the vertical distance from Z to the line containing XY is called its *height* (or "altitude", but "altitude" can often refer to the line segment from C to the base). If a triangle has base with length b and height h, then the area of the triangle is

$$A = \frac{1}{2}bh.$$

The above formula for area is likely the first learned by many students. (One easy proof is by dissecting and rearranging pieces—see steps 2 and 3 in Section 8.1.2.)

To use the formula $A = \frac{1}{2}bh$, one often has to first compute b and/or h separately. Different approaches to find b and h depend upon what is known about the triangle. For example, if all side-lengths of $\triangle XYZ$ are known, then any one its edges can be used as a base. To fully determine a triangle, at least one side length is needed, so choosing such a side as a base might be easiest.

If all side lengths are known and the triangle is not obtuse, one can use Pythagoras to find h as follows: suppose P lies on the base XY so that XPZ is a right angle, and then $|XP|^2 + h^2 = |XZ|^2$, $|PX|^2 + h^2 = |YZ|^2$ and $|XP| + |PY| = |XY|$—in which case, one can then also gets information about where P is. When a triangle is obtuse, the above idea still works, but the drawing is different and the equation $|XP| + |PY| = |XY|$ needs to be changed accordingly.

Another way to apply $A = \frac{1}{2}bh$ is when some angles are known (if two are known, then the third can also be found since the sum of interior angles of a triangle is 180°—see Exercise 82). For example, in $\triangle XYZ$, let $b = |XY|$. Then

$$h = |X| \sin(\angle YXZ) = |YZ| \sin(\angle XYZ),$$

and so if either of $|XZ|$ or $|YZ|$ is known, the area is easily found.

Without explicitly using trigonometry to find either b or h, there are other ways to find the area of a triangle. When only side lengths are known, Heron's formula (see Theorem 1.7.12) can be used. If a triangle is given in the plane with integer coordinates for vertices, then Pick's theorem can be applied (see Theorem 1.11.8).

Other ways to find the area of a triangle include linear algebra. If the coordinates of $X, Y, Z \in \mathbb{R}^2$ are known, the point P at the intersection of an altitude from Z and a vector containing XY is given by the projection formula

$$P = X + \operatorname{proj}_{\overrightarrow{XZ}}(\overrightarrow{XY})$$

and then $h = \|PZ\|$ is easily found.

Another way to compute the area of a triangle is by cross products. If $\mathbf{u} = (a, b, c)$ and $\mathbf{v} = (d, e, f)$ are vectors in $\mathbb{R}^3$, it is well-known that

$$\|\mathbf{u} \times \mathbf{v}\| = \|(bf - ec, -(af - dc), ae - bd)\| \tag{1.18}$$

is the area of the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$ because if θ is the angle from $\mathbf{u}$ to $\mathbf{v}$, then

$$\|\mathbf{u} \times \mathbf{v}\| = \sin(\theta)\|\mathbf{u}\| \cdot \|\mathbf{v}\|.$$

Using the cross product calculation and the fact that a triangle has one half the area of a parallelogram thereby formed, then using $\mathbf{u} = \overrightarrow{XY}$ and $\mathbf{v} = \overrightarrow{XZ}$, then

$$\operatorname{area}(\triangle XYZ) = \frac{1}{2}\|\overrightarrow{XY} \times \overrightarrow{XZ}\|.$$

To carry out the calculations for the area of a triangle whose points are given in $\mathbb{R}^2$, add a third coordinate 0, and use the above cross product. In this case, if, say $X = (x_1, x_2)$, $Y = (y_1, y_2)$, and $Z = (z_1, z_2)$, then $\overrightarrow{XY} = (y_1 - x_1, y_2 - x_2)$ and $\overrightarrow{XZ} = (z_1 - x_1, z_2 - x_2)$. By equation (1.18), $\|(a, b, 0) \times (d, e, 0)\| = \|(0, 0, ae - bd)\| = |ae - bd|$, and so

$$\operatorname{area}(\triangle XYZ) = \frac{1}{2}\left|(y_1 - x_1)(z_2 - x_2) - (y_2 - x_2)(z_1 - x_1)\right|.$$

Definition 1.7.11. The *semiperimeter* of a triangle is half the sum of its side lengths.

Theorem 1.7.12 (Heron's formula)**.** *A triangle with side lengths a, b, and c, and semiperimeter $s = \frac{1}{2}(a + b + c)$ has area $\sqrt{s(s-a)(s-b)(s-c)}$.*

Heron's formula is also written as

$$A = \frac{1}{4}\sqrt{(a + b + c)(-a + b + c)(a - b + c)(a + b - c)}. \tag{1.19}$$

There are many proofs of Heron's formula; see, e.g., [277] or [209] for a discussion.

The "isoperimetric inequality" for simple closed curves in the plane says that among all curves of fixed length, the circle has the largest area. There is a such an inequality when the curves are restricted to triangles.

Theorem 1.7.13 (Isoperimetric inequality for triangles.)**.** *Among all triangles with a fixed perimeter, the equilateral triangle has the greatest area.*

Proof: Let a, b, c be side lengths of a triangle and let $s = (a + b + c)/2$ be the semiperimeter. Letting A be the area of such a triangle, Heron's formula says

$$A^2 = s(s - a)(s - b)(s - c).$$

If the perimeter of the triangle is constant, then s is constant, so it remains to maximize $s(s-a)(s-b)(s-c)$ subject to to the constraint $a+b+c = 2s$. Recall that the AM-GM inequality for three terms says $(x + y + z)/3 \geq (xyz)^{1/3}$, with equality only when $x = y = z$. Applying this AM-GM inequality with $x = s - a$, $y = s - b$, $z = s - c$ gives

$$\frac{s}{3} = \frac{(s - a) + (s - b) + (s - c)}{3} \geq [(s - a)(s - b)(s - c)]^{1/3},$$

and so by Heron's formula,

$$\left(\frac{s}{3}\right)^3 \geq (s - a)(s - b)(s - c) = A^2/s,$$

which gives

$$A \leq \frac{s^2}{3^{3/2}},$$

with equality only when $s - a = s - b = s - c$ or $a = b = c$. □

The next theorem is folklore, and is given without proof.

Theorem 1.7.14. *Let T be a triangle with side lengths a, b, c, let $s = \frac{a+b+c}{2}$ be the semiperimeter, and let C be the circumcircle containing T. The radius of C (called the circumradius) is*

$$\frac{abc}{4\sqrt{s(s - a)(s - b)(s - c)}}.$$

See Exercise 110 for a Putnam problem that might be solved using Heron's formula. See Brahmagupta's formula (Theorem 1.9.7) for a similar formula for cyclic quadrilaterals.

Exercise 36. *Using Heron's formula, prove the theorem of Pythagoras.*

1.7.8 Theorems of Menelaus and Ceva

The following Theorem 1.7.15 and its converse (Theorem 1.7.17) are often collectively known as "Menelaus's theorem". The statements and proofs of these two theorems given here are based on the presentation given in [942].

Theorem 1.7.15 (Menelaus, $\sim$ 100 BC)**.** *Let $\triangle ABC$ be a triangle, and let X, Y, and Z be points (none of which are A, B, or C) on the lines $\overleftrightarrow{AB}$, $\overleftrightarrow{BC}$, and $\overleftrightarrow{AC}$ respectively. If the points X, Y, Z are collinear, then*

$$\frac{|AZ|}{|CZ|} \cdot \frac{|BX|}{|AX|} \cdot \frac{|CY|}{|BY|} = 1.$$

Proof: Suppose that X, Y, Z lie on a line ℓ. Either ℓ intersects $\triangle ABC$ twice or not at all. For the first case, suppose without loss of generality that X appears on the extended line $\overleftrightarrow{AB}$ but Y and Z are on proper sides as in Figure 1.1.

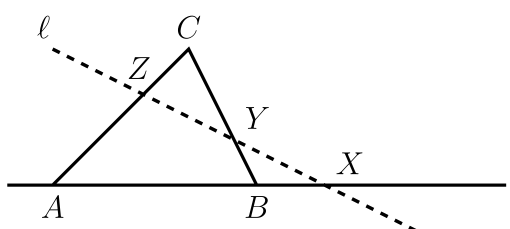


Figure 1.1: An example for Menelaus's theorem

Drop a perpendicular from each of A, B, and C to the line ℓ, and let P_A, P_B, P_C be the points on ℓ so that $AP_A \perp \ell$, $BP_B \perp \ell$, and $CP_C \perp \ell$.

Then there are three pairs of similar triangles, and since similar triangles have ratios of side lengths the same,

$$\triangle AP_AZ \sim \triangle CP_CZ \Rightarrow \frac{|AZ|}{|CZ|} = \frac{|AP_A|}{|AP_C|}; \tag{1.20}$$

$$\triangle BP_BX \sim \triangle AP_AX \Rightarrow \frac{|BX|}{|AX|} = \frac{|BP_B|}{|AP_A|}; \tag{1.21}$$

$$\triangle CP_CY \sim \triangle BP_BY \Rightarrow \frac{|CY|}{|BY|} = \frac{|CP_C|}{|BP_B|}. \tag{1.22}$$

The product of the three right hand terms of the equalities in (1.20), (1.21), and (1.22) is 1, and so

$$\frac{|AZ|}{|CZ|} \cdot \frac{|BX|}{|AX|} \cdot \frac{|CY|}{|BY|} = 1,$$

as desired.

The second case, when ℓ misses the triangle $\triangle ABC$ is similar, and is left to the reader. $\square$

Note that in the proof Theorem 1.7.15, the three perpendiculars could have been replaced by any three parallel lines from A, B, and C.

To streamline the proof of the converse to Theorem 1.7.15, a simple lemma is used; the proof is left to the reader.

Lemma 1.7.16. *Let A and B be distinct points in $\mathbb{R}^2$, and let α be a non-negative real number. Then there exists a unique point X on the segment AB so that $\frac{|AX|}{|BX|} = \alpha$. Also, if $\alpha < 1$, there is a second (unique) point X_1 on $\overleftrightarrow{AB}$ not in AB but on the ray extending from A not including B so that $\frac{|AX_1|}{|BX_1|} = \alpha$, and if $\alpha > 1$, then there exists a second point X_2 on $\overleftrightarrow{AB}$ not in AB but on the ray extending from B not including A so that $\frac{|AX_2|}{|BX_2|} = \alpha$.*

Theorem 1.7.17 (Menelaus's theorem converse)**.** *For a triangle $\triangle ABC$, let X, Y, and Z be respective points (none of which is A, B, C) on $\overleftrightarrow{AB}$, $\overleftrightarrow{BC}$, and $\overleftrightarrow{AC}$, where either one or three of X, Y, Z lie on the extended sides of $\triangle ABC$. Suppose that*

$$\frac{|AZ|}{|CZ|} \cdot \frac{|BX|}{|AX|} \cdot \frac{|CY|}{|BY|} = 1. \tag{1.23}$$

Then X, Y, and Z are collinear.

Proof: First consider the case where Y and Z are on proper sides BC and CA of $\triangle ABC$.

If ZY is parallel to AB, then by the transversal ratio theorem (Theorem 1.6.1), $\frac{|AZ|}{|CZ|} = \frac{|BY|}{|CY|}$ and so by (1.23), $\frac{|BX|}{|AX|} = 1$, which says X is the midpoint of AB, contradicting that X lies on the extended side. Thus, $\overleftrightarrow{AB}$ and $\overleftrightarrow{ZY}$ intersect.

So let X' be the point on $\overleftrightarrow{AB}$ so that $X' = \overleftrightarrow{ZY} \cap \overleftrightarrow{AB}$. By Menelaus's theorem (Theorem 1.7.15),

$$\frac{|AZ|}{|CZ|} \cdot \frac{|BX'|}{|AX'|} \cdot \frac{|CY|}{|BY|} = 1,$$

and so

$$\frac{|BX'|}{|AX'|} = \frac{|CZ|}{|AZ|} \cdot \frac{|BY|}{|CY|}.$$

By the assumption (1.23), the right side of the last equation is $\frac{|BX|}{|AX|}$, and so

$$\frac{|BX'|}{|AX'|} = \frac{|BX|}{|AX|}.$$

Call this last ratio α. Since neither X nor X' lie between A and B, by Lemma 1.7.16, the point X so that $\frac{|BX|}{|AX|} = \alpha$ is unique, and so $X = X'$, showing that X, Y, Z are collinear.

The case where X, Y, Z all lie on extended sides of $\triangle ABC$ is similar, and is left to the reader. $\square$

A *cevian* of a triangle is a line segment that joins a vertex of the triangle to a point on the opposite side. Cevians are named after the Italian mathematician and engineer Giovanni Benedetto Ceva (1647–1734). Examples of cevians are angle bisectors, medians, and (for acute triangles), altitudes. See Stewart's theorem (Theorem 1.7.25) for a general result on the length of a cevian. A famous result (preceding Stewart's theorem) is noted:

Theorem 1.7.18 (Ceva's theorem, 1678)**.** *For a triangle ABC, three cevians AY, BZ, and CX are concurrent iff*

$$\frac{|AZ|}{|CZ|} \cdot \frac{|BX|}{|AX|} \cdot \frac{|CY|}{|BY|} = 1.$$

Another version of Ceva's theorem holds for the cases where just one of X, Y, Z lie on a side of the triangle and the other two are on extended sides.

Exercise 37. *Prove Ceva's theorem in the case that AY, BZ, and CX are cevians.*

Ceva's theorem can be used to prove many popular results, including:

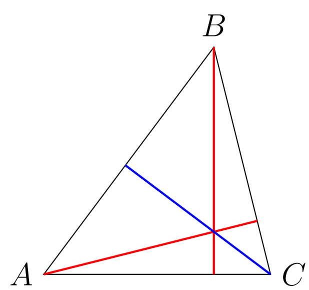
- Altitudes of a triangle are concurrent (see Lemma 1.7.19).
- Medians of a triangle are concurrent (see Lemma 1.7.20).
- Internal angle bisectors of a triangle are concurrent (see Lemma 1.7.23).

1.7.9 Altitudes are concurrent

In a triangle ABC, the *altitude* from A is a segment AX, where X lies on $\overleftrightarrow{BC}$ and AX is perpendicular to BC.

Lemma 1.7.19. *The three altitudes of a triangle are concurrent.*

Proof: Look at the two altitudes to points A and B (given in red in the following diagram), and let H be their intersection.



It remains to show that CH (and its extension to the blue segment) is perpendicular to AB. For this, it suffices to show (using dot products and vectors) that $(\mathbf{H}-\mathbf{C})\bullet(\mathbf{B}-\mathbf{A}) = 0$, or equivalently, that $\mathbf{H}\bullet(\mathbf{B}-\mathbf{A}) = \mathbf{C}\bullet(\mathbf{B}-\mathbf{A})$. Since H is on an altitude from A,

$$(\mathbf{H}-\mathbf{A})\bullet(\mathbf{B}-\mathbf{C}) = 0 \;\Leftrightarrow\; \mathbf{H}\bullet(\mathbf{B}-\mathbf{C}) = \mathbf{A}\bullet(\mathbf{B}-\mathbf{C}).$$

Similarly, since H is on an altitude from B,

$$(\mathbf{H}-\mathbf{B})\bullet(\mathbf{A}-\mathbf{C})=0 \ \Leftrightarrow\ \mathbf{H}\bullet(\mathbf{C}-\mathbf{A})=\mathbf{B}\bullet(\mathbf{C}-\mathbf{A}).$$

Adding these last two equalities gives

$$\mathbf{H}\bullet(\mathbf{B}-\mathbf{A})=\mathbf{A}\bullet(\mathbf{B}-\mathbf{C})+\mathbf{B}\bullet(\mathbf{C}-\mathbf{A})=\mathbf{A}\bullet\mathbf{B}-\mathbf{A}\bullet\mathbf{C}+\mathbf{B}\bullet\mathbf{C}-\mathbf{B}\bullet\mathbf{A}=(\mathbf{B}-\mathbf{A})\bullet\mathbf{C}$$

as desired. □

Another proof of Lemma 1.7.19 follows in a rather straightforward manner from Ceva's theorem (Theorem 1.7.18).

1.7.10 Medians are concurrent

In a triangle $\triangle ABC$, recall (Definition 1.7.2) that a *median* from A is a segment AX where X is the midpoint of BC.

Lemma 1.7.20. *All three medians of a triangle are concurrent at a point that divides each median in a ratio of 1/3 to 2/3.*

Proof: The proof is made easier by simply looking at the points of each median that determine the desired ratio and show that they all agree. Let $\triangle ABC$ have medians AX, BY, and CZ. Let G_1 be the point 2/3 from A to X. Then using vector notation,

$$\vec{G_1}-\vec{A}=\overrightarrow{AG_1}=\frac{2}{3}\overrightarrow{AX}=\frac{2}{3}\left(\vec{X}-\vec{A}\right)=\frac{2}{3}\left(\frac{1}{2}\left(\vec{B}+\vec{C}\right)-\vec{A}\right).$$

Adding $\vec{A}$ to each side shows $\vec{G_1}=\frac{1}{3}\left(\vec{B}+\vec{C}+\vec{A}\right)$. Repeating this idea for G_2 and G_3 (each a point 2/3 along a median) gives the same answer. □

Note: Another proof of Lemma 1.7.20 follows from Ceva's theorem (Theorem 1.7.18); an advantage of that proof is that one need not know to pick a point one third along a median. The point where all three medians cross is called the *barycenter* or *centroid* of the triangle.

Since points opposite one another from the median "balance", any triangle can balance along the axis containing the median. The centroid is the intersection of these three axes, and so any triangle of uniform thickness and density balances on just its centroid!

Simple project: Using some stiff cardboard (like the backing of a note pad), mark 3 points forming a triangle—for maximum effect, make the triangle fairly large. Draw the triangle and its medians (to find the midpoints use either a ruler or a compass). If the medians do not all intersect exactly in one point, recheck your lines. Now cut out the triangle and make a small pinhole at the intersection point. If your cuts are accurate, the triangle can be now made to balance at the tip of a pen or pencil. This "trick" was first shown to me by Professor H. K. Farahat in the summer of 1984, just before I decided to enroll in my first year undergraduate studies at University of Calgary. I had simply asked him what the center of triangle is, and he reached and got out his tools and showed me one center, the centroid.)

1.7.11 Angle bisectors are concurrent

An "angle bisector" is usually a line, ray, or a segment that cuts some particular angle in half. In a triangle $\triangle ABC$, the *angle bisector* of the angle at A is the segment from A to the point A^* on the opposite side BC so that AA^* bisects the angle at A.

The following observation is sometimes called the "angle bisector theorem", but because it is so simple, perhaps it is only a lemma.

Lemma 1.7.21 (Angle bisector lemma)**.** *In a triangle, any point on an angle bisector is equidistant from the two other sides.*

Proof: Let A, B, C be non-collinear points in the plane, and let A^* be the point on BC so that AA^* is the angle bisector in $\triangle BAC$ at A. PIck any point X on the bisector AA^*. Since the distance X to the other sides is measured orthogonally, let P be the point on AB be so that $XP \perp AB$, and let Q be the point on AC so that $XQ \perp AC$.

By SAA, triangles $\triangle AXP$ and $\triangle AXQ$ are congruent, and so $|XP| = |XQ|$. □

The proof of Lemma 1.7.21 shows that if $\overrightarrow{AB}$ and $\overrightarrow{AC}$ are rays, any point on the ray bisecting the angle at A is equidistant to both rays. Turning the proof of Lemma inside out also shows that if X is equidistant to two rays intersecting at A, then X lies on a bisector of the angle at A.

Another, less obvious result is also called the "angle bisector theorem":

Theorem 1.7.22 (Angle bisector theorem)**.** *In a triangle $\triangle ABC$, let D be a point on BC. Then AD is an angle bisector at A if and only if $\frac{|AC|}{|AB|} = \frac{|CD|}{|BD|}$.*

Proof: Let Z be the point on $\overrightarrow{BC}$ so that AZ is parallel to DC. Since ZB cuts parallel lines, $m\angle BAD = m\angle BZC$ and, since AC cuts parallel lines, by opposite interior angles, $m\angle ZCA = m\angle CAD$. Since AD is a bisector, $m\angle CAD = m\angle DAB$, and so $m\angle ZCA = m\angle CZA = m\angle CZB = m\angle BAD$. Thus $\triangle ZCA$ is isosceles and so $|AZ| = |AC|$.

Since $\triangle BAD$ and $\triangle BZC$ are similar,

$$\frac{|BZ|}{|AB|} = \frac{|BC|}{|BD|},$$

which gives

$$\frac{|ZA| + |AB|}{|AB|} = \frac{|CD| + |BD|}{|BD|}.$$

Then $\frac{|ZA|}{|AB|} = \frac{|CD|}{|BD|}$, and since $|ZA| = |AC|$, it follows that $\frac{|AC|}{|AB|} = \frac{|CD|}{|BD|}$, as desired.

To see the reverse implication, assume that

$$\frac{|AC|}{|AB|} = \frac{|CD|}{|BD|}, \tag{1.24}$$

and, as in the first part, let Z be the point on the ray $\overrightarrow{BA}$ so that $ZC \parallel AD$. Again, as in the first part, since ZC and AD are parallel, $m\angle BAD = m\angle BZC$ (and so $\triangle BAD$ and $\triangle BZC$ are similar since they also share the angle at B), and by opposite interior angles, $m\angle ZCA = m\angle CAD$. By similar triangles,

$$\frac{|ZB|}{|AB|} = \frac{|CB|}{|BD|},$$

and so

$$\frac{|ZA| + |AB|}{|AB|} = \frac{|CD| + |DB|}{|BD|},$$

from which it follows that

$$\frac{|ZA|}{|AB|} = \frac{|CD|}{|DB|}.$$

By the assumption (1.24), $\frac{|AC|}{|AB|} = \frac{|CD|}{|BD|}$, and so

$$\frac{|ZA|}{|AB|} = \frac{|AC|}{|AB|},$$

and so $|ZA| = |AC|$. Hence $\triangle AZC$ is isosceles, and thus $m\angle ZXA = m\angle CZA$, which is the same as $m\angle CAD$. Since $m\angle CZA = m\angle CZB = m\angle DAB$, conclude that $m\angle CAD = m\angle DAB$, and so AD is a bisector of $\triangle ABC$. □

Lemma 1.7.23. *The three angle bisectors of a triangle intersect in a common point.*

Exercise 38. *Prove Lemma 1.7.23.*

In the proof of Lemma 1.7.23, it was shown that the common intersection of the angle bisectors is equidistant from all edges of the triangle, and so is the center of a circle inscribed in the triangle, called the *incircle*; this "center" is called is the *incenter*, and the radius of the incircle is called the *inradius*. The following is folklore, given here without proof.

Theorem 1.7.24. *Let T be a triangle with side lengths a, b, and c and internal angles α, β, and γ. Then the inradius of T is*

$$r = \frac{abc}{2(a+b+c)}(\cos\alpha + \cos\beta + \cos\gamma - 1).$$

1.7.12 Stewart's theorem

The Scottish mathematician Matthew Stewart (1717–1785) proved a theorem about the lengths of cevians.

Theorem 1.7.25 (Stewart's theorem)**.** *In a triangle $\triangle ABC$, let AB, BC, CA have respective lengths c, a, b. Let $\alpha \in [0,1]$ and let A' be a point on side BC that divides BC into αa and $(1-\alpha)a$. If the length of the cevian AA' is d, then*

$$(1-\alpha)b^2 + \alpha c^2 = d^2 + \alpha(1-\alpha)a^2.$$

Exercise 39. *Prove Stewart's theorem by applying the Pythagoras theorem. Divide the proof into two cases, when the triangle is acute and when the triangle is obtuse.*

When $\alpha = \frac{1}{2}$, Stewart's theorem says that if AA' is a median with length d, then

$$d^2 = \frac{2b^2 + 2c^2 - a^2}{4}.$$

When AA' is an altitude, Stewart's theorem is confirmed by the Pythagoras theorem.

1.7.13 Morley's theorem

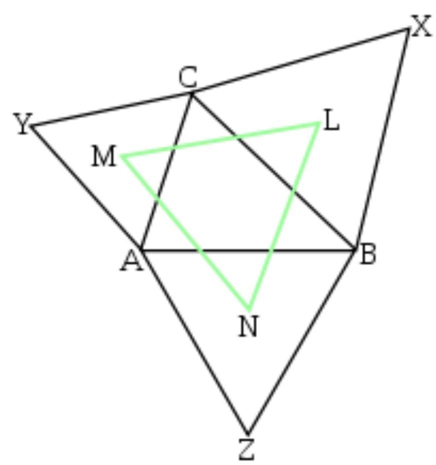
Theorem 1.7.26 (Morley's theorem)**.** *Let ABC be a triangle, and let D, E, and F be the first intersection points of the six interior angle trisectors (two for each interior angle). Then DEF is an equilateral triangle.*

It seems that most proofs of Morley's theorem are non-trivial. See [118, 44] for two proofs, both rather straightforward (the first based on expressions for $\sin(3x)$), although requiring a bit of work.

1.7.14 Napoleon's theorem

Theorem 1.7.27 (Napoleon's theorem)**.** *Let $\triangle ABC$ be any triangle. Erect equilateral triangles on all its edges, and let L, M, N be the centers of these equilateral triangles. Then $\triangle LMN$ is an equilateral triangle.*

Proof (using complex numbers):



Denote all the points as in the diagram. Each point in the plane can be associated with a complex number, which is denoted by the corresponding lower-case letter. Then it follows that $x - b = e^{-i\frac{\pi}{3}}(c - b)$ and $\ell - b = \frac{1}{3}((x - b) + (c - b))$. A simple computation implies

$$\ell = \left(\frac{1}{2} - \frac{1}{2\sqrt{3}}i\right) c + \left(\frac{1}{2} + \frac{1}{2\sqrt{3}}i\right) b.$$

Similarly, by symmetry,

$$m = \left(\frac{1}{2} - \frac{1}{2\sqrt{3}}i\right) a + \left(\frac{1}{2} + \frac{1}{2\sqrt{3}}i\right) c,$$

$$n = \left(\frac{1}{2} - \frac{1}{2\sqrt{3}}i\right) b + \left(\frac{1}{2} + \frac{1}{2\sqrt{3}}i\right) a.$$

Verifying that $\ell - n = e^{-\frac{\pi i}{3}}(m - n)$ completes the proof. $\square$

1.8 Pappus and Desargues

The following two theorems (of Pappus and Desargues, resp.) hold in the Euclidean plane; they also hold in some other geometries, but not all. In some geometries, they can not be proved and so these results may be held as axioms. One feature of these theorems is that they can be proved without any reference to distance, and so some say that these theorems are "projective".

Both can be proved for Euclidean geometries by different techniques; Pappus' theorem is not proved here, but two proofs of Desargues' theorem are given (for Euclidean geometries), one that uses vectors and another that uses Menelaus's theorem (Theorems 1.7.15 and 1.7.17). Two similar proofs are available for Pappus' theorem. For more on Pappus' theorem (with a proof using perspectivities), see the extensive article by Elena Anne Marchisotto [636].

Theorem 1.8.1 (Pappus)**.** *In the Euclidean plane, let A, B, and C be points on a line ℓ, and let A', B', and C' be points on a line m. Then the three intersection points $AB' \wedge A'B$, $AC' \wedge A'C$ and $BC' \wedge B'C$ are collinear. (See Figure 1.2.)*

The roles of the theorems of Desargues and Pappus have since become central in the analysis of axioms, other geometries, and algebra, particularly with regard to projective geometries (e.g., see Theorem 11.4.14), but I do not say much more on this here.

See [636] for the following affine version of Pappus's theorem (and diagrams).

Theorem 1.8.2 (Affine Pappus)**.** *If the vertices of a hexagon are alternately on two straight lines (excluding their intersection) and if two pairs of opposite sides are parallel, then so is the third pair.*

The following was not first found by Desargues (1591-1661), but it is apparently named after him in honour of his work on projective planes. [I cannot recall where I read this.]

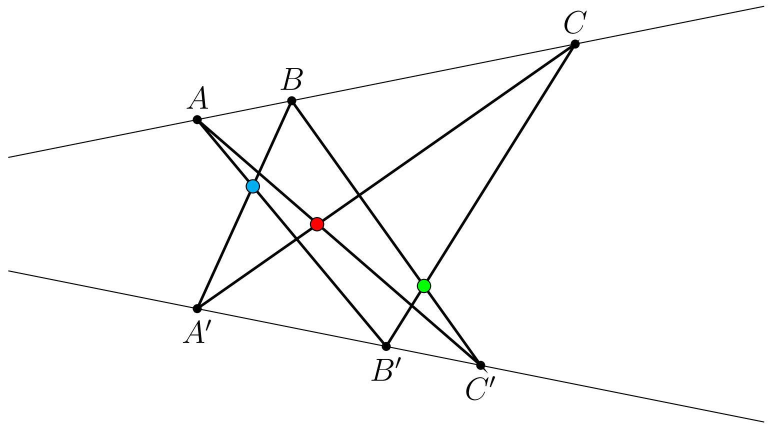


Figure 1.2: The Pappus configuration. The three coloured points are collinear.

Theorem 1.8.3 (Desargues' theorem for $\mathbb{E}^2$.)**.** *Let P, A, B, C, A', B', C' be points in the Euclidean plane so that P lies on each of $\overleftrightarrow{AA'}$, $\overleftrightarrow{BB'}$, and $\overleftrightarrow{CC'}$. Suppose that the following three intersection points exist: $X = \overleftrightarrow{AB} \wedge \overleftrightarrow{A'B'}$, $Y = \overleftrightarrow{AC} \wedge \overleftrightarrow{A'C'}$, and $Z = \overleftrightarrow{BC} \wedge \overleftrightarrow{B'C'}$. Then X, Y, Z are collinear. (See Figure 1.3.)*

Desargues' theorem also holds in finite projective planes that are coordinatized over some field, but does not necessarily hold for an arbitrary projective plane–see Theorem 11.4.14.

Proof of Theorem 1.8.3 using vectors: For any point $Q \in \mathbb{R}^2$, let $\mathbf{Q} = \overrightarrow{OQ}$ represent the corresponding vector. Recall that Lemma 1.5.2 says X, Y, Z are collinear if and only if there exist c_1, c_2, c_3 (not all zero) so that $c_1 + c_2 + c_3 = 0$ and (now using bold notation for the associated vectors $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$),

$$c_1\mathbf{X} + c_2\mathbf{Y} + c_3\mathbf{Z} = \mathbf{0}.$$

Since P lies on each of $\overleftrightarrow{AA'}$, $\overleftrightarrow{BB'}$, and $\overleftrightarrow{CC'}$, (and P is distinct from these six points) there exist constants α, β, γ (none of which are 0 or 1) so that

$$\begin{aligned}\mathbf{P} &= \alpha\mathbf{A} + (1-\alpha)\mathbf{A}' && (1.25)\\ &= \beta\mathbf{B} + (1-\beta)\mathbf{B}' && (1.26)\\ &= \gamma\mathbf{C} + (1-\gamma)\mathbf{C}'. && (1.27)\end{aligned}$$

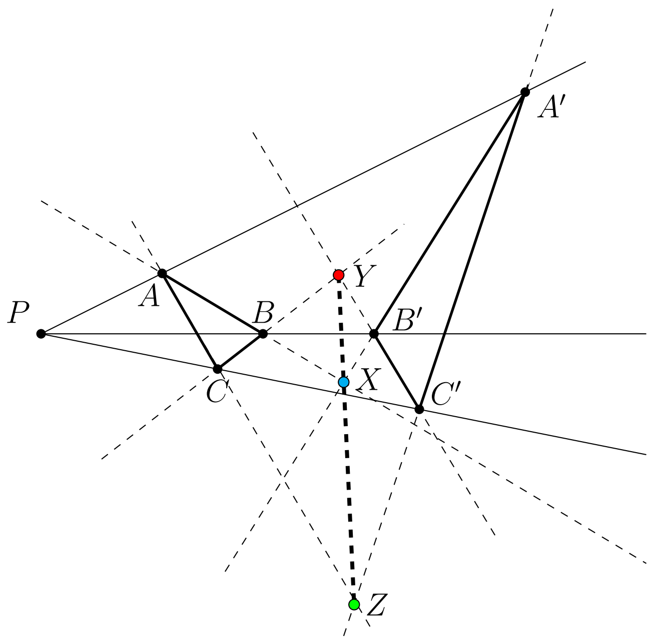


Figure 1.3: Desargues' configuration

Combining (1.25) with (1.26), then (1.26) with (1.27), and then (1.25) with (1.27), gives

$$\alpha\mathbf{A} - \beta\mathbf{B} = (1-\beta)\mathbf{B}' - (1-\alpha)\mathbf{A}'; \tag{1.28}$$

$$\beta\mathbf{B} - \gamma\mathbf{C} = (1-\gamma)\mathbf{C}' - (1-\beta)\mathbf{B}'; \tag{1.29}$$

$$\gamma\mathbf{C} - \alpha\mathbf{A} = (1-\alpha)\mathbf{A}' - (1-\gamma)\mathbf{C}'. \tag{1.30}$$

Since X lies on both $\overleftrightarrow{AB}$ and $\overleftrightarrow{A'B'}$, X is dependent upon both sides of (1.28); putting $\alpha\mathbf{A} - \beta\mathbf{B} + k\mathbf{X} = \mathbf{0}$, where $\alpha - \beta + k = 0$, shows $k = \beta - \alpha$. Thus,

$$\alpha\mathbf{A} - \beta\mathbf{B} = (\alpha - \beta)\mathbf{X}.$$

Similarly, from (1.29),

$$\beta\mathbf{B} - \gamma\mathbf{C} = (\beta - \gamma)\mathbf{Y},$$

and from (1.30),

$$\gamma\mathbf{C} - \alpha\mathbf{A} = (\gamma - \alpha)\mathbf{Z}.$$

Adding these last three equations shows

$$\mathbf{0} = (\alpha-\beta)\mathbf{X} + (\beta-\gamma)\mathbf{Y} + (\gamma-\alpha)\mathbf{Z}. \tag{1.31}$$

If, for example, $\alpha = \beta$, then the vectors $\mathbf{A} - \mathbf{B}$ and $\mathbf{A}' - \mathbf{B}'$ are parallel, and so X does not exist, so $\alpha \neq \beta$. Similarly, $\beta \neq \gamma$ and $\gamma \neq \alpha$. Noting that the coefficients in (1.31) sum to zero, X, Y, and Z are collinear. $\square$

Proof of Theorem 1.8.3 using Menelaus: (This proof is inspired by an exercise with hints in [497, 4D.7].) Menelaus's theorem is applied four times (Theorem 1.7.15 three times and Theorem 1.7.17 once).

Applying Menelaus's theorem to $\triangle PAB$ and points X, A', B':

$$\frac{|PB'|}{|BB'|} \cdot \frac{|BX|}{|AX|} \cdot \frac{|AA'|}{|PA'|} = 1. \tag{1.32}$$

Applying Menelaus's theorem to $\triangle PAC$ and points Z, A', C' :

$$\frac{|PC'|}{|CC'|} \cdot \frac{|CZ|}{|AZ|} \cdot \frac{|AA'|}{|PA'|} = 1. \tag{1.33}$$

Applying Menelaus's theorem to $\triangle PBC$ and points Y, B', C':

$$\frac{|PC'|}{|CC'|} \cdot \frac{|BY|}{|CY|} \cdot \frac{|BB'|}{|PB'|} = 1. \tag{1.34}$$

Multiplying reciprocals of (1.32) and (1.34) with (1.33) gives

$$\frac{|AX|}{|BX} \cdot \frac{|CY|}{|BY|} \cdot \frac{|CZ|}{|AZ|} = 1$$

and so by Theorem 1.7.17, the converse of Menelaus's theorem, applied to $\triangle AB'C'$, the points X, Y, Z are collinear. $\square$

1.9 Quadrilaterals

What is a quadrilateral? If A, B, C, D are distinct vertices (or points) in $\mathbb{R}^2$, a quadrilateral is a polygon formed by the four "sides" AB, BC, CD, and DA. A quadrilateral may be called "degenerate" if any three of its points are collinear. A non-degenerate quadrilateral is called *self-intersecting* if two non-consecutive sides intersect. In general, quadrilaterals are assumed to be non-degenerate and not self-intersecting unless otherwise stated; such quadrilaterals are also called "simple". If two sides of a quadrilateral are parallel, the quadrilateral is called a *trapezoid*. However, in many settings, the word "trapezoid" refers to quadrilaterals with only one pair of parallel sides. A quadrilateral is a *parallelogram* if and only if both pairs of opposite sides are parallel; a parallelogram with equal sides is a *rhombus*, and a rhombus with right angles is a square. See Figure 1.4 for images of types of quadrilaterals.

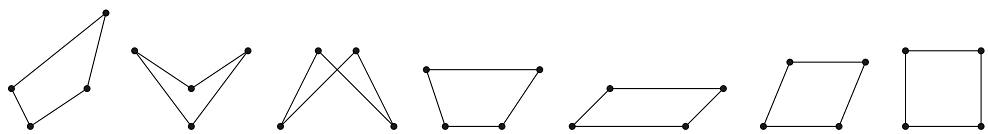


Figure 1.4: Quadrilaterals: convex, non-convex, self-intersecting, trapezoid, parallelogram, rhombus, and square.

Lemma 1.9.1. *The midpoints of the four sides of any convex quadrilateral are the vertices of a parallelogram.*

Exercise 40. *Prove Lemma 1.9.1.*

Lemma 1.9.2. *A quadrilateral is a parallelogram if and only if its diagonals bisect each other.*

Exercise 41. *Prove Lemma 1.9.2.*

Exercise 42. *Let α, β, γ, and δ be the angles of a quadrilateral taken in order. Show that $\cos(\alpha+\gamma) = \cos(\beta+\delta)$.*

Exercise 43. *Let $ABCD$ be a trapezoid where AB is parallel to DC (and BC is not parallel to AD). Let E be the midpoint of AB, F be the intersection of AC and DB, G be the midpoint of CD, and let H be the intersection of $\overleftrightarrow{AD}$ and $\overleftrightarrow{BC}$. Prove that E, F, G, and H are collinear.*

Theorem 1.9.3 (Parallelogram law)**.** *The sum of the squares of the lengths of four sides of a parallelogram in $\mathbb{R}^2$ is equal to the sum of the squares of the lengths of the two diagonals.*

Proof: The proof is given for any space with an inner product $\langle\cdot,\cdot\rangle$, of which the Euclidean inner product $\bullet$ is one.

Let $ABCD$ be a parallelogram. Put $\overrightarrow{AB} = \mathbf{x}$ and $\overrightarrow{BC} = \mathbf{y}$, the two vectors determining the parallelogram. The two diagonals of the parallelogram are

$\mathbf{x}+\mathbf{y}$ and $\mathbf{x}-\mathbf{y}$. Then the sum of the squares of the lengths of the four sides is

$$\begin{aligned}
2|AB|^2 + 2|BC|^2 &= 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 \\
&= 2\langle \mathbf{x}, \mathbf{x}\rangle + 2\langle \mathbf{y}, \mathbf{y}\rangle \\
&= \langle \mathbf{x}, \mathbf{x}\rangle + 2\langle \mathbf{x}, \mathbf{y}\rangle + \langle \mathbf{y}, \mathbf{y}\rangle + \langle \mathbf{x}, \mathbf{x}\rangle - 2\langle \mathbf{x}, \mathbf{y}\rangle + \langle \mathbf{y}, \mathbf{y}\rangle \\
&= \langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}\rangle + \langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y}\rangle \\
&= \|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2.
\end{aligned}$$

□

In the proof of Theorem 1.9.3, if the parallelogram is a rectangle, then $\langle \mathbf{x}, \mathbf{y}\rangle = 0$, in which case the equalities reduce to the Pythagoras theorem.

See, *e.g.*, Exercises 112, 115, 117 and 119 regarding quadrilaterals.

Definition 1.9.4. A *cyclic quadrilateral* is a quadrilateral whose vertices lie on a circle.

Usually, cyclic quadrilaterals are simple, *i.e.*, not self-intersecting. Recall that two angles are supplementary if their measures sum to π.

Lemma 1.9.5. *If $ABCD$ is a cyclic convex quadrilateral, then opposite interior angles are supplementary.*

Exercise 44. *Prove Lemma 1.9.5.*

See Theorem 1.12.8 for Miquel's theorem regarding intersecting circles and cyclic quadrilaterals.

Theorem 1.9.6 (Ptolemy's theorem)**.** *In a cyclic convex quadrilateral $ABCD$,*

$$|AC| \cdot |BD| = |AB| \cdot |CD| + |BC| \cdot |AD|.$$

Exercise 45. *Prove Ptolemy's theorem.*

For a simple application of Ptolemy's theorem, see Exercise 102.

The following is an extension of Heron's formula (see Theorem 1.7.12).

Theorem 1.9.7 (Brahmagupta, 620 AD)**.** *Let $ABCD$ be a cyclic quadrilateral with side lengths a, b, c, d. Then the area of $ABCD$ is*

$$\frac{1}{4}\sqrt{(a+b+c-d)(a+b-c+d)(a-b+b+c+d)(-a+b+c+d)}.$$

Heron's formula (see Theorem 1.7.12) follows from Brahmagupta's formula when one of a, b, c, d is zero (and the fact that all triangles have a circumcircle). The proof is omitted.

An even more general formula for the area of a quadrilateral is named after Carl Anton Bretschneider for his 1842 paper [145]; some authors give credit to Karl Georg Christian von Staudt or Strehlke in the same year. See the 1939 survey by Coolidge [212] for references to early proofs.

Theorem 1.9.8 (Bretschneider's formula)**.** *Let $ABCD$ be a quadrilateral (not necessarily cyclic), with side lengths a, b, c, d, and let α and γ be the angles at A and C. Also, let s be the semiperimeter, that is, $s = \frac{a+b+c+d}{2}$. The area of quadrilateral $ABCD$ is*

$$\sqrt{(s-a)(s-b)(s-c)(s-d) - abcd \cdot \cos^2\left(\frac{\alpha+\gamma}{2}\right)}.$$

In Bretschneider's formula, the other two (opposite) angles could also be used (see Exercise 42). For a short proof of Bretschneider's formula using dot products, see [914].

In a cyclic quadrilateral, by Lemma 1.9.5, the angles α and γ in Bretschneider's formula sum to π, and so Brahmagupta's formula follows from Bretschneider's.

1.10 Convexity

Convexity is studied more closely in Chapter 5; only the basic definition is given here. A set $C \subseteq \mathbb{R}^n$ is called *convex* if and only if for any two points $\mathbf{x}$ and $\mathbf{y}$ in C, for every $\lambda \in [0,1]$, the convex combination $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ is also in C. So if two points are in a convex set, then the straight line *segment* containing these two points is also in the set.

1.11 Polygons

1.11.1 Some basics

A polygon is said to be *triangulated* if the polygon is partitioned into triangles, each of which has vertices that are vertices of the polygon. For example, a convex n-gon can be triangulated by adding $n-2$ diagonals (chords) from one vertex.

There are two different ways to triangulate a square. A convex pentagon can be triangulated in exactly five ways (see Figure 1.5).

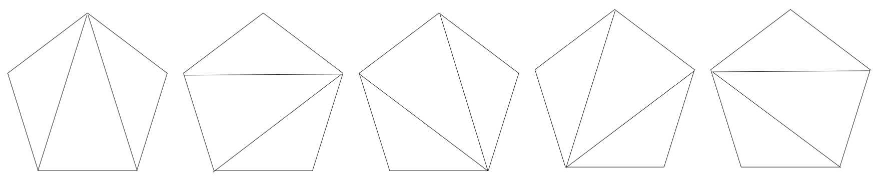


Figure 1.5: Five ways to triangulate a pentagon

For the moment, let t_n be the number of ways to triangulate a convex $(n+2)$-gon. Since a triangle is already "triangulated", $t_1 = 1$. As was just observed, $t_2 = 2$ and $t_3 = 5$. In 1751, Euler found (see [708] for historical references) the next few values $t_4 = 14$, $t_5 = 42$, $t_6 = 132$, $t_7 = 429$, and $t_8 = 1430$. Euler speculated that the formula describing this sequence is

$$t_n = \frac{1}{n+1}\binom{2n}{n}, \tag{1.35}$$

but was unable to prove it until 1759 (with the help of Goldbach and a critical result due to Segner).

Today, each number t_n is now called a "Catalan number", and Catalan numbers are now defined by the formula rather than "the number of ways to triangulate a convex $(n+2)$-gon".

Definition 1.11.1. For each integer $n \geq 0$, the nth Catalan number is

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

In early communications regarding these numbers, the notation C_n was not used (see Pak's website [709] for scans of original letters and early papers). Since Catalan published on such numbers eight decades after they were discovered, they are wrongly named. According to Igor Pak [708], the earliest use of the name "Catalan number" was by John Riordan (1903–1988) in *Math reviews* dated 1948 and 1964, and it was only after Riordan used the moniker in his 1968 text [752] did the name catch on.

There are many ways to write C_n; it is a simple exercise to show that for $n > 0$,

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

One of the keys to Euler finding the proof of (1.35), which says $t_n = C_n$, was the following recurrence relation for t_n found by Johann Andreas von Segner (1704–1777):

Theorem 1.11.2 (Segner, 1758 [783])**.** *For each $n \geq 1$, let t_n be the number of ways to triangulate a convex $(n+2)$-gon. Then for each n,*

$$t_{n+1} = t_n + t_1 t_{n-1} + t_2 t_{n-2} + \cdots + t_{n-1} t_1 + t_n. \tag{1.36}$$

Proof: Let P be a convex polygon on $n+3$ vertices, with vertices labelled (in cyclic order) $x, y, , v_0, v_1, v_2, \ldots, v_n$. Let $i \in \{0, 1, \ldots, n\}$ and consider the triangle $T_i = \triangle x v_i y$. If $i = 1$ T_1 lies on the outside of P, leaving a polygon on $n+2$ vertices to be further triangulated (which can be done in t_n ways). When $0 < i < n$, T_i separates P into two more polygons, say Q_i on the $i+2$ vertices $x, v_0, v_1, \ldots, v_i$ and R_i on the $n+3-(i+1) = n-i+2$ vertices $y, v_i, v_{i+1}, v_{i+2}, ldots, v_n$. Since Q_i can be triangulated t_i ways, and R_i can be triangulated in t_{n-i} ways, giving $t_i t_{n-i}$ triangulations that use T_i. Finally, when $i = n$, T_n lies on the outside, leaving a polygon with $n+2$ vertices, which can be triangulated in t_n ways. In any triangulation of P, exactly one of the T_i occur, so adding over all i gives the equation (1.36). □

To give a slightly more general form of equation (1.36), set $t_0 = 1$, which says there is only one way to "triangulate" a single edge. Then

$$t_{n+1} = \sum_{i=0}^{n} t_i t_{n-i}. \tag{1.37}$$

It is in this last form that Segner's recursion is often given (but with C_n replacing t_n).

Using Segner's recursion, Euler [323] (with help from Goldbach) was able to confirm his guess that $t_n = \frac{1}{n+1}\binom{2n}{n}$. The recursion was the key to using a generating functions to find the sequence (t_n). For those unfamiliar with the technique of generating functions, see, e.g., [753] or [754].

Theorem 1.11.3 (Euler, 1761 [323]). *Let $t_0 = 1$ and a sequence $(t_i)_{i \geq 0}$ of positive integers satisfiy the recursion $t_{n+1} = \sum_{i=0}^{n} t_i t_{n-i}$. Then $t_n = \frac{1}{n+1}\binom{2n}{n}$.*

Proof: First the generating function for the sequence is found.
CLAIM: The generating function for the sequence (t_i)

$$t(x) = \frac{1 - \sqrt{1-4x}}{2x}. \tag{1.38}$$

PROOF OF CLAIM: Put

$$t(x) = \sum_{i \geq 0} t_i x^i = 1 + x + 2x^2 + 5x^3 + 14x^4 + \cdots .$$

Then

$$\begin{aligned}(t(x))^2 &= \sum_{n \geq 0} \left(\sum_{i=0}^{n} t_i t_{n-i} \right) x^n \\ &= \sum_{n \geq 0} t_{n+1} x^n \qquad\qquad (by (1.37)) \\ &= \frac{1 - t(x)}{x},\end{aligned}$$

and so

$$x(t(x))^2 - t(x) + 1 = 0.$$

By the quadratic formula,

$$t(x) = \frac{1 \pm \sqrt{1-4x}}{2x}. \tag{1.39}$$

Using Newton's generalization of the binomial theorem

$$(1-4x)^{1/2} = 1 + \frac{1}{2}(-4x) + \frac{\frac{1}{2}(\frac{1}{2}-1)}{2!}(-4x)^2 + \cdots .$$

If the plus sign in (1.39) is used, $t(x) = \frac{1}{x} - 1 + \cdots$, which is impossible (see [754, p. 385] for details), so the minus sign is used, thereby finishing the proof of the the claim.

Since

$$\begin{aligned} t(x) &= \frac{1-\sqrt{1-4x}}{2x} \\ &= \frac{1}{2x}\left(1-\sum_{n\geq 0}\binom{\frac{1}{2}}{n}(-4x)^n\right) \\ &= \frac{1}{2x}\left(1-1-\sum_{n\geq 1}\binom{\frac{1}{2}}{n}(-4x)^n\right) \\ &= \frac{-1}{2x}\sum_{n\geq 1}\binom{\frac{1}{2}}{n}(-4x)^n \\ &= \frac{-1}{2}\sum_{n\geq 1}\binom{\frac{1}{2}}{n}(-4)^{n-1}x^{n-1}. \end{aligned}$$

For $n \geq 1$, the coefficient of x^n in above and in $t(x) = \sum_{i\geq 0} t_i x^i$ gives

$$\begin{aligned} t_n &= \frac{-1}{2}\binom{\frac{1}{2}}{n+1}(-4)^n \\ &= \frac{-1}{2}\frac{(\frac{1}{2})\cdot(\frac{1}{2}-1)(\frac{1}{2}-2)(\frac{1}{2}-3)\cdots(\frac{1}{2}-n)}{(n+1)!!}(-4)^n \\ &= \frac{(\frac{1}{2}-1)(\frac{1}{2}-2)(\frac{1}{2}-3)\cdots(\frac{1}{2}-n)}{(n+1)!}(-4)^n \\ &= \frac{(\frac{-1}{2})(\frac{-3}{2})(\frac{-5}{2})\cdots(\frac{1-2n}{2})}{(n+1)!}(-4)^n \\ &= \frac{(1)(3)(5)\cdots(2n-1)}{(n+1)!}\cdot 2^n \\ &= \frac{(2n)!}{(n+1)!n!} \\ &= \frac{1}{n+1}\binom{2n}{n} \\ &= C_n. \end{aligned}$$

Summarizing,

Theorem 1.11.4 (Euler, 1761 [323]). *For $n \geq 1$, the number of ways to triangulate a convex $(n+2)$-gon is the nth Catalan number*

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

Arbitrary (simple) polygons, not just convex polygons, can also be triangulated; to this end, the following is useful.

Lemma 1.11.5. *Every simple polygon (convex or not) has at least one diagonal lying completely inside the polygon.*

Proof: In any polygon, there exists at least one triple of vertices v_1, v_2, v_3 (taken in counterclockwise order) so that the interior angle at v_2 is less than π. Fix such a triple. If segment $\overline{v_1v_3}$ lies entirely inside P, then $\overline{v_1v_3}$ is a diagonal as desired. So suppose that $\overline{v_1v_3}$ is not entirely contained in P. Then $\triangle v_1v_2v_3$ contains additional points of P. Of these points, choose the one w so that $m\angle v_1v_2w$ is smallest. Since w is inside the triangle, w does not lie on the ray $\overrightarrow{v_1v_2}$, so w is visible from v_2, and so $\overline{v_2w}$ is the desired diagonal. □

Exercise 46. *Using the result in Lemma 1.11.5, prove that any simple n-gon can be triangulated with diagonals between vertices that lie inside the n-gon, producing $n-2$ triangles.*

Exercise 47. *Prove that in any triangulation of a simple polygon (using only vertices of the polygon), there exists at least one triangle with two sides forming edges of the polygon.*

Exercise 48. *Prove that the vertices of a triangulated n-gon can be coloured with three colours so that no two vertices of the same colour are connected by an edge.*

Exercise 49. *Suppose that $n \geq 1$ points are in the interior of some square. Prove that the square can be divided into $2n+2$ triangles with vertices chosen from the n given points and the four vertices of the square.*

Exercise 50. *Suppose that 7 points are placed inside a unit square. Show that there exists a triangle with area at most $\frac{1}{16}$, whose vertices are chosen from these 7 and the square's vertices.*

Finding placements of points in a square that maximize the area of a smallest triangle is called "Heilbronn's problem", which is discussed in Section 8.8.

Exercise 51. *Prove that if a polygon P is convex and contained in the polygon Q, the perimeter of P is shorter than the perimeter of Q.*

1.11.2 Interior angles

The following has a simple proof by mathematical induction. (A more general result is asked for in Exercise 46.)

Lemma 1.11.6. *A convex n-gon can be partitioned into $n-2$ triangles using diagonals of the n-gon.*

Since the sum of angles in a triangle is π, a useful result follows:

Corollary 1.11.7. *The sum of all interior angles of a convex n-gon is $(n-2)\pi$.*

1.11.3 Regular polygons

A polygon is called *regular* if and only if all interior angles are the same and all edge lengths are the same. From Corollary 1.11.7, it follows that the interior angle of the regular n-gon is $\frac{(n-2)\pi}{n}$. For example, the interior angle of a regular pentagon is $\frac{3}{5}\pi$, which is $108°$.

1.11.4 Pick's theorem

For present purposes, a *lattice point* is a point $(x, y) \in \mathbb{R}^2$ in the real cartesian plane whose coordinates x, y are integers. In other words, a lattice point is an element of $\mathbb{Z}^2$.

To calculate the area of an arbitrary polygon might be very cumbersome, however if the polygon has vertices that are lattice points, then finding its area is nearly trivial by the noted 1899 result of Georg Alexander Pick (1859–1942) [724]. For a simple (non-intersecting) polygon P on lattice points, let $I(P)$ be the number of lattice points on the interior of P, and let $B(P)$ be the number of lattice points occurring on the boundary of P.

Theorem 1.11.8 (Pick's theorem, 1899 [724])**.** *Let P be a polygon whose vertices are lattice points. Then the area of P is*

$$A(P) = I(P) + \frac{B(P)}{2} - 1. \tag{1.40}$$

Pick's theorem is discussed in many popular sources, *e.g.*, [42, p. 17], [220], [373, p. 215], [431], [687], [826, pp. 96–98], and [918, pp. 83–4]. Pick's theorem also occurs as an exercise in many textbooks; *e.g.*, [762, 19, p. 292] contains a solution.

Proof outline: Pick's theorem has many proofs, but (at least) one is by induction, taking into account the following cases: (i) when P is a simple rectangle with sides parallel to the axes; (ii) when P is a right triangle with two legs parallel to the axes; (iii) when P is any triangle (by first surrounding P with a rectangle, then subtracting the area of the outside right triangles and rectangles thereby formed); (iv) when P is an arbitrary simple n-gon by induction on n (using Lemma 1.11.5, either by splitting the polygon into two pieces or by adjoining a triangle).

See Figure 1.6 for the idea behind step (iii); to measure the area of triangle ABC, calculate the area of the rectangle $BDEF$ and subtract off the area of the three outer triangles, each being a right triangle. □

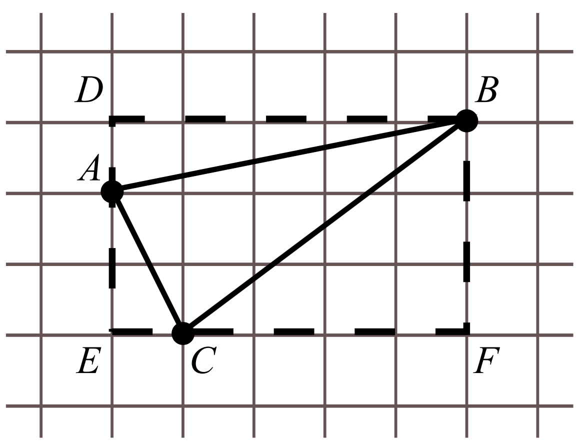


Figure 1.6: Finding area of arbitrary triangle

Example 1.11.9. *In Figure 1.7, one can verify that the area is 24 by finding 16 squares inside and the rest is a collection of simple right angle triangles. On the other hand, there are 17 interior points (yellow) and 16 border points (black), and* $17 + \frac{16}{2} - 1 = 24$*, so equation (1.40) is verified.*

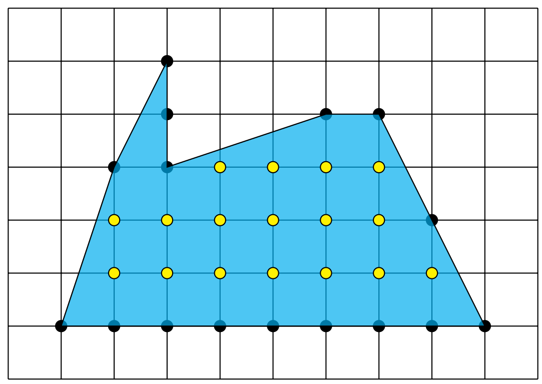


Figure 1.7: Pick's theorem: $I(P) = 17$, $B(P) = 16$; area $=24$.

1.12 Circles

1.12.1 Basics

A *circle* in $\mathbb{R}^2$ is defined to be the set (or locus) of points equidistant from some given point called the center of the circle. The distance between the center and points on the circle is called its *radius*, usually denoted by r.

Lemma 1.12.1 (Subtended angles)**.** *Let C be a circle with center O, and diameter AB. Let X be a point on the circle. Then $m\angle OAX = \frac{1}{2} \cdot m\angle XOB$.*

Proof: Since $|OA| = |OX|$, the angles $\angle OAX$ and $\angle AXO$ agree, and so $m\angle XOB = \pi - m\angle AOX = 2m\angle XAO = 2m\angle AXO$. □

Corollary 1.12.2. *Any two inscribed angles in a circle subtending the same arc are equal (to half of the angle of the arc).*

Proof: Let C be a circle with center O and let P and Q be points on C determining an arc (of size less than π). Let X be another point on C outside of the arc PQ, and let Y be the point antipodal to X (so XY is a diameter). There are two cases:

If Y is on the arc PQ, then

$$\begin{aligned} m\angle PXQ &= m\angle PXY + m\angle YXQ \\ &= \frac{1}{2}m\angle POY + \frac{1}{2}m\angle YOQ \qquad \text{(by Lemma 1.12.1)} \\ &= \frac{1}{2}(m\angle POY + m\angle YOQ) \\ &= \frac{1}{2}m\angle POQ. \end{aligned}$$

If Y is outside of the arc PQ, say, closer to P, then

$$\begin{aligned} m\angle PXQ &= m\angle PXQ - m\angle QXY \\ &= \frac{1}{2}m\angle POY - \frac{1}{2}m\angle QOY \qquad \text{(by Lemma 1.12.1)} \\ &= \frac{1}{2}(m\angle POY - m\angle QOY) \\ &= \frac{1}{2}m\angle POQ. \end{aligned}$$

So for a particular P and Q, all angles subtending PQ are the same. $\square$

One direction in the following lemma follows from Corollary 1.12.2, but both directions can be proved directly using Pythagoras and coordinate geometry (see Exercise 86).

Lemma 1.12.3. *An angle subscribed in a semi-circle is a right angle. In other words, if AB is a diameter of a circle, and X is another point on the circle, then $m\angle AXB = \pi/2$. Also, if $\triangle AXB$ is a right angled triangle with $m\angle AXB = \pi/2$, then A, X, and B lie on a circle with AB as a diameter.*

A consequence of Lemma 1.12.3 is that if a convex quadrilateral has opposite corners with right angles, then all four points of the quadrilateral lie on the same circle (with diameter formed by the other two points). Recall from Definition 1.9.4 that quadrilateral whose points lie on a circle is a called a *cyclic quadrilateral*. (See Lemma 1.9.5 and Theorems 1.9.6 and 1.9.7 for more on cyclic quadrilaterals. For an open problem in Ramsey theory for cyclic quadrilaterals, see Conjecture 18.2.16.)

Lemma 1.12.4. *If C_1, C_2, and C_3 are circles in the plane with non-collinear centers, then there is a most one point where all three circles intersect.*

Proof: Suppose that C_1, C_2, C_3 share two points, say X and Y. Then XY is a chord of each circle, and the center of a circle is on a line that is a perpendicular bisector of any chord, so the three centers of the circles are collinear. $\square$

Exercise 52. *Show that if at least four points are in $\mathbb{R}^3$, not all on one plane, then there exists a circle containing exactly three of the points.*

1.12.2 Power of a point

Theorem 1.12.5 (Power of an inside point)**.** *Let A,B,C,D be points in order on a circle. If the intersection of the two chords AC and BD is a point P inside the circle, then $|AP| \cdot |PC| = |BP| \cdot |PD|$.*

Proof: Since both $\angle CAD$ and $\angle CBD$ subtend the same arc (namely, CD), by Corollary 1.12.2, these angles agree. Similarly, $\angle BCA$ and $\angle BDA$ also agree. So triangles APD and BPC are similar. Since corresponding sides of similar triangles are proportional,

$$\frac{|AP|}{|DP|} = \frac{|PB|}{|PC|},$$

and cross multiplying finishes the proof. $\square$

Theorem 1.12.6 (Power of an outside point)**.** *Let A,B,C,D be points in order on a circle. If the lines containing the chords AB and CD intersect in a point P outside the circle, then $|AP| \cdot |BP| = |CP| \cdot |PD|$.*

1.12.3 Miquel's theorems

The next theorem is due to the French mathematician Auguste Miquel (1816–1851), a theorem that does not seem to appear in many geometry books.

Theorem 1.12.7 (Miquel, 1838)**.** *Let X, Y, Z be vertices of a triangle, let X', Y', Z' be arbitrary points on $\overline{YZ}, \overline{ZX}, \overline{XY}$ respectively. Let C_1 be the circumcircle of $\triangle XY'Z'$; let C_2 be the circumcircle of $\triangle X'YZ'$, and let C_3 be the circumcircle of $\triangle X'Y'Z$. Then C_1, C_2, and C_3 share a common point.*

See [432] for a short proof of Theorem 1.12.7.

The next theorem is sometimes called *Miquel's six circle theorem* (or his four circle theorem); see [715, p. 424] for one proof using algebraic geometry. See [694, p. 352] for some history of this theorem; apparently, this theorem is due to Jakob Steiner, but was first published by Miquel.

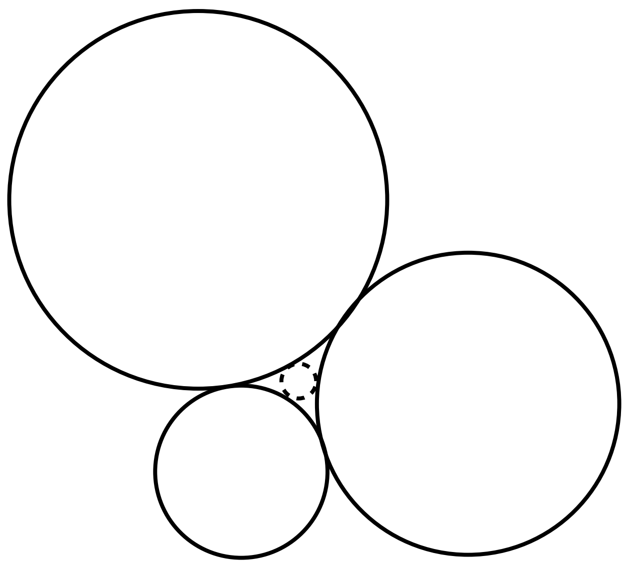
Theorem 1.12.8. *Let C_1, C_2, C_3, C_4 be circles in the plane $\mathbb{E}^2$ so that*

- *C_1 intersects C_2 in points P_1 and Q_1;*
- *C_2 intersects C_3 in points P_2 and Q_2;*
- *C_3 intersects C_4 in points P_3 and Q_3;*
- *C_4 intersects C_1 in points P_4 and Q_4.*

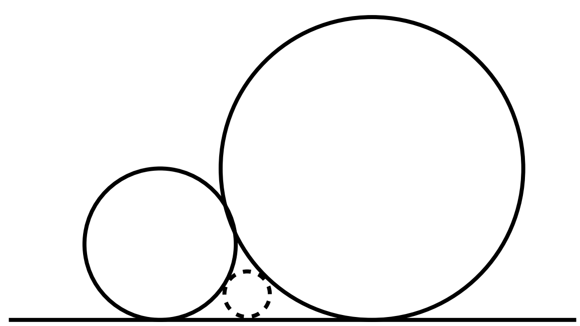
If P_1, P_2, P_3, P_4 lie on a circle, then also Q_1, Q_2, Q_3, Q_4 lie on a circle.

1.12.4 Apollonius circles

If three circles of arbitrary radii are drawn so the each pair of circles is tangent, how does one find a fourth circle between the three circles and tangent to all three? In the following diagram, the sought after fourth circle is dashed.



Can one find a fourth circle surrounding the first three and tangent to each? How does one find a third circle that is tangent to two circles and a line (as in the following diagram)? This 2-circle-1-line problem can be seen as a special case of the three circle problem, where one circle as infinite radius.



Apollonius of Perga (Greek, 3rd century BC) studied these and related problems (but there are no historical records of his conclusions or proofs).

For details regarding the history of such problems, see [628], the article on which this section is based.

It is not known if Apollonius solved the three circle problem, but in 1643, Descartes found the radius of the fourth circle (although the proof he provided was incorrect).

Theorem 1.12.9 (Descartes, 1643). *Let three circles of radii r_1, r_2, r_3 be pairwise tangent. If r_4 is the radius of a fourth circle touching the three, then*

$$\frac{1}{r_1^2} + \frac{1}{r_2^2} + \frac{1}{r_3^2} + \frac{1}{r_4^2} = \frac{1}{2}\left(\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} + \frac{1}{r_4}\right)^2.$$

According to [628], in the 1990s, Allan Wilkes and Colin Mallows of AT&T used the equation above (and complex numbers) to also compute the center of the missing circle in the three-circle problem, not just the radius.

See [628] for "Apollonious gaskets", patterns made by consecutively filling in tangent circles. Apollonious also wrote on conics and pursuit problems; however, this volume does not "persue" these ideas.

1.12.5 The nine-point circle and Feuerbach's theorem

This section is based on the exposition found in Kay's *College geometry* [532, pp. 17–19]. Please see this reference for proofs and more details (the same variables are used here).

To each triangle, there is a circle associated with the triangle containing nine particular points. The following theorem was apparently known by Brianchon and Poncelet in 1821 (sorry, I don't have the references), but also appears in Feuerbach's 1922 book [333], which also includes a further remarkable result.

Theorem 1.12.10 (Nine-point circle). *Let $\triangle ABC$ have orthocenter (the intersection of the altitudes) H. Then the following nine points lie on one circle (see Figure 1.8):*

(a) The three midpoints L, M, N of the sides of the triangle;
(b) The feet D, E, F of the altitudes on the sides;
(c) The midpoints X, Y, Z of the segments HA, HB, and HC.

The following are properties of the nine-point circle (again, see [532] for details) for $\triangle ABC$:

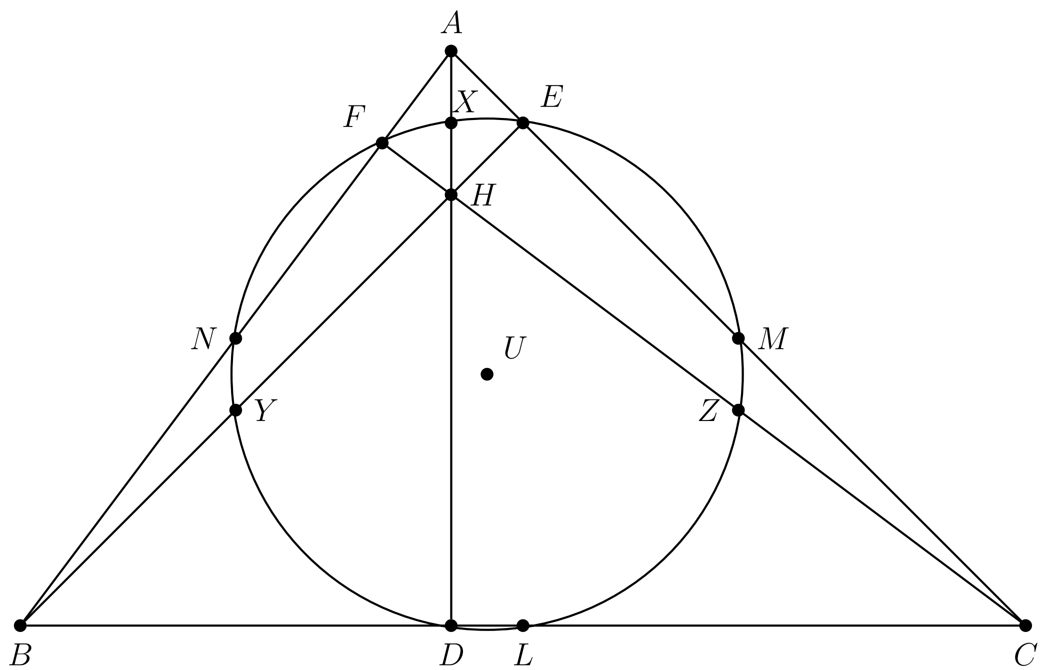


Figure 1.8: The nine-point circle for triangle $\triangle ABC$

- The nine-point circle has center U (see Figure 1.8), which is the midpoint of MY or NZ.
- The radius of the nine-point circle is half of the radius of the circumcircle of $\triangle ABC$.
- If O is the center of the circumcircle, H is the orthocenter (intersection of altitudes), and G is the centroid (intersection of medians), then O, H, and G lie on a line (called the *Euler line*). The point U (the center of the nine-point circle) also lies on the Euler line.

One more property of the nine-point circle is called Feuerbach's theorem, due to Karl Wilhelm Feuerbach (1800–1834). (See [716] for more on Feuerbach.) To each triangle $\triangle ABC$, there corresponds four circles, called "tri-tangent" circles, each of which is tangent to all three of $\overleftrightarrow{AB}$, $\overleftrightarrow{BC}$ and $\overleftrightarrow{AC}$ (as in Figure 1.9).

Theorem 1.12.11 (Feuerbach, 1822 [333])**.** *The nine-point circle for a triangle is tangent to all four tri-tangent circles.*

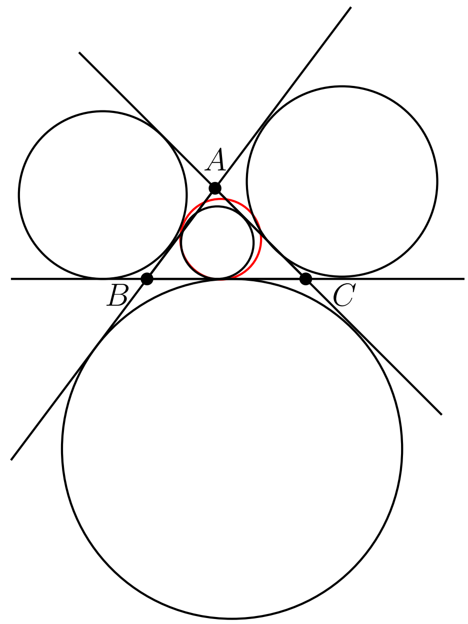


Figure 1.9: Feuerbach's theorem, nine-point circle in red

1.13 Isometries and symmetries

1.13.1 Basic definitions

Definition 1.13.1. An *isometry* between subsets X and Y in a metric space is a distance preserving function $f : X \to Y$. An isometry is also called a *rigid motion*.

Since an isometry is distance preserving, an isometry is one-to-one (injective). For a metric space S, is every isometry $f : S \to S$ surjective?

Not always. For example, look at the infinite dimensional real space X whose elements are sequences $\mathbf{x} = (x_1, x_2, x_3, \ldots)$ satisfying $\sum_{i=1}^{\infty} x_i^2 < \infty$. Two sequences $(x_1, x_2, \ldots)$ and $(y_1, y_2, \ldots)$ have distance defined by $\left(\sum_{i=1}^{\infty}(y_i - x_i)^2\right)^{1/2}$. The function $f : X \to X$ defined by $f((x_1, x_2, \ldots)) = (0, x_1, x_2, \ldots)$ is distance preserving, and so is an isometry, but not a surjection. For a second example, consider the (discrete) metric space on $\mathbb{Z}$, where the distance metric is defined by $d(x, x) = 0$ and for $x \neq y$, $d(x, y) = 1$.

Then the function $f : \mathbb{Z} \to \mathbb{Z}$ defined by $f(k) = 2k$ is an isometry, but is not surjective (it misses all odd integers).

The following theorem can be found as an exercise in [274, p. 314], where suggested steps and a reference to an early statement of the theorem is given; equivalent theorems appear in other topology texts, as well.

Theorem 1.13.2 (Freudenthal–Hurewicz, 1936 [352])**.** *If S is a compact metric space, then every isometry on S is surjective, and hence bijective.*

Both $\mathbb{E}^2$ and $\mathbb{E}^3$ are compact, and so any isometry of the plane is bijective. For the Euclidean plane, a geometric argument (that doesn't invoke topology) also shows that an isometry on the plane is bijective.

Theorem 1.13.3. *Let $f : \mathbb{E}^2 \to \mathbb{E}^2$ be an isometry. Then f is also onto, and so is a bijection.*

Proof: The proof given here appears in [890]; I do not know its origin. For each $P \in \mathbb{E}^2$, it remains to show the existence of some $X \in \mathbb{E}^2$ with $f(X) = P$. In fact, the following proof says how to find X. The idea is to create four circles, three of which are forced to contain the image of some particular point X (identified partway through the proof); since P also belongs to these three circles, the image of the new-found X is indeed P.

Let $Q = f(P)$, and let $|PQ| = r$. If $r = 0$, then $f(P) = P$, in which case, P is in the range of f, so let $r > 0$. Let C_1 be the circle centered at P with radius r (so it contains Q). Let $R = f(Q) = f^2(P)$. By distance preserving, $|PQ| = |f(P)f(Q)| = |QR|$, so R lies on a circle centered at Q with radius r; call this circle C_2. If $R = P$, then $f(Q) = P$, putting P in the range of f, so assume $R \neq P$.

The circles C_1 and C_2 intersect in two points; let $A \in C_1 \cap C_2$ be the one closest to R. If $R = A$, since $\triangle PQA$ is equilateral, and f is angle preserving, then $f(A) = P$, again putting P in the range of f. So assume that $R \neq A$. Let $|PR| = s$, and let C_3 be the circle centered at R with radius s (so it passes through P).

Let $B = f(A)$. Since f is distance preserving, $r = |PA| = |f(P)f(A)| = |QB|$, B is on C_2. Since $\triangle APQ$ is equilateral, so is $\triangle f(A)f(P)f(Q) = \triangle BQR$, also with side length r. (Since $|QA| = r$, $|f(Q)f(A)| = |RB| = r$.) If $B = P$, then $f(A) = P$, showing that P is in the range of f, so suppose that $B \neq P$.

Let X be the point on C_1 so that $\triangle QAX$ is congruent with $\triangle RBP$, with $\angle QAX = \angle RBP$. This can be done by first picking X on C_1 so that $|XA| = |PB|$. Since f preserves angles, $\angle XAQ = \angle PBR$. Since $r = |AQ| = |BR|$, by SAS, the triangles $\triangle XAQ$ and $\triangle PBR$ are congruent, and so $|XQ| = |PR|$. By distance preserving, $|f(X)R| = |XQ|$, which is $|PR|$, and so $f(X)$ lies on C_3 (which also contains P).

By distance preserving, $|AX| = |Bf(X)|$, and so $f(X)$ is on a circle (call it C_4) of radius $|AX| = |PB|$ centered at B (which also contains P).

Since $r = |XP| = |f(X)f(P)| = |f(X)Q|$, it follows that $f(X)$ is also on C_2 (which also contains P).

So $f(X)$ is on each of circles C_2, C_3, and C_4. The original point P is also on these three circles, and since these three circles do not have collinear centers (why?), by Lemma 1.12.4, $f(X) = P$. $\square$

Since isometries preserve distance, isometries preserve shapes of triangles, and hence preserves angles. An isometry from a vector space (with a dot product) to itself also preserve dot products (using $\mathbf{u} \bullet \mathbf{v} = \|\mathbf{u}\| \cdot \|\mathbf{v}\| \cos\theta$). It is also not difficult to see that isometries preserve collinearity and "in-betweenness".

If there is a bijective isometry between two objects X and Y in a metric space, X and Y are said to be *congruent* (or *isometric*); it is not difficult to verify that this congruence relation is an equivalence relation.

The inverse of a bijective isometry is still an isometry, and the composition of isometries is again an isometry. Recall that composition of functions is also associative.

Definition 1.13.4. For a set $X \subseteq \mathbb{E}^2$, a *symmetry* of X is a bijective isometry from X to X.

Definition 1.13.5. A *group* is a pair $(G, *)$, where G is a set and $+$ is a binary operator on G satisfying four axioms:

(i) G is closed under $*$ (i.e., $\forall g, h \in G, g * h \in G$).

(ii) The operation $*$ is associative (i.e., $\forall a, b, c \in G$, $(a * b) * c = a * (b * c)$).

(iii) There exists an identity $e \in G$ (so that $\forall g \in G, g * e = e * g = g$).

(iv) G contains inverses ($\forall g \in G \exists h \in G$ so that $gh = hg = e$).

Groups are often written as either additive groups or multiplicative groups (where the binary operation is viewed as addition or multiplication). A group

$(G, *)$ is called *abelian* if and only if for every $g, h \in G$, $g*h = h*g$. Common examples of abelian groups are $(\mathbb{Z}, +)$ or $(\mathbb{R}\backslash\{0\}, \cdot)$. The *symmetric group* S_n is the collection of all permutations of some set of n elements, where the operation is composition. The *alternating group* A_n is the subgroup of S_n containing only even permutations (an even permutation is a product of an even number of transpositions).

The next lemma follows from observations already made above.

Lemma 1.13.6. *For any set $X \subseteq \mathbb{R}^2$, the collection of all symmetries of X is a group, denoted Sym(X).*

For example, if X is a set of 3 points forming an equilateral triangle, then Sym(X) is S_3, the permutation group on 3 elements. Similarly, a regular tetrahedron has S_4 as its symmetry group. A regular n-gon has symmetries that form what is called the *dihedral group*, generated by two elements g, h where g has order n (corresponding to a rotation by one point) and h has order 2 (corresponding to a reflection about the central vertical axis). A theorem, sometimes referred to as "Leonardo's theorem" (see [641, p. 67] or [922]) says that any finite group of isometries in the plane is either a cyclic group or a dihedral group. A less trivial example is the truncated icosahedron, (the Archimedean solid with soccer-ball shape; see Figure 3.6) which has A_5 as its symmetric group.

Since the focus of this section is only to identify symmetries in the plane, a more complete discussion about groups and symmetries is not given here. For more on groups of symmetries and patterns in the plane, see [431] or [641]. Much of the work surrounding groups of symmetries of polyhedra seems to have been accomplished by chemists studying crystallography (see, e.g., [38], [210], [793, Ch. 10], or many of the works by Coxeter).

1.13.2 Symmetries as matrix transformations

Common symmetries of (or in) the plane are translations, rotations, and reflections, and compositions thereof. A *glide reflection*, is the composition of a translation and a reflection (both along the same line). If X is finite or bounded, then translations and glide reflections are not included in symmetries of X. Symmetries can be classified by matrix computations.

Definition 1.13.7. Let V and W be (real) vector spaces. A function $f : V \to W$ is called a *linear transformation* if and only if for any two vectors

$\mathbf{u}, \mathbf{v} \in V$ and any scalar $k \in \mathbb{R}$,

$$f(\mathbf{u}+\mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v}), \quad \text{and} \quad f(k\mathbf{u}) = kf(\mathbf{u}).$$

The following is a standard result in an early course on linear algebra.

Theorem 1.13.8. *For an integer $n \geq 2$, a linear transformation in $\mathbb{E}^n$ is a matrix transformation. In other words, if $T : \mathbb{E}^n \to \mathbb{E}^n$ is linear, then there exists a $n \times n$ (real) matrix A so that for every $\mathbf{v} \in \mathbb{E}^n$ (written as a column vector),*

$$T(\mathbf{v}) = A\mathbf{v},$$

and every such matrix transformation is a linear transformation.

Proof: For convenience, the proof given here is only for $n = 2$; the general proof is analogous. Let $T : \mathbb{E}^2 \to \mathbb{E}^2$ be linear, and consider the images of the two standard unit vectors. If $T\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} a \\ b \end{bmatrix}$ and $T\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} c \\ d \end{bmatrix}$, then for any $\mathbf{u} = \begin{bmatrix} x \\ y \end{bmatrix}$,

$$\begin{aligned} T(\mathbf{u}) &= T\left(x\begin{bmatrix} 1 \\ 0 \end{bmatrix} + y\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \\ &= xT\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) + yT\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \\ &= x\begin{bmatrix} a \\ b \end{bmatrix} + y\begin{bmatrix} c \\ d \end{bmatrix} \\ &= \begin{bmatrix} xa+yc \\ xb+yd \end{bmatrix} \\ &= \begin{bmatrix} a & c \\ b & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}, \end{aligned}$$

and so T is a matrix transformation. Proving that a matrix transformation is linear is a direct verification of the two equalities required by a linear transformation (see Definition 1.13.7). □

The matrix for a linear transformation T is called the *standard matrix* for T (sometimes denoted $[T]$). The reader can verify that a linear transformation with standard matrix A is invertible if and only if A is invertible (in which case the transformation is injective).

Theorem 1.13.9. *Let θ be any angle. In $\mathbb{E}^2$, a rotation counterclockwise by θ about the origin is a linear transformation with matrix*

$$A = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}.$$

Proof: Let R_θ denote the rotation. It suffices to see where a point on the unit circle goes (other points are some scalar multiple of a unit circle point). Let $\mathbf{u}$ be a point on the unit circle, say for some angle γ, put $\mathbf{u} = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\gamma \\ \sin\gamma \end{pmatrix}$. Then by equalities (1.5) and (1.4),

$$\begin{aligned} R_\theta(\mathbf{u}) &= \begin{pmatrix} \cos(\gamma+\theta) \\ \sin(\gamma+\theta) \end{pmatrix} \\ &= \begin{pmatrix} \cos(\gamma)\cos(\theta) - \sin(\gamma)\sin(\theta) \\ \cos(\gamma)\sin(\theta) + \sin(\gamma)\cos(\theta) \end{pmatrix} \\ &= \begin{pmatrix} x\cos(\theta) - y\sin(\theta) \\ x\sin(\theta) + y\cos(\theta) \end{pmatrix} \\ &= \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \end{aligned}$$

finishing the proof. □

If a transformation moves the origin, then it is not a linear transformation. For reference, some common linear transformations in the plane are in Figure 1.10.

Remark 1.13.10. *In the literature, often R_θ denotes the rotation (in counterclockwise direction) about the origin by an angle of θ and certain variations are used for the rotation about an arbitrary point. The notation ρ_ℓ is often used to indicate a reflection about a line ℓ.*

Without going into too much detail, For each $n \geq 3$, all of the techniques using matrices also apply to linear transformations in $\mathbb{R}^n$. For example, in $\mathbb{R}^3$, the matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Transformation	**Standard matrix**
The identity map.	$A = I_2$.
Rotation (ccw) by angle θ about origin.	$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$.
Reflection about the y-axis.	$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$
Reflection about the line $y = x$.	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
Reflection about the line $y = mx$.	$\frac{1}{1+m^2}\begin{bmatrix} 1-m^2 & 2m \\ 2m & m^2-1 \end{bmatrix}$
Reflection about $y = mx$, $m = \tan\theta$.	$\begin{bmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{bmatrix}$
Reflection through the origin.	$\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$
Dilation with expansion factor k.	$\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}$
Projection onto the y-axis.	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$
Projection onto a line $y = mx$	$\frac{1}{1+m^2}\begin{bmatrix} 1 & m \\ m & m^2 \end{bmatrix}$
Shear in the x direction by factor of k.	$\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$

Figure 1.10: Some linear transformations with associated standard matrices

corresponds to a dilation in the z direction and

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

corresponds to a rotation about the x-axis.

The reader might also note that an eigenvector for a particular matrix A can correspond to an axis of rotation.

Definition 1.13.11. A transformation $f : \mathbb{E}^2 \to \mathbb{E}^2$ is called an *affine transformation* if and only if there exists a (translation vector) $\mathbf{t} \in \mathbb{E}^2$ and a 2×2 matrix $A = A_f$ with real entries so that for every $\mathbf{v} \in \mathbb{E}^2$,

$$f(\mathbf{v}) = \mathbf{t} + A\mathbf{v}.$$

In many settings, an affine transformation in the plane is one so that for every $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, $\lambda \in [0, 1][$, $f((1-t)(\mathbf{u} + t\mathbf{v}) = (1-t)f\mathbf{u} + tf(\mathbf{v})$. In fact, it can be shown that an affine transformation is one that preserves midpoints (the proof uses continuity and dyadic fractions). Affine transformations need not preserve length or angles. An affine transformation can take circles to ellipses.

Example 1.13.12. *Let T be the transformation in $\mathbb{E}^2$ defined by reflecting about a line $y = mx + b$ (where $b \neq 0$). Then for any point $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$,*

$$T\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}\right) = \frac{1}{m^2+1}\begin{bmatrix} 1-m^2 & 2m \\ 2m & m^2-1 \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \frac{1}{m^2+1}\begin{bmatrix} -2mb \\ 2b \end{bmatrix}.$$

Thus reflecting about a line not containing the origin is not a linear transformation, but it is an affine transformation.

It may seem strange that in the Euclidean plane, rotations about the origin are linear, but reflections about any line not containing the origin are not linear (an easy check is to observe that such reflections move the origin, and hence are not linear). Rotations about a point other than the origin also move the origin:

Example 1.13.13 (Rotation about an arbitrary point)**.** *Let T be the transformation of the plane obtained by rotation about a point $(p, q) \neq (0, 0)$ by an angle θ. To compute the image of any point, first translate by subtracting the vector giving the center, rotate, then translate back. (Some call such a process a "conjugation"; algebraically, it corresponds to compositions of functions of the form $f^{-1}gf$.) Then the image of any point $\begin{bmatrix} x \\ y \end{bmatrix}$ can be computed by*

$$T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}\begin{bmatrix} x-p \\ y-q \end{bmatrix} + \begin{bmatrix} p \\ q \end{bmatrix}.$$

Written in the form $T(\mathbf{x}) = A\mathbf{x} + \mathbf{t}$, the above expression becomes

$$T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} -p\cos(\theta) + q\sin(\theta) + p \\ -p\sin(\theta) - q\cos(\theta) + q \end{bmatrix}.$$

To verify with a simple example, suppose that the point $(x, y) = (3, 4)$ is rotated about the point $(p, q) = (2, 5)$ counterclockwise by $90°$. Simple geometry shows that the image is $(3, 6)$, the same numbers obtained by the above calculations. It is not difficult to verify that the expression above gives that the only point that is fixed is the center of the rotation. Furthermore, such a rotation is distance preserving, bijective, and invertible (just rotate in the opposite direction).

Another affine transformation is a *glide reflection*, a translation along a line, followed by a reflection about that same line. Glide reflections are also distance preserving.

Exercise 53. *For each of the above transformations, give a drawing demonstrating the effect of moving a letter "F" inscribed in the basic unit square. It may help to first find where the corners of the square go, and then determine the orientation of the F after.*

The proof of the following simple lemma can be verified directly, or by using matrix representations given above.

Lemma 1.13.14. *Reflections (about any line or point), rotations about any point, and translations are all distance preserving functions. Furthermore, each of these is invertible, and so are bijections.*

Of the transformations given above (or in Figure 1.10), rotations, reflections, and translations are distance preserving; those that are not distance preserving include projections, dilations, and shears. Projections are also not injective (and so not invertible).

1.13.3 Bijective distance preserving functions in $\mathbb{R}^2$

A few facts about compositions of bijective distance preserving functions (also called isometries, symmetries, or rigid motions) in $\mathbb{E}^2$ are reviewed here for reference; some proofs are left as exercises. Sorry, some original sources are not identified yet (but these all appear in popular literature). The next few results follow from simple geometry or by using matrix transformations.

Lemma 1.13.15. *Let ℓ and m be parallel lines in the plane. Let f be the composition of first a reflection in ℓ, then a reflection in m. Then f is a translation. Furthermore, any translation by a vector $\mathbf{v}$ is a composition of two reflections about any two parallel lines orthogonal to $\mathbf{v}$ separated by a distance of $\frac{1}{2}\|\mathbf{v}\|$.*

Exercise 54. *Prove Lemma 1.13.15.*

Lemma 1.13.16. *Let ℓ and m be lines in the plane that intersect where the angle from ℓ to m is θ. Then the reflection in ℓ followed by the reflection in m is a rotation of angle 2θ about the intersection point. Furthermore, any rotation is a product of reflections about two intersecting lines.*

Exercise 55. *Prove Lemma 1.13.16.*

Exercise 56. *Show that a translation in the plane followed by a rotation (by an angle not a multiple of 2π) about some point is a rotation. Similarly, show that a rotation followed by a translation is a rotation. Hint: By Lemmas 1.13.15 and 1.13.15, there is some choice for the pairs of reflecting lines that produce each translation or rotation.*

Lemma 1.13.17. *Let $f : \mathbb{E}^2 \to \mathbb{E}^2$ be an isometry. Then f preserves angles and dot products. To be specific, if θ is the angle $\angle ABC$ at B, then θ is also the angle of the angle $\angle f(A)f(B)f(C)$. For $\mathbf{u}, \mathbf{v} \in \mathbb{E}^2$, $f(\mathbf{u}) \bullet f(\mathbf{v}) = \mathbf{u} \bullet \mathbf{v}$.*

Exercise 57. *Prove Lemma 1.13.17.*

Many students have likely heard that every symmetry of the plane can be formed by a combination of translations, reflections, and rotations. In Theorems 1.13.23 and 1.13.24 below, two more precise classifications of symmetries are given.

One approach to classifying symmetries is by using well-known properties of linear transformations.

Lemma 1.13.18. *If f is an isometry of the plane satisfying $f(\mathbf{0}) = (\mathbf{0})$, then f is a linear transformation.*

Proof: Let f be an isometry of $\mathbb{E}^2$ that fixes the origin. Let $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$ and put $\mathbf{w}_1 = f(\mathbf{e}_1)$ and $\mathbf{w}_2 = f(\mathbf{e}_2)$. Since f preserves distances (and f fixes $\mathbf{0}$), $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$. By Lemma 1.13.17, f preserves angles and dot products, so $\mathbf{w}_1$ and $\mathbf{w}_2$ are orthogonal; in other words, $\{\mathbf{w}_1, \mathbf{w}_2\}$ is also an orthonormal basis for $\mathbb{E}^2$.

Claim: If $\mathbf{u} = (a, b)$, then $f(\mathbf{u}) = a\mathbf{w_1} + b\mathbf{w_2}$, and for any $k \in \mathbb{R}$, $f(k\mathbf{u}) = kf(\mathbf{u})$.

Proof of claim: Since f preserves dot products, $a = \mathbf{u} \bullet \mathbf{e}_1 = f(\mathbf{u}) \bullet \mathbf{w}_1$, which is the first coordinate of $f(\mathbf{u})$ with respect to the basis $\{\mathbf{w}_1, \mathbf{w}_2\}$. The similar result holds for b, ending the proof of the claim.

Let $\mathbf{u} = (a, b)$ and $\mathbf{v} = (c, d)$ (with respect to the standard basis). Then

$$\begin{aligned}
f(\mathbf{u}+\mathbf{v}) &= (f(\mathbf{u}+\mathbf{v}) \bullet \mathbf{w}_1)\,\mathbf{w}_1 + (f(\mathbf{u}+\mathbf{v}) \bullet \mathbf{w}_2)\,\mathbf{w}_2 \\
&= ((\mathbf{u}+\mathbf{v}) \bullet \mathbf{e}_1)\,\mathbf{w}_1 + ((\mathbf{u}+\mathbf{v}) \bullet \mathbf{e}_2)\,\mathbf{w}_2 && \text{(by Lemma 1.13.17)} \\
&= (a+c)\mathbf{w}_1 + (b+d)\mathbf{w}_2 \\
&= a\mathbf{w}_1 + b\mathbf{w}_2 + c\mathbf{w_1} + d\mathbf{w}_2 \\
&= f(\mathbf{u}) + f(\mathbf{v}) && \text{(by claim),}
\end{aligned}$$

and for any $k \in \mathbb{R}$,

$$\begin{aligned}
f(k\mathbf{u}) &= (f(k\mathbf{u}) \bullet \mathbf{w}_1)\,\mathbf{w}_1 + (f(k\mathbf{u}) \bullet \mathbf{w}_2)\,\mathbf{w}_2 \\
&= ((k\mathbf{u}) \bullet \mathbf{e}_1)\,\mathbf{w}_1 + ((k\mathbf{u}) \bullet \mathbf{e}_2)\,\mathbf{w}_2 && \text{(by Lemma 1.13.17)} \\
&= k[(\mathbf{u} \bullet \mathbf{e}_1)\mathbf{w}_1 + (\mathbf{u} \bullet \mathbf{e}_2)\mathbf{w}_2] \\
&= k[a\mathbf{w}_1 + b\mathbf{w}_2] \\
&= kf(\mathbf{u}) && \text{(by claim).}
\end{aligned}$$

So f is linear. □

Corollary 1.13.19. *For any symmetry $f : \mathbb{E}^2 \to \mathbb{E}^2$, there is a vector $\mathbf{t}$ and a 2×2 matrix A so that for any point $\mathbf{v} \in \mathbb{E}^2$ (written as a column vector),*

$$f(\mathbf{v}) = \mathbf{t} + A\mathbf{v}.$$

Corollary 1.13.20. *Any isometry of the plane is affine (and hence bijective).*

A square matrix A is called *orthogonal* if and only if $A^{-1} = A^T$. If A is orthogonal, A^TA is the identity matrix, and so columns of A are pairwise orthogonal and each column has norm 1. Also, $A^TA = I$ implies that $\det(A) = \pm 1$.

Lemma 1.13.21. *Let A be a 2×2 matrix over the reals, and let the associated linear transformation $T_A : \mathbb{E}^2 \to E^2$ be defined by, for every $\mathbf{x} \in \mathbb{E}^2$ (viewed as a column matrix), $T_A(\mathbf{x}) = A\mathbf{x}$. If T_A is distance preserving, then A is orthogonal.*

Orthogonal 2×2 matrices give rise to only two types of symmetries in the plane, as seen in the following well-known lemma.

Lemma 1.13.22. *Let A be a 2×2 orthogonal matrix (i.e., $A^T A = I_2$). Then the linear transformation $T_A : \mathbb{R}^2 \times \mathbb{R}^2$ defined by $T_A(\mathbf{u}) = A\mathbf{u}$ is either a rotation (about the origin) or a reflection (about some line through the origin).*

Proof: Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then

$$A^T A = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \cdot \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{bmatrix},$$

and so $a^2+c^2 = 1 = b^2+d^2$ and $ab+cd = 0$. Let $\mathbf{u} = \begin{pmatrix} a \\ c \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} b \\ d \end{pmatrix}$.

Since $a^2 + c^1 = 1$, let θ be so that $a = \cos\theta$ and $c = \sin\theta$. Also, since $b^2 + d^2 = 1$, let ψ be so that $b = \cos\psi$ and $d = \sin\psi$. Since $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, $\psi = \theta \pm \frac{\pi}{2}$.

Case 1: Suppose that $\psi = \theta - \frac{\pi}{2}$. Then by the equalities (1.5) and (1.4),

$$\mathbf{v} = \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} \cos(\theta - \frac{\pi}{2}) \\ \sin(\theta - \frac{\pi}{2}) \end{pmatrix} = \begin{pmatrix} \sin\theta \\ -\cos\theta \end{pmatrix}.$$

In this case, $A = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$. Using the right triangle with side lengths $2m$, m^2-1, m^2+1, and $\theta = \tan^{-1}(\frac{2m}{m^2-1})$ shows $A = \frac{1}{m^2+1}\begin{bmatrix} 1-m^2 & 2m \\ 2m & m^2-1 \end{bmatrix}$, which is a reflection matrix for the reflection about $y = mx$. Another way to see that this matrix is a reflection matrix is to observe (as noted in Figure 1.10) that for the reflection in the line $y = mx$ and $m = \tan\theta$, then the matrix is $\begin{bmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{bmatrix}$, so replace θ by 2θ.

Case 2: Suppose that $\psi = \theta + \frac{\pi}{2}$. Then by the equalities (1.5) and (1.4),

$$\mathbf{v} = \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} \cos(\theta + \frac{\pi}{2}) \\ \sin(\theta + \frac{\pi}{2}) \end{pmatrix} = \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \end{pmatrix}.$$

In this case, $A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$, which is the rotation matrix for a rotation counterclockwise about the origin by the angle θ. $\square$

In the above proof, another way to confirm that Case 1 gives rise to a reflection about some line is to find the eigenvectors of A.

Putting together Lemmas 1.13.18, 1.13.21, and 1.13.22 gives the following classification of isometries (and symmetries) of the plane.

Theorem 1.13.23. *Any isometry of the plane is a composition of a translation (including the zero translation) and one of either a reflection or a rotation.*

Since a translation is a composition of two reflections (see Lemma 1.13.15), and a rotation is a composition of two reflections (see Lemma 1.13.16), Theorem 1.13.23 says that a symmetry of the plane is the composition of at most four reflections. Another very different proof of Theorem 1.13.23 gives a result that may seem stronger than Theorem 1.13.23:

Theorem 1.13.24. *Any symmetry of the plane is the composition of either two or three reflections.*

One proof of Theorem 1.13.24 is by considering the sets that are fixed (e.g., if a single point is fixed, the isometry is a rotation about that point, and if a line is fixed (pointwise), the isometry is a reflection about that line). For proof details, see [610, p. 24ff]. Also, a theorem from [610, Thm 1.4.8] gives another classification, which is based on Theorem 1.13.24.

Theorem 1.13.25. *Any isometry of the plane is either a translation, rotation, or glide reflection.*

By Exercise 56, Theorem 1.13.25 includes compositions of a translation and a rotation.

By Corollary 1.13.20, isometries of the plane are affine transformations. Included for interest's sake is a generalization to any normed space (not necessarily with an inner product), but where the assumption of bijectivity is added (the proof is omitted).

Theorem 1.13.26 (Mazur–Ulam, 1932 [650]). *If U and V are real normed vector spaces and f is a bijective isometry from U to V, then f is affine.*

Exercise 58. *Is there a subset of the real line that is isometric to a proper subset of itself? Is there such a set that is bounded?*

Exercise 59. *Show that for each set S in the real line there is at most one point $p \in S$ so that S and $S \setminus \{p\}$ are isometric.*

Exercise 60. *Is there a bounded subset of the plane that is isometric to a proper subset of itself?*

Exercise 61. *Does there exist a bounded set $S \subseteq \mathbb{R}^2$ and a $p \in S$ so that S is isometric to $S \setminus \{p\}$?*

Exercise 62. *Let $n \geq 2$ be a positive integer, and put $\theta = 2\pi/n$. For $k = 1, \ldots, n$, define points $P_k = (k, 0)$ in the xy-plane. Let R_k be the map that rotates the plane counterclockwise by the angle θ about the point P_k. Let R denote the map obtained by applying, in order, R_1, then R_2, ..., then R_n. For an arbitrary point (x, y), find and simplify the coordinates of $R(x, y)$.*

1.14 Diameter and width of a bounded set

1.14.1 Diameter

What is "the diameter" of a set? The diameter of a circle with radius r is $2r$. It would also be fair to say that the diameter of a sphere with radius r is $2r$, as well. What is the diameter of a cube?

Definition 1.14.1. For any positive integer d and set $X \subset \mathbb{R}^d$, the *diameter of X* is

$$\operatorname{diam}(X) = \sup_{\mathbf{u},\mathbf{v} \in X} \|\mathbf{u} - \mathbf{v}\|,$$

if it exists, and if it does not, write $\operatorname{diam}(X) = \infty$.

If a set X is closed and bounded, the supremum in Definition 1.14.1 exists.

Note: It is common to say that a chord through the center of a circle is "a diameter" (see Section 16.3), but here, a diameter is a length in $[0, \infty]$ and is unique.

By definition, the diameter of the unit square is $\sqrt{2}$. For a set X with diameter D, how small of a sphere contains X?

Exercise 63. *In $\mathbb{R}^2$, show that an equilateral triangle with side length s has diameter s and is contained in a circle of radius $\frac{s}{\sqrt{3}}$.*

Theorem 1.14.2 (Jung, 1901 [513], [514])**.** *Let n be a positive integer and X be a closed and bounded set in $\mathbb{R}^n$ with diam$(X) = D$. Then X is contained in a sphere of radius*

$$r \leq D\sqrt{\frac{n}{2(n+1)}}.$$

A regular simplex (on $n+1$ points) shows that the bound given in Theorem 1.14.2 is optimal.

1.14.2 Width and the Reuleaux triangle

If S is some bounded object in two dimensions, the "width of S" is measured with respect to some direction. For some vector $\mathbf{v} \in \mathbb{R}^2$, the width of S in direction $\mathbf{v}$ is the minimum distance between parallel lines orthogonal to $\mathbf{v}$ so that the parallel lines contain S. For example, the width of the unit square can vary from 1 to $\sqrt{2}$, depending on the direction in which width is measured. Similarly, widths of higher dimensional objects can be defined.

A circle has constant width (in any direction, its width is its diameter). However, there are other convex shapes of constant width. The most famous of these is called the "Reuleaux triangle", so named after a German engineer Franz Reuleaux (1829–1905). The Reuleaux triangle is easily constructed by a compass (see Figure 1.11): Let $\triangle ABC$ be equilateral with side length s. From each of the vertices A, B, and C, construct an arc of radius s through the other two points of the triangle.

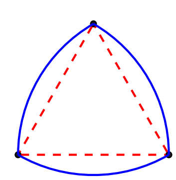


Figure 1.11: Constructing a Reuleaux triangle

The same idea also constructs "Reuleaux polygons", starting with (for n odd) a regular n-gon. Among all 2-D shapes with some constant width s, the Reuleaux triangle has the least area.

Exercise 64. *Find the area of the Reuleaux triangle with constant width s.*

The Reuleaux triangle has found many applications (some studied by, e.g., Leonardo da Vinci or Leonhard Euler). For example, the Reuleaux triangle can be used to construct drill bits (that can cut a square hole!). See Figure 1.12 for a model used to demonstrate that opposite sides of a square remain in contact with the shape.

Figure 1.12: Reuleaux triangle model, bloodwood and walnut, DSG 2017

Another famous occurrence of the Reuleaux triangle is (approximately) in the Wankel engine (found, for example, in the Mazda RX-7), famous for its capacity to endure higher RPM. Other remarkable applications produce linkages that advance motion picture film in little jumps (see [553, p. 21]) or other rotary to reciprocal paths (apparently, Cornell University Library has a collection of such linkages, although I have not researched this).

The Reuleaux triangle is an example that shows checking for constant width is not sufficient to guarantee a circular shape (which might have contributed to the Challenger disaster; see [377]).

The Reuleaux triangle also has 3-dimensional analogues. The Museum of Math in New York City has a display (see Figure 1.13) where a boat can "float" across a river of such objects in a smooth ride.

Figure 1.13: 3D shapes of constant diameter, Museum of Math, New York City, photo taken 25 Dec. 2016, dsg

Many sets of constant diameter 3D shapes are also available commercially; some are based on tetrahedra. For more on the history, mathematics, and applications of the Reuleaux triangle, see [215], [377] or [594].

Another constant width curve is found in [826, p. 151], and can be described as follows: start with an equilateral triangle $T = \triangle ABC$ and extend each of the edges an equal amount on both sides (so the three lines, if extended, divide the plane into 7 parts, the triangle and 6 others surrounding). On each of these lines, mark a point some constant distance away from the vertices of the original triangle. From each vertex, draw two arcs, one small on one side and the opposite side, one larger so that all six form a smooth curve.

1.15 Golden ratio

1.15.1 Definition and basic properties

Given a finite segment AC, where does one put an additional point B between A and C so that

$$\frac{|AC|}{|AB|} = \frac{|AB|}{|BC|}?$$

If one assumes that $|AB| = 1$, and $\tau = |AC|$, then the above expression becomes

$$\tau = \frac{1}{\tau - 1}.$$

Cross-multiplying gives $\tau^2 - \tau - 1 = 0$, and solving gives $\tau = \frac{1 \pm \sqrt{1+4}}{2}$. Since $\tau > 0$,

$$\tau = \frac{1+\sqrt{5}}{2},$$

which is approximately 1.618. The proportion τ is called *the golden ratio* (sometimes the variable ϕ is used). This ratio is also called the "golden mean", "golden section", or the "divine proportion".

There are very many results regarding the golden ratio and geometry; for example, see [234] for a collection. Some of these results are given as exercises at the end of this section.

The golden ratio has been a central topic in mathematics for centuries. In 1498 or so, Luca Pacioli finished *De divina proportione* (which was illustrated by Leonardo da Vinci). Using scribes, he made three copies (two of which still exist). In 1509, woodcuts were made and the book was printed. For a modern collection of facts regarding the golden ratio and its history, see the book [615]; only a brief introduction is given here.

The golden ratio also comes up in the worlds of art and architecture (for example, in the work of the architect Le Corbusier). However, many of the occurrences of τ are mythical (*e.g.*, in the *Mona Lisa* or the pyramids); see [640] for discussion.

1.15.2 Golden ratio and Fibonacci numbers

The golden ratio is related to the Fibonacci numbers (a fact known by Kepler (1571–1630) in 1611—see [150]), but it is not clear who first wrote about this relationship (see also [233]).

Definition 1.15.1. The Fibonacci numbers $F_0, F_1, F_2, F_3, \ldots$, are defined recursively by $F_0 = 0$, $F_1 = 1$, and for $n \geq 2$, $F_n = F_{n-1} + F_{n-2}$.

(See Section 8.11.3 for some basic properties of Fibonacci numbers. Also see the books [433] and [603] for identities using Fibonacci numbers.)

Theorem 1.15.2. *Let $F_0, F_1, F_2, \ldots$ denote the Fibonacci numbers. Then $\lim_{n\to\infty} \frac{F_{n+1}}{F_n} = \tau$, the golden ratio.*

Proof: For the moment, suppose that $\lim_{n\to\infty} \frac{F_{n+1}}{F_n}$ exists; call it L. Since $F_{n+1} = F_n + F_{n-1}$, for $n \geq 2$,

$$\frac{F_{n+1}}{F_n} = 1 + \frac{F_{n-1}}{F_n} = 1 + \frac{1}{\frac{F_n}{F_{n-1}}}.$$

Observing that $\lim_{n\to\infty} \frac{F_{n+1}}{F_n} = \lim_{n\to\infty} \frac{F_n}{F_{n-1}}$, taking limits of the previous equation gives

$$L = 1 + \frac{1}{L}$$

and so $L^2 = L + 1$. Since $L > 0$, solving this last equation shows $L = \tau$. $\square$

1.15.3 Golden rectangle and golden triangles

A rectangle whose sides have lengths in ratio τ is called a *golden rectangle*. A golden rectangle can be constructed using only a compass and straightedge as follows (see Figure 1.14, or for more details, see Section 1.16). Draw a square with vertices $(0,0),(0,2),(2,0),(2,2)$. Draw a circle centered at $(0,1)$ that passes through $(2,2)$ (and so has radius $\sqrt{5}$). This circle intersects the x-axis at $1+\sqrt{5}$. The rectangle with vertices $(0,0),(0,2),(1+\sqrt{5},0),(1+\sqrt{5},2)$ is golden.

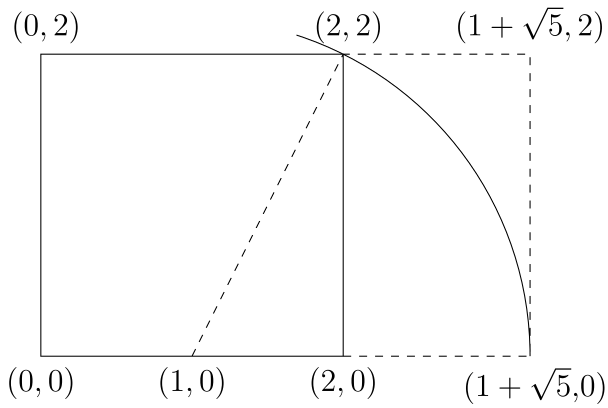


Figure 1.14: Constructing a golden rectangle

An isosceles triangle with edge lengths that are in ratio τ is called a golden triangle. There are two types of golden triangle (both isosceles—see Figure 1.15), one similar to the triangle with side lengths $\tau,\tau,1$ (an acute triangle) and the other with $\tau,1,1$ (an obtuse triangle).

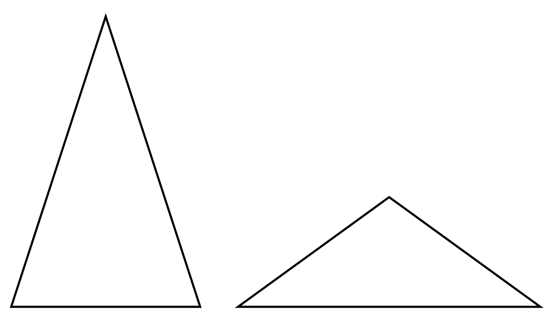


Figure 1.15: The two golden triangles

The former is often called "the golden triangle" and the second is sometimes called a "gnomon"; the pair of these triangles have been called "Robinson triangles".

The interested reader might confirm that the small angle of each golden triangle is 36° (see Exercise 66) and the five outer triangles in a pentagram (a 5-star) are (acute) golden triangles of type $\tau : \tau : 1$ (see Exercise 68). Also see Section 1.16.2 for a compass-straightedge construction of an acute golden triangle, and Section 1.16.2 for the appearance of an obtuse golden rectangle in a pentagon.

A standard exercise (see Exercise 70) shows how both golden triangles of different shape are found in one acute golden triangle.

1.15.4 Penrose tilings

A famous occurrence of golden triangles is in what are called *Penrose tiles* (studied by Sir Roger Penrose in the 1970s), only briefly introduced here. There are two Penrose tiles, each formed by joining two golden triangles (see Figure 1.16: the "kite" consists of two acute triangles joined on a long side, and the "dart" consists of two obtuse golden triangles joined on a short side.

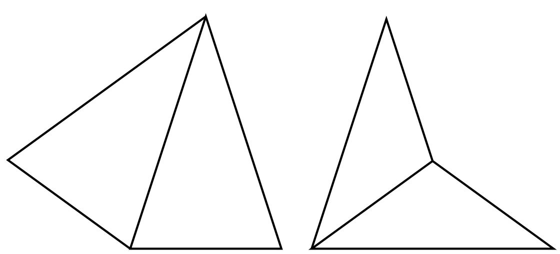


Figure 1.16: Kite and dart, the two Penrose tiles

It was proved that a set of tiles using only these two shapes can be made to tile the plane so that the resulting pattern (called a "Penrose tiling") has no translational symmetries (the pattern is non-periodic) but has 5-fold (pentagonal) symmetry and is, in a sense, self-similar to larger dilations of itself.

Observe that a kite and dart together can make up a rhombus, but the Penrose tilings forbid such subpatterns. See Figure 1.17 for an example (that is based on code I found at StackExchange by "Herbert" on 2 July 2012, available at `https://tex.stackexchange.com/questions/61437/penrose-tiling-in-tikz?`, last accessed 21 April 2018).

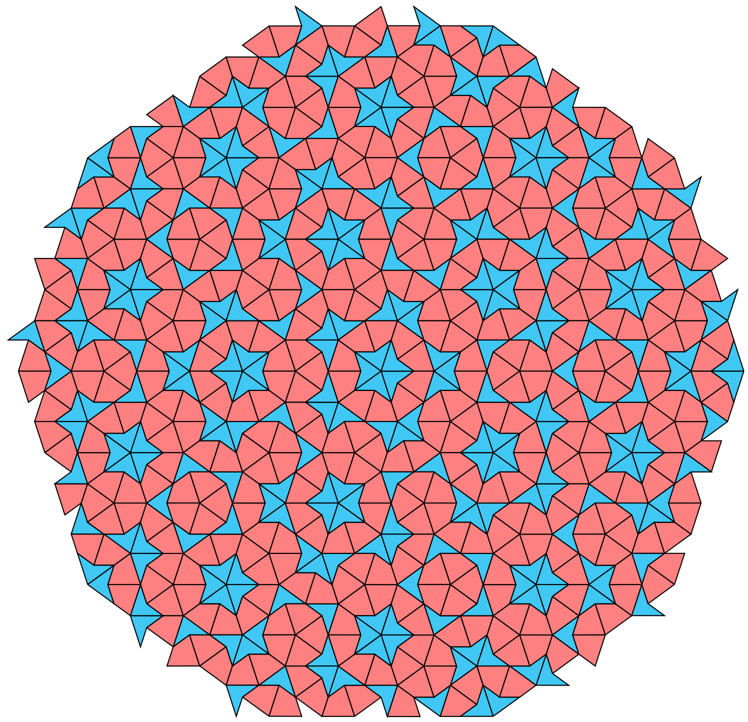


Figure 1.17: A Penrose tiling (based on code from StackExchange)

The problem of finding non-periodic tilings of the plane using only a limited number of different tiles has a long history (see [376] or [430] for references). In brief, the first set of tiles used for non-periodic tiles contained 20426 types of tile (based on a square), was found in 1964. The number of tiles was then brought down to 104, then to 92 by Donald Knuth in 1968. In 1971, this number was reduced to 6 by Raphael Robinson. In 1974, Penrose produced a set of 6 tiles based on pentagons rather than squares. Penrose finally arrived at the dart and kite, which were popularized by Gardner in 1977 (see [376]). Penrose tilings (with dart and kite) have been studied widely, and also occur in many works of art and architecture (e.g., in the entrance floor of the Andrew Wiles building at Oxford University). Using differently coloured Penrose tiles, a variety of patterns are possible (subject to rules that forbid two adjacent kites forming a rhombus); e.g., see Figure 1.18 (which is not a Penrose tiling).

Figure 1.18: Playing with Penrose tiles, model by dsg, 2003

1.15.5 Kepler triangles

A triangle is called a *Kepler triangle* if the ratio of its side lengths are in geometric progression. The only right triangles that are Kepler triangles have side lengths in ratio $1 : \sqrt{\tau} : \tau$ (see Exercise 72).

1.15.6 Golden crystal

In February 2004, I designed a polyhedron whose faces consist of three basic shapes—the golden rectangle and the two golden triangles (see Figure 1.19).

This polyhedron uses one golden rectangle (the base), two golden triangles of type $1 : \tau : \tau$ and six of type $1 : 1 : \tau$. It has bilateral symmetry, and is the simplest (fewest faces) such polyhedron (I recall proving this but I no longer have these calculations). I humbly call this polyhedron "the golden crystal".

1.15.7 Exercises with the golden ratio

Exercise 65. *Suppose that three identical golden rectangles are given, each centered at the origin* $(0, 0, 0)$*, one in each of the three planes formed by the axes. (See Figure 1.20.) Show that the twelve vertices of these three rectangles form the vertices of a regular icosahedron.*

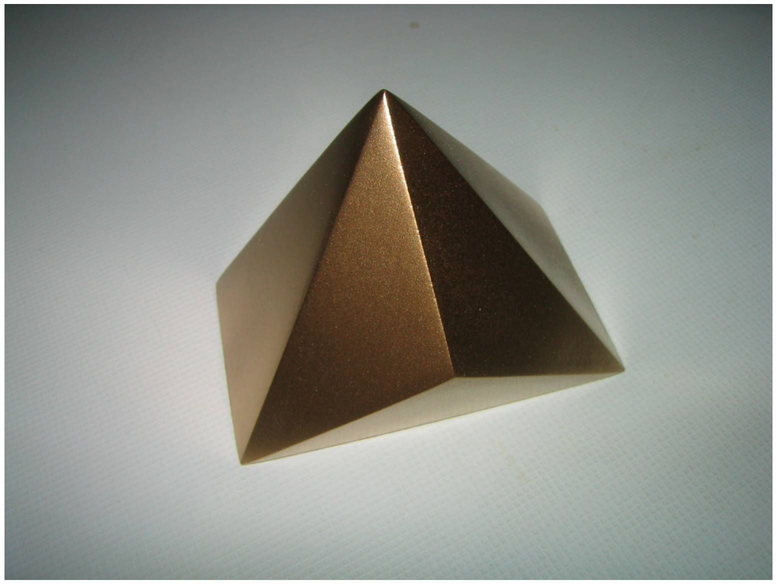


Figure 1.19: The golden crystal, by DSG, 2004

Exercise 66. *Find the angles in each of the two golden triangles.*

Exercise 67. *Show that if a regular pentagon has side lengths 1, then the length of the diagonals is the golden ratio τ, and two intersecting diagonals are cut with the same ratio.*

Exercise 68. *Show that each of the five outer triangles in a pentagram are golden.*

Exercise 69. *Show that the golden triangle with sides in ratio $\tau : \tau : 1$ is the only triangle whose angles are in the ratio 2:2:1.*

Exercise 70. *Let T be a golden triangle of type $\tau : \tau : 1$. Show that with one cut, T can be dissected into two golden triangles, one of each type.*

Exercise 71. *Let C be a disk of radius R. A smaller disk of C' radius r is removed from C so that C' is tangent to C (producing the blue shape below, sometimes called a "golden earring").*

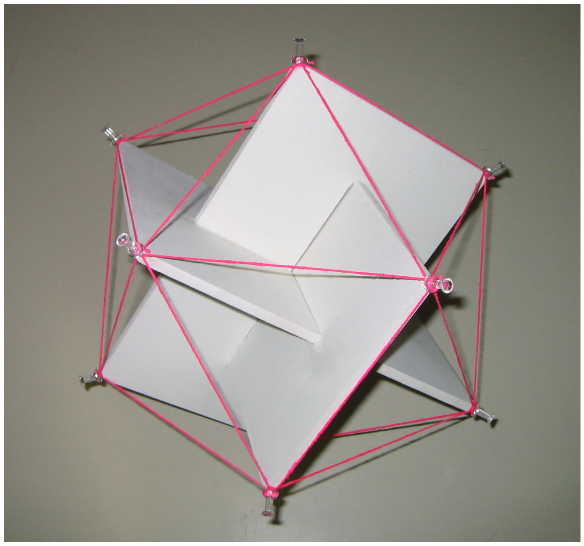
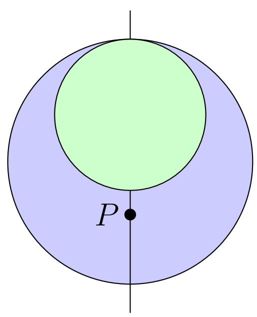


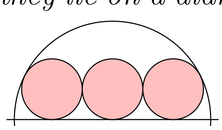
Figure 1.20: Three intersecting golden rectangles



Prove that if $R/r = \tau$, then the center of gravity P of the remaining (blue) shape occurs at the edge of the removed disk.

Another result similar to that in Exercise 71 for squares was given by Lord [621].

Exercise 72. *Show that if the sides of a right triangle are in geometric progression, then the triangle is similar to one with side lengths $1, \sqrt{\tau}, \tau$.*

Exercise 73. *Suppose that three identical circles with radius r are placed side-by-side, and they lie on a diameter of a smallest circle with radius $R > r$ containing them.*  *Show that $\frac{R}{2r} = \tau$.*

1.16 Compass and straightedge constructions

1.16.1 Introduction

In this section, a “compass and straightedge” construction is a drawing in the plane where only a compass (a device used for drawing circles, not the magnetic direction finding type—see Figure 1.21) and straightedge are used; straightedges with markings on them are not allowed.

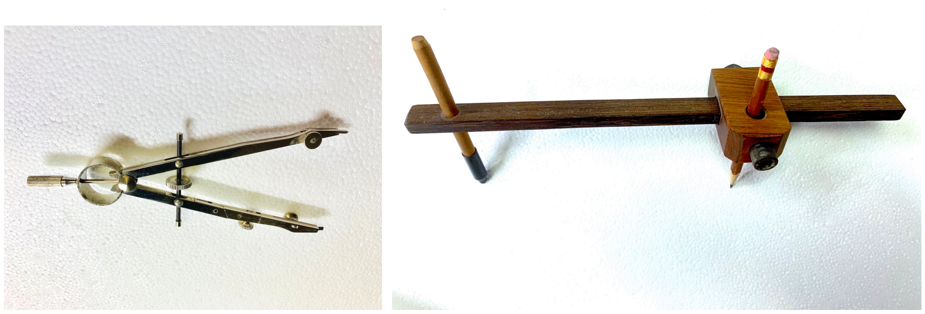


Figure 1.21: Geotec compass and homemade beam compass (kingwood)

From the time of Euclid (or earlier), compass-straightedge constructions were studied. In Section 1.16.2, many of the classical (or “standard”) compass-straightedge constructions are given. Some simple compass and straightedge constructions also appear in in other chapters. For example, constructions are given for the square root of a number (Exercise 193), the Reuleaux triangle (see Figure 1.11), and a golden rectangle (see Section 1.15.3, and Figure 1.14). For more on compass-straightedge constructions, see, e.g., [215, pp. 145–152], [265, Problems 7, 34, 35, 36], [521], [649] or [911, pp. 162–177].

The following result may seem surprising:

Theorem 1.16.1 (Mohr, 1672 [662] and Mascheroni, 1797 [643])**.** *Any compass-straightedge construction can be made by compass alone.*

(Mohr first proved this, but his proof went undiscovered until after Mascheroni had published his proof.) For more on constructions that use only compasses, see [215, pp.146–152], [479] or [571]. Coxeter [220, p. 79] also gives five exercises on compass only constructions.

Using only a compass and a straightedge (with no markings on it), surprisingly many geometric (or algebraic) constructions are possible, including finding square roots of integers, bisecting angles, or multiplying two real numbers. Again, see Section 1.16.2.

There are three problems from ancient times (circa 430 BC, Greece) for which no compass-straightedge constructions have been found:

- "Doubling a cube": construct a cube with twice the volume of any given cube (the "Delian problem");
- "Squaring a circle": construct a square with the area of any given circle;
- "Trisecting an angle": subdivide any angle into three equal angles.

Another famous problem from the ancients is:

- Decide which regular polygons have a compass-straightedge construction.

The Greeks knew how to construct regular n-gons for $n = 3, 4, 5, 6$ (proofs are in Section 1.16.2), but could not decide on the regular heptagon (7 sides). (Of course, they knew how to construct a regular $2n$-gon from an n-gon, so they could do, e.g., $n = 8, 10, 12, 16, \ldots$.)

In 1796, Carl Friedrich Gauss (1777–1855) found that a construction for the heptagon did not exist, but then found a construction for the regular 17-gon (when he was only 19). In fact, Gauss found much more. To state his result, a definition is needed. A *Fermat number* is one of the form $2^{2^m} + 1$. The first few Fermat numbers are $F_0 = 3$, $F_1 = 5$, $F_2 = 17$, $F_3 = 257$, and $F_4 = 65537$. In 1630, Fermat wondered when a number of the form $2^k + 1$ is prime, and showed that k must be a power of 2 (i.e, the number must be a Fermat number). Fermat conjectured that all Fermat numbers are prime, but in 1732, this was disproven by Euler, who showed that F_5 is divisible by 641. Recent literature indicates that the only known prime Fermat numbers (called Fermat primes) are the first five found by Fermat himself; it is not known if infinitely many Fermat primes exist.

Theorem 1.16.2 (Gauss, 1796 [381, SS365–6], Wantzel 1837 [912])**.** *For integers $n \geq 3$, a regular n-gon is compass-straightedge constructible if and only if n is of the form*

$$n = 2^r p_1 p_2 \cdots p_s,$$

where $r \geq 0$ and the p_is are distinct Fermat primes.

The sufficiency condition on n in Theorem 1.16.2 was proved by Gauss (when he was 19 years old), and he also claimed, without a complete proof that the condition is necessary; the necessity was proved by Wantzel (1814–1848). (Wantzel's proof went basically unknown for nearly a century; see [627] for a historical perspective.)

Since there are only 5 known Fermat primes, there are only 31 regular n-gons with an odd number of sides that are compass-straightedge constructible (and $n = 7$ is not a Fermat prime, so a regular heptagon is not constructible).

Getting back to one of the first three problems, note that for some particular angles trisecting is easy (e.g., 90°), but in general, no trisection construction is known for an arbitrary angle. Often, proving that a particular construction is impossible relies on facts from algebra; such ideas are not discussed here. Such work began with Descartes (1596–1650), who showed that many geometric problems can be translated into algebraic problems. Vandermonde (1735–1796) and Lagrange (1736–1813) contributed to this area, as well.

The work of Évariste Galois (1811–1832) in algebra implicitly led to the (negative) solutions to the three problems above, as well as many other such compass-straightedge questions. Galois essentially created what are now called "field theory" and "Galois theory", topics beyond the scope of these notes. Essentially, there is a correspondence between field extensions with solutions to quadratic equations and compass-straightedge constructions. For more details, see any of [37, p. 80], [216, Ch. 10], [761], [769, pp. 129–138], or [830, Ch. 19].

In 1837, Pierre Wantzel (1814–1848) [912] used field extensions to show that the Delian problem was not solvable with a compass-straightedge construction (see, e.g., [842] for details and more history), as well as the fact that general angles cannot be trisected (by showing that $\pi/3$ can not be). Since Lindeman proved that π is transcendental, $\sqrt{\pi}$ is not compass-straightedge constructible, and so squaring the circle is impossible; for details see, e.g., [44]

1.16.2 Classical compass-straightedge constructions

In the steps for the following constructions, new markings are often made in red. For brevity, a diagram may represent more than just one step. Also, some constructions can be attained by a different sequence of steps than given here; the reader is invited to experiment with possibly simpler sequences.

Addition or subtraction

Given two segments AB and CD with respective lengths r and s, use a straightedge to extend AB to a line ℓ. Put the compass centered at B with radius s and scribe an arc to the right on ℓ, intersecting at a point X. Then AX has length $r + s$. If $r > s$ and the arc is scribed to the left from B, then AX has length $r - s$.

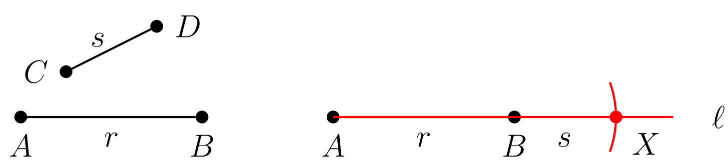


Figure 1.22: Addition

Constructing a perpendicular bisector

Let segment AB be given. Set the compass to any radius larger than half the distance from A to B. Draw arcs, one from A, one from B, that intersect in two points X and Y. The (dashed) line $\overleftrightarrow{XY}$ is the perpendicular bisector of AB.

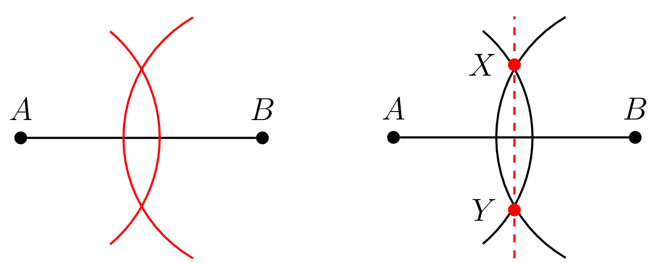


Figure 1.23: Constructing perpendicular bisector of AB

Dropping a perpendicular from a point to a line

See Figure 1.24. Let P be a point and ℓ a line not containing P. Set the compass at a distance slightly larger than the shortest distance from P to ℓ and scribe an arc form P intersecting ℓ in two points A and B Then apply the construction for the perpendicular bisector of AB (dashed red line), which then passes through P as desired. So the angle 90° is constructible.

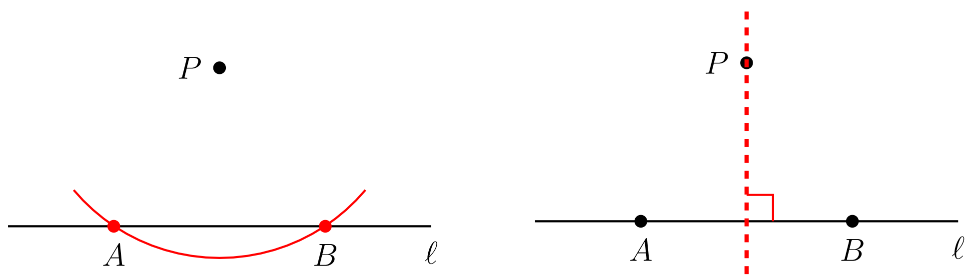


Figure 1.24: Dropping a perpendicular from a point to a line

Line through a point parallel to a given line

See Figure 1.25. Let ℓ be a line and P a point not on ℓ. The goal is to construct a line m through P parallel to ℓ. By the construction in Figure 1.24, drop a perpendicular k from P perpendicular to ℓ, and then putting the compass centered at P, draw arcs on k on either side of P; call these two points A and B. Now apply the construction for the perpendicular bisector of AB, producing the desired line m, which is parallel to ℓ.

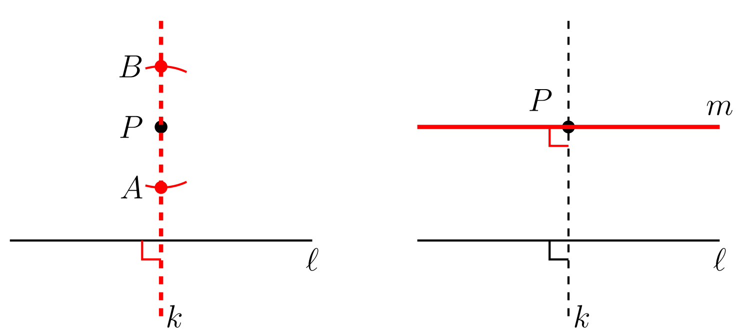


Figure 1.25: The line through a point P parallel to a given line ℓ

Replicating an angle

Suppose that some angle θ is given, say by $\angle XYZ$. How does one reproduce θ in another drawing? Suppose that ℓ is a line containing a point P; the goal is to construct θ at P with respect to ℓ.

As in Figure 1.26, set the compass to some "reasonable" distance and centered at Y, scribe an arc intersecting both YX and YZ, producing points A on YX and B on YZ. Leaving the compass at this same radius, scribe an arc centered at P and crossing ℓ at point Q.

Now set the point of the compass at B, and set its radius to $|AB|$ (so the pencil passes through A). With this radius, scribe an arc centered at Q intersecting the previous arc at some point R. Then $\angle RPQ$ has measure θ, as desired.

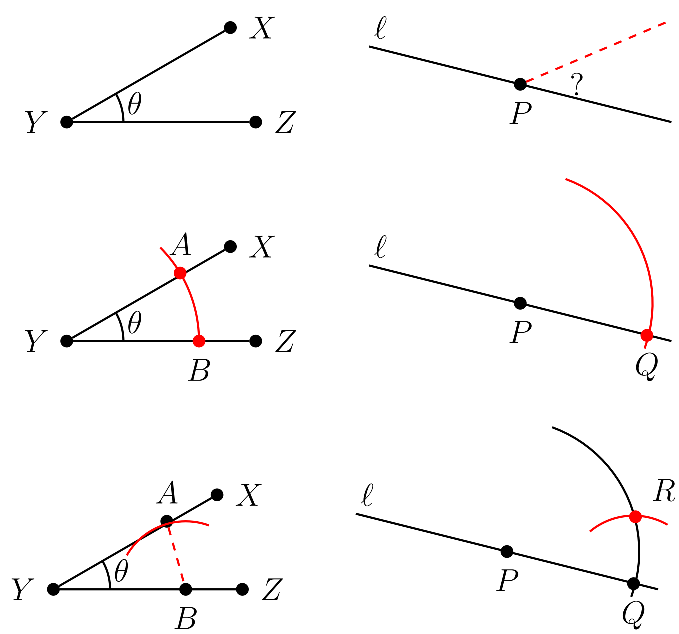


Figure 1.26: Replicating an angle

Bisecting an angle

To bisect an angle of less than 180°, see Figure 1.27. Set the compass to any "reasonable" distance and scribe two vertices equidistant from the vertex where the angle is, getting points A and B. Setting the compass to any radius greater than half the distance from A to B, draw arcs from each of A and B. These two arcs intersect in two points, which thereby (with a straightedge) form the angle bisector. (If one repeats this with a slightly different radius, the two points obtained still lie upon the bisector line.) Since

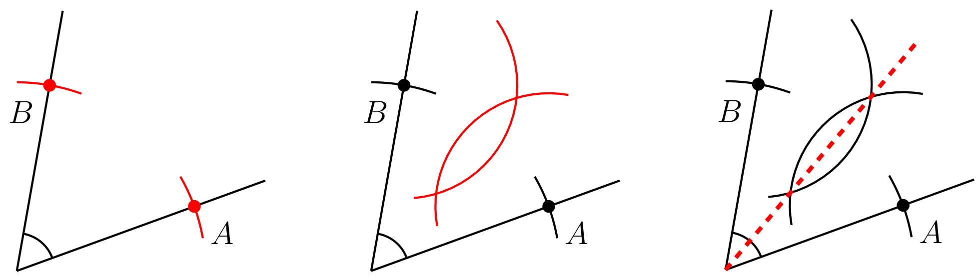


Figure 1.27: Bisecting an angle

the angle 90° is compass-straightedge constructible, by bisecting, so is 45°, but 45° has its own trivial construction (the diagonal of a square).

Multiplication

For any two positive real numbers x and y, if lengths 1, x, y have been constructed, then the number xy can be compass-straightedge constructed (as the length of some segment). Note that the given unit length is crucial to this method. For simplicity, let $x > 1$ (the case $x = 1$ is trivial, and the case $x < 1$ works the same as this construction, but with a slightly different drawing).

As in Figure 1.28, let AB be a segment of length x and AC be a segment of length y, so the two segments are joined at one end (and make an acute angle, say). With the compass, find the point P on AC so that $|AP| = 1$, and use the straightedge to draw the segment PB. Using the construction of a line through a point parallel to a given line (see Figure 1.25), construct the line ℓ through C parallel to PB. (The line ℓ can also be found by translating the angle at $\angle APB$.) Then ℓ intersects the ray $\overrightarrow{AB}$ at a point Q. The desired length is $|AQ| = xy$.

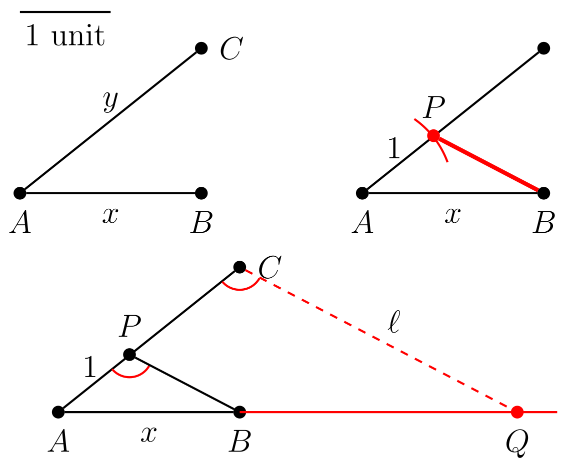


Figure 1.28: Multiplication

Constructing an equilateral triangle

Let AB have length s; set the compass to radius s and scribe two intersecting arcs, one from A and one from B. These two arcs intersect in a point C. Using a straightedge, draw AC and BC, thereby forming the equilateral triangle $\triangle ABC$.

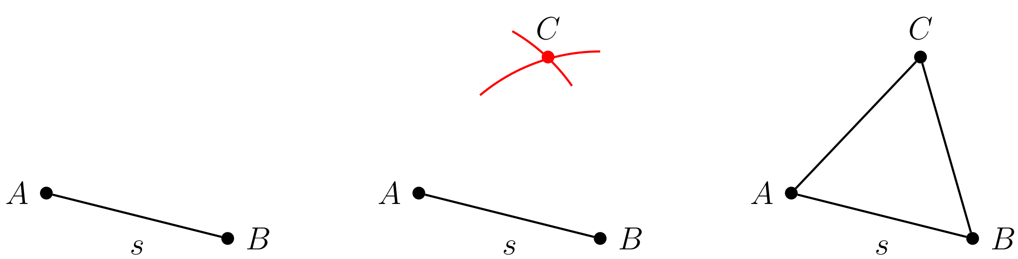


Figure 1.29: Constructing an equilateral triangle of given side length

So the angle of 60° is "constructible", and since bisecting an angle is also constructible (see Figure 1.27), so is 30° and 15°.

Constructing a regular hexagon

See Figure 1.30. The following constructs a regular hexagon with a specified side length s. Let O be any point on the plane, and scribe a circle centered at O of radius s. Let P be any point on the circle (then OP has length s). Leaving the compass set at radius s, center the compass at P and scribe an arc intersecting the circe at, say, Q. Then move the center of the compass to Q and scribe another arc, intersecting the circle at, say, R.

Continue to scribe six arcs moving the compass around to create points P, Q, R, S, T, U, the vertices of a regular hexagon (the vertices S, T, and U are also antipodal from P, Q, and R, and so can be found with just the straightedge touching O, too).

Using a straightedge, construct segments PQ, QR, RS, ST, TU, and UP, the sides of the desired hexagon.

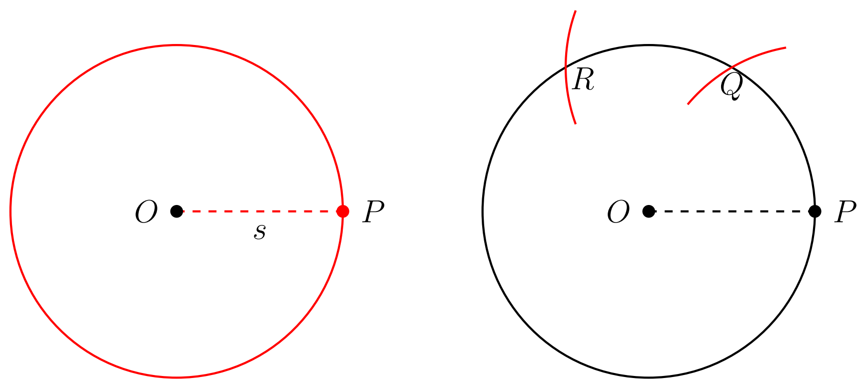


Figure 1.30: Constructing a regular hexagon with given side length s

Observe that finding the vertices of either an equilateral triangle or a regular hexagon can be found without the use of a straightedge!

Constructing a golden rectangle of given height

See Figure 1.31. Let $\tau = \frac{1+\sqrt{5}}{2}$ denote the golden ratio. A rectangle is golden if the ratio between two consecutive sides is τ. Suppose that a golden rectangle of height h is to be found.

Draw a segment AB with length h. Dropping perpendiculars to each of A and B, with compass set to h, find corners of the square $ABCD$. Using the construction to perpendicular bisect, find the midpoint M of AD. Setting the compass to the length $|AC|$, scribe an arc from C downward until it crosses the line $\overleftrightarrow{AD}$ at point F.

Construct the perpendicular to AE which crosses BC at E. The rectangle $ABEF$ is the desired golden rectangle.

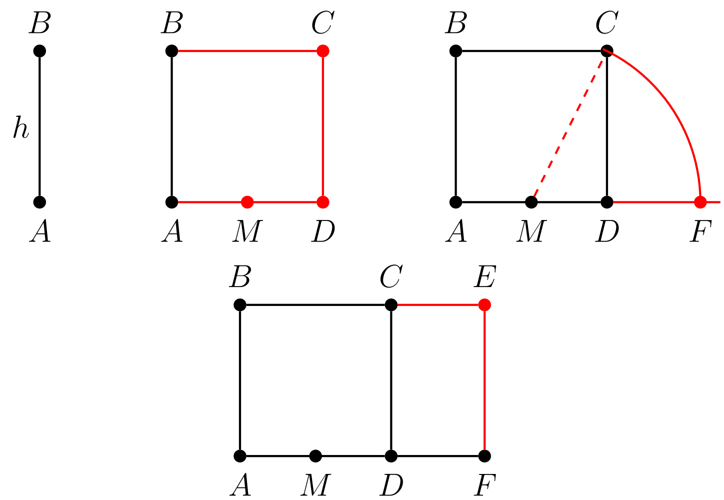


Figure 1.31: Constructing a golden rectangle with given height h

Constructing an acute golden triangle

Recall from Section 1.15.3 that an acute golden triangle is one whose sides are in the ratios $\tau : \tau : 1$, where $\tau = \frac{1+\sqrt{5}}{2}$ is the golden ratio.

To construct an acute golden triangle with a given base AB with length s, first perform the construction of the golden rectangle (see Figure 1.31) where s is replaced by the height, and then set the compass to the length of the rectangle (which is then $\tau \cdot s$). As in Figure 1.32, draw two intersecting arcs, one from A and one from B; the point of intersection is then the third point C of the desired golden triangle $\triangle ABC$.

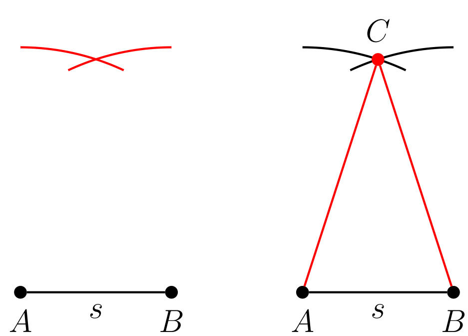


Figure 1.32: Constructing an acute golden triangle with given base

The interior angles 36° and 72° are therefore constructible, and hence (by bisecting angle as in Figure 1.27) so are 18° and 9°.

Constructing a regular pentagon

See Figure 1.33 for one construction of a regular pentagon with a given side length. Construct (as in Figure 1.32) an acute golden triangle $\triangle ABD$, with $|AB| = s$, $|AD| = |BD| = \tau \cdot s$.

Set the compass to radius $s = |AB|$ and from A, B and D, scribe intersecting arcs, giving points C and E, as in Figure 1.33; then $ABCDE$ is the required regular pentagon with side length s.

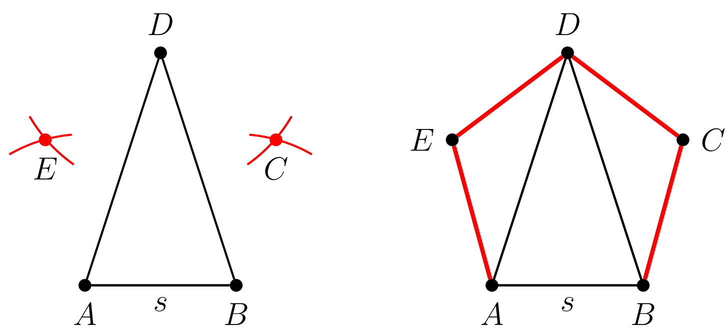


Figure 1.33: Constructing a regular pentagon with given side length

Observe that the two triangles added at the end are both obtuse golden triangles (can you prove this?), and so a regular pentagon can be decomposed (with just two cuts) into three golden triangles. For another construction of a regular pentagon, see [707].

Dividing a segment into n equal lengths

See Figure 1.34. Let AB be a segment and let $n > 1$ be an integer. Draw a line ℓ from A (perpendicular to AB is fine, but at angle works), and set the compass to some small radius, call it 1 unit. Centering the compass at A, make an arc on ℓ, and call this intersection P_1. Move the compass to P_1 and draw a second arc across ℓ, say at P_2. Continue until equally spaced points $P_1, \ldots, P_n$ are constructed.

Using a straightedge, draw P_nB. For each $i = 1, \ldots, n-1$, construct (as in Figure 1.25) a line parallel to P_nB passing through AB. These $n-1$

lines partition AB into n equal pieces. Note that although a segment can be

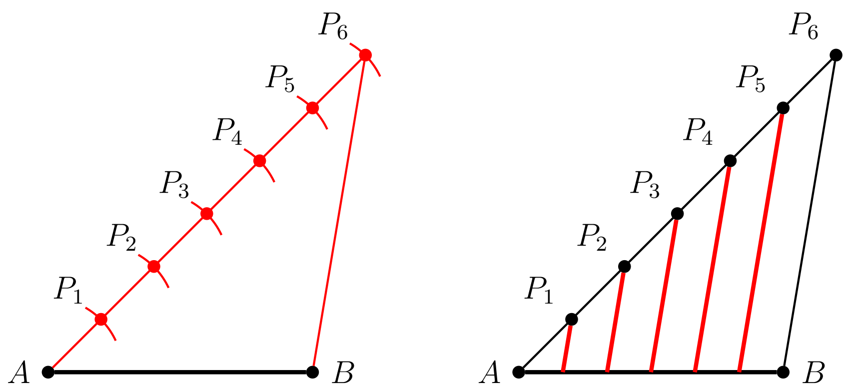


Figure 1.34: Dividing a segment into 6 equal parts

trisected, this in no way implies that angles (in general) can be trisected.

1.16.3 Exercises for compass-straightedge constructions

Exercise 74. *Let α and β be positive real numbers with $\alpha > \beta$. If angles of measures α and β are constructible, show that angles of measures $\alpha + \beta$ and $\alpha - \beta$ are also constructible.*

Exercise 75. *Give a compass-straightedge construction to bisect an angle with measure θ where $180° < \theta < 360°$.*

The next exercise is simple since a tangent to a circle is always at right angles to the corresponding radius (

Exercise 76. *For any circle C and a point P on C, give a compass-straightedge construction of a line ℓ tangent to C at P.*

Exercise 77. *Find a compass-straightedge construction for the obtuse golden triangle.*

Exercise 78. *For positive real numbers a, b, c with $a + b > c$, suppose that segments of lengths a, b, and c have been constructed (or given). Find a compass-straightedge construction for a triangle with side lengths a, b, c respectively.*

Exercise 79. *Prove that if $\theta < 90°$ is an angle whose degree measure is a multiple of 3, then θ is compass-straightedge constructible.*

Exercise 80. *Let x and y be positive real numbers. If segments of length 1, x, and y are given, find a compass-straightedge construction for $\frac{x}{y}$.*

1.17 Inversion with respect to a circle and the relation to linkages

The topic of "inversion with respect to a circle" (also called "reflecting in a circle" is a popular topic in many geometry and popular math texts (e.g., [215, pp. 140–145], [656, pp. 39–45], [801, Ch. 6]), often as a precursor to hyperbolic geometry or other non-Euclidean geometry. An inversion, as described below, is an example of a transformation that is not linear.

Suppose that a circle of radius r and its center O are given. To each point P not on C (not including O!), there is a unique point P' on $\overrightarrow{OP}$ so that $|OP| \cdot |OP'| = r^2$. The point P' is called the inverse of P with respect to C.

To construct such a P', first consider the case when $|OP| > r$, i.e., when P is outside the circle C (the remaining case, given below, is similar). The following construction (see Figure 1.35) can be achieved with only a compass. Setting the compass to radius $|OP|$, from P, draw an arc, intersecting C in points R and S. Setting the compass to radius r, draw an arc from each of R and S, intersecting at O and a second point P', which is on OP. To see that P' has the desired property, use similar triangles. By construction, $m\angle PRO = m\angle POR = m\angle POS = m\angle PSO$ and $m\angle ROP' = m\angle RP'O = m\angle SOP' = m\angle SP'O$. Hence, $\triangle POR$ is similar to $\triangle ROP'$, and so $\frac{|OP|}{|OR|} = \frac{|OR|}{|OP'|}$, from which it follows that $|OP| \cdot |OP'| = |OR|^2 = r^2$, as desired.

Now consider the case when P is inside C. See Figure 1.36. (The construction given here is reversible so that it works for outside points as well, however requires both compass and straightedge.) First find the line through P perpendicular to OP, which intersects C in two points X and Y. The two tangents to C at X and Y intersect in the desired point P' (actually, only one tangent is needed, since it intersects $\overrightarrow{OP}$).

Since $m\angle OXP = m\angle XP'P$, triangles $\triangle XOP$ and $\triangle OXP'$ are similar. Hence $\frac{|OP|}{|OX|} = \frac{|OX|}{|OP'|}$, and so $|OP| \cdot |OP'| = |OX|^2 = r^2$, as desired. □

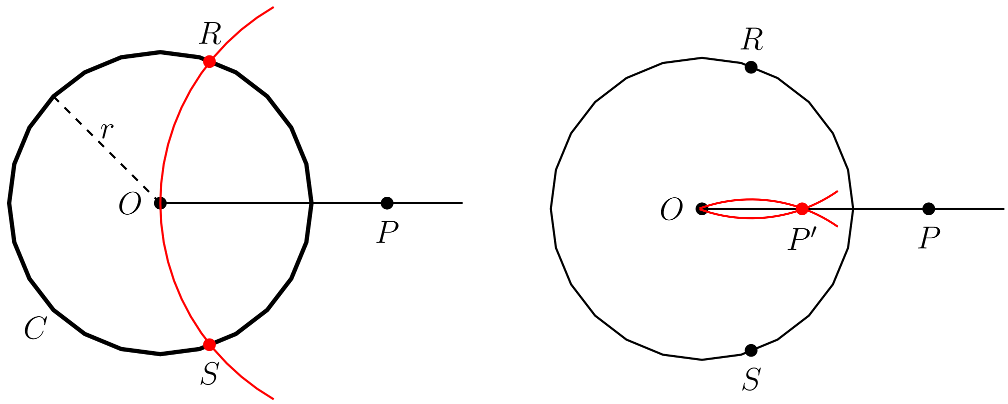


Figure 1.35: Inversion with respect to the circle C for P outside C

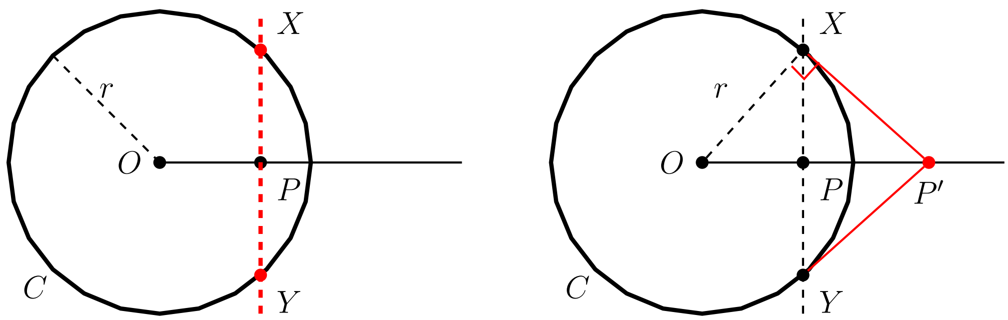


Figure 1.36: Inversion with respect to the circle C for P inside C

As mentioned above, the point P' is called the "inverse of P with respect to C. For any point P in the "punctured plane" $\mathbb{E}^2 \backslash \{O\}$, let f be defined by $f(P) = P'$; the function f is called the *inversion* with respect to C". If one adds a point x at infinity (that all lines contain), then f can be extended by $f(O) = x$; the result is sometimes called the "inversive plane".

An inversion f with respect to some C satisfies a number of remarkable properties. For a proof of the following theorem, see, e.g., [215, pp. 142–144] .or [54, pp. 134–137].

Theorem 1.17.1. *Let C be a circle centered at O, and let f be the inversion with respect to C. Then*

(i) For each P on C, $f(P) = P$;

(ii) *f takes a line through O (actually, a "punctured line") to a line through O;*

(iii) *f takes a line not containing O to a circle through O;*

(iv) *f takes a circle through O to a line not through O;*

(v) *f takes a circle not through O to a circle not through O.*

(vi) *f preserves angles.*

Circle inversion can be presented in terms of complex numbers. For an introduction to these and related ideas, please see, e.g., *Indra's pearls: the vision of Felix Klein* [680, p. 54ff]. [I view this book as a wonderful story of non-Euclidean geometry by experts (I only wish that I could write so well) with incredible illustrations. The epilog contains a historical summary of the subject; such details are seldom given here since my expertise is limited.]

Represent the point (a, b) in the plane by the complex number $z = a + bi$. Recall that $\overline{z} = a - bi$ is the complex conjugate.

Exercise 81. *Let C be a circle in the real plane centered at A (viewed as a complex number) with radius r. Show that the transformation of the real plane defined by $T(z) = A + \frac{r^2}{\overline{z-A}}$ takes a point z to its inverse with respect to C.*

The fact that an inversion can take a circle to a line has a physical manifestation in linkages. For a short treatise on linkages, see [668]. For a review of work on linkages, see the article by Joseph Malkevitch [631] that explains how many famous mathematicians have worked in the area, including early work by Kempe [541] (famous for his work on the 4 colour conjecture) and Watt (famous for steam engines). Two linkages looked at here are found in [215, 155–158]; also see [325] for more history.

A central problem in mechanics was to convert circular motion to linear motion (or linear to circular). For example, what kind of linkage will convert the motion of a cylinder to that of a crankshaft? Such problems became more relevant with the invention of the steam engine and other machines.

In 1784, James Watt produced a three bar linkage that converted circular motion to approximately linear motion (instead, it actually produces a lemniscate which, in certain ranges closely approximates a straight line); see Figure 1.37.

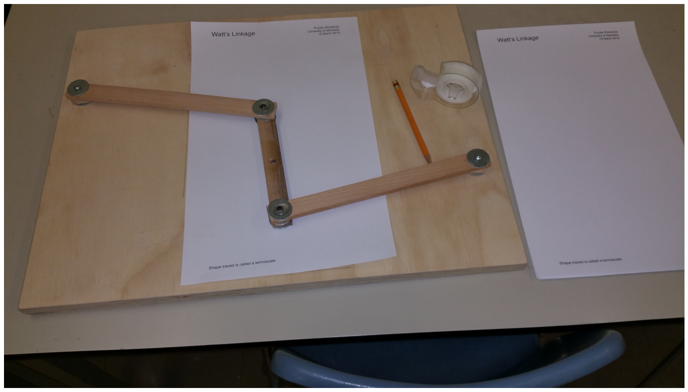


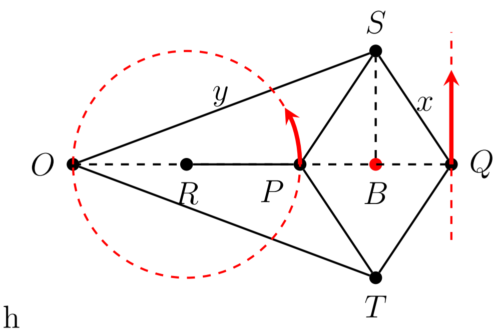
Figure 1.37: Watt's linkage, pencil goes in hole in central bar, DSG 2016

Charles-Nicolas Peaucellier (1832–1913) was a lieutenant in the French army with a background in engineering. [Some sources say that he was a captain, perhaps an old Encyclopedia Britannica article? Some references [215, p. 155] say that he was a naval officer.]

In 1864, Peaucellier solved the problem with a seven bar linkage that translates rotary motion to (exactly) linear motion. His mechanism is now often called the "Peaucellier cell". Independently, Yom Tov Lipman Lipkin (1846–1876), a student of Chebyshev, also found the same linkage in 1871 or so.

Figure 1.38: Peaucellier's linkage, mahogany and wenge, DSG 2016

In Figure 1.38, the dark center bolt moves in a circular fashion, and the right-most bolt travels in a straight line (orthogonal to the front table edge);



h

Figure 1.39: Peaucellier's linkage

I have since replaced the right-most bolt with a device to hold a pencil, but I have no photo of that yet. The base clamps to any table top, thereby fixing the position of the two pivots on the left.

As mentioned above, the math behind Peaucellier's linkage can be explained using circle inversion. Based on [215], let Figure 1.39 represent the linkage.

Let the four rods PS, SQ, QT and TP have length x, and let the two rods OS and OT have length y (which is larger than x). For convenience, let R be drawn precisely half-way between O and P, where P is the point that moves in a circle (dotted circle) passing through O. (In fact, R can be drawn anywhere equidistant from O and P.) To be shown is that if P is rotated about R, then Q travels in a straight line.

Letting B be the foot of the perpendicular from S onto OQ,

$$\begin{aligned}
|OP| \cdot |OQ| &= (|OB| - |PB) \cdot (|OB| + |PB|) \\
&= |OB|^2 - |PB|^2 \\
&= (y^2 - |SB|^2) - (x^2 - |SB|^2) \qquad \text{(by Pythagoras)} \\
&= y^2 - x^2,
\end{aligned}$$

which is constant. Putting $r = \sqrt{y^2 - x^2}$, the calculations show that

$$|OP| \cdot |OQ| = r^2,$$

and so P and Q are inverse points with respect to the circle of radius r centered at O. By property (iv) of Theorem 1.17.1, Q then travels in a straight line. $\square$

In 1874, Harry Hart (1848–1920) invented another linkage that converted rotation to linear movement, one that only used 5 rods—see Figure 1.40. (Chebyshev previously thought he proved no such linkage exists [668].)

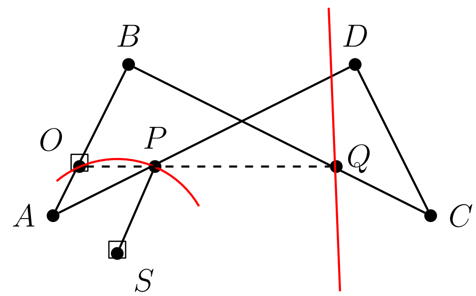


Figure 1.40: The first Hart's linkage with 5 rods

In a model, the pivot points O and S are fixed The point O is fixed on the rod AB, the point P is fixed on AD, and the point Q is fixed on BC, and each of O, P, S divide their rod in the same proportion. The point P is rotated about S (going through O). Rods AB and CD have the same length, and rods BC and AD have the same length.

With elementary geometry, one can prove that P and Q are inverse points, and therefore, when P moves along the circle containing O, Q travels in a straight line. (See [215, p.157] for details.)

A year later, Hart published a second linkage, called the "A-frame", that achieved the same goal. (Various animations are available online.) The two are called the "Hart's inversors".

1.18 More exercises in plane geometry

Many of the exercises given in this section ask to prove a famous result. References and further discussion of such a result might only be mentioned in the solution or subsequent comments. As is often the case, a simple exercise might serve as an introduction to a large class of problems, any one of which might ordinarily be given as "standard" in a geometry text.

1.18.1 Exercises with triangles

Exercise 82. *Prove that the sum of the measures of (interior) angles in a triangle is π (or 180 degrees).*

Exercise 83. *In a triangle $\triangle ABC$, let D be the midpoint of AB, and let E be a point on AC. Prove that DE is parallel to BC if and only if E is the midpoint of AC.*

Exercise 84. *In a triangle $\triangle ABC$, let D, E, and F be the respective midpoints of AB, AC, and BC. Show that $\triangle DEF$ is similar to $\triangle ABC$. Show also that the four triangles $\triangle DEF$, $\triangle ADE$, $\triangle EFC$, $\triangle BDF$ partition $\triangle ABC$ into four congruent triangles.*

Exercise 85. *Two triangles $\triangle ABC$ and $\triangle DEF$ are similar with sides being integer lengths. Suppose that $|AB| = |DE|$ and $|AC| = |DF|$. If the remaining two sides BC and EF differ in length by 774, what are the possible lengths of the sides?*

The next exercise is an old classic (mentioned in Lemma 1.12.3):

Exercise 86. *Let C be a circle with two points X, Y so that XY is a diameter. Let Z be another point on C. Prove that $\angle XZY$ is a right angle.*

Exercise 87. *Among all triangles inscribed in the unit circle, find those with maximum perimeter.*

Exercise 88. *For real numbers x, y, z and angles α, β, γ satisfying $\alpha+\beta+\gamma = \pi$, show that*

$$x^2 + y^2 + z^2 \geq 2(yz\cos\alpha + zx\cos\beta + xy\cos\gamma).$$

Exercise 89. *Consider an equilateral triangle with side length n, drawn with a grid on it forming unit equilateral triangles whose sides are parallel to the large triangle (as in Figure 1.41). Prove that the number of triangles that can be counted in such a figure is*

$$\left\lfloor \frac{n(n+2)(2n+1)}{8} \right\rfloor.$$

For example, when $n = 3$, there are 13 triangles and for $n = 4$, there are 27.

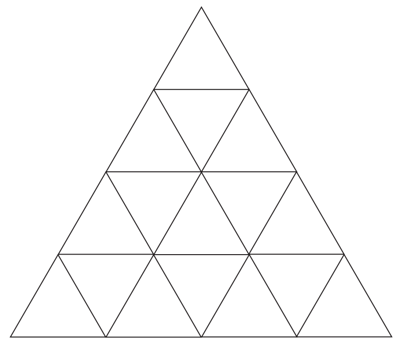


Figure 1.41: Count the triangles; $n = 4$

Exercise 90. *For $n \geq 1$, suppose that $3n$ points in a plane are given in general position (no three on a line). Prove that these points form the vertices of n mutually disjoint triangles.*

Exercise 91. *Suppose that a pizza is in the shape of an equilateral triangle, and is to be shared by you and your partner. You let your partner pick any spot on the pizza, and you can then cut the pizza with one straight cut passing through the designated spot. You then get first choice between the two pieces. Show that you can always get $\frac{5}{9}$ of the pizza, but cannot guarantee more. Show also, that if the pizza is square, you can only guarantee half for yourself.*

Exercise 92. *Suppose that a square with area S_1 is covered by a triangle with area S_2. Show that $2S_1 \leq S_2$.*

Exercise 93. *Let $a_1, \ldots, a_7$ be real numbers in the open interval $(1, 13)$. Prove that there exist three of these a_i's that are side-lengths of a (non-trivial) triangle.*

Exercise 94. *Let C be a circle. Find a red-blue colouring of points on C so that no three of the points forming a right triangle are coloured the same.*

Exercise 95. *Each point of the perimeter P of an equilateral triangle is coloured either red or blue. Prove that there exists a right triangle whose three vertices are points in P with the same colour.*

The next problem shows that if two medians in a triangle have same lengths, then the triangle is isosceles.

Exercise 96. *Suppose that in $\triangle ABC$, medians BY and CZ have equal lengths. Show that $|AB| = |AC|$.*

Exercise 97. *Let ABC be a triangle. Let D be on AB and E be on AC so that*

- *CD bisects $\angle BCA$,*
- *BE bisects $\angle ABC$, and*
- *$|AD| = |AE|$.*

Prove that $\triangle ABC$ is isosceles.

The next exercise is a classic (but some say that it is a cruel exercise).

Exercise 98. *Show that in a triangle, if two angle bisectors have the same length, then the triangle is isosceles.*

Exercise 99. *Let T be an equilateral triangle. Find all points P in the interior of T so the sum of the three distances from P to sides of T is minimum, and find this minimum distance.*

Comments on Exercise 99: The result in Exercise 99 is called "Viviani's theorem", named after Vincenzo Viviani (1622–1703), a student of Torricelli, an assistant of Galileo, and later, court mathematician to the Grand Duke Ferdinando II de' Medici in Florence. Viviani's theorem says that the sum of the distances is constant!

A version of Viviani's theorem is true if one allows points outside the triangle when negative distances are implemented (inside distance, positive, and outside distance, negative). See the article by Polster [727] for a different proof of Viviani's theorem and more references. Viviani's theorem was presented as a puzzle in Dudeney's famous book *536 Puzzles and curious problems* [272, No. 284, p. 95]. The puzzle asked to find the point that minimized the distances, and the solution given only said that the total distance is constant, equal to the altitude, but no proof was offered (and "Viviani's theorem" was not referenced!). The same puzzle, but with dimensions added, "Equidistant meeting points" occurred in Graham's book *Ingenious mathematical problems and methods* [402, Prob. 31] (Graham was an engineer).

There also, no mention of Viviani is given, but some (interesting?) alternative proofs due to readers of "DIAL" are discussed (some of which are far more complicated than the proof given in the solutions).

Is there a generalization of Viviani's theorem for general triangles, not just equilateral ones? After only a brief experimentation, the reader can find a non-equilateral triangle for which the sum of distances is not constant. However, a little something can be said about the sum of distances from a particular point, as given in Exercise 100 (also see the Erdős–Mordell inequality in Theorem 1.18.1 below).

Exercise 100. *Let T be an acute triangle with circumradius R and inradius r, and let P be the center of the circumcircle for T. If x, y, z are the distances from P to the each side of T, show that $x + y + z = R + r$. (Distances need to be "signed", depending upon whether or not the circumcenter is inside or outside the triangle.)*

Comment on Exercise 100: The result is called "Carnot's theorem", after the French mathematician (and Napoleon's Minister of War, and co-founder of Ecole Polytechnique) Lazare Nicolas Marguérite Carnot (1753–1823). See [484, p. 25]. Thanks to Andriy Prymak [736] for this reference. □

Another result, somewhat related to Viviani's theorem and Carnot's theorem, was posed as a problem by Erdős in 1935 [295] and solved two years later by Louis Mordell (the famous number theorist, 1888–1972) and David Francis Barrow (1888–1970). Both proofs by Mordell and Barrow appeared in the same article, however Barrow proved a stronger statement, which is stated separately below in Theorem 1.18.2.

Theorem 1.18.1 (Mordell, 1937 [667]). *Let P be a point inside a triangle $\triangle ABC$, and let PD, PF and PG be perpendiculars drawn from P to the sides of $\triangle ABC$. Then*

$$|PA| + |PB| + |PC| \geq 2(|PD| + |PE| + |PF|).$$

Mordell's proof relies on the inequality given in Exercise 88, which says for real numbers x, y, z and angles α, β, γ satisfying $\alpha + \beta + \gamma = \pi$:

$$x^2 + y^2 + z^2 \geq 2(yz \cos \alpha + zx \cos \beta + xy \cos \gamma).$$

Theorem 1.18.2 (Barrow, 1937 [61]). *Let P be a point inside a triangle $\triangle ABC$, and let D, E', and F' be points on $\triangle ABC$ so that PD', PE', and PF' are bisectors of the angles at A, B, and C, respectively. Then*

$$|PA| + |PB| + |PC| \geq 2(|PD'| + |PE'| + |PF'|).$$

The points D', E', F' in Theorem 1.18.2 are at least as far away from P than the points D, E, F in Theorem 1.18.1, and so Barrow's result is a strengthening of Mordell's.

Exercise 101. *Show that equality holds in Theorem 1.18.1 if and only if P is the center of an equilateral triangle.*

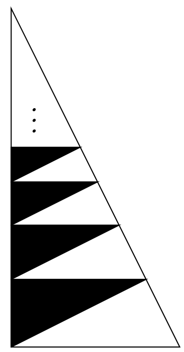
The inequality in Theorem 1.18.1 is now called "the Erdős–Mordell inequality". Proofs simpler than the original were given by Kazarinoff [533] in 1957, and by Bankoff [49] in 1958. Yet another proof was given in 2001 by Lee [596]. In 2007, Alsina and Nelsen [25] gave a "visual" proof. In 1961, Lenhard [606] gave an extension of the Erdős–Mordell inequality to convex polygons, given with two proofs, one using Ptolemy's theorem (Theorem 1.9.6). A strengthening of the inequality (that uses the circumcircle of $\triangle ABC$) was given by three mathematicians from Vietnam [236] in 2016 (see also [637] for a weighted version and an elegant application).

Exercise 102. *An equilateral triangle ABC is inscribed in a circle, and an arbitrary point M is chosen on the arc BC. Prove that $|MA| = |MB| + |MC|$.*

Exercise 103. *Let $\triangle ABC$ be a scalene triangle with longest side AB. Let P and Q be points on AB such that $|AQ| = |AC|$ and $|BP| = |BC|$. Show that the circumcenter of $\triangle CPQ$ is equal to the incenter of $\triangle ABC$.*

This next exercise appeared in *The Guardian*, a British newspaper, but I do not know its origin.

Exercise 104 ([644]). *Consider a right triangle with hypotenuse length 5, with height twice the length of the base. As in the figure below, black "teeth" are marked off by lines that meet alternately the hypotenuse and height at right angles (continuing all the way up). What fraction of the triangle is black?*

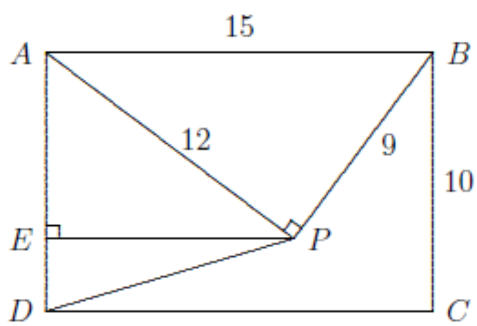


Exercise 105. *Let ABC be an acute triangle. Let points P and Q lie on the side BC so that and $\angle PAB = \angle BCA$ and $\angle CAQ = \angle ABC$. Let M a point so that P is a midpoint of AM and let N be the point so that Q is the midpoint of AN. Prove that the lines $\overleftrightarrow{BM}$ and $\overleftrightarrow{CN}$ intersect on the circumcircle of $\triangle ABC$.*

Exercise 106. *If two sides of an isosceles triangle have length 1, what is the length of the third side that maximizes the area of the triangle? [Hint: Calculus is not required.]*

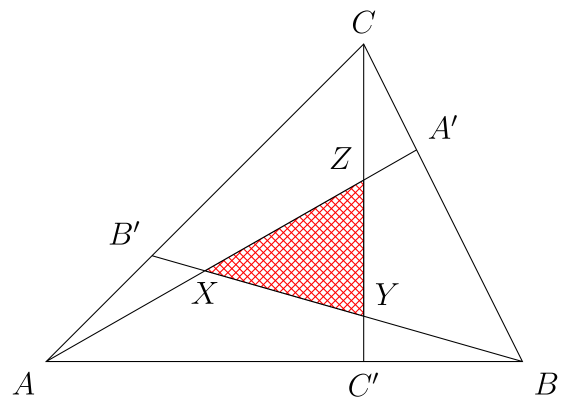
Exercise 107. *Let S be a finite set of points in the plane so that for any $A, B, C \in S$, the area of $\triangle ABC$ is at most 1. Show that there exists a triangle of area 4 that (together with its interior) contains the set S.*

Exercise 108. *In the rectangle $ABCD$, sides AD and CD have lengths* 10 *and* 15*, respectively. The point P lies inside the rectangle, and the lengths of AP and BP are, respectively,* 12 *and* 9*. Prove that $\triangle APD$ is isosceles.*



Exercise 109. *On a triangle $\triangle ABC$, put points A', B', C' on the sides so that B' is one third the way from A to C, A' is one third the way from*

C to B, and C' is one third the way from B to A. The lines AA', BB', CC' form a triangle $\triangle XYZ$ (where $X = AA' \wedge BB'$, $Y = BB' \wedge CC'$ and $Z = CC' \wedge AA'$):



Prove that area$(\triangle XYZ) = \frac{1}{7}$area$(\triangle ABC)$.

Comments on Exercise 109: This problem appeared in the Hugo Steinhaus book *Mathematical snapshots* [826, p. 9] along with a "graphical proof" consisting of a diagram using seven triangles all congruent to $\triangle XYZ$ (not all inside $\triangle ABC$). The references Steinhaus gave for this problem are Dudeney's book *Amusements in mathematics* [271, p. 27] (but I looked there and could not find it) and [657] for the "graphical proof" (I have not seen this paper, but Math Reviews verifies this reference contains a proof). A similar problem (called "The inside triangle") was given in [402, Prob. 52], where points are one quarter along the way of each side. Included in the solutions given in [402] are many different approaches; I recommend having a look at these, in particular, one by Howard D. Grossman from New York, where it is generalized from fourths to one over any integer. Another "standard" solution to Exercise 109 uses inclusion–exclusion and Menelaus's theorem (Theorem 1.7.15).

The triangle with cevians given in Exercise 109 is sometimes called "Feynman's triangle", since, according to Cook and Wood [211], Richard Feynman was once introduced to the problem at a dinner party after a colloquium at Cornell University, and at first, he could not prove it (but later did—he and his friends apparently found four proofs). Cook and Wood go on to give a number of different proofs. Later, in the same journal, Michael de Villiers

[255] gave a slightly more general version and a version for parallelograms (see also the website [256] of De Villiers for more information).

The result in Exercise 109 is a special case of a more general theorem by John Edward Routh (1831–1907):

Theorem 1.18.3 (Routh, 1896 [771]). *On a triangle $\triangle ABC$, let C' be the point on AB dividing AB in a ratio $1 : \lambda$, let A' be the point on BC dividing BC in a ratio $1 : \mu$, and let B' be the point on CA dividing CA in a ratio $1 : \nu$.*

The lines AA', BB', CC' form a triangle $\triangle XYZ$ (where $X = AA' \cap BB'$, $Y = BB' \cap CC'$ and $Z = CC' \cap AA'$). Then

$$\mathit{area}(\triangle XYZ) = \frac{(\lambda\mu\nu - 1)^2}{(\lambda\mu + \lambda + 1)(\mu\nu + \mu + 1)(\nu\lambda + \nu + 1)} \cdot \mathit{area}(\triangle ABC).$$

Note that when $\lambda = \mu = \nu = 2$, Routh's theorem solves the problem given in Exercise 109.

There are many proofs of Routh's theorem, starting with the original [771, p. 82] (which I have not seen). In the literature, it has been suggested that Theorem 1.18.3 might have been discovered earlier by Steiner (1796–1863), (but I have not verified the details), and so Theorem 1.18.3 is sometimes (e.g., see [639]) called the Steiner–Routh theorem.

Coxeter [220, pp. 211, 219–220] gives a (long) proof, based on barycentric coordinates, that shows the result even when λ, μ, ν can take on negative values (so, e.g., when A' is on the line $\overleftrightarrow{BC}$ outside the segment $\overline{BC}$). Coxeter also shows that both theorems of Ceva (Theorem 1.7.18) and Menelaus (Theorem 1.7.15) are equivalent to Routh's theorem. In the popular book *College geometry*, [532, pp. 205–207] David Kay gives a (long) synthetic proof. Routh's theorem is "affine invariant", so it suffices to show the theorem for a specific triangle, so picking, say, an isosceles right triangle, makes certain calculations easier—this fact was used in a proof by Melzak [654, pp. 7–9]. Ivan Niven [686] gives a proof using coordinate geometry. (To verify that my drawing above was correct, I computed intersection points, and then area using cross products; I believe that such a calculation can be generalized to a proof, although such a proof might involve too much notation as to make things readable. Niven's proof might also be along these lines.) Pedoe [715, Ex. 11.8, p. 55] also posed Routh's theorem as an exercise (with hints) with vectors.

In 1981, Murray Klamkin and Andy Liu [550] gave a short paper in *Crux* containing the above references and three more proofs of Routh's theorem, the first two of which used a vector approach. The third proof, which they call "synthetic" is short and sweet. For recent (2015, 2017) papers on Routh's theorem, see [638] (where a proof for the 3-dimensional version is given) and [639]. (These two papers also give many other references than are given here.)

Exercise 110. *For $i = 1, 2$ let T_i be a triangle with side lengths a_i, b_i, c_i, and area A_i. Suppose that $a_1 \leq a_2$, $b_1 \leq b_2$, $c_1 \leq c_2$, and that T_2 is an acute triangle. Does it follow that $A_1 \leq A_2$?*

For the next question, recall Definition 1.15.1 for the description of Fibonacci numbers.

Exercise 111. *Let $F_0 = 0$, $F_1 = 1$, $F_2 = 1$, $F_3 = 2$, $F_4 = 3$, $F_5 = 5$, ..., denote the Fibonacci numbers. For any $n \geq 1$, construct a right-angle triangle with legs F_nF_{n+3} and $2F_{n+1}F_{n+2}$. Show that the hypotenuse is $F_{n+1}^2 + F_{n+2}^2$. Also, show that the area of the triangle is the product of the four Fibonacci numbers $F_nF_{n+1}F_{n+2}F_{n+3}$.*

For example, when $n = 4$, the four Fibonacci numbers are 3, 5, 8, 13, giving legs of length 39 and 80, with hypotenuse $5^2 + 8^2 = 89$ and area 1560.

1.18.2 Exercises with quadrilaterals

Exercise 112. *Let $ABCD$ be a quadrilateral with $|AB| = 5$, $|BC| = 17$, $|CD| = 5$, and $|DA| = 9$. Can BD have integer length? If so, which one(s)?*

Exercise 113. *A rectangle with sides a and b is circumscribed by another rectangle of area m^2. Determine all possible values of m in terms of a and b.*

Exercise 114. *Show that if two squares with perimeters a and b are enclosed in a square with perimeter c so that their interiors do not overlap, then $a + b \leq c$.*

Note that the same result in Exercise 114 holds if one replaces squares with circles. Beck and Bleicher [66] showed that a similar result holds only if the shapes are regular polygons or curves of constant width.

Problems related to Exercise 114 were studied by Erdős and Graham [306] in 1975. For any $k \in \mathbb{Z}^+$, it is possible to put k^2 non-overlapping squares in a unit square where the total perimeter of the k^2 squares is $4k$ (just use squares forming a $k \times k$ grid). It was conjectured by Erdős in the 1930s that if k^2+1 squares are packed in a unit square, the total perimeter still remains at most $4k$. The case $k = 1$ is essentially Exercise 114 with $c = 1$. The case $k = 2$ was solved by D. J. Newman (oral communication, see [306]), but as of 1975, the conjecture remained unproved for $k > 2$. If $f(\ell)$ denotes the maximum sum of perimeters of ℓ squares packed into a unit square; so Erdős's conjecture is that $f(k^2 + 1) = 4k$.

One of the main results in the Erdős–Graham paper is that $f(k^2 + O(k^{7/11})) > 4k$, and this is done using all squares of the same size. For convenience, the problem was scaled so that all the small squares become unit squares and the large square had arbitrary size. For each real $\alpha > 0$, put $W(\alpha)$ to be the difference in area between a square of area α and the maximum area covered by a packing with unit squares. Erdős and Graham [306] showed that $W(\alpha) = O(k^{7/11})$ by finding a packing with the unit squares slightly tilted.

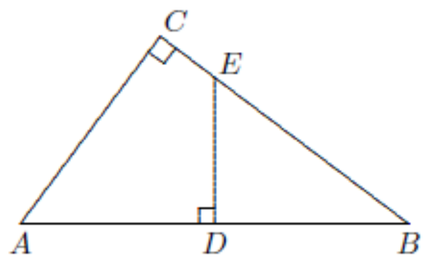
Exercise 115. *Let $ABCD$ be a quadrilateral with $\angle ADB = \angle ACB$. Show that $\angle ABD = \angle ACD$.*

Exercise 116. *Show that a quadrilateral $ABCD$ is cyclic (its vertices lie on a circle) if and only if $\angle DAB$ and $\angle ADC$ are supplementary.*

Hint for Exercise 116: From [497, p.55], for the "if" part, show that D lies on the unique circle through A, B, and C. □

Exercise 117. *Let $ABCD$ be a quadrilateral and let P be a point in its interior. Draw four line segments from P to each of the midpoints of the sides of $ABCD$, thereby partitioning the quadrilateral into four smaller quadrilaterals. Two of these smaller quadrilaterals that do not share a common border have a total area half that of $ABCD$.*

Exercise 118. *In the figure $|AB| = 20$, $|AC| = 12$, $|AD| = |DB|$, and $\angle ACB$ and $\angle ADE$ are right angles. Find the area of the quadrilateral $ADEC$.*



Exercise 119. *Prove that a parallelogram is a rhombus if and only if its diagonals are perpendicular.*

1.18.3 Exercises with polygons

Exercise 120. *Let $ABCDE$ be a regular pentagon and M be a point in its interior such that $m\angle MBA = m\angle MEA = 42°$. Prove that $m\angle CMD = 60°$.*

Exercise 121. *Let $k \in \mathbb{Z}^+$. Show that any convex $(3k+1)$-gon cannot be divided by some of its diagonals into k pentagons.*

Exercise 122. *Let H be a regular heptagon (7 sides) on vertices (in order) $P_1, \ldots, P_7$. Let Q be the midpoint of the chord P_1P_4. Show that the area of the quadrilateral $P_1P_2P_3P_4$ is equal to the area of the quadrilateral $QP_5P_6P_7$.*

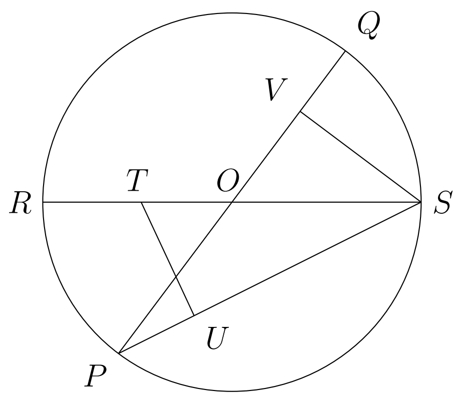
Exercise 123. *Show that any convex polygon with area 1 is contained in a rectangle of area 2.*

Exercise 124. *How many of the regular polygons have interior angles that when measured in degrees are integers?*

Exercise 125. *Let P be a convex polygon inscribed in a circle. Triangulate P by drawing chords (see Section 1.11). Then the sum of the radii of the incircles for the triangles is the same, no matter how the triangulation is done.*

1.18.4 Exercises with circles

Exercise 126. *Let C be a circle centered at O. As in the following diagram, let RS and PQ be diameters, and let T be the point on RS that bisects RO. Let U be the point on PS so that $\angle TUS$ is a right angle and let V be the point on OQ so that $\angle OVS$ is a right angle. Prove that $|OT| \cdot |PV| = |PU| \cdot |PS|$.*



Exercise 127. *Show that any two squares of side-length 0.9 inside a circle of radius 1 must overlap.*

Exercise 128. *Suppose that a point A is covered by each of six congruent disks in the plane. Show that at least one of the disks contains the center of some of the remaining disks.*

Exercise 129. *Let C_1 be a circle of radius 1 and let C_2 be a circle of radius 4. Suppose that C_1 and C_2 are tangent (externally). Suppose that ℓ is one of the two lines tangent to both circles. If C_3 is a circle tangent to both C_1 and C_2 and tangent to one of ℓ, what is the radius of C_3?*

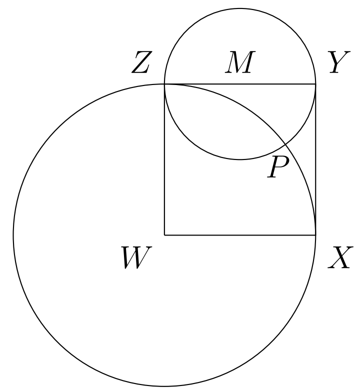
Exercise 130. *Let $(x_1, \frac{1}{x_1}), (x_2, \frac{1}{x_2}), (x_3, \frac{1}{x_3}), (x_4, \frac{1}{x_4})$ be four distinct points on the hyperbola $xy = 1$, and suppose that these four points lie on a circle. Prove that $x_1x_2x_3x_4 = 1$.*

Exercise 131. *A point $(x, y) \in \mathbb{R}^2$ is called rational if both x and y are rational. What is the maximum number of rational points on a circle in $\mathbb{R}^2$ with its center not rational?*

Exercise 132. *What is the largest possible radius of a circle contained in a 4-dimensional hypercube of side length 1?*

Exercise 133. *Consider a semicircle with AB as its base (diameter) on the x-axis, with $A = (-x, 0)$ and $B = (x, 0)$. In the first quadrant, there are four points, C,D,E,F on the semicircle so that the distances from each to both A and B are integers; also AB has integer length. What is the smallest such value of x?*

Exercise 134. *Let $WXYZ$ be a square with side length 4. Let M be the midpoint of YZ. Let C_1 be a circle of radius 2 centered at M, and C_2 be a circle of radius 4 centered at W. Suppose that C_1 and C_2 intersect at points Z and P.*



What is the distance from P to WZ? What is the distance from P to WY?

The next exercise seems to be a classic. [Rob Craigen showed me this in 2003 or so.]

Exercise 135. *Suppose that three circles C_1, C_2, C_3 of different sizes are drawn in the plane with no circle entirely inside another. For each pair of circles C_i, C_j, draw the two external tangent lines to both circles (use the outer two tangents, not the ones that cross between the circles). Since the circles are of different sizes, these two tangent lines meet in a point; call this point $X_{i,j}$. Show that $X_{1,2}$, $X_{1,3}$ and $X_{2,3}$ are collinear.*

See also Section 1.15 for two more problems on circles (regarding the golden ratio).

1.18.5 Miscellaneous geometric problems in the plane

Exercise 136. *Show that if two unit-segments in the plane are disjoint, then at least one pair of endpoints are at distance greater than one.*

Exercise 137. *Some n points are on the plane, where no two distances between points are the same. Connect each point to its nearest point with a straight line segment. Prove that the resulting figure does not contain:*

(a) *any closed polygon;*

(b) *any intersecting segments;*

(c) *any point with six or more neighbours.*

Exercise 138. *Show that at most eight non-overlapping unit squares can touch another unit square.*

For a discussion regarding the number of unit cubes touching one unit cube, see [370] (the answer seems to be 24, given by Robert S. Holmes, a particle physicist).

Exercise 139 (Folklore)**.** *Let $\ell_1, \ldots, \ell_n$ be lines in the plane. Show that some two of these lines form an angle of at most $\frac{\pi}{n}$.*

Exercise 140 (Folklore)**.** *Prove that* $\tan(1°)$ *(one degree) is irrational. Hint: Relate this to* $\tan(30°)$ *or some such irrational.*

Chapter 2

Graphs

Graphs are a central tool in polyhedral geometry and combinatorial geometry; this chapter gives a brief review of terminology and basic results.

2.1 Some basics

In mathematics, there are two types of "graph". The *graph* of a function $f : X \to Y$ is the collection of pairs $\{(x, f(x)) : x \in X\}$, so many functions can be encoded as a picture of points in the Cartesian plane (for example, a parabola is a graph of a quadratic function whose domain is the set of reals).

The second type of "graph" is a relational structure. These graphs are often used to encode properties of geometric structures (like polyhedra) or networks. Such "graphs" come in many varieties, including simple graphs, multigraphs, hypergraphs, and directed graphs.

A "graph" defined here is what some authors call a "simple" graph. This section gives only a very brief review of some basic concepts in graphs—if only to establish basic notation, definitions, and mention some major theorems for later use. Many results in this section do not have proofs or references, as they are all standard in a first year course in discrete mathematics or a course in combinatorics. In this section, I have copied liberally from my *Notes on advanced graph theory* [435]. For more details, see any of the more popular graph theory textbooks (e.g., [117], [125], [184], or [919]).

Much of the terminology for graphs comes from geometry, particularly from polyhedra. Certain relationships between graphs and combinatorial geometry have produced major theorems in both areas.

Definition 2.1.1. A graph is an ordered pair $G = (V, E)$, where $V = V(G)$ is a non-empty set and $E = E(G)$ is a set of (unordered) pairs from V. Elements of V are called vertices and elements of E are called edges.

In the above definition, E is a set, and so any unordered pair of vertices occurs in E at most once (in graph theory jargon, this says that there are no multiple edges). Also, since edges are pairs, there is no "edge" of the form $\{x, x\}$ (which is called a called a loop). Graphs without multiple edges or loops are called *simple*; graphs with multiple edges or loops are called *multigraphs* or *pseudographs*. Unless otherwise needed, all graphs here are considered to be simple.

Graphs can often be described most easily by a drawing, where vertices are represented by dots (or small circles) and edges are denoted by arcs or line segments joining vertices. The positions of the vertices and the lines joining vertices can be given in any number of ways.

One notation that is common in graphs is the following. Let S be a set and k be a positive integer, and define

$$[S]^k = \{T \subseteq S : |T| = k\}.$$

Using this notation, a graph is a non-empty set V together with edge set $E \subseteq [V]^2$.

Two vertices x, y in a graph G are said to be *adjacent* if and only if $\{x, y\} \in E(G)$. An edge $e = \{x, y\}$ is said to be *incident* with both vertices x and y, (and conversely, the vertices are incident with the edge).

Two graphs $G = (V, E)$ and $H = (W, F)$ are *isomorphic* if and only if there exists a bijection $\sigma : V \to W$ so that $\{x, y\} \in E$ iff $\{\sigma(x), \sigma(y)\} \in F$. If G is isomorphic to H, write $G \cong H$. The relation "$\cong$" is an equivalence relation. Isomorphic graphs can have different drawings, but non-isomorphic graphs always have different drawings, so when one speaks of "a graph G", one is really talking about an equivalence class of graphs.

Unless specific labels on vertices are considered, when one says "the graph G", one is usually talking about any graph from the equivalence class containing G. In other words, a graph G can be drawn or presented in many different ways (see Figure 2.1).

A graph $G = (V, E)$ is said to be *complete* if and only if E consists of all (unordered) pairs from V, that is, $E = [V]^2$. A complete graph on n vertices is denoted by K_n, which has $\binom{n}{2}$ edges. Thus, any (simple) graph G on n vertices has at most $\binom{n}{2}$ edges.

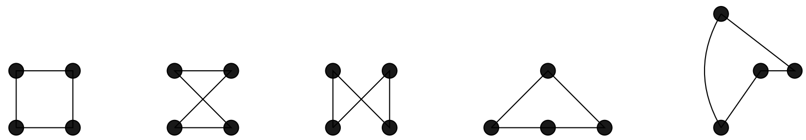


Figure 2.1: Five isomorphic graphs.

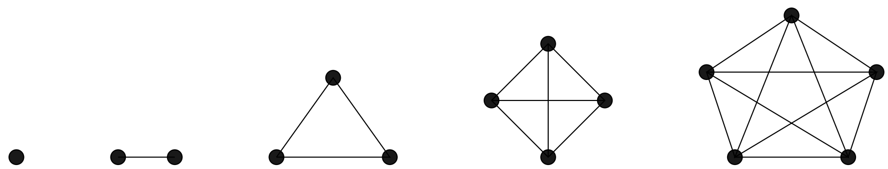


Figure 2.2: Small complete graphs: K_1, K_2, K_3, K_4, K_5

For positive integers m and n, the notation mK_n is occasionally used to denote m vertex disjoint copies of K_n.

A graph H is a (weak) subgraph of G if $V(H) \subseteq V(G)$ and $E(H) \subseteq (E(G) \cap [V(H)]^2)$ and is an *induced subgraph* if $E(H) = (E(G) \cap [V(H)]^2)$. If G is a graph, a *spanning subgraph* of G is a subgraph H with $V(H) = V(G)$. If the term "subgraph" is used, it is usually taken to mean "weak subgraph". A graph H is often said to be a subgraph of G if there exists a subgraph F of G so that $H \cong F$.

The *complement* of graph G, denoted $\overline{G}$, is the graph on $V(G)$, where for every $x, y \in V(G)$, $\{x, y\} \in E(\overline{G})$ if and only if $\{x, y\} \notin E(G)$.

A graph G is called *bipartite* if and only if there exists a partition $V(G) = A \cup B$ (where A and B are non-empty and $A \cap B = \emptyset$) of its vertex set so that every edge in G is of the form $\{x, y\}$, where $x \in A$, $y \in B$. The sets A and B are called *partite sets*. For example, K_3 is not bipartite. The *complete bipartite graph* $K_{a,b}$ is a graph with a partition $V(K_{a,b}) = A \cup B$, where $|A| = a$, $|B| = b$, and $E(K_{a,b}) = \{\{a, b\} : a \in A, b \in B\}$. Thus, $K_{a,b}$ has $a + b$ vertices and ab edges. (See Figure 2.4 for an example.) Many authors draw bipartite graphs with the partite sets beside each other (as in Figure 2.3). and some draw bipartite graphs with one partite set above the other (as in Figure 2.4).

If $G = (V, E)$ is a graph, a subset $X \subset V$ is called an *independent set* if and only if the subgraph of G induced by X has no edges. In other words, if $[X]^2$ denotes the 2-element subsets of X, then $X \subseteq V$ is *independent* if and only if $[X]^2 \cap E = \emptyset$. Note that the partite sets in a bipartite graph are

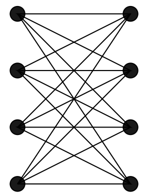


Figure 2.3: *Standard drawing of $K_{4,4}$.*

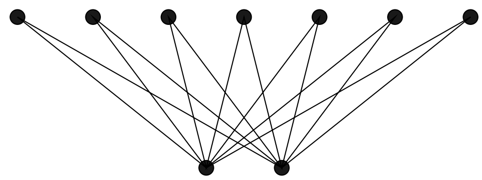


Figure 2.4: $K_{2,7}$*, drawn vertically.*

independent sets.

For a graph G and a vertex $x \in V(G)$, the *neighbourhood of* x in G is

$$N(x) = N_G(x) = \{y \in V(G) : \{x, y\} \in E(G)\}.$$

For any vertex $x \in V(G)$, the *degree* of x is the number of edges containing x. [This definition also holds for multigraphs and hypergraphs.] The degree of a vertex x in G is denoted by $\deg_G(x)$ or $d_G(x)$, or when clear, simply $d(x)$ or $\deg(x)$. For simple graphs, $d_G(x) = |N_G(x)|$. A vertex x with $\deg(x) = 0$ is said to be an *isolated vertex*.

Since each edge of a graph contributes to counting two degrees, the following lemma has a direct proof, but also has an easy inductive proof.

Lemma 2.1.2 (Handshaking lemma)**.** *For any graph G,*

$$\sum_{x \in V(G)} \deg(x) = 2|E(G)|.$$

The smallest degree in a graph G is denoted by $\delta(G)$ and the largest degree by $\Delta(G)$. If every vertex in a graph G has degree k, then G is called k*-regular* (in which case $\delta(G) = \Delta(G)$).

For a non-negative integer m, a *walk* of length m in a graph G is a sequence of vertices (not necessarily distinct) $w_0, w_1, w_2, \ldots, w_m$ so that for each $i = 0, \ldots, m-1$, $\{w_i, w_{i+1}\} \in E(G)$. A walk on vertices $w_0, w_1, \ldots, w_m$

is called *closed* if and only if $w_0 = w_m$. A *trail* is a walk with no edge repeated and a *path* is a walk with no vertex repeated (and hence no edge repeated). A path with k edges (and hence $k+1$ vertices) is denoted by P_k, and is said to be of length k. [Note: this terminology varies, as many use P_{k+1} to denote a path with k edges.]

In a graph, a *cycle* is a closed walk on at least three vertices with no vertex repeated (except the first and last). Since no vertex in a cycle is repeated, neither can any edge be repeated, so a cycle is an example of a closed trail. The cycle with k vertices is denoted by C_k. A graph with no cycles is called *acyclic*. The *girth* of a graph G is the length of a shortest cycle in G (if there are no cycles, the girth is said to be infinite) and is often denoted by $g(G)$, or simply girth(G).

A graph is *connected* if and only if there is a path between every pair of vertices. A *component* of a graph is a maximal connected subgraph.

Definition 2.1.3. A *tree* is a connected acyclic graph.

By a simple induction argument, a tree on n vertices has $n-1$ edges. In a tree, between any two vertices, there exists a unique path joining them (use a simple proof by contradiction—if there are two paths, get a cycle). Another simple proof (by induction) shows that all trees are bipartite.

2.2 Colouring graphs

Definition 2.2.1. For a positive integer k and a graph G, a *good vertex k-colouring* (or *proper vertex k-colouring*) of G is a function $c : V(G) \to \{1, 2, \ldots, k\}$ so that for any $\{x, y\} \in E(G)$, $f(x) \neq f(y)$. A graph is called vertex *k-colourable* iff there exists a good k-colouring of G.

Elements in the set $\{1, 2, \ldots, k\}$ above are called colours; in fact, any k-element set can form the set of colours; when k is small, often the colours are given names like "red" and "blue". For each $i \in \{1, 2, \ldots, k\}$, the set of vertices $c^{-1}(i)$ is called the ith colour class. Each colour class forms an independent set of vertices.

Definition 2.2.2. For a graph G, the *chromatic number* of G, denoted $\chi(G)$, is the least k for which G is k-colourable.

For example, for $n \geq 1$, $\chi(K_n) = n$, $\chi(C_{2n+1}) = 3$, and $\chi(\overline{K_n}) = 1$. Also, for $n \geq 2$, $\chi(C_{2n}) = 2$. All non-trivial bipartite graphs have chromatic number 2, and so for any tree T on at least two vertices, $\chi(T) = 2$.

There also many theorems for colouring *edges* of graphs; here is a brief introduction.

Definition 2.2.3. For any graph (or multigraph), define the *edge-chromatic number* $\chi'(G)$ (or *chromatic index*) to be the minimum number k of colours so that there exists a colouring $c : E(G) \to [k]$ so that no two incident edges (at a vertex) get the same colour.

Since there exists a vertex with $\Delta(G)$ edges incident,

$$\Delta(G) \leq \chi'(G). \tag{2.1}$$

The edge-chromatic number is always one of only two possible values:

Theorem 2.2.4 (Vizing,1964 [902])**.** *For any simple graph G,*

$$\Delta(G) \leq \chi'(G) \leq \Delta(G) + 1.$$

For a proof of Vizing's theorem, see any of many popular texts, including, e.g., [117, pp. 153–154].

2.3 Planar graphs

A graph G is called *planar* if and only if it can be drawn in the Euclidean plane with no edges crossing; such a drawing is called a *plane drawing* (or *planar embedding*) of G. For example, trees and cycles are planar. The complete graph K_4 is also planar (see Figure 2.5).

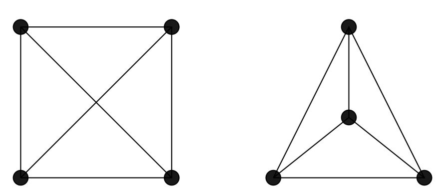


Figure 2.5: The standard drawing of K_4 and a plane drawing

It is known that both $K_{3,3}$ and K_5 are not planar (see Theorems 2.3.4 and 2.3.5).

One large class of planar graphs are "graphs of polyhedra". If P is a polyhedron, form the graph of P to be the graph $G = (V, E)$ where V is the set of vertices of P and two vertices are adjacent in G if and only if there is an edge of P joining these two vertices. The graph of a polyhedron is seen to be planar by considering a type of "projection". Imagine that the edges of a polyhedron are elastic. Stretching out the edges of one face, the resulting figure can be laid flat with no crossing edges.

The five platonic solids (tetrahedron, cube, octahedron, icosahedron, dodecahedron, see Figure 2.6) give rise to regular (all vertex degrees the same) graphs.

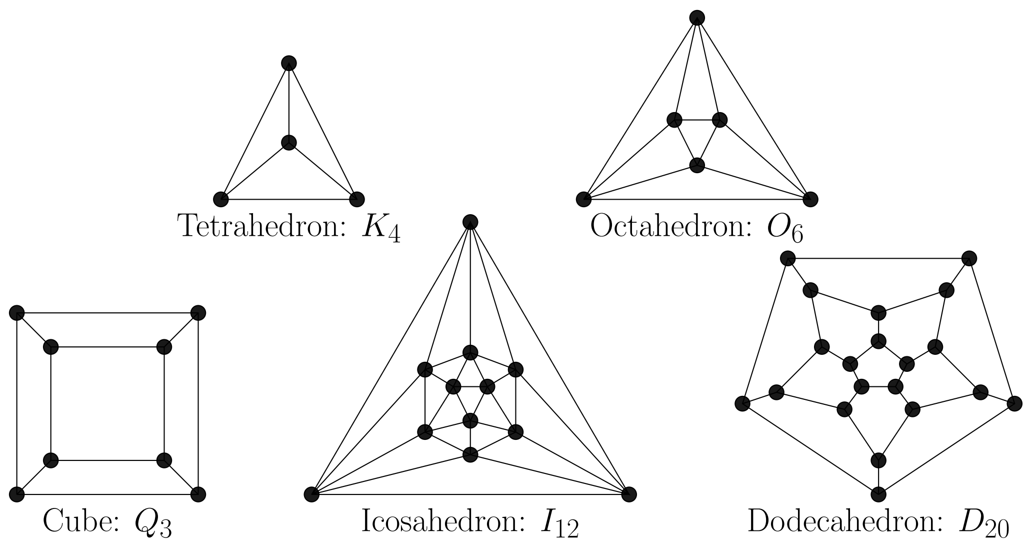


Figure 2.6: Planar graphs of the 5 Platonic Solids

If G has a plane drawing in $\mathbb{R}^2$, deleting all points in the drawing leaves connected regions called *faces* of G; the outside infinite region is counted as a face.

Theorem 2.3.1 (Euler's formula for planar graphs). *Let G be a connected planar graph, with v vertices and e edges. Then in any plane drawing of G, the number f of faces is the same, namely $f = e + 2 - v$, or more commonly written,*

$$v + f = e + 2. \tag{2.2}$$

Proof: The proof given here is by induction on the number of edges. For each $e \geq 0$, let $S(e)$ be the statement that any connected planar graph with e edges and v vertices has precisely $e + 2 - v$ faces.

Base step: When $e = 0$, the only connected planar graph is a single vertex. In this case, $v = 1$, and there is only 1 face. Then $S(0)$ says this graph has $e + 2 - v = 0 + 2 - 1 = 1$ face, which is true. As an additional check, when $e = 1$, the only connected planar graph is a single edge on $v = 2$ vertices in which case there is only 1 face. Thus $S(1)$ says $1 = 1 + 2 - 2$, which is correct. So both $S(0)$ and $S(1)$ hold.

Induction step: Let $e \geq 2$, and suppose that $S(e - 1)$ is true. Let G be a planar graph with e edges, and let $v = |V(G)|$.

If G is a tree (connected, acyclic), then G has only one face and $v - 1$ edges, in which case $e + 2 - v$ becomes $v - 1 + 2 - v = 1$, verifying $S(e)$.

Next, suppose that G is not a tree; since G is connected, G contains a cycle. Let $\{x, y\}$ be an edge on a cycle. Produce the subgraph $G' \subseteq G$ on $V(G)$ by deleting the edge $\{x, y\}$ (but do not delete the vertices). In some plane drawing of G, let f be the number of faces. In G', the two faces in G on either side of $\{x, y\}$ are joined in G'. So G has one more face than G'. Since G' has $e - 1$ edges, by $S(e - 1)$, the number of faces in G' is $(e - 1) + 2 - v$, and so G has $e + 2 - v$ edges, confirming $S(e)$ and completing the inductive step.

By mathematical induction, for each $e \geq 0$, $S(e)$ holds. □

Theorem 2.3.2. *If a planar graph has v vertices, f faces, e edges, and k components, then $v + f = e + k + 1$.*

Proof: The proof is by induction on k. Since any graph has a non-empty vertex set, $k \geq 1$.

Base step: The base case $k = 1$ is true by Euler's formula for connected planar graphs (Theorem 2.3.1).

Inductive step: Let $\ell \geq 1$ and suppose that the formula holds for a planar graph with ℓ components. Consider a planar graph G with $\ell + 1$ connected components, say $C_0, C_1, \ldots, C_\ell$. Let G' be the graph obtained from G by removing C_0. If G' has v' vertices, e' edges and f' faces, then by the induction hypothesis,

$$v' - e' + f' = \ell + 1.$$

If the single component C_0 has v_0 vertices, e_0 edges and f_0 faces, then by Euler's formula (for a connected planar graph),

$$v_0 + f_0 = e_0 + 2.$$

Then G has $v = v' + v_0$ vertices, $e = e' + e_0$ edges, and $f = f' + f_0 - 1$ faces (since the infinite face of G' is a face of C_0), and hence

$$\begin{aligned} v - e + f &= (v_0 - e_0 + f_0) + (v' - e' + f') - 1 \\ &= 2 + (\ell + 1) - 1 && \text{(by Euler's formula and IH)} \\ &= (\ell + 1) + 1, \end{aligned}$$

proving that the formula holds for graphs with $\ell + 1$ components, completing the inductive step.

Therefore, by mathematical induction, for every $k \geq 1$, the result holds for all graphs with k components. □

Exercise 141. *Does there exist a planar graph on $v = 6$ vertices with $e = 6$ edges and $f = 3$ faces? If you can not give an example, prove why no such example exists.*

The next lemma gives an upper bound on the number of edges in a planar graph, and is used in many other theorems regarding planar graphs.

Lemma 2.3.3. *If G is a connected planar graph on $v \geq 3$ vertices with e edges, then*

$$e \leq 3v - 6. \tag{2.3}$$

Proof: Let G be a connected planar graph on $v \geq 3$ vertices, with e edges and f faces. The number k of edge-face incidences satisfies $k \leq 2e$ since each edge is adjacent to at most two faces. On the other hand, since every face has at least 3 edges, $3f \leq k$. Thus

$$3f \leq 2e. \tag{2.4}$$

Multiplying Euler's formula by 3 gives $3v + 3f = 3e + 6$, and using (2.4) yields

$$3v + 2e \geq 3e + 6,$$

from which the result follows. □

Exercise 142. *Show that Lemma 2.3.3 holds even for planar graphs that are not necessarily connected but whose every component has at least 3 vertices. Does this result extend to simply* any *planar graph?*

Theorem 2.3.4. *The complete graph K_5 is not planar.*

Proof: In K_5, there are $\binom{5}{2} = 10$ edges, so put $v = 5$, and $e = 10$; these values violate Lemma 2.3.3. (Another proof is also available—draw C_5 as a planar cycle and examine the possible placement of the other five edges.) $\square$

Is the graph $K_{3,3}$ (see Figure 2.7) planar? (The graph $K_{3,3}$ is sometimes called the "Thomsen graph".)

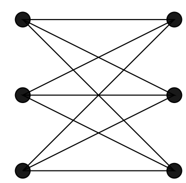


Figure 2.7: Standard drawing of $K_{3,3}$

Theorem 2.3.5. *The complete bipartite graph $K_{3,3}$ is not planar.*

Proof: Suppose, in hopes of a contradiction, that $K_{3,3}$ is planar, with $v = 6$, $e = 9$. Euler's formula $v + f = e + 2$ then becomes $6 + f = 9 + 2$ and so $f = 5$. Since $K_{3,3}$ is bipartite, any shortest cycle has length at least 4, and therefore each face has at least 4 edges. Counting face-edge incidences, the equation analogous to (2.4) is $4f \leq 2e$, which yields $20 \leq 18$, a contradiction. [Note: Again, a direct proof is also possible; draw C_6 and try to place the remaining three edges to produce $K_{3,3}$.] $\square$

Exercise 143. *Find (with proof) a formula analogous to (2.4) for planar graphs whose smallest cycle length (girth) is g. For simplicity, assume that each edge lies on some cycle.*

Exercise 144. *Let G be a connected planar (simple) graph with $v \geq 3$ vertices, e edges and girth g. Suppose that every edge of G is on at least one cycle. Prove that*

$$e \leq \frac{g(v-2)}{g-2}.$$

The result in Exercise 144 also applies to planar graphs having edges that are not on any cycle (like bridges or tree branches attached), but a little more care is needed in the proof for these cases.

Lemma 2.3.6. *Every planar graph contains a vertex of degree at most 5.*

Proof: Let G be planar. One need only consider when G is connected. The proof is by contradiction. If every vertex has degree at least 6, then by the handshaking lemma (Lemma 2.1.2) and Lemma 2.3.3,

$$6v \leq \sum_{x \in V(G)} \deg(x) = 2e \leq 2(3v - 6),$$

from which the desired contradiction is obtained. □

Exercise 145. *Show that a connected planar graph G on at least 3 vertices has at least 3 vertices with degree at most 5.*

Exercise 146. *Show that a connected planar graph G on at least 4 vertices has at least 4 vertices with degree at most 5.*

Exercise 147. *Either find a planar graph with 8 vertices, 10 faces, and 19 edges, or prove that one does not exist.*

Exercise 148. *Either find a planar graph that has each vertex incident with 4 faces, and each face has four edges, or prove that such a graph does not exist.*

The next two theorems characterize planar graphs. The first was proved by Kazimierz Kuratowski (1896–1980), a Polish topologist (see [184, p. 237] for a discussion of other authors who also proved the same theorem, some perhaps earlier).

Recall that a subdivision of a graph G is one obtained by inserting vertices (of degree 2) into edges of G. (Technically, if one does not insert any new vertices, still say that G is a subdivision of G.)

Theorem 2.3.7 (Kuratowski, 1930 [583])**.** *A graph G is planar if and only if G contains no subgraph that is a subdivision of either $K_{3,3}$ or K_5.*

The next theorem was proved by Klaus Wagner (1910–2000). An *edge-contraction* in a graph G is achieved by identifying two adjacent vertices and removing any multiple edges or loops thereby formed. A graph H is a *minor* of G if and only if H can be obtained by a sequence (in any order) of edge removals, vertex removals, or edge-contractions in G.

Theorem 2.3.8 (Wagner, 1937 [908])**.** *A graph G is planar if and only if G contains no minor isomorphic to either $K_{3,3}$ or K_5.*

2.4 Colouring planar graphs

Using Lemma 2.3.6, the next exercise has almost an immediate solution by induction on the number of vertices of a planar graph.

Exercise 149. *Show that any planar graph G is 6-colourable, i.e., $\chi(G) \leq 6$.*

The famous "four colour conjecture" for map colourings said that any map with contiguous countries can be coloured with four colours so that neighbouring countries are coloured differently. This conjecture is now a theorem, first proved in 1977 by Appel, Haken, and Koch (see [31], [32] or see [929] for a description of the problem and its history); it was proved again in 1997 by Neil Robertson, Daniel Sanders, Paul Seymour, and Robin Thomas [756] with far fewer cases. Another substantial (but earlier) reference for the four colour theorem (4CT) is a book by Oystein Ore [691].

If G is a planar graph with a specific plane drawing, its *planar dual*, denoted G^*, is the graph whose vertices are regions (faces) of G and two vertices are adjacent if and only if their respective regions share a common edge. Another definition for the dual of planar graph allows repeated edges: the *multigraph dual* of a planar graph G is a multigraph, where there is an edge between vertices in the dual for every edge on the border of two faces in G. However, a proper colouring for one type of dual is also a proper colouring of the other kind, so this difference in definitions is not restrictive when it comes to colouring. In either definition, the dual of a planar graph is again planar.

The result in the next exercise is an observation that was likely known before Euclid, but from a geometric point of view.

Exercise 150. *Verify that if the platonic solid graphs listed in Figure 2.6 are represented by T, O, C, I, D, then the respective duals are T, C, O, D, I respectively.*

One can verify the result in Exercise 150 in 3 dimensions, too, where faces and vertices merely interchange roles. For example, the faces and vertices of a cube have the same pattern of incidences as the vertices and faces (respectively) of the octahedron, and so are called (geometric) duals of one another.

If a planar graph has two different plane drawings, the multigraph duals of each need not be isomorphic, as seen in the following example, which can be found in [18, p. 265].

Exercise 151. *Show that the images in Figure 2.8 are different plane drawings of the same planar graph. Using the alternate definition of planar dual that gives a multigraph (one edge in the dual for every edge in G that separates two regions), show that these two drawings have non-isomorphic multigraph duals.*

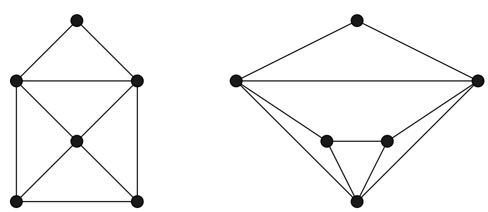


Figure 2.8: Isomorphic graphs with different plane drawings and different mulitigraph duals.

A proper map-colouring of (the faces of) some planar graph G corresponds to a proper vertex colouring of G^*. Using duals and the 4 colour theorem gives:

Theorem 2.4.1 (4CT for planar graphs)**.** *If G is a planar graph, then $\chi(G) \leq 4$.*

Exercise 152. *Let G be a connected planar graph with all vertex degrees even. Prove that for any planar drawing of G, the faces can be 2-coloured so that faces sharing a common edge receive different colours.*

Exercise 153. *Let G be a connected planar (simple) graph with a drawing so that every face is a triangle, and $\chi(G) = 3$. Show that the faces can be 2-coloured so that faces sharing a common edge are coloured differently.*

Exercise 154. *Find an example of a triangulated planar graph (every face is a triangle) that is not 3-colourable.*

For more on colouring planar graphs, see Section 15.5.

Chapter 3

Three dimensional geometry

This chapter contains only a very few basic facts; many of these facts are common knowledge in grade school, some follow easily by techniques learned in a first year linear algebra course (with vector spaces), and some are mentioned in a first year calculus course. Many formulae for volume can be derived using simple integrals. Hence, the coverage here is very selective and is given with the intent of only being a reminder of elementary facts.

3.1 Basic volume formulae

The volume of a $\ell \times w \times h$ rectangular box (a parallelepiped with all right angles) is ℓwh. In general, the volume of a parallelepiped is given by the scalar triple product (see Lemma 23.2.6; these calculations are repeated in Section 3.4.2).

The volume of a (right) circular cylinder with radius r and height h is $\pi r^2 h$. In fact, the cylinder need not be a right cylinder, as long as the height is measured perpendicular to the base. This result is a special case of a more general result: let S be a subset of xy-plane in $\mathbb{R}^3$, and for some fixed $\mathbf{v} \in \mathbb{R}^3$, define the cylinder

$$C = \{\mathbf{s} + \epsilon\mathbf{v} : \mathbf{s} \in S, 0 \leq \epsilon \leq 1\}.$$

Then the volume of C is the area of S times the altitude of C (measured perpendicular to S, that is, the length of the projection of $\mathbf{v}$ onto a normal of S). The proof of this fact follows by embedding S is in the xy-plane and integrating with respect to the z coordinate (Cavalieri's principle).

In the above description of a cylinder, if S is a polygon, the resulting cylinder is called a *prism* and if $\mathbf{v}$ is orthogonal to S, the shape is a right prism (often, the word "prism" denotes a right prism).

The volume of a simple cone (its vertex lies directly above the center of the base) whose base is a circle with radius r and height h is $\frac{1}{3}\pi r^2 h$. In fact, with some simple calculus, one can prove that for any planar set A with area a, and any vertex off the plane at height h (measured at right angles to the plane containing A), the volume of the cone thereby generated is $\frac{1}{3}ah$. Although it is not needed elsewhere in this text, the interested reader can verify that the area of a simple cone with radius r and height h is

$$\pi r^2 h + \pi r\sqrt{r^2 + h^2}.$$

(The first term above is the area of the base; the second term is the "lateral area", which is found by cutting the side, and laying it flat, forming a "sector" of a circle of radius $\ell = \sqrt{r^2 + h^2}$. This "sector" occupies a $\frac{2\pi r}{2\pi \ell} = \frac{r}{\ell}$ portion of the circle, and so has area $\frac{r}{\ell} \cdot \pi \ell^2 = \pi r \ell$.)

Using simple calculus (given in Theorem 3.4.3) a sphere with radius r has volume $\frac{4}{3}\pi r^3$. Calculus also easily shows that a sphere with radius r has area $4\pi r^2$.

I learned the remarkable fact in the next exercise in a lecture from David Ford at Emory University in 1991.

Exercise 155. *Let A be a cup formed by a solid cylinder of height r and radius r and removing a cone (with the same height and radius). Let H be half of a ball with radius r. First show that the A and H have the same volume. Then show that both A and H can be balanced on a rod (joining the bottom center of A and the north pole of H) at the rod's midpoint because both have identical cross section areas.*

The following is also an exercise in calculus, which is left to the reader.

Theorem 3.1.1. *Let S be a unit sphere in $\mathbb{R}^3$ centered at the origin. For $z_0 \in [0, 1]$, the area of the spherical cap*

$$\{(x, y, z) \,|\, x^2 + y^2 + z^2 = 1,\ z \geq z_0\}$$

is $2\pi(1 - z_0)$.

3.2 Volumes of revolution

The following is sometimes called Pappus's theorem; however, it may not be his most famous theorem (see Theorem 1.8.1).

Theorem 3.2.1 (Pappus's theorem)**.** *The volume of a solid of revolution is equal to the product of the area of the revolving region times the distance through which the center of mass is rotated.*

The interested reader might confirm Pappus's theorem for the volume of torus:

Exercise 156. *Let $R \geq r > 0$ be real numbers and let $C = \{(x, 0, z) : (x - R)^2 + z^2 = r^2\}$ be a circle in the xz-plane. Rotating C about the z-axis gives a torus with volume $2\pi R \cdot \pi r^2$. Prove this result using either integrals (either by washers or by cylindrical shells) or by Pappus's theorem (Theorem 3.2.1).*

3.3 Polyhedra

3.3.1 Introduction

There seems to be no standard definition of a polyhedron (plural, polyhedra). The word "polyhedron" comes from Greek for "many base", perhaps with the idea that a polyhedron is a shape with flat sides (faces) that can be used as a base. However, many polyhedra are not convex (they could have indentations or holes), and so special care is needed when "polyhedron" is meant to include these situations.

One definition of "polyhedron" might be: a solid shape in 3 dimensions whose faces are (simple) polygons.

For example, the five Platonic solids (tetrahedron, cube, octahedron, icosahedron, dodecahedron), pyramids, cylinders with polygonal bases (prisms), antiprisms, Archimedean solids (there are 13— see Table 3.3.3 for complete list), and parallelepipeds and the rhombic triacontahedron (see Figure 3.3.1) are all convex polyhedra.

There are countless other polyhedra, many with special names and properties. For example, see George Hart's website "Virtual polyhedra" [459] for numerous examples.

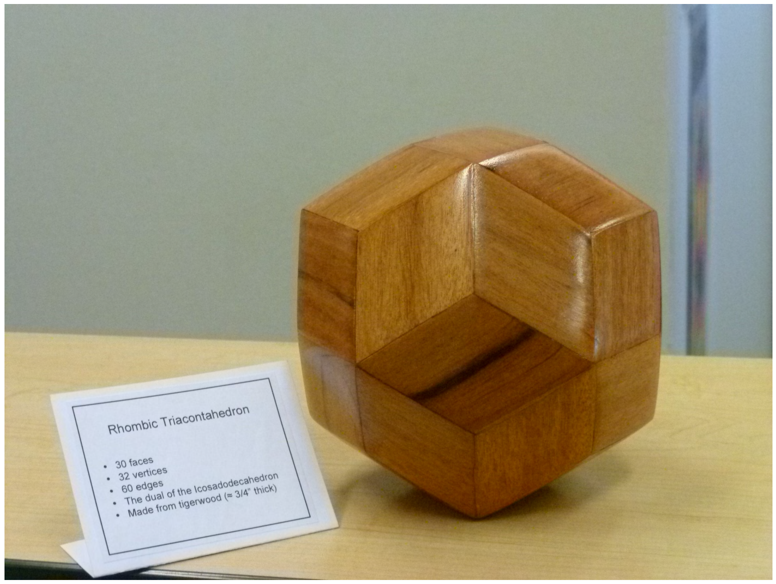


Figure 3.1: Rhombic triacontahedron, made from tigerwood, DSG, 2006?

An L-shaped trominoe is not convex, nor is a prism whose base is a star. The Kepler–Poinsot solids (see Figure 3.2) are also not convex. Some authors allow polyhedra to have faces that intersect, as in the great dodecahedron where it can be seen as pentagons cutting through each other. However, for present purposes, the great dodecahedron can be seen simply as a (non-convex) polyhedron with triangular faces.

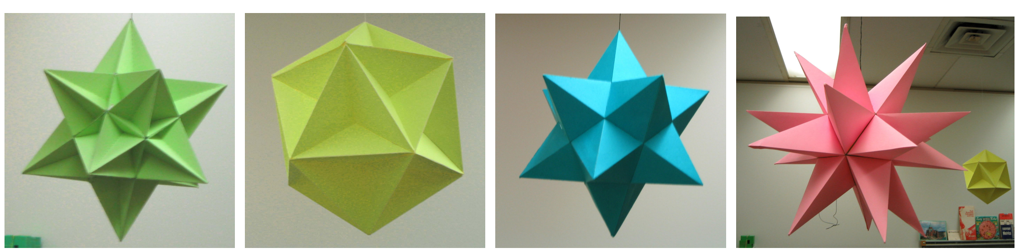


Figure 3.2: The Kepler–Poinsot solids: great icosahedron, great dodecahedron, small stellated dodecahedron, great stellated dodecahedron (paper models by DSG, 2003)

If $G = (V, E)$ is a graph, a *spanning subgraph* of G is a subgraph H with $V(H) = V(G)$. Recall that a *tree* is a connected acyclic graph. A *spanning* *tree* of a graph G is simply a spanning subgraph that is a tree. Any connected graph has a spanning tree.

Theorem 3.3.1 (Barnette, 1966 [60])**.** *The graph of a convex polyhedron has a spanning tree with maximum degree 3.*

See also the *Handbook of convex geometry* [421].

A connected graph is called *k-connected* if and only if the deletion of any fewer than k vertices does not disconnect the graph. (So trees are 1-connected but not 2-connected, and cycles of length at least 4 are 2-connected.)

Theorem 3.3.2 (Steinitz, 1922 [827])**.** *The graphs of (3-dimensional) convex polyhedra are precisely the 3-connected planar graphs.*

For more on Steinitz's theorem, see [940, Ch. 4].

One class of polyhedra that is most easily dealt with includes those that are homeomorphic to a sphere—in other words, those that can be continuously deformed into a sphere. Polyhedra that are excluded from this class include those with "holes" or faces that have a hole (a face with hole is often called *annular* since an annulus is a ring with a hole.).

Two famous polyhedra with holes are the Császár polyhedron, invented by the Hungarian Ákos Császár in 1949 [229], and the Szilassi polyhedron found by the Hungarian Lajos Szilassi sometime before 1983 (see [853], where the polyhedron is discussed, only citing his 1983 paper [852] in Hungarian—popular sources put the year at 1977, but I have not confirmed this.) Both are given in Figure 3.3. These two polyhedra have attracted much attention in literature (see, e.g., [525]).

The Császár polyhedron has 7 vertices, 14 triangular faces, and every pair of vertices is joined by an edge; so there are $\binom{7}{2} = 21$ edges. The graph of the Császár polyhedron is K_7 and so shows one way to embed K_7 in a torus. (For more on embedding K_7 in a torus, see also [114], [626] and Figure 17.9.) The Szilassi polyhedron has 7 hexagonal faces, and any two faces share a common edge. These two polyhedra are duals of one another. The graph of the Szilassi polyhedron is the Heawood graph (see Figure 11.3), the incidence graph of the Fano plane (see Figure 11.2), embedded in a torus. The Szilassi polyhedron and the tetrahedron are the only two known polyhedra whose each face shares an edge with every other face.

Theorem 3.3.3 (Euler's formula)**.** *If a polyhedron with no holes or annular faces has v vertices, e edges, and f faces, then*

$$v + f = e + 2.$$

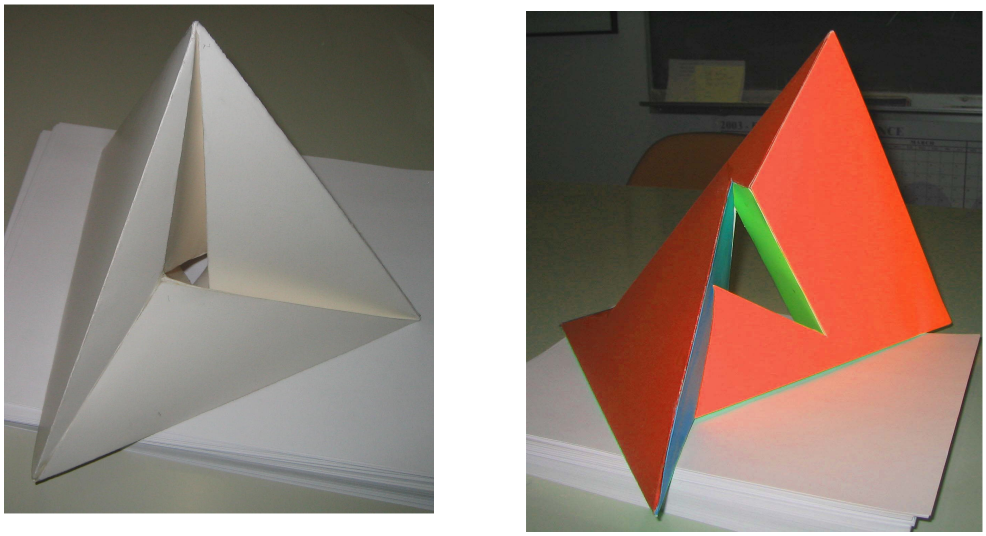


Figure 3.3: The Császár and Szilassi polyhedra

Proof: There are many proofs of Euler's formula for polyhedra. The simplest proof is to form a planar graph for each polyhedron, and then apply Theorem 2.3.1, the analogous formula for planar graphs. This graph theory approach is not only simple, it answers a slightly more general problem since not every planar graph gives rise to a polyhedron (*e.g.*, a planar graph can have vertices of degree 0, 1, or 2, and graphs of polyhedra have no such vertices).

Inductive proofs using actual polyhedra can be tricky, since in such proofs, it seems that all simple polyhedra need to be inductively constructed. There are two approaches that seem to work. The base case for each is the tetrahedron.

First let P be a polyhedron with only triangular faces. The number of edges per face is 3, and each edge is on 2 faces, and so counting all edge-triangle pairs, $2e = 3f$. Multiplying Euler's formula by 3 and making this replacement shows that it suffices to prove for triangle polyhedra (also called *deltahedra*), that $e = 3v-6$, and this can be done by induction. For example, adding a new vertex to the center of any face, and joining it to the three corners of the triangle, increases the number of vertices by 1, the number of edges by 3, and the number of faces by 2, and so each side of $v + f = e + 2$ is increased by 1+2=3. But there are other ways to create a triangulated polygon; an analysis is required for each case. To get the formula for a general polyhedron, subdivide each face into triangles, and observe that again Euler's formula is preserved upon each successive subdivision; induction from the triangulated polyhedron back to the general polyhedron does the rest.

For a second geometric approach, use "truncation", a process of cutting off a corner of a polyhedron to get, if possible, a polyhedron with one less vertex (and so a proof is by induction on the number of vertices). For truncation to work one needs a small observation. Suppose that v is a vertex of a polyhedron P, and let $v_1, \ldots, v_d$ be those $d \geq 3$ vertices for which $\{v, v_i\}$ is an edge of P. If P is convex, then the v_i's can be shifted so that they all lie on one plane (while preserving all combinatorial properties of P). Then removal of v together with all edges touching v produces another polyhedron P' with $d-1$ fewer faces (d old faces are gone, and one new face with vertices $v_1, \ldots, v_d$ is added), d fewer edges, and 1 fewer vertex, and so if Euler's formula is true for P', then it is true for P. $\square$

Euler's formula for polyhedra was also generalized to higher dimensions (see Section 6.6). The assumption in Theorem 3.3.3 that the polyhedron need be homeomorphic to a sphere is not always required. For example, consider a square block with a square hole through it; in this case, $v = 16$, $f = 10$, and $e = 24$ (however, two of these faces are annuli).

There is also a formula for polyhedra with holes or annular faces (see Exercise 172). For example, for any polyhedron homeomorphic to a torus,

$$v + f = e.$$

For example, the Császár polyhedron (see Figure 3.3) has 7 vertices, 14 triangular faces, and 21 edges; its dual, the Szilassi polyhedron, has 14 vertices, 7 hexagonal faces, and 21 edges.

See Section 17.3.9 for an embedding of K_7 in a torus with $v = 7$, $e = 21$, and $f = 14$ (that section also contains some references for embedding graphs on surfaces like the torus).

3.3.2 Platonic solids

A *Platonic solid* is a convex polyhedron whose every face is the same regular polygon, and at each vertex, the same number of polygons meet. Some authors call a Platonic solid a *regular solid* or a *perfect solid*. It was known from 2000 BC in neolithic Scotland that five such solids exist (a set made from stone was found), and these five solids were known to Plato and Euclid. The names of these solids are tetrahedron, octahedron, cube, icosahedron, and dodecahedron (in order of size when all have same edge length). See Figure 3.4.

Figure 3.4: Tetrahedron, octahedron, cube, icosahedron, and dodecahedron; open faced models of the Platonic solids by DSG, first attempt, roughly based on drawings by da Vinci from Pacioli's *De divina proportione*

Using Euler's formula (Theorem 3.3.3) one can show that only five Platonic solids exist.

Theorem 3.3.4. *There are only five platonic solids, namely the tetrahedron, octahedron, icosahedron, cube, and dodecahedron.*

Proof: Consider a polyhedron P formed by regular s-gons, $(s \geq 3)$ with $d \geq 3$ such s-gons meeting at a vertex. There are only five choices for the pair (s, d), and this fact is seen by looking at how polygons can surround a point in the plane.

If $s = 3$, then since 6 equilateral triangles can meet at a point in the plane, $d \in \{3, 4, 5\}$ are the only possibilities.

If $s = 4$, since 4 squares surround a point in the plane, $d = 3$ is the only choice.

If $s = 5$, only three regular pentagons (each with internal angle $108°$) can surround a point (with a little room to spare), so $d = 3$ is the only possibility.

Thus, $(s, d) \in \{(3, 3), (3, 4), (3, 5), (4, 3), (5, 3)\}$. The amazing (to me at least!) thing is that for each pair (s, d), there is precisely one polyhedron satisfying the constraints. This is proved by Euler's formula and some "double counting" as follows: Let P have v vertices, f faces, e edges, (with s sides per face, and d s-gons meeting at a vertex). Note that d is also the number of edges that meet at any vertex (called the *degree* of a vertex).

Count the number of pairs (vertex, edge), where the vertex and edge are incident. (Such a pair is often called a "flag", but this terminology is not used here.) Counting from the vertices, since each vertex is incident with d edges, there are vd pairs. Counting from the edges, since each edge is incident with two vertices, there are $2e$ such pairs (this is a special case of the "handshaking lemma", Lemma 2.1.2). Thus,

$$vd = 2e. \tag{3.1}$$

Similarly, count the pairs (edge, face), where the edge is on the border of the face. Since each edge is on the border of two faces, the number is $2e$. Since each face has s edges on its border (sometimes this number s is called a "face-degree"), the number of such pairs is fs. Thus,

$$fs = 2e. \tag{3.2}$$

Putting $v = \frac{2e}{d}$ and $f = \frac{2e}{s}$ into Euler's formula $v + f = e + 2$ (see Theorem 3.3.3),

$$\frac{2e}{d} + \frac{2e}{s} = e + 2.$$

Multiplying through by $\frac{1}{2e}$,

$$\frac{1}{d} + \frac{1}{s} = \frac{1}{2} + \frac{1}{e}.$$

The key idea is that for each pair (s, d), the number e of edges is uniquely determined. In fact, one can solve for e:

$$\frac{1}{e} = \frac{s + d}{sd} - \frac{1}{2} = \frac{2s + 2d - sd}{2sd},$$

and so

$$e = \frac{2sd}{2s + 2d - sd}.$$

Case 1: When $s = 3$ and $d = 3$, get $e = 6$; by equation (3.1), get $v = 4$ and by equation (3.2), get $f = 4$. These parameters agree with those for the tetrahedron.

Case 2: When $s = 3$ and $d = 4$, get $e = 12$; by equation (3.1), get $v = 6$ and by equation (3.2), get $f = 8$. These parameters agree with those for the octahedron.

Case 3: When $s = 3$ and $d = 5$, get $e = 12$; by equation (3.1), get $v = 12$ and by equation (3.2), get $f = 20$. These parameters agree with those for the icosahedron.

Case 4: When $s = 4$ and $d = 3$, get $e = 12$; by equation (3.1), get $v = 8$ and by equation (3.2), get $f = 6$. These parameters agree with those for thecube.

Case 5: When $s = 5$ and $d = 3$, get $e = 12$; by equation (3.1), get $v = 20$ and by equation (3.2), get $f = 12$. These parameters agree with those for the dodecahedron.

To prove that each of these shapes is unique (for the given v, f, e, d, s), one need only experiment with putting together pieces and retaining convexity. Note that convexity is critical, since otherwise, one could put a dent in the icosahedron. □

Name	v	f	e	s	d
Tetrahedron	4	4	6	3	3
Cube	8	6	12	4	3
Octahedron	6	8	12	3	4
Icosahedron	12	20	30	3	5
Dodecahedron	20	12	30	5	3

Table 3.1: Table for Platonic solids

It might be interesting to observe that four vertices of a cube form a regular tetrahedron, as in Figure 3.5.

Note that the embedding from Figure 3.5 can be done in only two ways. Another lesser known fact is that eight vertices of a regular dodecahedron determine vertices of a cube (see Exercise 165 for the number of ways this can be done). Can you see how?

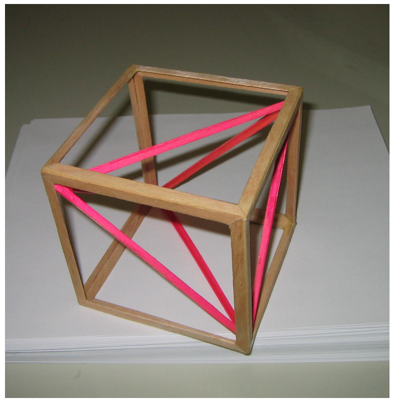


Figure 3.5: A tetrahedron in a cube, made from coffee stir sticks, 2003

3.3.3 Archimedean solids

An Archimedean solid is a convex polyhedron whose faces are all regular polygons (all with the same edge length), not all the same, so that at each vertex, the pattern of polygons surrounding the vertex is the same. It is generally said that there are 13 Archimedean solids, but two of these (the snub solids) also have different mirror reflections (called "enantiomorphs"). See Figure 3.6 for pictures of each (only one version of each of the snub solids) and Table 3.3.3 for details on each (given in the same order as in Figure 3.6).

To prove that only 13 Archimedean solids exist, one can repeat the ideas in the proof of Theorem 3.3.4, but first by considering how different polygons can surround a vertex. The calculations might take a few pages, but they are elementary.

3.3.4 Dihedral angles

An angle formed by two planes (or faces of a polyhedron) is called a *dihedral* angle. Since two planes form two (supplemental) angles, there are two dihedral angles, and usually the one on the "interior" of a particular shape is examined. Among the Platonic and Archimedean solids, only the tetrahedron has dihedral angles less than $\pi/2$. One can often compute such angles by employing dot products or by simple geometry.

Exercise 157. *Use dot products to find the dihedral angles between faces of a regular tetrahedron.*

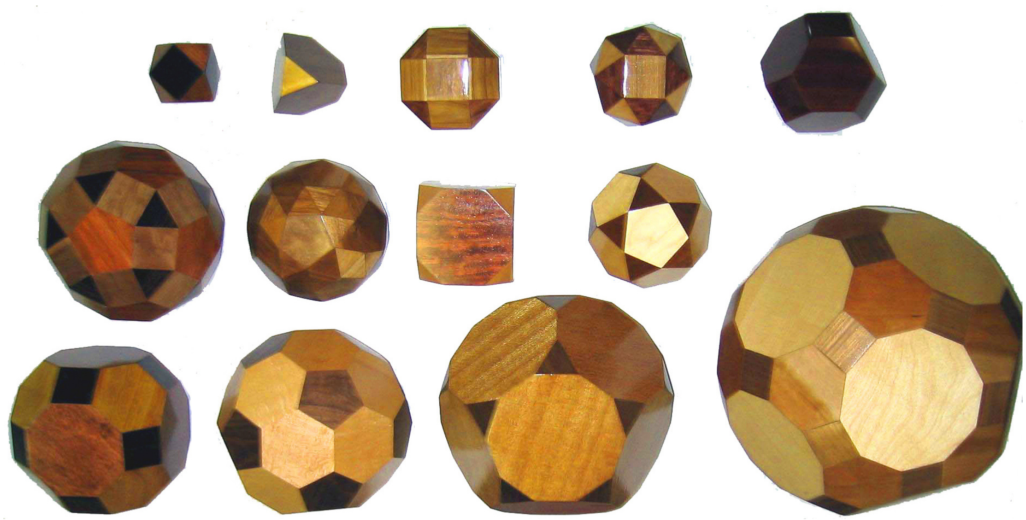


Figure 3.6: The Archimedean solids; hollow wood by DSG, 2003

It is well-known that one can bisect a cube into two (identical)pieces with one slice that forms a regular hexagon–see Figure 3.3.4.

As in Figure 3.8, if the cube is a unit cube situated on the three standard axes, such a slice is by using a plane with normal vector $(1,1,1)$, in which case the vertices of the desired hexagon are (in order):

$$\left(\frac{1}{2},0,1\right),\ \left(1,0,\frac{1}{2}\right),\ \left(1,\frac{1}{2},0\right),\ \left(\frac{1}{2},1,0\right),\ \left(0,1,\frac{1}{2}\right),\ \left(0,\frac{1}{2},1\right).$$

Exercise 158. *As in Figure 3.8, cut a cube in half with a single plane, where the intersection of the plane and the cube is a regular hexagon. Show that the dihedral angle between a face of the cube and the cutting plane is* $\cos^{-1}\left(\frac{1}{\sqrt{3}}\right) \sim 54.73$ *degrees.*

I needed to compute the angle in Exercise 158 to make a cover for a tire sandbox I made for our son; see Figure 3.3.4. (The tire is 73.2 inches across, 19.2 inches deep, and the cube has 4 foot edge lengths; this was constructed in 2020.)

The following lemma appears in [525], giving a formula (without proof) for computing the dihedral angle between two identical isosceles triangles. In fact, this lemma can be used in many more situations.

Name	v	e	f	face types	d
Cuboctahedron (Dymaxion)	12	24	14	8(3), 6(4)	4
Truncated tetrahedron	12	18	8	4(3), 4(6)	3
(Lesser) Rhombicuboctahedron	24	48	26	8(3), 18(4)	4
Snub cube	24	60	38	32(3), 6(4)	5
Truncated octahedron (Mecon)	24	36	14	6(4), 8(6)	3
Lesser rhombicosidodecahedron	60	120	62	20(3), 30(4), 12(5)	4
Snub dodecahedron	60	150	92	80(3), 12(5)	5
Truncated cube	24	36	14	8(3), 6(8)	3
Icosidodecahedron	30	60	32	20(3), 12(5)	4
Great rhombicuboctahedron (Truncated cuboctahedron)	48	72	26	12(4), 8(6), 6(8)	3
Truncated icosahedron	60	90	32	12(5), 20(6)	3
Truncated dodecahedron	60	90	32	20(3), 12(10)	3
Great rhombicosadodecahedron (Truncated icosidodecahedron)	120	180	62	30(4), 20(6), 12(10)	3

Table 3.2: Table for Archimedean solids

Lemma 3.3.5. *Let $\triangle ABC$ and $\triangle ABD$ be congruent isosceles triangles with $\|AB\| = \|AC\| = \|AD\|$ (sharing the edge AB) and $\|BC\| = \|BD\|$, where $\angle CBD$ is β. Let $\alpha = m\angle BAC = m\angle BAD$ (see Figure 3.10). Then the dihedral angle θ between these two triangles satisfies*

$$\cos(\theta) = \frac{2\cos(\beta) - 1 + \cos(\alpha)}{1 + \cos(\alpha)}.$$

Proof: Let P be on AB so that $\theta = \angle CPD$, that is, $\angle APC$ and $\angle APD$ are right angles. For any two points X, Y in 3-space, let $\mathbf{XY}$ denote the vector from X to Y, and let the norm of such a vector $\|\mathbf{XY}\|$ be written simply as $|XY|$. For simplicity, assume that $|PC| = |PD| = 1$. Using the standard inner product for vectors in $\mathbb{R}^3$,

$$\begin{aligned}
\cos(\theta) &= \mathbf{PC} \bullet \mathbf{PD} \\
&= (\mathbf{PB} + \mathbf{BC}) \bullet (\mathbf{PB} + \mathbf{BD}) \\
&= |PB|^2 + \mathbf{PB} \bullet \mathbf{BD} + \mathbf{BC} \bullet \mathbf{PB} + \mathbf{BC} \bullet \mathbf{BD} \\
&= |PB|^2 - \mathbf{BP} \bullet \mathbf{BD} - \mathbf{BC} \bullet \mathbf{BP} + \mathbf{BC} \bullet \mathbf{BD} \\
&= |PB|^2 - |BP| \cdot |BD| \cos\left(\frac{\pi}{2} - \frac{\alpha}{2}\right) - |BC| \cdot |BP| \cos\left(\frac{\pi}{2} - \frac{\alpha}{2}\right)
\end{aligned}$$

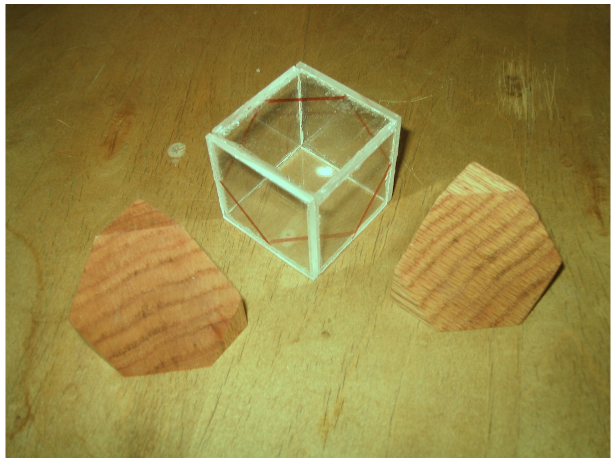


Figure 3.7: Bisecting a cube with a regular hexagon, DSG, 2003?

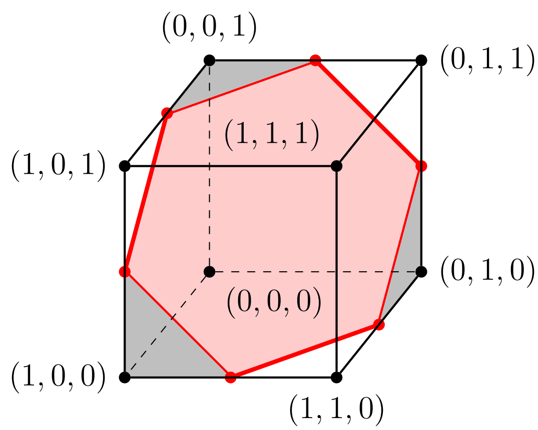


Figure 3.8: Coordinates for slicing the cube

$$\begin{aligned}
&\quad + |BC| \cdot |BD| \cos(\beta) \\
&= |PB|^2 - |BP| \cdot |BD| \sin\left(\frac{\alpha}{2}\right) - |BC| \cdot |BP| \sin\left(\frac{\alpha}{2}\right) \\
&\quad + |BC|^2 \cos(\beta) \\
&= |PB|^2 - |BP| \cdot |BD| \sin\left(\frac{\alpha}{2}\right) - |BD| \cdot |BP| \sin\left(\frac{\alpha}{2}\right) \\
&\quad + |BC|^2 \cos(\beta)
\end{aligned}$$

Figure 3.9: Sandbox with bisected cube cover, DSG, 2020

$$= |PB|^2 - 2|BP| \cdot |BC| \sin\left(\frac{\alpha}{2}\right) - |BC|^2 \cos(\beta).$$

Since $m\angle BCP = \frac{\alpha}{2}$, it follows that $|BP| = |BC| \sin(\frac{\alpha}{2})$ and $\cos(\frac{\alpha}{2}) = \frac{1}{|BC|}$. Hence

$$\begin{aligned}
\cos(\theta) &= |BC|^2 \sin^2\left(\frac{\alpha}{2}\right) - 2|BC|^2 \sin^2\left(\frac{\alpha}{2}\right) + |BC|^2 \cos(\beta) \\
&= |BC|^2 \left[\cos(\beta) - \sin^2\left(\frac{\alpha}{2}\right)\right] \\
&= \frac{1}{\cos^2\left(\frac{\alpha}{2}\right)} \left[\cos(\beta) - \sin^2\left(\frac{\alpha}{2}\right)\right] \\
&= \frac{2}{1+\cos(\alpha)} \left[\cos(\beta) - \frac{1-\cos(\alpha)}{2}\right] \\
&= \frac{2\cos(\beta) - 1 + \cos(\alpha)}{1+\cos(\alpha)},
\end{aligned}$$

as desired. □

Exercise 159. *Use Lemma 3.3.5 to find the (internal) dihedral angle formed by adjacent faces of an icosahedron. (Recall that each vertex of an icosahedron is surrounded by 5 equilateral triangles.)*

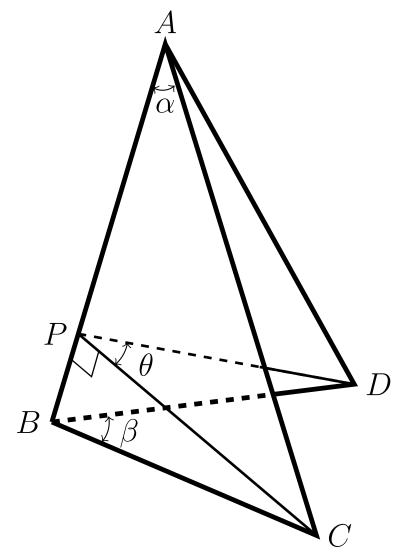


Figure 3.10: Dihedral angle θ between isosceles triangles

Note that the result in Exercise 157 is also obtained by a simple application Lemma 3.3.5 with $\alpha = \beta = \frac{\pi}{3}$.

3.3.5 Descartes' deficiency

In the Euclidean plane, if a number of line segments meet a point x, the angles determined sum to 2π. However, in a polyhedron, if (polygonal) faces meet at a vertex x adding all the angles given by the polygons meeting at x sum to less than 2π. For example, in the tetrahedron, three equilateral triangles meet at a vertex, and the sum of angles at a vertex is π, somewhat less than 2π. (The closer a vertex is to being "flat", the smaller this difference is.) For a polyhedron P, define the *deficiency* (or defect) of a vertex to be 2π minus the sum of the angles surrounding the vertex. Define the *total deficiency* of the polyhedron P to be the sum of the deficiencies of all vertices in P.

Theorem 3.3.6 (Descartes, 1630 [253])**.** *If a polyhedron P is homeomorphic to a sphere (i.e., can be continuously deformed into a sphere), then the total deficiency of P is 4π.*

I am not aware of Descartes' original proof, but I found it interesting that Descartes' deficiency result implies Euler's formula $v + f - e = 2$ (see

Theorem 3.3.3) for polyhedra (that are homeomorphic to a sphere). [This result was taught to me by Hayri Ardal, a former student of Veso Jungic at SFU, while in Veso's car in Burnaby, perhaps in 2008 or so.]

Lemma 3.3.7. *For polyhedra homeomorphic to a sphere, Descartes' deficiency result is equivalent to Euler's formula.*

Proof: Let P be a polyhedron (homeomorphic to a sphere) with vertex set V, edge set E, and face set F; put $v = |V|$, $e = |E|$, and $f = |F|$. For a face $S \in F$, let $E(S)$ denote the set of edges of S. Then by Theorem 3.3.6,

$$\begin{aligned}
4\pi &= \text{deficiency of } P \\
&= \sum_{x \in V} \text{deficiency of } x \\
&= \sum_{x \in V} (2\pi - \text{sum of angles at } x) \\
&= 2\pi v - \sum_{x \in V} \text{sum of angles at } x \\
&= 2\pi v - \text{sum of all face angles in } P \\
&= 2\pi v - \sum_{S \in F} \text{sum of face angles in } S \\
&= 2\pi v - \sum_{S \in F} (|E(S)| - 2)\pi && \text{(by Corollary 1.11.7)} \\
&= 2\pi v - \left(\sum_{S \in F} (|E(S)|\pi \right) + 2f\pi \\
&= 2\pi v - 2e\pi + 2f\pi && \text{(each edge occurs in two faces),}
\end{aligned}$$

and dividing all terms by 2π shows $2 = v - e + f$. Reversing the steps above shows that Descartes' result follows from Euler's formula. $\square$

3.4 Volume in higher dimensions

3.4.1 Definition of volume

What is the definition of "volume"? Is "volume" only defined for three dimensions? Even if "volume" is defined only for objects in $\mathbb{R}^3$, what is its precise definition? One might start to define volume by insisting that the volume of a $1 \times 1 \times 1$ cube (called a "unit cube") is 1. Already, there is a difficulty. Suppose $C \subset \mathbb{R}^3$ is a unit cube, and let C' be the unit cube with all points with rational coordinates deleted. What is the volume of C'? What about the cube that contains only rational coordinates? Since the number of points in the "rational cube" is countable, and the number of points in a "real unit cube" is uncountable, one might expect that the rational cube has volume 0, leaving C' still with volume 1.

One can think up many complicated sets inside the unit cube, and in many cases, it might be hard to determine the volumes of these sets. In fact, the volume of complicated sets is left to an area of mathematics called "measure theory", an area that invokes generalizations of integration. (A "volume" in three dimensions is an example of a "measure" on subsets of $\mathbb{R}^3$.) Rather than going into such details, suppose that for "simple" sets (like cubes, spheres, and other convex bodies) some function called "volume" is to be defined. What properties would one want or expect volume to have? If these properties can be listed, a definition of "volume" might be generalized to any dimension.

It seems reasonable that "volume" depends on dimension; for example, the volume of (2-dimensional) square in $\mathbb{R}^3$ is 0, but the measure (area) of a $r \times r$ square in $\mathbb{R}^2$ is r^2. Another requirement of volume is that it is a non-negative number, perhaps infinity. Suppose that for each positive integer d, a mathematician tries to define a function $\text{vol}_d : \mathbb{R}^d \to [0, \infty]$ (called d-dimensional volume). The notation "vol_d" is chosen here because V is often used for other things, like vector spaces or vertex sets.

Natural requirements for vol_d might include:

- The functions vol_1, vol_2, vol_3 correspond to the standard concepts of length, area, and volume in dimensions 1, 2, and 3, respectively.

- For each dimension d, if

  $$C = [0,1]^d = \{(x_1, x_2, \ldots, x_d) : \forall i, 0 \leq x_i \leq 1\} \subset \mathbb{R}^d$$

  denotes the unit cube, then $\text{vol}_d(C) = 1$.

- If C and C' are isometric (there is a distance preserving bijection between them), then $\text{vol}_d(C) = \text{vol}_d(C')$.

- If $C \subset \mathbb{R}^d$, $k \in \mathbb{R}$ and C' is defined by multiplying just one coordinate of points in C by k, then $\mathrm{vol}_d(C) = |k| \cdot \mathrm{vol}_d(C')$.

- Applying the previous property to all coordinates, if $C \subset \mathbb{R}^d$, then for any $k \in [0, \infty)$, and a dilation (or contraction if $k < 1$)
$$kC = \{k\mathbf{x} : x \in C\}$$
has $\mathrm{vol}_d(kC) = k^d \cdot \mathrm{vol}_d(C)$.

- If $C \subset \mathbb{R}^d$, then for each $n > d$, $\mathrm{vol}_n(C) = 0$.

- Corresponding to the familiar notation that the volume of a prism is the area of its base times its perpendicular height, if $C \subseteq \mathbb{R}^d$ is a body with volume defined and a new body D is created by moving a copy of C through $\mathbb{R}^{d+1}$ along a vector $\mathbf{v} \in \mathbb{R}^{d+1}$ then $\mathrm{vol}_{d+1}(D)$ is $\mathrm{vol}_d(C)$ times the length of $\mathbf{v}$ projected onto a vector (in $\mathbb{R}^{d+1}$) normal to the hyperplane containing C.

The "volumes" calculated in the next sections indeed satisfy the above requirements for some basic (convex) shapes at least. For more detailed introduction to volumes of convex bodies (usually symmetric), see, e.g., [45].

3.4.2 Parallelepipeds

Recall from elementary geometry that if $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^2$ with angle $\theta \leq \pi/2$ between them, the area of the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$ is
$$A = \|\mathbf{u}\| \cdot \sin(\theta)\|\mathbf{v}\|.$$
If $\mathbf{u} = (a, b)$ and $\mathbf{v} = (c, d)$, viewing these two vectors in $\mathbb{R}^3$ given by $(a, b, 0)$ and $(c, d, 0)$, then the above equation agrees with the size of the cross product
$$A = \|\mathbf{u} \times \mathbf{v}\| = |ad - bc| = \left| \det \begin{bmatrix} a & c \\ b & d \end{bmatrix} \right|.$$

Recall (see Section 23.2.3) that for two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^3$, the cross product $\mathbf{u} \times \mathbf{v}$ is a vector orthogonal to both $\mathbf{u}$ and $\mathbf{v}$, and (by calculations similar to above), $\mathbf{u} \times \mathbf{v}$ has length equal to the area of the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$. Also recall if θ is the angle determined between vectors $\mathbf{a}$ and $\mathbf{b}$, the dot product satisfies (in any dimension!)

$$\mathbf{a} \bullet \mathbf{b} = \|\mathbf{a}\| \cdot \|\mathbf{b}\| \cos\theta. \tag{3.3}$$

Calculating the volume of a parallelepiped in three dimensions was examined in Section 23.2.4, but is repeated here for convenience. Let $\mathbf{u} = (u_1, u_2, u_3)$, $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$ be three linearly independent vectors in $\mathbb{R}^3$, and let P be the parallelepiped determined by $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$.

As in the above calculations, the "base" generated by $\mathbf{u}$ and $\mathbf{v}$ has area $\|\mathbf{u} \times \mathbf{v}\|$. For convenience, put $\mathbf{n} = \mathbf{u} \times \mathbf{v}$ (a vector normal to the base). Let h be the height of P (measured on a perpendicular from the base). By Cavelieri's principle (or a simple argument that chops off ends and moves them around), the volume of P is h times the area of the base.

Then h is the length of the projection of w onto $\mathbf{n}$. If $\mathbf{w}$ forms an angle of θ with $\mathbf{n}$, then $h = \|\mathbf{w}\| \cos\theta$. Thus, the volume of P is

$$\begin{aligned} h \cdot \|\mathbf{n}\| &= \|\mathbf{w}\| \cdot \cos\theta \cdot \|\mathbf{u} \times \mathbf{v}\| \\ &= \mathbf{w} \bullet (\mathbf{u} \times \mathbf{v}) && \text{(by (3.3))}, \end{aligned}$$

also called the *scalar triple product*. In the above calculations, it was tacitly assumed that θ is in $(0, \pi/2)$, and since the cross product satisfies $\mathbf{u} \times \mathbf{v} = -(\mathbf{v} \times \mathbf{u})$, take the absolute value of the above expression (and rearrange the dot product) to give the volume of P:

$$\mathrm{vol}(P) = |(\mathbf{u} \times \mathbf{v}) \bullet \mathbf{w}|.$$

Choosing the vectors in a different order also gives the expression $|\mathbf{u} \bullet (\mathbf{v} \times \mathbf{w})|$, another common way to write the volume. In any case, one can verify that the volume of P is found by a determinant,

$$\mathrm{vol}(P) = \left| \det \begin{bmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{bmatrix} \right|.$$

Recall that interchanging two columns of a matrix A gives a matrix with determinant $-\det(A)$, so the absolute value can be dispensed with by re-ordering or re-naming the vectors.

Suppose now that a parallelepiped P in four dimensions is given by vectors (u_1, u_2, u_3, u_4), (v_1, v_2, v_3, v_4), (w_1, w_2, w_3, w_4) and (x_1, x_2, x_3, x_4). If the four-dimensional volume of P is defined by

$$\mathrm{vol}_4(P) = \left| \det \begin{bmatrix} u_1 & v_1 & w_1 & x_1 \\ u_2 & v_2 & w_2 & x_2 \\ u_3 & v_3 & w_3 & x_3 \\ u_4 & v_4 & w_4 & x_4 \end{bmatrix} \right|,$$

then it can be shown that this volume respects the requirements given in Section 3.4.1. In fact, with a little more work, it can be shown that the similar definition of vol_d applied to parallelepipeds of any dimension (just put the vectors as columns in a matrix) also behaves as expected.

To be slightly more formal, what is being used above is the definition of a "determinant function". (See, e.g., [480, Ch. 5] for more details.) Let R be any commutative ring with identity, and let $R^{n\times n}$ denote the set of all $n \times n$ matrices with entries from R. A function $f : R^{n\times n} \to R$ is called n-linear if and only if for each $i = 1, \ldots, n$, f is a linear function of the ith row when the other $n-1$ rows are fixed. The function f is called *alternating* if and only if whenever two rows of an $n \times n$ matrix A are switched, producing a matrix B, then $f(B) = -f(A)$, and if A has a two identical rows, then $f(A) = 0$. A function $f : R^d \to R$ is called a *determinant function* if and only if f is n-linear, alternating, and $f(I_d) = 1$. It can be shown that there is only one determinant function, called "det", and that for any two $n \times n$ matrices, $\det(AB) = \det(A)\det(B)$. Also note that a function that is n-linear in its rows is also n-linear in its columns, and so the above definition of a determinant can also be made with "columns" replacing "rows". Then the volume of a parallelepiped as defined above is the absolute value of the determinant function, or simply the determinant function if the columns (generating vectors) are ordered appropriately.

Some authors even define the determinant in terms of volume! This equivalence between determinants and volume is used here (without proof) in Section 9.3 when discussing volumes of cells in higher dimension lattices of points.

3.4.3 Spheres and balls

Recall that the Euclidean norm of a point or vector $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ is
$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}.$$

Definition 3.4.1. Let $d \geq 1$, $\mathbf{x} \in \mathbb{R}^d$, and let r be a non-negative real. Define
$$B^d(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| \leq r\}$$
to be the d-dimensional ball with radius r centered at $\mathbf{x}$. Any 2-dimensional ball is called a *disk*. The boundary of a ball in d-dimensions is called a $(d-1)$-dimensional sphere, denoted by
$$S^{d-1}(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| = r\}.$$

Let $B^d(r)$ denote any ball of radius r (no matter where its center is) and let B^d denote a unit ball $B^d(1)$. Similarly, let any sphere in $\mathbb{R}^d$ with radius r be denoted by $S^{d-1}(r)$ and let S^{d-1} denote a unit sphere.

The length of an interval $[-r, r] \subset \mathbb{R}$ is $2r$, so interpreting this as "1-dimensional volume", this says that $\mathrm{vol}_1(B^1(r)) = 2r$. The "surface" of this 1-dimensional "ball" is just two points, which have no "volume", so one might write $\mathrm{vol}_1(S^0(r)) = 0$. Going to 2 dimensions, the area of a circle with radius r is πr^2, so write $\mathrm{vol}_2(B^2(r)) = \pi r^2$; the surface of this "ball" is a circle whose length is $2\pi r$, so one might write $\mathrm{vol}_1(S^1(r)) = 2\pi r$, and since a circle has no area, write $\mathrm{vol}_2(S^1(r)) = 0$.

In three dimensions, the volume of a ball with radius r is $\frac{4\pi r^3}{3}$, so write $\mathrm{vol}_3(B^3(r)) = \frac{4\pi r^3}{3}$, and since the area of a sphere with radius r is $4\pi r^2$, write $\mathrm{vol}_2(S^2(r)) = 4\pi r^2$. Since a sphere has no thickness, and hence no volume, write $\mathrm{vol}_3(S^2(r)) = 0$.

The reader with some calculus background might have noticed that
$$\frac{d}{dr}\left[\pi r^2\right] = 2\pi r,$$
and
$$\frac{d}{dr}\left[\frac{4\pi r^3}{3}\right] = 4\pi r^2.$$
This says that for circles, the derivative of area is length, and for spheres, the derivative of volume is area. This connection between the "volume" of a ball

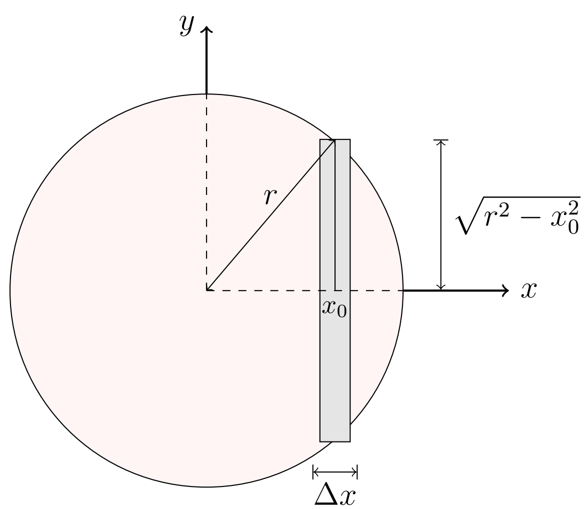
and the “size” of the corresponding boundary is demonstrated by looking at the corresponding integrals that define area and volumes.

To find the area of a circle by using integrals, a few popular methods work, including the “vertical slicing” technique, the “sector method” (close to what Archimedes used) and the “shell” method (which requires that you know the length of a circle). To find the volume of a sphere, again these three types of derivation are available, but the “sector” method works after changing to some other coordinate system (like polar coordinates), and the “shell” method requires that one knows the area of a sphere.

The idea behind the slicing technique seems to be the simplest, and is also a simple method to generalize to higher dimensions. Before looking at circumference of a circle or area of a sphere, consider the two proofs for area of a circle and volume of a sphere (with notation used above, these are $\mathrm{vol}_2(B^2(r))$ and $\mathrm{vol}_3(B^3(r))$).

Theorem 3.4.2. *The area of the disk $B^2(r)$ is πr^2.*

Proof: Let $r > 0$ be a radius, and let a disk $B^2(r)$ be centered on the x-y axis. Slice the circle into vertical strips with some small width Δx. For a value $x_0 \in [-r, r]$, the vertical strip at x_0 is approximated with a rectanlge of height $2\sqrt{r^2 - x_0^2}$ and width Δx:



Adding up the area of these strips (and letting $\Delta x \to 0$) gives the integral

$$I = \int_{x=-r}^{x=r} 2\sqrt{r^2 - x^2}\, dx.$$

To evaluate this integral, let $x = r\sin\theta$, in which case $dx = r\cos\theta$; also, when $x = -r$, $\theta = -\pi/2$, and when $x = r$, $\theta = \pi/2$. So

$$\begin{aligned} I &= \int_{-\pi/2}^{\pi/2} 2\sqrt{r^2 - r^2\sin^2\theta} \cdot r\cos\theta\, d\theta \\ &= 2\int_{-\pi/2}^{\pi/2} r\sqrt{1-\sin^2\theta} \cdot r\cos\theta\, d\theta \\ &= 2r^2 \int_{-\pi/2}^{\pi/2} |\cos\theta| \cdot \cos(\theta)\, d\theta. \end{aligned}$$

Since $\cos\theta$ is positive in the domain $[-\pi/2, \pi/2]$, the absolute value signs can be dropped; also by symmetry, integrate from 0 to $\pi/2$ and double (this is not necessary, but it sometimes makes evaluating easier) to get

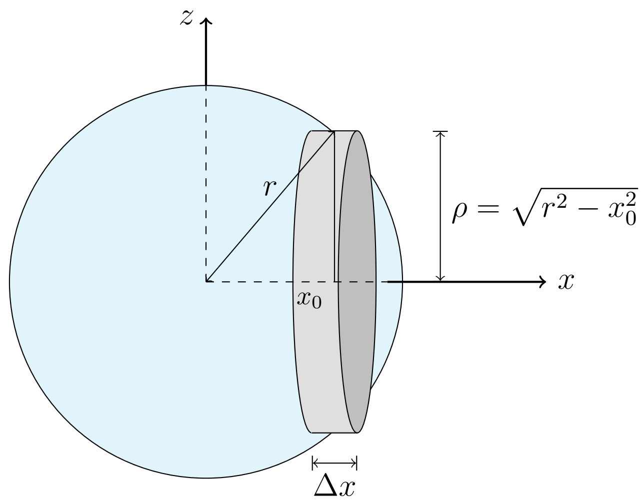
$$\begin{aligned} I &= 4r^2 \int_0^{\pi/2} \cos^2\theta\, d\theta \\ &= 4r^2 \int_0^{\pi/2} \frac{1+\cos(2\theta)}{2}\, d\theta \\ &= 2r^2 \int_0^{\pi/2} [1+\cos(2\theta)]\, d\theta \\ &= 2r^2 \left(\theta + \frac{\sin(2\theta)}{2}\right)\Bigg|_{\theta=0}^{\theta=\pi/2} \\ &= 2r^2[(\pi/2+0) - (0+0)] \\ &= \pi r^2. \end{aligned}$$

□

Following the same slicing idea, the volume of a sphere is also easy to calculate (in fact, it is easier because the double-angle formula for cosine is not required to crack the resulting integral).

Theorem 3.4.3. *For $r > 0$, the volume of a ball of radius r in $\mathbb{R}^3$ is $\frac{4\pi r^3}{3}$.*

Proof: Let $r > 0$, and position a ball $B^3(r)$ at the origin. For $x_0 \in [-r, r]$, consider a slice of the ball (parallel to the yz-plane) centered at x_0 with width Δx. This slice is a pie shape (sometimes called a "disk"), with volume approximately equal to the area of a circle times Δx, where the circle has radius $\rho = \sqrt{r^2 - x_0^2}$. The following sketch can be interpreted as looking from the negative y-axis (the pie actually has a curved rim, but is drawn as cylinder for clarity):



By the theorem for the "volume" of $B^2(\rho)$, the area of this circle is $\pi(\sqrt{r^2 - x_0^2})^2$, and so the volume of this slice is approximately $\pi(r^2 - x_0^2)\Delta x$. Summing the volume of all disks and taking limits as $\Delta x \to 0$ gives the volume of $B^3(r)$ to be

$$\begin{aligned}\int_{-r}^{r} \pi(r^2 - x^2)\,dx &= 2\pi \int_0^r (r^2 - x^2)\,dx \\ &= 2\pi \left(r^2 x - \frac{x^3}{3}\right)\Bigg|_{x=0}^{x=r} \\ &= 2\pi \left[\left(r^3 - \frac{r^3}{3}\right) - (0 - 0)\right] \\ &= \frac{4\pi r^3}{3}.\end{aligned}$$

□

The techniques are now in place to compute the volume of higher dimensional balls. For example, in $\mathbb{R}^4$, let x, y, z, w be the four axes. A ball of radius r is given by the equation $x^2 + y^2 + z^2 + w^2 = r^2$, and so a "slice" at some fixed x_0 has the equation $y^2 + z^2 + w^2 = r^2 - x_0^2$, the equation of a 3-dimensional ball with radius $\sqrt{r^2 - x_0^2}$. In a similar manner for higher dimensions, arrive at the recursion (for $n \geq 2$)

$$\mathrm{vol}(B^n(r)) = \int_{-r}^{r} \mathrm{vol}(B^{n-1}(\sqrt{r^2 - x^2}))\, dx, \tag{3.4}$$

where, the volume on the left is n-dimensional volume vol_n and the volume in the integrand is the volume vol_{n-1} in $n-1$ dimensions.

For example, the volume of the 4-dimensional ball is

$$\begin{aligned}\mathrm{vol}_4(B^4(r)) &= \int_{-r}^{r} \mathrm{vol}_3(B^3(\sqrt{r^2 - x^2}))\, dx \\ &= \int_{-r}^{r} \frac{4\pi}{3}(r^2 - x^2)^{3/2}\, dx \\ &= \frac{8\pi}{3}\int_{0}^{r} (r^2 - x^2)^{3/2}\, dx\end{aligned}$$

As in the area of a circle calculations, put $x = \sin\theta$, and get

$$\mathrm{vol}(B^4(r)) = \frac{\pi^2 r^4}{2}. \tag{3.5}$$

Starting with the above cases, and the recursion (3.4) an inductive proof verifies that the volume of the unit ball in $\mathbb{R}^d$ is

$$\mathrm{vol}(B^d) = \frac{\pi^{\lfloor d/2 \rfloor} 2^{\lceil d/2 \rceil}}{\prod_{i:\ 0 \leq 2i < d} (d - 2i)}. \tag{3.6}$$

The expression in (3.6) takes on different kinds of forms for odd and even dimensions (because in the even cases, the square root signs vanish, leaving only a polynomial). Figure 3.11 gives volumes of a ball in the first few dimensions.

Surprisingly, the volume of the unit ball goes down after 5 dimensions, and approaches 0 as the dimension tends to infinity (this is an exercise).

One more fact that might be interesting is that the volume of larger dimensional spheres is concentrated near the surface of the sphere; this can be shown by integrating with spherical shells (which is also used to show that the derivative of a volume in one dimension is the surface “area” in the previous one). It is left as an exercise for the reader to derive formulae for

d	2	3	4	5	6	7	8	9	10
$\mathrm{vol}(B^d(r))$	πr^2	$\frac{4\pi}{3}r^3$	$\frac{\pi^2}{2}r^4$	$\frac{8\pi^2}{15}r^5$	$\frac{\pi^3}{6}r^6$	$\frac{16\pi^3}{105}r^7$	$\frac{\pi^4}{24}r^8$	$\frac{32\pi^4}{945}r^9$	$\frac{\pi^5}{120}r^{10}$
$\sim \mathrm{vol}(B^d(1))$	3.14	4.19	4.93	5.26	5.17	4.72	4.06	3.30	2.55

Figure 3.11: Volume of balls in higher dimensions

d	2	3	4	5	6	7	8	9	10
$\mathrm{vol}(S^{d-1}(r))$	$2\pi r$	$2\pi^2 r^2$	$2\pi^2 r^3$	$\frac{8\pi^2}{3}r^4$	$\pi^3 r^5$	$\frac{16\pi^3}{15}r^6$	$\frac{\pi^4}{3}r^7$	$\frac{32\pi^4}{105}r^8$	$\frac{\pi^5}{12}r^9$

Figure 3.12: Area of sphere in dimension d is a $(d-1)$-dimensional volume

the surface areas of spheres given in Figure 3.12. (Note also that the surface area of the unit ball decreases past 7 dimensions.)

There is also a strong connection between the volumes in high dimensions and the normal distribution, in part (see, e.g., [337] for details) because of the limit

$$\lim_{n\to\infty} \cos^n\left(\frac{t}{\sqrt{n}}\right) = e^{-\frac{t^2}{2}},$$

but such a discussion leads away from present goals.

3.5 Exercises in 3D

3.5.1 Exercises with volume

Exercise 160. *Find the volume of a regular tetrahedron with side length s.*

Exercise 161. *Find the volume of the region of points (x, y, z) such that*

$$(x^2 + y^2 + z^2 + 8)^2 \le 36(x^2 + y^2).$$

3.5.2 Exercises with polyhedra

Exercise 162. *Let N denote a positive integer and consider the sphere $x^2 + y^2 + z^2 = N$. For which N is there a regular tetrahedron with vertices on the sphere?*

Exercise 163. *Show that if a polyhedron has only triangular faces (such is called a deltahedron), then the number of faces plus the number of edges is a multiple of 5.*

Exercise 164. *In how many different ways can a dodecahedron be placed on the table (so that one corner touching the table is closest to you).*

Exercise 165. *A cube can be inscribed in a dodecahedron in five different ways. Describe the intersection of these five cubes.*

Exercise 166. *Use Euler's formula to show that no convex polyhedron can have 7 edges.*

Exercise 167. *Let P be a simple polyhedron (no holes, no annular faces). Show that there are two faces with the same number of edges.*

Exercise 168. *Let P be a (convex) polyhedron, and let P' be the polyhedron formed by slicing (truncating) off a small tip at each vertex. Let v', f', e' denote the number of vertices, faces, and edges, respectively, of P'. (i) If one of v', f', or e' is 11, find the two possibilities for P. (ii) If one of v', f', or e' is 13, find the four possibilities for P.*

Exercise 169. *Let P be a convex polyhedron where every vertex is incident with exactly three edges. If each face of P is either pentagonal or hexagonal, show that there are exactly 12 pentagonal faces.*

Remark: The Archimedean solid known as the truncated icosahedron (soccer ball shape) is an example of a polyhedron with only (regular) hexagons and pentagons.

If the vertices of the truncated icosahedron represent carbon atoms and the edges represent bonds, this polyhedron represents the chemical structure of the molecule C_{60}, sometimes referred to as the "buckyball molecule" or one of the "fullerenes", named after Richard Buckminster Fuller (1895–1983), the developer of the geodesic dome. (Thanks to K. Gunderson for the graphic in Figure 3.13.) Other similar carbon molecules exist, all with precisely 12

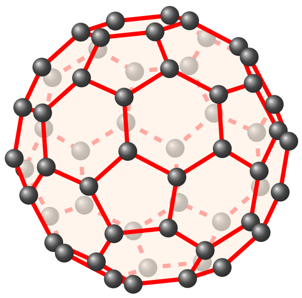


Figure 3.13: The molecule C_{60}.

pentagons and with the rest of the faces being hexagons. The 1985 discovery of C_{60} led to the 1996 Nobel Prize in chemistry (awarded to Harold Kroto, Robert Curl, and Richard Smalley). Mathematicians might be interested to know that algebraic graph theory helped to confirm the structure of C_{60} (see, e.g., [191]).

Exercise 170. *On each of the 20 faces of a regular icosahedron a non-negative integer is written, and the sum of the integers on all faces is 39. Show that there are two faces sharing a vertex that have the same integer written on them.*

The next exercise extends the ideas used in Exercise 170.

Exercise 171. *On each face of a truncated icosahedron is a positive integer. The integers on the (twelve) pentagonal faces sum to 25, and the integers on the (twenty) hexagonal faces sum to 39. Show that two faces sharing a vertex have the same integer.*

Exercise 172. *Find a formula similar to Euler's formula (see Theorem 3.3.3) that takes into account polyhedra with some faces that are annuli, or a polyhedra with holes in them.*

Comment on Exercise 172: If there are no annulus faces, the answer is discussed in [67]. Check your answer for two shapes: a small cube sitting in the center of a face of a larger cube, and a cube with a square hole through two opposite sides. If a convex polyhedron has v vertices, f faces, and e edges, then Euler's formula says $v + f = e + 2$. This formula holds for any polyhedron with no holes and faces with no holes, as long as the polyhedron can be continuously deformed into a sphere. Now suppose that of the faces, a of them are annular, and if there are h holes through the polyhedron, then

$$v + f = e + 2 + a - 2h.$$

Verify this for some obvious examples; for example, if a square hole is put through the center of a face of a cube all the way to the other side, $a = 2$, $h = 1$, $v = 16$, $f = 10$, $e = 24$. However, if, for example, one of these annular faces is divided into four pieces by edges going to the corners of the cube, then $a = 1$, $h = 1$, $v = 16$, $f = 13$, and $e = 28$.

Exercise 173. *For polyhedral graphs in the real projective plane, is it true that $v + f = e + 1$?*

The angle between two adjacent faces in a polyhedron is called the *dihedral* angle. In general, it can be quite difficult to find such angles. Using simple geometry or linear algebra, some dihedral angles are fairly simple to compute.

Exercise 174. *Find the exact dihedral angle in a regular tetrahedron (formed by four equilateral triangles).*

Exercise 175. *Find the exact (interior) dihedral angles in an icosahedron.*

3.5.3 Exercises with spheres

Exercise 176. *Three spheres intersect in a point P, but no line containing P is tangent to all three spheres. Prove that the spheres intersect in an additional common point.*

Exercise 177. *Let S be a sphere. Show that for any five points on S, there exists a closed hemisphere (a hemisphere, together with an equator) that contains at least four of these five points.*

Exercise 178. *Let H be the unit hemisphere $\{(x, y, z) : x^2 + y^2 + z^2 = 1, z \geq 0\}$, C the unit circle $\{(x, y, 0) : x^2 + y^2 = 1\}$, and P the regular pentagon inscribed in C. Determine the surface area of that portion of H lying over the planar region inside P, and write your answer in the form $A \sin \alpha + B \cos \beta$, where A, B, α, β are real numbers.*

Chapter 4

Other problems with geometric aspects

4.1 The art gallery problem

In architecture, "lines of sight" often need careful consideration. Considering a tourist's view through a museum or placement of lights are two examples where "visibility" issues arise. Similar issues arise in monitoring and security, where the placement of security stations, guards, or cameras is important.

In 1973, Victor Klee asked (see [194]) the following:

Question 4.1.1 (Art gallery problem)**.** *For any art gallery whose floor plan is a polygon on n vertices, how many guards (or 360° degree cameras) need be placed on the floor (or ceiling) so that every point on the walls of the gallery is in plain view of at least one guard (or camera)?*

Let $g(n)$ be the minimum number of guards required to guard the interior of any n-gon. It is not difficult to check that $g(3) = g(4) = g(5) = 1$. To see that $g(6) \leq 2$, use a diagonal (which exists by Lemma 1.11.5) to split a hexagon into two areas, one of which is a quadrilateral, and apply $g(3) = g(4) = 1$. This solution to the case $n = 6$ might be seen as a prelude to a general inductive step for a solution to the art gallery problem (if one could only guess a formula for $g(n)$). In 1975, an inductive proof was found where the formula for $g(n)$ is simpler than one might expect.

Theorem 4.1.2 (Chvátal, 1975 [194])**.** *If an art gallery has walls forming a polygon with $n \geq 3$ sides, then $\lfloor n/3 \rfloor$ guards can be placed so that all areas* *of the gallery are guarded. Furthermore, this number is minimal, that is, for each $n \geq 3$, there is a floor plan of a gallery that requires $\lfloor n/3 \rfloor$ guards.*

The original proof of the the first statement in Theorem 4.1.2 was a delicate inductive argument, breaking off a piece of the polygon and reattaching it. A simpler proof (given below) was given by Steve Fisk [336] in 1978; Fisk's proof also relies on induction (and some observations made in some earlier exercises given here regarding triangulations of polygons). For another exposition, see O'Rourke's book [693] on computational geometry.

Proof of first statement in Theorem 4.1.2: Let P be a polygon on $n \geq 3$ points. By drawing diagonals between vertices (by Exercise 46, say), triangulate P into $n-2$ triangles. By Exercise 48, the vertices of this triangulation can coloured in three colours so that each triangle receives all three colours. Put guards at all vertices using the least used colour. This shows that $g(n) \leq \lfloor n/3 \rfloor$. □

Exercise 179. *For $n > 6$, show that $\lfloor n/3 \rfloor$ guards are necessary in Theorem 4.1.2.*

Chvátal's paper was entitled "A combinatorial problem in the plane", and so it is not a reach to say that the art gallery problem is one in combinatorial geometry. The art gallery problem is now a central focus in computational geometry. For example, see *Discrete and computational geometry* [254] by Devadoss and O'Rourke, or the survey article by Urrutia [892] in *The handbook of computational geoemtry.* (See also [692].) These references contain many related results, including those on computational aspects of guard placement, generalizations to higher dimensions, and to results for the art gallery problem for polygonal regions with holes (which might correspond to pillars or other obstructions).

4.2 Probability and geometry—three examples

For some basic references regarding probability, see Gorroochurn's *Classic problems in probability* [401]. For a detailed discussion and references for the Buffon needle problem, and Sylvester's four point problem, see [787]. The first two examples here (both solved by Buffon) were inspired by Ellenberg's book [293] *How not to be wrong: the power of mathematical thinking*. Some facts mentioned in his book are repeated here.

4.2.1 *Franc-carreau*

In the game of *franc-carreau*, a round coin is dropped onto a floor with square tiles, and bets are placed on whether or not the coin lies on a crack or not. Even though the name of the game might suggest it, the coin was not a franc since francs were not in circulation at that time; instead, the coin used was the *écu*. According to [293], *franc-carreau* means roughly, squarely in a square, but *carreau* means tile, so a more likely translation is "fair tile" or "fair square".

In 1733, Georges-Louis Leclerc, Comte de Buffon (1707–1788) (now just referred to as Buffon) analyzed this game of chance with some simple geometry. The 27-year old Buffon was trying to get into the Royal Academy of Sciences in Paris, and so presented his results to the Academy. He was a naturalist and aristocrat with mathematical tendencies (he later wrote volumes on natural history [167], that also contained a recap of some of his math—see [293] for more on his history).

Buffon's [166] argument was quite simple. If the coin has radius r, and the tile has side length $s > 2r$, then as long as the coin's center lies within a smaller square of side length $s - 2r$, (see Figure 4.1) then the coin does not reach the edge of the tile.

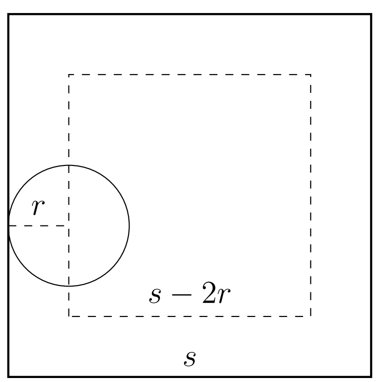


Figure 4.1: One tile and a coin for *franc-carreau*

So the chance that the coin does not touch an edge of a tile is the area of the inner square divided by the larger, namely $(s-2r)^2/s^2$. So if you win a coin whenever your coin does not cross a line and you lose your coin if it does, then if the game is to be fair, $s=(4+\sqrt{2})r$ is needed.

Realizing that this result was not significant enough to be elected to the Academy (at least as a mathematician), he then suggested (and solved) a harder problem, where the nice symmetry of a coin is replaced by another shape—the topic of the next section—for which he is often cited. Buffon also produced results on curves with constant width; such results might be interesting to study, but such topics are only touched upon in this book.

4.2.2 Buffon needle problem

Problem 4.2.1 (Buffon needle problem). *Suppose that a needle of length ℓ is dropped onto a floor covered with strips of wood, each strip d wide. What is the probability p that the needle touches or crosses one of the cracks between the wood strips?*

In 1754 (see [167]), in his "Essai d'arithméque morale", Buffon proposed the above problem, and showed that $p=\frac{2\ell}{d\pi}$ using calculus (and cycloid curves!). In the case where $\ell=d$, $p=\frac{2}{\pi}$.

In 1860, Joseph-Émile Barbier [57] gave a much simpler solution to the Buffon needle problem. Suppose that $d=\ell=1$, that is, the width of each slat is the length of the (unit length) needle. The key idea used is the linearity of expectation, that is, if X and Y are random variables, then

$$\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y].$$

The expected number of crossings of the needle is computed. Consider the event that the needle lands perpendicular to the slats and touches both lines; this occurs with probability 0. The needle crosses just one crack with probability p, and the needle does not touch a crack with probability $1-p$. So the expected number of crossings is $2\cdot 0+1\cdot p+0\cdot(1-p)=p$.

Now consider two needles N_1 and N_2, joined together to form a long needle N. The expected number of crossings for N is the expected number of crossings for N_1 plus the expected number of crossings for N_2, (which are both p), so the expected number of crossings for N is $2p$. Even if N is bent in the middle, the expected number of crossings is the same.

With the same argument, if a polygon is drawn with perimeter n, the expected number of crossings for the polygon is np. (This is easily confirmed when the lengths of the sides are rational, but with a limit argument, also holds for any real length n.)

In particular, when the polygon approaches that of a circle with diameter 1 (which has circumference π), the expected number of crossings is πp. But the expected number of crossings circles with diameters being the width of the slats is 2 (do you see why?), so $p = \frac{2}{\pi}$. □

In 1812, Pierre Simon de Laplace (1749–1827) revisited the Buffon needle problem [588, pp. 359–362], where the floor is instead tiled with rectangles with dimension $c \times d$. The probability of a length ℓ needle crossing a line is then $\frac{\ell^2}{\pi cd}$. For more details and references, please see [787].

4.2.3 Sylvester's four point problem

There are many problems in geometry that ask for the probability of a certain event (like the Buffon needle problem), but any serious discussion of such problems is not appropriate for this text. However, a famous question due to Sylvester might be natural considering the questions regarding k-holes given in Section 7.6. In an 1864 *Educational Times*, Sylvester wrote (quoting from [721]):

Question 4.2.2 (Sylvester, 1864 [845])**.** *Show that the chance of four points forming the apices of a reentrant quadrilateral is 1/4 if they be taken at random in an indefinite plane, but $1/4 + d^2 + x^2$, where d is a finite constant and x a variable quantity, if they be limited by an area of any magnitude and of any form.*

Many different answers were given (see [721] for references), including 1/4 by Cayley and Sylvester, 1/2 by DeMorgan, 1/3 by Wilson, and $35/12\pi^2 \sim .296$ by Woolhouse [933] (and two more answers, including 3/8). Since no appropriate measure of "random" in the plane seemed to be found, answers differed according to the type of convex space in which points are selected. Sylvester used a triangular domain and Woolhouse used a circle. For a history of this problem, see the survey by Pfiefer [721]. There have been many articles written on this topic; for example, see also one by Imre Bárány [54].

Couched in a slightly different manner, one can ask for the probability that four random points in the plane determine four faces? This reformulation might be motivated by asking a more general question: what is the likelihood that n points in $\mathbb{R}^d$ are in convex position (no one contained in the convex hull of any others)? For more on this question, see, e.g., Valtr's article [895]. Also see, e.g.,, a 2017 article by Beermann and Reitzner [71] for more references regarding random polytopes (polytopes are discussed here in Chapter 6).

In Section 17.3.8, work by Scheinerman and Wilf [777] is briefly looked at, relating the four point problem to rectilinear crossing numbers of graphs.

4.3 Frieze and wallpaper patterns

4.3.1 Frieze patterns

A *frieze* is a horizontal strip near or at the top of a wall (or across the top of a pillar), usually decorated with either by a scene of characters, or with some abstract pattern. Often, in such architectural patterns, a certain portion is repeated for the length of the frieze.

For some set S of points in the plane, recall that a symmetry of S is a distance preserving bijection from S to itself. For present purposes, a frieze pattern (or simply, a frieze) is a pattern between parallel horizontal lines that has at least one symmetry that is a horizontal translation. Repeating a translation then shows that frieze patterns are infinite. Many of the architectural friezes have a repeating pattern, and it is only those that are of interest here; from any such finite example, consider the repetition going infinitely far to both the left and right, and with an imaginary center line exactly midway between the upper and lower bounding lines.

In addition to translations, there are four other basic types of symmetries for frieze patterns.

- Rotation: Rotate the entire picture 180 degrees (about some central point).
- Up-down (UD) reflection: Reflect the picture top to bottom about the horizontal center line of the frieze.
- Left-right (LR) reflection: Reflect left-to-right about some vertical line.
- Glide reflection: Combine a translation with a reflection about the center line.

See Figure 4.2 for a chart that shows only 7 combinations are possible (a similar chart also appears in [641, p. 83]). The full binary tree is not necessary since some collections of symmetries generate the same group. For example, translations, rotations, and glide reflections generate the same group as translations, LR reflections and glide reflections. Similarly, if some frieze has translations, both types of reflection, then it trivially has glide reflections.

There are many notations for the various groups, but here I follow Conway's (but not with his nicknames for each).

F1: Translations only. [e.g., ...FFFFFF...]

F2: Translations and glide reflections. [pbpbpb]

F3: Translations, up-down reflection (and so also glide reflection). [BBBBBB]

F4: Translations and rotation by 180°. [SSSSSS]

F5: Translations, glide reflections and (LR reflection OR rotation). [△ ▽ △ ▽ △▽]

F6: Translations and LR reflection. [AAAAAAA]

F7: Translations, rotation by 180°, up-down reflection, LR reflection, glide reflection. [IIIIII, OOOOOO]

4.3.2 Wallpaper patterns

What kind of patterns are available for a periodic colouring of the plane (repeating patterns for wallpaper)? All such patterns can be classified into 17 "types" according to the symmetries available. This was first proved by Evgraf Fedorov [331] in 1891 and later (independently) by George Pólya [730] in 1924. As in the case of frieze patterns, possible symmetries are translations, rotations (but now by angles 60°, 90°, 120°, 180°, in other words, rotations of orders 6,4,3, and 2, not just 2 as in frieze patterns), reflections, and glide reflections.

See Figure 4.3 for Polya's sketch, with his signature, from [730]; the sketch was downloaded from http://www.alt.mathematik.uni-mainz.de/Members/ruddat/gaz/polya/view (Institut für Mathematik, Johannes Gutenberg Universität, Mainz, accessed 27 April 2018).

Wallpaper groups are discussed in many books; for example, see [641, Ch. 11] for wonderfully illustrated examples.

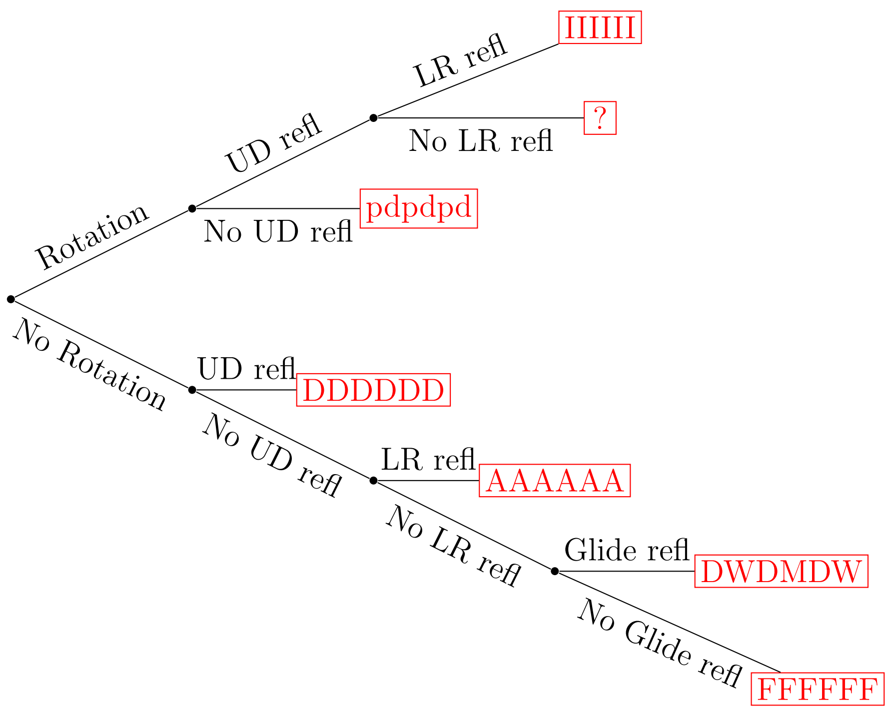


Figure 4.2: Frieze pattern flow chart

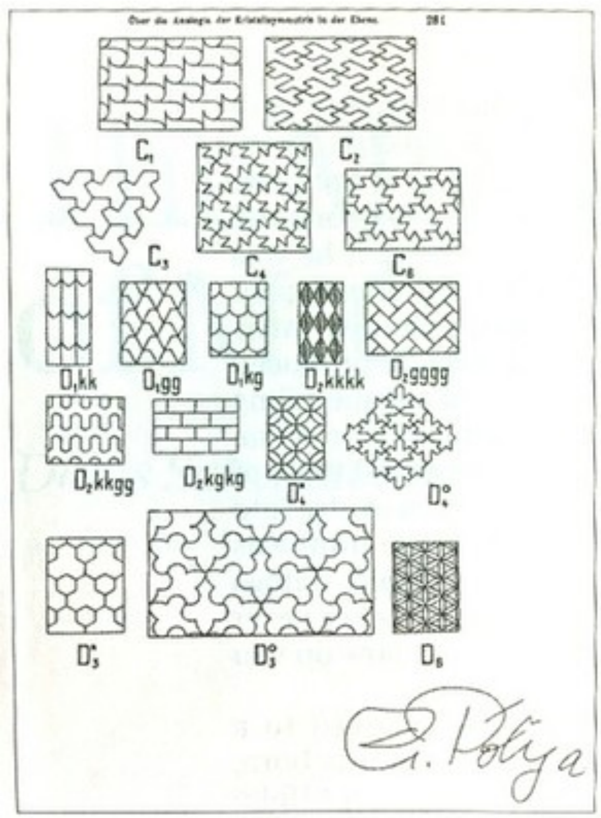


Figure 4.3: Polya's 1924 sketch of 17 patterns for wallpaper groups

4.4 Pursuit problems

Exercise 180. *A square field with side length 100m has a horse at each corner. All horses begin to walk at the same speed, each walking directly towards the horse on its right. Eventually, they all meet at the middle. How far did each horse walk?*

Exercise 181. *In Exercise 180, what is the path of each horse?*

Comments on Exercises 180 and 181: The problem of four horses is in a class of problems called "pursuit problems", and such problems have a rich history. The literature abounds with examples where horses are replaced by dogs, bugs, or even flies, and when there are three, four, or more animals. Such examples are often called "cyclic pursuit problems". The problems are also generalized to cases where the starting positions are vertices of a triangle (not just equilateral triangles) or larger n-gons, and even in cases where the animals pursue at different speeds.

I first found this problem in [542, Prob. C7, p. 22], (where men played the role of the above horses), but later found more references in more popular literature. For example, Steinhaus [826, pp. 126–149] gave a detailed discussion of pursuit problems for two ships, circles of Apollonius, spirals, as well as the cyclic pursuit for four dogs; the three dog problem was also given for equilateral triangles. (The first edition of Steinhaus's book was 1950, but this was likely also in the original Polish version from a few years earlier). Exercise 181 is given in many books on differential equations; for example, see Bateman's 1918 book [63, pp. 8–10], Ince's 1944 book [495, p. 71], or Lester Ford's 1955 book [340, Q36, p. 58] for "Four flies on a card table..." (this exercise does not seem to be in Ford's first edition of 1933). In 1957, Martin Gardner also popularized Exercise 180 as "the four bug problem". Wells [918, pp. 201–202] gives a very brief introduction to pursuit curves, including the four dog problem and a pursuit problem along an ellipse.

A popular 1971 article by Klamkin and Newman [551] gives generalizations of the problem to three or more bugs, along with more references. The case for three bugs starting at vertices of an arbitrary triangle with different speeds is completely solved [551, Prob. III] using complex numbers, along with other generalizations (an open problem for vertices of a trapezoid is also given). The Klamkin–Newman article says "The three bug problem can be traced back to H. Brocard [1] in 1877." The reference given in [551] was "1. Nouv. Corres. Math., 3 (1877), pp. 175, 280." However, this reference is not entirely correct (the same seems to have been copied by other sources, as well). Thanks to Karen Gunderson and interlibrary loans, it turns out that the relevant pages are 175–176, where two problems are given. The first, problem 250, on p. 175 is by H. Brocard [146], which says (thanks to K. R. Gunderson for the translation from French):

> 250. A moving object runs around a circumference with a constant angular velocity of w. Another moving object, departing from the center, moves toward the first with a speed of b. Give the differential equation of the 'curve of pursuit'.** (H. Brocard)

The volume where Brocard's problem was published was edited by Eugéne Catalan, who commented (again, translated by KRG):

> (**) One of the first examples of the 'curve of pursuit' is that which was presented in the celebrated "Mémoire de Dubois-Aymé: De la courbe que décrit un chien courant aprés san Maitre" [On the curve determined by a dog running after his master].
>
> The illustrious Sturm generalized this problem a lot (Annales de Gergonne, vol. 13). (E. C.)

Brocard was a French army officer who (according to [918]) in 1875 studied the possible points in the three dog problem (for an arbitrary triangle) where the dogs could meet (now called "Brocard points"). The "illustrious Sturm" that Catalan mentions is M. Ch. Sturm, and Aymé was a student at Ecole Polytechnique in 1796 (and later an engineer on an expedition to Egypt in 1798). The volume 13 was written in 1822–1823. The Memoire mentioned was with Bigeon, "Mémoiire sur les développées des courbes planes", Paris, Malher, 1829.

The three dog problem was known to Brocard, so it might seem odd that Lucas [625] was the person to record the problem right after Brocard's problem (again, thanks to KRG for the translation):

> 251. Three dogs are placed on the three vertices of an equilateral triangle; they run after one another. Which curve is given by each of them? (É. Lucas)

Cyclic pursuit problems (and logarithmic curves) are also discussed in the Eli Maor's 1994 book *e: The story of a number* [635]. The four bug problem has even been used in interviews for hiring in some large companies (see the New York Times blog post [30] by Gary Antonick). According to [551], the general (non-trivial!) three bug problem appeared on the Cambridge Mathematics Tripos exam of 1871 set by R. K. Miller. See also [720] for a detailed solution for the four bug problem, as well as the generalization to n bugs and an n-gon and [779] for more references. In *The Penguin dictionary of curious and interesting geometry* [918, pp. 21–22], David Wells gives an introduction to Brocard points.

Exercise 182. *A cop chases a robber into a circular swimming pool, and the robber iis now treading water at the center, and the cop is at the edge. The cop will not enter the pool. If the robber can reach a point on the edge of pool before the cop does, the robber gets away. The robber can swim one quarter as fast as the cop can run (around the perimeter). Is there a strategy so the robber can make it to the edge of the pool and get away? Prove your answer.*

According to Littlewood [614, pp. 114–117], the following problem was invented by Richard Rado in the late 1930s.

Problem 4.4.1 (Rado)**.** *A lion and a man in a closed circular arena each can run with the same maximum speed. What strategy should the lion use to catch the man?*

Apparently, Rado used "Christian" instead of "man", but Littlewood called the problem "LION AND MAN". For more on this problem, see [118]. Let L, M, and O denote the points of the lion, the man and the center of the arena. Littlewood mentions that if L always runs directly at M (giving a pursuit curve), that time of capture can be infinite, so the lion would need another strategy for a successful meal.

To give a finite capture time, the lion might use the following strategy: keep L on OM; in other words, stay on the radius given by M. Since the possible angular velocity for the lion is greater than that for the man, the lion can close the distance. Even if the man runs along the outside circle of the arena, the lion can continue to gain on the man, eventually catching him. (For calculations regarding this approach, see [614, pp. 114–115].)

However, in 1952, Besicovitch observed that moving in straight line segments (instead of around the boundary) can make the time required for the lion infinite! Details are in both *Littlewood's Miscellany* [614] and, for example, [118]. Here is a brief explanation. Let $c > 0$ be a "small" constant (how small, yet to be determined). The man runs along a "polygonal" path $M_0 - -M_1 - -M_2. - -...$ At each step, when the man is at M_i, the man runs a distance $d_i = ci^{-3/4}$ to M_{i+1} in the direction perpendicular to OM_i, into the half disk given by OM not containing L. The total distance the man (or lion) runs is infinite (providing the path given stays inside the arena).

4.5 Pigeonhole problems in geometry

Recall that one form of the pigeonhole principle (PHP) states that for positive integers r, k, if $rk + 1$ objects are sorted into r holes, then one hole contains at least $k + 1$ objects. This section contains a few classic applications of the PHP to geometric shapes. The PHP is used in a few other places in this text, as well (for example, in Exercise 121, the second solution to Exercise 167, or the five points on a sphere problem in Exercise 177).

Exercise 183. *Let T be an equilateral triangle with side length 1. Show that if ten points are selected on or inside T, then there are two points whose distance apart is at most $\frac{1}{3}$.*

Exercise 184. *If a target is an equilateral triangle with side length 2, and the target is hit by 17 arrows, what is the minimum distance between arrows?*

Exercise 185. *Let T be a triangle with unit area. Show that if 9 points are chosen from the edges or interior of T, there always exist three points forming a triangle of area at most $\frac{1}{4}$.*

Exercise 186. *Ten points are placed on a circular disk of diameter 5. Show that there are two points with a distance apart of at most 2.*

Exercise 187. *Show that if 51 points are in a square of unit side length, then some three lie in a disk of radius $\frac{1}{7}$.*

Exercise 188. *In a 35×36 rectangle, 55 points are placed. Show that some two of these points have distance at most $\sqrt{50}$.*

4.6 Parabolas and hyperbolas

Exercise 189. *Parabolas $y = a_1x^2 + b_1x + c_1$ and $y = a_2x^2 + b_2x + c_2$ have foci at F_1 and F_2 respectively, and are tangent to each other at T. Prove that F_1, F_2, and T are collinear.*

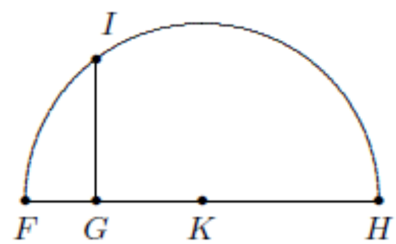
Exercise 190. *Find the least possible area of a convex set in the plane that intersects both branches of the hyperbola $xy = 1$ and both branches of the hyperbola $xy = -1$.*

4.7 Miscellaneous exercises

Exercise 191 ([231, 17])**.** *Consider the letter T as drawn with just two straight segments (no serifs), where the horizontal stroke on the top is the same length as the vertical handle. Is it possible to draw disjoint T's so that every rational on the x-axis is the bottom point of such a T? (They may be, of course, different sizes.)*

Exercise 192. *Several kids are playing on the beach, each with a water gun. All pairwise distances between kids are different. Each kid shoots their nearest neighbour. If all kids get wet, what is their position?*

Exercise 193. *In "La Géométrie", Descartes gives the following geometric construction of a square root: "If the square root of GH is desired, I add, along the same straight line, FG equal to unity; then bisecting FH at K, I describe the circle FIH about K as center, and draw from G a perpendicular and extend it to I, and GI is the required root." Assuming, as in the figure below, that $GH > 1$, prove that the length of GI is the required root.*



Chapter 5

Convexity theorems

5.1 Definitions

This is only an introduction to some of the many now standard theorems regarding convexity. For information on convex sets not given here, see nearly any of the many books titled *Convex sets...*, (*e.g.*, [594]). An introduction to more modern work on convex sets was written by Keith Ball [45].

Throughout this section, n is a positive integer and $\mathbb{R}^n$ is endowed with the usual metric (in which case, this space is often denoted by $\mathbb{E}^n$, the n-dimensional Euclidean space). If $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, then $\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$. The distance between points (or vectors) $\mathbf{x}$ and $\mathbf{y}$ is $\|\mathbf{x}-\mathbf{y}\|$.

Recall that a "linear combination" of points $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in $\mathbb{R}^n$, is any expression of the form

$$\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_k\mathbf{x}_k, \tag{5.1}$$

where each $\lambda_i \in \mathbb{R}$. The linear combination (5.1) is called an *affine combination* if and only if $\sum_{i=1}^k \lambda_i = 1$, and is called a *convex combination* if and only if $\sum_{i=1}^k \lambda_i = 1$ and each $\lambda_i \geq 0$.

A set $C \subseteq \mathbb{R}^n$ is called *affine* if and only if for any two points $\mathbf{x}$ and $\mathbf{y}$ in C, for every $\lambda \in \mathbb{R}$, the affine combination $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ is also in C. Intuitively, this says that if two points are in an affine set, the entire (infinite) straight line containing these two points is also in the set. One can show that the intersection of finitely many affine sets is again affine.

A set $C \subseteq \mathbb{R}^n$ is called *convex* if and only if for every $\mathbf{x}, \mathbf{y} \in C$ and for every $\lambda \in [0,1]$, the convex combination $\lambda\mathbf{x}+(1-\lambda)\mathbf{y}$ is also in C. Intuitively, this says that if two points are in a convex set, then the straight line *segment* containing these two points is also in the set. The following exercise is nearly trivial.

Exercise 194. *For $k \geq 2$, suppose that $C_1, \ldots, C_k \in \mathbb{R}^n$ are convex sets. Show that the intersection $\cap_{i=1}^{k} C_i$ is also convex.*

The solution of Exercise 194 shows also that the intersection of arbitrarily many convex sets is again convex. Similarly, the intersection of affine sets is again affine.

An affine set is, by definition (restricting the λ's) convex, however a convex set need not be affine (e.g., a straight line segment is convex, but is not affine). An affine set in $\mathbb{R}^n$ is sometimes called an affine *space* (for reasons made apparent below).

A linear subspace of $\mathbb{R}^n$ is closed under arbitrary finite linear combinations. Similar statements are true for convex and affine subspaces of $\mathbb{R}^n$.

Exercise 195. *Let $n \in \mathbb{Z}^+$. Let $C \subseteq \mathbb{R}^n$ be a convex set. Prove by induction on $m \geq 2$, that if $\mathbf{x}_1, \ldots, \mathbf{x}_m \in C$, then for any $\alpha_1, \ldots, \alpha_m \in [0,1]$ satisfying $\sum_{i=1}^{m} \alpha_i = 1$, the convex combination*

$$\alpha_1 \mathbf{x}_1 + \cdots + \alpha_m \mathbf{x}_m$$

is also in C. Repeat the exercise for affine sets.

It is not difficult to check that every plane, line, or point in $\mathbb{R}^3$ is an affine space. In fact, these are the only affine spaces in $\mathbb{R}^3$, as is stated in the next theorem.

Theorem 5.1.1. *Let $A \subset \mathbb{R}^n$ be an affine space. Then there exists a linear subspace (containing the origin) $W \subseteq \mathbb{R}^n$ and a vector (or point) $\mathbf{v} \in \mathbb{R}^n$ so that*

$$A = \mathbf{v} + W = \{\mathbf{v} + \mathbf{w} : \mathbf{w} \in W\}.$$

Proof: (The proof given here is based on [594, p. 14].) Pick $\mathbf{x} \in A$, and put

$$B = -\mathbf{x} + A = \{-\mathbf{x} + \mathbf{a} : \mathbf{a} \in A\}.$$

Then $A = \mathbf{x} + B$. To prove the theorem, it suffices to prove that B is indeed a (linear) subspace of $\mathbb{R}^n$. Observe that $\mathbf{0} \in B$ (by picking $\mathbf{a} = \mathbf{x} \in A$). Since

$B \subseteq \mathbb{R}^n$, it suffices to prove only that B is closed; so it remains to prove that for $k \in \mathbf{R}$, and $\mathbf{v_1}, \mathbf{v_2} \in B$, then $\mathbf{v_1} + k\mathbf{v}_2 \in B$.

For $i = 1, 2$, let $\mathbf{a}_i \in A$ satisfy $\mathbf{v}_i = -\mathbf{x} + \mathbf{a}_i$. Then

$$\begin{aligned}\mathbf{v_1} + k\mathbf{v}_2 &= (-\mathbf{x} + \mathbf{a}_1) + k(-\mathbf{x} + \mathbf{a}_2)\\ &= -\mathbf{x} + \mathbf{a}_1 - k\mathbf{x} + k\mathbf{a}_2\\ &= -\mathbf{x} + k\mathbf{a}_1 + k\mathbf{a}_2 - k\mathbf{x} + \mathbf{a}_1 - k\mathbf{a}_1\\ &= -\mathbf{x} + k(\mathbf{a}_1 + \mathbf{a}_2 - \mathbf{x}) + (1-k)\mathbf{a}_1\\ &= -\mathbf{x} + k\left(2\left[\frac{1}{2}\mathbf{a}_1 + \frac{1}{2}\mathbf{a}_2\right] - \mathbf{x}\right) + (1-k)\mathbf{a}_1.\end{aligned}$$

To see that this above expression is indeed an element of B, consecutively deduce (by noting that the coefficients add to 1) certain terms on the right are elements of A. Since $\frac{1}{2} + \frac{1}{2} = 1$,

$$\left[\frac{1}{2}\mathbf{a}_1 + \frac{1}{2}\mathbf{a}_2\right] \in A,$$

and so (with coefficients 2 and -1),

$$2\left[\frac{1}{2}\mathbf{a}_1 + \frac{1}{2}\mathbf{a}_2\right] - \mathbf{x} \in A.$$

Finally, since $k + (1-k) = 1$,

$$k\left(2\left[\frac{1}{2}\mathbf{a}_1 + \frac{1}{2}\mathbf{a}_2\right] - \mathbf{x}\right) + (1-k)\mathbf{a}_1 \in A.$$

So,

$$\mathbf{v_1} + k\mathbf{v}_2 \in -\mathbf{x} + A = B,$$

concluding the proof. □

When W is a (linear) subspace and $\mathbf{v}$ is a vector, a set of the form $\mathbf{v} + W$ in Theorem 5.1.1 is called a *translate* of W, a *shifted linear space*, an *affine space*, an *affine variety*, or a *flat*. The dimension of an affine space $A = \mathbf{v} + W$ is defined to be the dimension of W.

In linear algebra, if V is a vector space of dimension d, any linear subspace (containing the origin) of dimension $d-1$ is called a *hyperplane*, and often the word "hyperplane" is reserved only for linear subspaces. However, in the theory of convex sets, it is common to widen this definition.

Definition 5.1.2. An affine subspace $W \subseteq \mathbb{R}^n$ is called a *hyperplane* if W has dimension $n-1$.

Definition 5.1.3. The dimension of any set $S \subseteq \mathbb{R}^n$ is defined to be the minimum dimension of an affine subset containing S.

Recall from linear algebra, that vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ in a vector space V (over the reals) are called *linearly independent* if and only if for any scalars $c_1, \ldots, c_k \in \mathbb{R}$, the equation

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \tag{5.2}$$

is satisfied only when $c_1 = \cdots = c_k = 0$. If there exist c_i's not all zero satisfying the above equation, the vectors are *linearly dependent.*

If there exist $c_1, \ldots, c_k \in \mathbb{R}$ not all zero but with $\sum_{i=1}^{k} c_i = 0$ satisfying (5.2), then $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are called *affinely dependent*, and are *affinely independent* if the only c_i's satisfying $\sum_{i=1}^{k} c_i = 0$ and (5.2) are all zeros. Recall from linear algebra that any $n+1$ vectors in $\mathbb{R}^n$ are linearly dependent.

Lemma 5.1.4. *Let $n \in \mathbb{Z}^+$. Any $n+2$ vectors in $\mathbb{R}^n$ are affinely dependent.*

Proof: Let $\mathbf{x}_1, \ldots, \mathbf{x}_{n+2} \in \mathbb{R}^n$. The $n+1$ vectors $\mathbf{x}_2 - \mathbf{x}_1$, $\mathbf{x}_3 - \mathbf{x}_1$, $\ldots$, $\mathbf{x}_{n+2} - \mathbf{x}_1$ are linearly dependent, so let $\lambda_2, \ldots, \lambda_{n+2} \in \mathbb{R}$ (not all zero) satisfy

$$\lambda_2(\mathbf{x}_2 - \mathbf{x}_1) + \lambda_3(\mathbf{x}_3 - \mathbf{x}_1) + \cdots + \lambda_{n+2}(\mathbf{x}_{n+2} - \mathbf{x}_1) = \mathbf{0}.$$

Rewriting,

$$-\left(\sum_{i=2}^{n+2} \lambda_i\right)\mathbf{x_1} + \lambda_2\mathbf{x}_2 + \lambda_3\mathbf{x}_3 + \cdots + \lambda_{n+2}\mathbf{x}_{n+2} = \mathbf{0},$$

the coefficients sum to 0, and so this equality shows the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{n+2}$ are affinely dependent. $\square$

From the proof of Lemma 5.1.4, observe that by multiplication of an appropriate constant, any one of the $n+2$ vectors can be written as an affine combination of the remaining ones.

Definition 5.1.5. For a set $S \subseteq \mathbb{R}^n$, define the *convex hull* of S, denoted $\mathrm{conv}(S)$, to be the intersection of all convex sets containing S.

It is quite common to refer to the convex hull of a set S by simply $[S]$. The convex hull of a set S is the "smallest" convex set containing all of S. If C is a convex set, then $\text{conv}(C) = C$. By Exercise 195, the convex hull of S is the set of all convex linear combinations of points in S; for later reference, this fact is identified as a lemma:

Lemma 5.1.6. *A set S is convex if and only if every convex combination of finitely many points in S is also in S.*

Theorem 5.1.7. *Let S be a set and let T consist of all (finite) convex linear combinations of points in S. Then $conv(S) = T$.*

Proof: By definition, $S \subseteq T$. By Lemma 5.1.6, applied to the convex set $\text{conv}(S)$, $T \subseteq \text{conv}(S)$. It remains to show that $\text{conv}(S) \subseteq T$, and for this, it suffices to show that T is convex.

Suppose that $\mathbf{x}$ and $\mathbf{y}$ are points in T, that is, they are convex combinations of the form

$$\mathbf{x} = \sum_{i=1}^{r} \alpha_i \mathbf{x}_i \quad \text{and} \quad \mathbf{y} = \sum_{j=1}^{s} \beta_i \mathbf{y}_j,$$

where all the $\mathbf{x}_i$'s and $\mathbf{y}_j$'s are in S, each α_i and β_i are in $[0, 1]$ and $\sum_{i=1}^{r} \alpha_i = \sum_{i=1}^{s} \beta_i = 1$.

Then for any $\lambda \in [0, 1]$,

$$\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} = \sum_{i=1}^{r} \lambda\alpha_i \mathbf{x}_i + \sum_{j=1}^{s} (1 - \lambda)\beta_i \mathbf{y}_j$$

is a linear combination of points in S, where the sum of the coefficients is

$$\sum_{i=1}^{r} \lambda\alpha_i + \sum_{j=1}^{s} (1 - \lambda)\beta_i = \lambda \sum_{i=1}^{r} \alpha_i + (1 - \lambda) \sum_{j=1}^{s} \beta_i = \lambda(1) + (1 - \lambda)(1) = 1.$$

Thus T is convex, and so $\text{conv}(S) \subseteq T$. □

Any hyperplane H can be described as the set of points $\mathbf{x} \in \mathbb{R}^d$ satisfying an equation of the form $\mathbf{u} \bullet (\mathbf{x} - \mathbf{p}) = 0$, where $\mathbf{u}$ is a vector orthogonal to H and $\mathbf{p}$ is some point on H. Each hyperplane cuts $\mathbb{R}^d$ into two so-called *half-spaces* H^+ and H^-, convex open sets given by $H^+ = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{u} \bullet (\mathbf{x} - \mathbf{p}) > 0\}$ and $H^- = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{u} \bullet (\mathbf{x} - \mathbf{p}) < 0\}$.

For a proof of the following well-known theorem using "cones", see [427, pp. 31–32].

Theorem 5.1.8. *For $d \in \mathbb{Z}^+$, the convex hull of any finite set $A \subseteq \mathbb{R}^d$ is the intersection of all closed half-spaces in $\mathbb{R}^d$ containing A.*

Definition 5.1.9. The *affine hull* of a set X, denoted $\text{aff}(X)$, is the intersection of all affine spaces containing X.

Equivalently, the affine hull of a set X is the collection of all affine combinations of points in X. For example, the affine hull of any two distinct points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ is the line

$$L = \{\lambda\mathbf{x} + (1-\lambda)\mathbf{y} : \lambda \in \mathbb{R}\}.$$

5.2 Three main theorems for convexity

Theorem 5.2.1 (Carathéodory, 1907 [176])**.** *If $S \subseteq \mathbb{R}^n$, then every $\mathbf{x} \in$ conv(S) can be expressed as a convex linear combination of at most $n+1$ points from S.*

Proof: The idea of the proof is to suppose that some $\mathbf{x}$ is a convex combination of more than $n+1$ points in S, and from this combination, derive that one of these points can be discarded. Repeating this step sufficiently many times produces $\mathbf{x}$ as a combination of (at most) $n+1$ points in S.

Let $p \geq n+2$ and suppose that $\mathbf{x} \in \text{conv}(S)$ is a convex combination of points $\mathbf{x}_1, \ldots, \mathbf{x}_p$ from S. In other words, there exist $\lambda_1, \ldots, \lambda_p \in [0,1]$ with $\sum_{i=1}^p \lambda_i = 1$ so that

$$\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_p\mathbf{x}_p = \mathbf{x}. \tag{5.3}$$

Note that one can assume (if needed) that all the λ_is are non-zero (and so positive) because otherwise, fewer points are needed for the convex combination.

By Lemma 5.1.4, the $\mathbf{x}_i$s are affinely dependent, so let $\alpha_1, \ldots, \alpha_p \in \mathbb{R}$, not all zero, be so that $\sum_{i=1}^p \alpha_i = 0$ and

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_p\mathbf{x}_p = \mathbf{0}. \tag{5.4}$$

Multiplying the equation (5.4) through by $\frac{\lambda_1}{\alpha_1}$ and then subtracting from (5.3) eliminates the first term, leaving

$$\left(\lambda_2 - \frac{\lambda_1}{\alpha_1}\alpha_2\right)\mathbf{x}_2 + \cdots + \left(\lambda_p - \frac{\lambda_1}{\alpha_1}\alpha_p\right)\mathbf{x}_p = \mathbf{x}. \tag{5.5}$$

Using the facts that $\sum_{i=1}^{p} \lambda_i = 1$ and $\sum_{i=1}^{p} \alpha_i = 0$, it is not difficult to verify that the coefficients of the left side of (5.5) sum to 1. To show a convex combination, it suffices to have all coefficients positive. For this, observe that at least one of the α_i's is positive, and so one only need assume that

$$\frac{\lambda_1}{\alpha_1} = \min\left\{\frac{\lambda_i}{\alpha_i} : \alpha_i > 0\right\},$$

and this assumption can be made by relabelling. Then each of the coefficients in (5.5) is positive, and so (5.5) is a convex combination with fewer than p points. □

Another proof (see [118, p. 88]) of Carathéodory's theorem uses the following theorem due to Radon.

Theorem 5.2.2 (Radon's theorem, 1921 [739]). *Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ be a set of points in $\mathbb{R}^n$. If $r \geq n+2$, then S can be partitioned into two disjoint sets $S = S_1 \cup S_2$ so that $conv(S_1) \cap conv(S_2) \neq \emptyset$.*

Proof: Suppose $r \geq n+2$. By Lemma 5.1.4, there exist $\lambda_1, \ldots, \lambda_r$, not all zero, with $\sum_{i=1}^{r} \lambda_i = 0$ and $\sum_{i=1}^{r} \lambda_i \mathbf{x}_i = \mathbf{0}$. Since the λ_i's sum to zero, some are positive, and some are negative; without loss of generality, let $k \in \{1, \ldots, r-1\}$ be so that $\lambda_1, \ldots, \lambda_k$ are non-negative and $\lambda_{k+1}, \ldots, \lambda_r$ are all negative. Again since their sum is zero,

$$\lambda_1 + \cdots + \lambda_k = -(\lambda_{k+1} + \cdots + \lambda_r).$$

Letting $s = \lambda_1 + \cdots + \lambda_k > 0$, for each $i = 1, \ldots, r$ put $\alpha_i = \frac{\lambda_i}{s}$; then $\sum_{i=1}^{k} \alpha_i = 1$, $\sum_{i=k+1}^{r} -\alpha_i = 1$, and

$$\sum_{i=1}^{k} \alpha_i \mathbf{x}_i = \sum_{i=k+1}^{r} -\alpha_i \mathbf{x}_i,$$

a vector expressed as a convex combination of $S_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and as a convex combination of $S_2 = \{\mathbf{x}_{k+1}, \ldots, \mathbf{x}_r\}$. □

Exercise 196. *Let $d \in \mathbb{Z}^+$, and let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{d+2}\} \subset \mathbb{R}^d$. Prove that there exists a point $\mathbf{y} \in \mathbb{R}^d$ so that for any subcollection of $d+1$ of the points in X, y is a convex linear combination of these $d+1$ points.*

The following, perhaps surprising, theorem was published by Edward Helly (1884–1943) in 1923 [463], but appeared earlier in a paper published by Radon in 1921. Helly discovered this theorem and told Radon of it in 1913; however, Radon was delayed in publishing by joining the Austrian army that year, getting wounded by Russians, taken prisoner to Siberia, and only finding his way back to Vienna two years after the war ended.

Theorem 5.2.3 (Helly's theorem, $\leq$ 1923 [463])**.** *For $r > n \geq 1$, if convex sets $C_1, C_2, ..., C_r$ in $\mathbb{R}^n$ have the property that any $n+1$ of them share a common point, then some point is contained in all of the sets.*

By a compactness argument (see, *e.g.* [450, p. 60]) the number of convex sets in Helly's theorem may also be infinite (I think that the C_is need to be compact).

Using Exercise 196, one can show that Helly's theorem follows (see [116, Ex. 2, p. 86]); Helly's theorem also follows using Radon's theorem:

Exercise 197. *Prove Helly's theorem by induction on r, using Radon's theorem.*

For more references and a survey of applications of Helly's theorem, see the article by Danzer, Grünbaum, and Klee [235].

Exercise 198. *Let $k, \ell \in \mathbb{Z}^+$ satisfy $2 \leq k \leq \ell$. Let $\mathcal{S}$ be a family of segments on a line so that for any ℓ segments in $\mathcal{S}$, some k of these have a non-empty intersection. Show that there is a partition of $\mathcal{S}$ into $\ell - k + 1$ classes so that the segments in each class have non-empty intersection.*

Exercise 199. *Let $k, \ell \in \mathbb{Z}^+$ satisfy $2 \leq k \leq \ell$. Let $\mathcal{A}$ be a family of arcs on a circle so that for any ℓ arcs in $\mathcal{A}$, some k of these have a non-empty intersection. Show that there are at most $\ell - k + 2$ points on the circle so that each arc in $\mathcal{A}$ contains at least one of these points.*

Exercise 200. *Let $k \geq 2$ and let $R_1, \ldots, R_k$ be rectangles in $\mathbb{R}^2$ whose sides are parallel to the coordinate axes, where any two rectangles intersect (e.g., see Figure 5.1).*

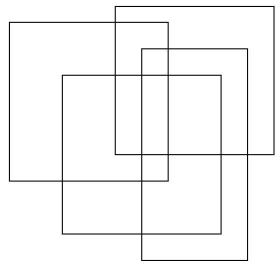


Figure 5.1: Four rectangles, any two of which intersect

Show that all rectangles have a common intersection.

Exercise 201. *Let $\mathcal{R}$ be a family of rectangles in $\mathbb{R}^2$ whose sides are parallel to the coordinate axes. If $\mathcal{R}$ contains at most n disjoint rectangles, show that there are at most $\binom{n+1}{2}$ points so that each rectangle in $\mathcal{R}$ contains at least one of these points.*

Exercise 202. *Let $\mathcal{R}$ be a family of "parallel rectangles" (e.g., assume that all rectangles are oriented to that edges are either vertical or horizontal). If for each rectangle $R \in \mathcal{R}$, there are at most m rectangles in $\mathcal{R}$ disjoint from R, show that there are at most $m + 1$ points so that each rectangle in $\mathcal{R}$ includes at least one of these points.*

Exercise 203. *Let K be a convex set in $\mathbb{R}^d$. For $n \geq d+1$, let $C_1, C_2, \ldots, C_n$ be subsets of $\mathbb{R}^d$ such that the intersection of any $d + 1$ of them contains a translation of K. Prove that the intersection of all C_is contains a translation of K. [Hint: Apply Helly's theorem.]*

5.3 Three more convexity theorems

Here are statements (without proofs) of three more theorems. There are many more versions of some of these (see, *e.g.*, [645]).

Theorem 5.3.1 (Fractional Helly, Katchalski–Liu, 1979 [529])**.** *For any $d \in \mathbb{Z}^+$ and any $\alpha > 0$, there exists $\beta = \beta(d, \alpha) > 0$ so that for any $n > d + 1$ convex sets $S_1, \ldots, S_n$ in $\mathbb{R}^d$, if there exist $\alpha\binom{n}{d+1}$ index sets I so that $\cap_{i \in I} S_i \neq \emptyset$, then there exists a point contained in at least βn of the S_is.*

Kalai (and others) proved that in the above statement, one could take $\beta = 1 - (1-\alpha)^{1/(d+1)}$. Other information regarding the fractional Helly theorem can be found in [645, p.197].

The original form of the following theorem is slightly different from the one presented here (which comes from [645, p. 199]); the original did not have the condition that S intersects each set in precisely one point, and $\mathbf{0}$ is replaced by an arbitrary point—Sergei Tsaturian has pointed out that the version presented here fails for small examples, a fact not noted in the source I used.

The following result was published by Imre Bárány in 1982.

Theorem 5.3.2 (Colourful Carathéodory, Bárány [53]). *Let $M_1, \ldots, M_{d+1}$ be finite point sets in $\mathbb{R}^d$, so that for each i, conv(M_i) contains the origin. Then there exists $S \subseteq \cup_{i=1}^{d+1} M_i$ with $|S| = d+1$ so that for each i, $|M_i \cap S| = 1$ and $\mathbf{0} \in$ conv(S).*

The next theorem (and its proof) was published by Tverberg [888] in 1966, and then he gave a simpler proof [889] in 1981; many other versions have appeared since, including a coloured version (see [645, pp. 203–205] for details).

Theorem 5.3.3 (Tverberg, 1966 [888]). *Let $d, r \in \mathbb{Z}^+$. If $A \subseteq \mathbb{R}^d$ contains at least $(d+1)(r-1)+1$ points, then there exist pairwise disjoint subsets $A_1, \ldots, A_r$ of A so that $\cap_{i=1}^{r}$conv$(A_i) \neq \emptyset$.*

In fact, in Theorem 5.3.3, one may assume that the sets A_i partition A.

Exercise 204. *Show why Tverberg's theorem generalizes Radon's theorem.*

For more problems and references surrounding Tverberg's theorem, see Gil Kalai's blog [519].

Chapter 6

Convex polytopes

6.1 Introduction

A "convex polytope" is a generalization of "convex polyhedron" to arbitrarily many dimensions. The theory of convex polytopes includes a broad array of topics; only a brief introduction is given here. Some central aspects of convex polytopes are not discussed here, including representation of polytopes by Gale diagrams, complex polytopes, the simplex method, or packing problems. For a more thorough understanding of convex polytopes, the reader is hereby highly recommended to see Grünbaum's now classic book, *Convex polytopes* [427]. Other introductions to convex polytopes include [120], [221], [222], [645], [868], and [940].

There are two equivalent definitions for a convex polytope; here is the first one:

Definition 6.1.1. For a positive integer d, a *convex d-polytope* P is the convex hull of some finite subset V of $\mathbb{R}^d$.

A second definition of a convex polytope depends on hyperplanes. Recall from Definition 5.1.2 that in $\mathbb{R}^n$, a hyperplane is an *affine* subspace of dimension $n-1$ and each hyperplane divides $\mathbb{R}^n$ into two half-spaces.

Definition 6.1.2. A subset $P \subset \mathbb{R}^d$ is a convex d-polytope if and only if P is the intersection of closed half-spaces in $\mathbb{R}^d$ containing P.

By Theorem 5.1.8, Definition 6.1.2 is equivalent to Definition 6.1.1.

If $P = \mathrm{conv}(V)$, the set $V = V(P)$ is called the vertex set of P. Throughout this document, "polytope" means "convex polytope" (i.e., for some d, a convex d-polytope). A *subpolytope* of a polytope P is the convex hull of some subset of $V(P)$.

The following four facts hold; however, the proof of each may be easier using one definition over the other:

- The intersection of a polytope with an affine subspace is again a polytope.
- The intersection of two polytopes is again a polytope.
- The projection of a polytope is again a polytope.
- If P and Q are polytopes in $\mathbb{R}^d$, then so is their "Minkowski sum"

$$P + Q = \{\mathbf{p} + \mathbf{q} : \mathbf{p} \in P, \mathbf{q} \in Q\}.$$

For any polytope $P \subset \mathbb{R}^d$, a *face* of P is either the empty set, P itself, or the intersection of P with a supporting (touching, but not cutting through) affine space (often called a "hyperplane", but with arbitrary dimension).

Definition 6.1.3. Let P be a d-polytope. A j-face of P is a face with affine dimension j. A (-1)-face is the empty set, 0-faces are vertices, 1-faces are edges, and $(d-1)$-faces are called *facets*.

In case notation is needed, let $\mathcal{F}_j(P)$ denote the set of j-faces of a polytope P; one might write $\mathcal{F}_0(P) = V(P)$ to denote vertices of P.

For $j = -1, 0, 1, \ldots, d-2$, each j-face of a d-polytope is contained in a $(j+1)$-face (proof to be added), and hence the faces of a polytope P are ordered by inclusion, yielding the *face lattice* of P, denoted $\mathcal{L}(P)$. For more on such lattices, see Section 21.1.

Definition 6.1.4. Two polytopes are called *combinatorially equivalent* if and only if they have isomorphic face lattices.

The following lemma has two proofs; both can be found in, for example, [426, p.33]. It gives a property of convex polytopes that is often implicitly used, but rarely referred to explicitly.

Lemma 6.1.5. *Let P be a polytope. If F_1 is a face of P and F_2 is a face of F_1, then F_2 is a face of P.*

A *k-simplex* in $\mathbb{R}^d$ is the convex hull of $k+1$ affinely independent points. (Sometimes only the point set is called a simplex.)

Definition 6.1.6. A d-polytope is *simplicial* if each of its proper faces is a simplex.

(Note: Some define a polytope to be simplicial if each of its facets is a simplex, but since a face of a simplex is again a simplex, Lemma 6.1.5 shows that this definition is equivalent to the original.) For example, any tetrahedron is a simplex and so is simplicial. An octahedron is simplicial.

There are a number of simple [sic] results regarding simplicial polytopes; if P is a simplicial polytope, then:

- Every proper face of P is simplicial.
- Every facet has d vertices.
- Every k-face has $k+1$ vertices.

In fact (see, *e.g.*, [940, pp. 65–66]), each of the above three conditions is equivalent to being simplicial.

It might be of interest that every simplicial polytope is combinatorially equivalent to one with integer coordinates (see [940, p. 66]).

6.2 Polytopes on a moment curve

For a positive integer d, the set

$$M_d = \{(t, t^2, \ldots, t^d) : t \in \mathbb{R}\} \subset \mathbb{R}^d$$

is called the standard *moment curve* (of order d). There are other examples of so-called moment curves—for example, the trigonometric moment curve in even dimensions is

$$\{(\cos\theta, \sin\theta, \cos 2\theta, \sin 2\theta, \ldots, \cos m\theta, \sin m\theta) : 0 \le \theta \le 2\pi\},$$

studied by Carathéodory [177]. See the recent article by Barvinok [62] on "the symmetric moment curve"

$$\{(\cos\theta, \sin\theta, \cos 3\theta, \sin 3\theta, \ldots, \cos(2k-1)\theta, \sin(2k-1)\theta) : 0 \le \theta \le 2\pi\},$$

and a review of some connections between moment curves and polytopes. (Barvinok [62] also briefly references the connection between moment curves and symmetric polytopes and their use in sparse signal transmission.) Only a few basic ideas are given here for moment curves and simplicial polytopes.

Denote points on the moment curve M_d by $\mathbf{x}(t) = (t, t^2, \ldots, t^d)$. If $n \geq d+1$, and $t_1 < t_2 < \cdots < t_n$ are real numbers, then define the d-polytope

$$C_d(t_1, t_2, \ldots, t_n) = [\{\mathbf{x}(t_1), \mathbf{x}(t_2), \ldots, \mathbf{x}(t_n)\}].$$

If P is such a polytope for some choice of $t_1 < t_2 < \cdots < t_n$, whenever the vertex set of P is written in a list, say $V(P) = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, it is understood that the labelling of the $\mathbf{x}_i$s respect the ordering of the t_is, that is, if for each i, $\mathbf{x}_i = \mathbf{x}(t_i)$, then $i < j$ if and only if $t_i < t_j$. Although the terminology is not standard, for the moment (no pun intended), one might call $C_d(t_1, t_2, \ldots, t_n)$ a *moment curve d-polytope*.

Before looking at the next theorem, the following exercise (found in, e.g., [923, p. 81]) might be enlightening:

Exercise 205. *Show that no four points in* $\{(t, t^2, t^3) : t \geq 0\} \subseteq \mathbb{R}^3$ *are coplanar.*

Theorem 6.2.1. *Let* $d \in \mathbb{Z}^+$*, and for* $n \geq d+1$*, let* $t_1 < t_2 < \cdots < t_n$ *be real numbers. Any* $d+1$ *vertices of* $C_d(t_1, t_2, \ldots, t_n)$ *are affinely independent.*

Proof: Let $t_0^*, t_1^*, \ldots, t_d^*$ be an ordered subsequence of $t_1, t_2, \ldots, t_n$.

The set $\{\mathbf{x}(t_0^*), \ldots, \mathbf{x}(t_d^*)\}$ is affinely independent if and only if $\sum_{i=0}^{d} \lambda_i \mathbf{x}(t_i^*) = \mathbf{0}$ and $\sum_{i=0}^{d} \lambda_i = 0$ imply that each $\lambda_i = 0$. For clarity, first rewrite these equations in the following form:

$$\begin{aligned}
\sum_{i=0}^{d} \lambda_i &= 0; \\
\sum_{i=0}^{d} \lambda_i t_i^* &= 0; \\
\sum_{i=0}^{d} \lambda_i (t_i^*)^2 &= 0; \\
&\vdots
\end{aligned}$$

$$\sum_{i=0}^{d} \lambda_i (t_i^*)^d = 0.$$

These $d+1$ equations can be put in matrix form:

$$\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ t_0^* & t_1^* & t_2^* & \dots & t_d^* \\ (t_0^*)^2 & (t_1^*)^2 & (t_2^*)^2 & \dots & (t_d^*)^2 \\ \vdots & \vdots & \vdots & & \vdots \\ (t_0^*)^d & (t_1^*)^d & (t_2^*)^d & \dots & (t_d^*)^d \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The above matrix on the left is a Vandermonde matrix with determinant $\prod_{i<j}(t_j^* - t_i^*) > 0$, and so the homogeneous system above has only the trivial solution with each $\lambda_i = 0$. □

Corollary 6.2.2. *For any real numbers $t_1 < t_2 < \cdots < t_n$, the d-polytope $C_d(t_1, t_2, \ldots, t_n)$ is simplicial.*

Proof: Examine any facet $F = \text{conv}(V')$ of $C_d(t_1, t_2, \ldots, t_n)$, that is, F has affine dimension $(d-1)$. If $|V'| > d$, then by Theorem 6.2.1, V' contains $d+1$ affinely independent points, that is, F has affine dimension d, a contradiction. So $|V'| \leq d$ (in fact, $|V'| = d$, since it takes at least d points to have affine dimension $d-1$). Since any $d+1$ points are affinely independent, so are any d points; in particular, V' is an affinely independent set, that is, $\text{conv}(V')$ is a simplex. □

Definition 6.2.3. A polytope P is said to satisfy *Gale's evenness condition* (GEC) with respect to a linear ordering $v_1 < v_2 < \ldots < v_n$ of $V(P)$ if for any $X \subset V(P)$, if $[X]$ is a facet of P then any two distinct vertices in $V(P)\backslash X$ are separated by an even number of vertices in X.

For a polytope P, if there exists a linear ordering of $V(P)$ that satisies GEC, then say P is a *Gale polytope*.

Exercise 206. *Let P be the unit cube polytope on vertices $V(P) = \{0, 1\}^3$. Find an ordering of $V(P)$ confirming that P is a Gale polytope. Is any higher dimensional cube also Gale?*

For relations between the above exercise and applications to randomized simplex algorithms, see, for example, [380], where slightly deformed cubes (called Klee-Minty cubes) are used.

Theorem 6.2.4. *For any choice of real numbers $t_1 < t_2 < \cdots < t_n$, the d-polytope $C_d(t_1, t_2, \ldots, t_n)$ satisfies Gale's evenness condition (with respect to the natural ordering given by $\mathbf{x}(t_i) < \mathbf{x}(t_j)$ if and only if $i < j$).*

Proof: Let $t_1 < t_2 < \cdots < t_n$ be fixed and put $P = C_d(t_1, t_2, \ldots, t_n)$ and let $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be the vertex set of P where for each i, $\mathbf{v}_i = \mathbf{x}(t_i)$.

By Theorem 6.2.1, any d points in V are affinely independent, and so any facet of P has precisely d points.

Let $t_1^* < t_2^* < \ldots < t_d^*$ be a subsequence of the t_i's and put $X = \{\mathbf{x}(t_1^*), \mathbf{x}(t_2^*), \ldots, \mathbf{x}(t_d^*)\}$. The points of X are affinely independent and so are contained in a unique affine hyperplane. Writing $\prod_{i=1}^{d}(t - t_i^*) = \sum_{i=0}^{d} \gamma_i t^i$, it is not too difficult to see that

$$H = \{w \in \mathbb{R}^d : w \bullet (\gamma_1, \ldots, \gamma_d) + \prod_{i=1}^{n}(-t_i^*) = 0\}$$

is the hyperplane containing $\mathbf{x}(t_1^*), \mathbf{x}(t_2^*), \ldots, \mathbf{x}(t_d^*)$, since for each i,

$$(t_i^*, (t_i^*)^2, \ldots, (t_i^*)^d) \bullet (\gamma_1, \ldots, \gamma_d) + \prod_{i=1}^{n}(-t_i^*) = 0.$$

Examine the function $f(t) = \prod_{i=1}^{d}(t - t_i^*)$. Observe that for each i, $f(t_i^*) = 0$ (corresponding to where the moment curve intersects H), $f(t)$ is non-zero elsewhere, and when t passes through one of the t_i^*'s, $f(t)$ changes sign. Two points, $\mathbf{x}(t_i)$ and $\mathbf{x}(t_j)$ of $C_d(t_1, t_2, \ldots, t_n)$ lie on the same side of H if and only if $f(t_i)$ and $f(t_j)$ are the same sign. So $\mathbf{x}(t_i)$ and $\mathbf{x}(t_j)$ in $V \backslash X$ lie on the same side of H iff there are an even number of roots of $f(t) = 0$ between t_i and t_j, that is, there are an even number of points in X which are between $\mathbf{x}(t_i)$ and $\mathbf{x}(t_j)$.

So, since $[X]$ has affine dimension $(d - 1)$, $[X]$ is a facet if and only if all points of $V \backslash X$ lie on one side of H, (that is, H is a supporting hyperplane) and this condition holds if and only if between any two points of $V \backslash X$ there are an even number of points in X. $\square$

6.3 Cyclic polytopes

Since Theorem 6.2.4 completely describes the facets of the moment d-polytope with n vertices, it then follows that for any choice of $t_1 < t_2 < \cdots < t_n$ and $s_1 < s_2 < \cdots < s_n$, the two moment polytopes $C_d(t_1, t_2, \ldots, t_n)$ and $C_d(s_1, s_2, \ldots, s_n)$ are combinatorially equivalent. Hence, the notation $C(n, d)$ denotes any element of the equivalence class of moment d-polytopes on n vertices.

Definition 6.3.1. A d-polytope P on n vertices is called *cyclic* if and only if P is combinatorially equivalent to $C(n, d)$.

Theorem 6.3.2 (see [102])**.** *A convex polytope P is cyclic if and only if P is simplicial and Gale.*

Why are these creatures called "cyclic"? The answer lies in the following lemma.

Lemma 6.3.3. *Let d be even and let P be a cyclic polytope with vertex set $V = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. If $X \subset V$ with $|X| = d$ satisfies GEC with respect to the ordering $\mathbf{x}_1 < \mathbf{x}_2 < \cdots < \mathbf{x}_n$, then X satisfies GEC with respect to the ordering $\mathbf{x}_2 \preceq \mathbf{x}_3 \preceq \cdots \preceq \mathbf{x}_n \preceq \mathbf{x}_1$.*

Proof idea: Suppose that $X \subset V$ with $|X| = d$ satisfies GEC with respect to $<$, that is, $(X, <)$ is so that X separates (in $<$) any two elements in $V \backslash X$ by an even number of vertices in X. Now suppose that in $(X, <)$, $1 \leq i < j \leq n$ are so that $A = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i\} \subset X$ with $\mathbf{x}_{i+1} \notin X$ and $B = \{\mathbf{x}_j, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_n\} \subset X$ with $\mathbf{x}_{j-1} \notin X$. (Note: A or B might be empty, that is, such an i or j might not exist.) By GEC, $|X \backslash (A \cup B)|$ is even, and since d is even, then $|A \cup B|$ is even as well. Now check the possible ways for A and B to sit in $(X, \preceq)$ and to verify that GEC still holds with respect to $\preceq$ as well. Then apply Theorem 6.3.2. $\square$

A natural question one might ask is: are there cyclic d-polytopes that are not moment d-polytopes?

Definition 6.3.4. Two polytopes are *geometrically equivalent* if and only if all respective subpolytopes are combinatorially equivalent.

As an example, let P be the (3-dimensional) octahedron, and let Q be a polytope found by slightly perturbing a vertex or two of P. Then P and Q are combinatorially equivalent; however, they are not geometrically equivalent (proof left as an exercise). It turns out that in even dimensions, cyclic polytopes are geometrically equivalent to moment polytopes, (maybe see [104]) and [102]?); however in odd dimensions, there are other cyclic polytopes.

The following result by Shemer [790, Thm. 2.12] is quoted in [101, Lemma 3].

Theorem 6.3.5 (Shemer, 1982)**.** *If d is even, any d-subpolytope of a cyclic d-polytope is also cyclic.*

In odd dimensions, however, subpolytopes of a cyclic d-polytope need not be cyclic:

Example 6.3.6. *Let $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_6\} \subset \mathbb{R}^3$ where $\mathbf{v}_1 = (0, 0, 1)$, $\mathbf{v}_2 = (2, 1, 0)$, $\mathbf{v}_3 = (1, 2, 0)$ $\mathbf{v}_4 = (-1, 2, 0)$, $\mathbf{v}_5 = (-2, 1, 0)$, and $\mathbf{v}_6 = (0, 0, -1)$. Put $P = [V]$.*

The facets of P are precisely those triples that satisfy Gale's evenness condition, namely $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_6\}$, $\{\mathbf{v}_1, \mathbf{v}_5, \mathbf{v}_6\}$,$\{\mathbf{v}_1, \mathbf{v}_i, \mathbf{v}_{i+1}\} : i = 2, 3, 4,$, and $\{\mathbf{v}_i, \mathbf{v}_{i+1}, \mathbf{v}_6\} : i = 2, 3, 4$. Note also that P is combinatorially equivalent to $C(6, 3)$. Let $V' = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4, \mathbf{v}_5\}$ and $P' = [V]$. The vertices $\mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4, \mathbf{v}_5$ form a facet of P' that is not a simplex, and hence P' is not cyclic.

Theorem 6.3.7 (Sturmfels, 1987 [835])**.** *A d-polytope P and each of its d-subpolytopes is cyclic if and only if there is a moment curve polytope that is geometrically equivalent to P.*

So if one restricts to cyclic polytopes whose every subpolytope is cyclic, then one can use the moment curve to completely analyze them.

In 1970, McMullen [651] proved that among all d-polytopes with n vertices, for each $j = 1, \ldots, d - 1$, cyclic polytopes have the maximum number of j-faces. (This result was conjectured by Motzkin in 1957.) For convex polyhedra (3-polytopes) with v vertices, e edges and f faces, recall that by Lemma 2.3.3, $e \leq 3v - 6$, and putting this back into Euler's formula gives $f \leq 2v - 4$. So for polyhedra, the number of both 1-faces and 2-faces is linear in n. However, for cyclic d-polytopes on n vertices, the number of faces is on the order of $n^{d/2}$.

Theorem 6.2.4 says something only about when a collection of d vertices in a cyclic polytope form a facet; what about for faces in general? Here is some preparation:

Given an ordered set of vertices $V = \{\mathbf{x}_1 < \mathbf{x}_2 < \cdots < \mathbf{x}_n\}$, say $Y \subset V$ is *contiguous* if $Y = \{\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_j\}$ for some i and j satisfying $1 < i \leq j < n$ and $Y \cap \{\mathbf{x}_{i-1}, \mathbf{x}_{j+1}\} = \emptyset$. A set $Z \subset V$ is called an *endset* if either for some i, $Z = \{\mathbf{x}_1, \ldots, \mathbf{x}_i\}$ and $\mathbf{x}_{i+1} \notin Z$, or if for some i, $Z = \{\mathbf{x}_i, \ldots, \mathbf{x}_n\}$ and $\mathbf{x}_{i-1} \notin Z$. Any $V' \subset V$ can thus be uniquely written

$$V' = Z_1 \cup Y_1 \cup \cdots \cup Y_a \cup Z_2,$$

where each Z_i is an end-set or is empty, and each Y_i is contiguous. Say that a contiguous set $Y \subset V$ is even [odd] if and only if $|Y|$ is even [resp. odd].

Theorem 6.3.8 (Shephard, 1968 [791])**.** *Let P be a cyclic d-polytope with linearly ordered vertex set $V(P)$ and let $0 \leq j \leq d-1$. Let $V' \subset V(P)$, $|V'| = j+1$, and write V' as the disjoint union of end-sets and contiguous sets*

$$V' = Z_1 \cup Y_1 \cup Y_2 \cup \cdots \cup Y_n \cup Z_2.$$

Then V' is a j-face of P if and only if at most $d-j-1$ of the contiguous sets Y_i are odd.

Note that in the case $j = d-1$, Shephard's theorem says that if P is a cyclic d-polytope, a subset $V' \subset V(P)$ with $|V'| = d$ is the vertex set of a facet if and only if there are no odd contiguous sets Y_i in the above representation, that is, all contiguous sets in V' are even. This says that any two vertices not in V' are separated by an even number of of vertices in V', and so Shephard's theorem implies Gale's evenness condition.

Perhaps of independent interest is work on the number of hyperplanes required to cut each face of a cyclic polytope; see [103] for bounds on such numbers that arise from Gray codes and partitioning of boolean lattices. [Comment: Many of the notes in this chapter were written while I was studying polytopes under the very patient Ted Bisztriczky so that I might help with [103].]

6.4 Neighbourly polytopes

Definition 6.4.1. A d-polytope P is called *k-neighbourly* if and only if every set of k vertices in $V(P)$ is the vertex set of a face of P.

The next four results are given by Grunbaum [426, pp. 122–123].

Theorem 6.4.2. *If P is a k-neighbourly d-polytope, then every k vertices of P are affinely independent.*

Proof: Let P be k-neighbourly, and let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\} \subset V(P)$. In hopes of contradiction, assume that X is an affinely dependent set, with $\mathbf{x}_d \in \text{aff}(\{\mathbf{x}_1, \ldots, \mathbf{x}_{d-1}\})$. Put $X' = \{\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}\}$. For some $w \notin X'$, let $W = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}, w\}$. Since P is k-neighbourly, $\text{conv}(W)$ is a face of P and the vertex set of $\text{conv}(W)$ is precisely W, that is, $\mathbf{x}_d \notin W$.

Since a face is the intersection of a hyperplane with P, in any such hyperplane H with $\text{conv}(W) \subset H$, then $\text{aff}(W) \subset H$, and since $\mathbf{x}_d \in \text{aff}(X') \subseteq \text{aff}(W)$, it follows that $\mathbf{x}_d \in H$. Since $\mathbf{x}_d$ is contained in every such H, $\mathbf{x}_d$ is a vertex of $\text{conv}(W)$, a contradiction. □

If F is a $(k-1)$-face of a d-polytope P, then it has at least k vertices. Theorem 6.4.2 says that if P is k-neighbourly, then any k vertices are affinely independent, so $(k-1)$-faces contain at most k vertices. Hence, if P is k-neighbourly, every $(k-1)$-face of P is a $(k-1)$-simplex.

Theorem 6.4.3. *If P is k-neighbourly, and $1 \leq k^* \leq k$, then P is also k^*-neighbourly.*

Proof: Let $1 \leq k^* \leq k$ and let P be k-neighbourly. Any set K of k^* vertices sits in some $(k-1)$-simplex S which is a face of P, and so (by Lemma 6.1.5) there is a face F with $V(F) = K$. □

Theorem 6.4.4. *If P is a k-neighbourly d-polytope and $V' \subset V(P)$ with $|V'| > k$, then $\text{conv}(V')$ is k-neighbourly.*

Theorem 6.4.5. *If P is a d-polytope which is k-neighbourly for some $k > \lfloor d/2 \rfloor$, then P is a d-simplex.*

Proof: Suppose that P is a d-polytope which is k-neighbourly for some $k > \lfloor d/2 \rfloor$. Assume, in hopes of contradiction, that P is not a simplex, *i.e.*, $|V(P)| > d + 1$, and let X be a set of $d + 2$ vertices in $V(P)$.

By Radon's theorem (Theorem 5.2.2) there exist disjoint sets B and C so that $B \cup C = X$ and $\text{conv}(B) \cap \text{conv}(C) \neq \emptyset$. Without loss of generality, assume that B is the smaller set, that is, $|B| \leq \frac{d+2}{2} = \frac{d}{2} + 1 \leq k$. The fact that $\text{conv}(B) \cap \text{conv}(C) \neq \emptyset$ implies that for every supporting hyperplane H which contains B, $\text{conv}(C) \cap H \neq \emptyset$. A supporting hyperplane touches some extremal points of a subset of vertices, and so for every supporting hyperplane H that contains B, $C \cap H \neq \emptyset$. Thus, every hyperplane supporting B contains another point of X, and so B is not the vertex set of a face.

On the other hand, since $|B| \leq k$, and P is k-neighbourly, Theorem 6.4.3 yields that $\mathrm{conv}(B)$ is a face, the desired contradiction. □

By virtue of Theorem 6.4.5, if a polytope P is not a simplex, then P can be at most $\lfloor d/2 \rfloor$-neighbourly.

Definition 6.4.6. A d-polytope is called *neighbourly* if it is $\lfloor d/2 \rfloor$-neighbourly.

Motzkin conjectured (see [362]) that all neighbourly polytopes are combinatorially equivalent to cyclic polytopes; this is now known to be false (see [362], [790]), however the following holds:

Theorem 6.4.7 (Shephard, 1968 [791]). *Cyclic d-polytopes are neighbourly.*

Proof: Let P be a cyclic d-polytope and let X be a set of $\lfloor d/2 \rfloor$ vertices in $V(P)$. The affine dimension of $\mathrm{conv}(X)$ is $\lfloor d/2 \rfloor - 1$, so X is the vertex set of a $(\lfloor d/2 \rfloor - 1)$-face if and only if X has at most $d - (\lfloor d/2 \rfloor - 1) - 1 = \lceil d/2 \rceil$ odd contiguous subsets; however, X only has $\lfloor d/2 \rfloor$ vertices, and so has at most $\lfloor d/2 \rfloor \leq \lceil d/2 \rceil$ odd contiguous subsets. Hence X is a face. □

It may be of interest to note that Bárány [52] used neighbourly polytopes and a theorem of Gale to prove Kneser's conjecture, which says that if the n-sets from a $(2n + k)$-set are partitioned into $k + 1$ classes, then one of the classes contains two disjoint n-sets. (The theorem of Gale [361] used says that in $\mathbb{R}^d$, there exist $\lfloor d/2 \rfloor$-neighbourly polytopes with any number of vertices.)

6.5 Two more classes of polytopes

Two more types of polytopes are easily described and have relations to many other areas of mathematics.

Definition 6.5.1. A 0/1 polytope P in $\mathbb{R}^d$ is one whose every vertex is a 0-1 vector; so P is the convex hull of vertices in the unit cube Q^d.

To describe one particular 0/1 polytope, a common definition is recalled: A *permutation matrix* is a square 0-1 matrix with one 1 in each row and one 1 in each column. The *Birkhoff polytope* has its vertices being the permutation matrices in $M_{n\times n}(\{0,1\})$.

Definition 6.5.2. A matrix $A = (a_{i,j})$ is called doubly stochastic if each $a_{i,j} \geq 0$, and all entries in any row sum to 1 and all entries in any column sum to 1.

So a permutation matrix is a doubly stochastic matrix. It follows from the definition that a doubly stochastic matrix is square, since the sum of all entries counted according to rows or according to columns is the same. The following theorem can be found in many graph theory texts since one proof relies on Hall's matching theorem:

Theorem 6.5.3 (König, 1935 [569]; Birkhoff, 1946 [97]; von Neumann, 1953 [903])**.** *Any doubly stochastic matrix is a convex combination of permutation matrices.*

So Theorem 6.5.3 says that the set of all $n \times n$ doubly stochastic matrices is the polytope formed by the convex hull of all $n \times n$ permutation matrices. These polytopes are also called *assignment polytopes* or *polytopes of doubly stochastic matrices*. See [940, p. 20–21] for more references (including the relation to the Travelling Salesman Problem and other linear programming problems).

The theory of cyclic polytopes is rather rich, perhaps because they are so easy to describe in an organized way. Another kind of polytope is also somewhat easy to see. In various texts, "cross-polytopes" are defined differently. Letting $\mathbf{e}_i$ denote a standard basis vector in $\mathbb{R}^d$, some define the d-dimensional cross-polytope to be the convex hull of $\{\mathbf{e}_1, -\mathbf{e}_1, \ldots, \mathbf{e}_d, -\mathbf{e}_d\}$. So, for example, when $d = 3$, the polytope is simply the octahedron. Another way to define the cross-polytope is the set of all $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ satisfying $\sum_{i=1}^{d} |x_i| \leq 1$. Slightly more general, it can be defined by the convex hull of d convex segments, each the same length, and any two orthogonal, and all intersecting at their midpoints. Cross-polytopes are very well-studied.

Exercise 207. *Are cross-polytopes simplicial?*

6.6 Euler's formula for polytopes

Euler's formula $v + f = e + 2$ for polyhedra (Theorem 3.3.3) has a generalization to polytopes (given below). According to Peter M. Gruber (1941–2017), an Austrian expert in convex and discrete geometry (see [422] or [423]), Schläfi found this generalization in the mid 1800s, but his proof was not quite complete; there were later other proofs, but the first elementary proof of this formula is due to Hadwiger in 1955.

Theorem 6.6.1. *Let P be a convex polytope in $\mathbb{E}^d$, where for each $i = 0, 1, \ldots, d-1$, P has f_i i-dimensional faces. Then*

$$f_0 - f_1 + \cdots + (-1)^{d-1} f_{d-1} = 1 + (-1)^{d-1}. \tag{6.1}$$

When $d = 3$, equation (6.1) says $v - e + f = 2$.

Remark 6.6.2. *Schläfi developed a notation for regular "polytopes" (generalizing regular polyhedra) that extends to not-necessarily-convex regular polyhedra; this notation is used extensively by many geometers; see, e.g., [218]. An entire chapter could be written about polyhedra and their Schläfi symbols.*

Chapter 7

Polygons and early Ramsey theory

7.1 Ramsey's theorem

Only a brief exposure to Ramsey's theorem is given here, as only some basic results are needed. For more on Ramsey's theorem, see [410].

Definition 7.1.1. For positive integers n, m, k, r, let

$$n \longrightarrow (m)^k_r$$

mean that for any colouring $\Delta : [n]^k \to [r]$, there exists $A \in [n]^m$ so that Δ is constant on $[A]^k$.

The Ramsey arrow notation $n \longrightarrow (m)^k_r$ is due to Erdős and Rado [314].

Theorem 7.1.2 (Ramsey, 1930 [745] (finite version))**.** *For any $m, k, r \in \mathbb{Z}^+$, there is a smallest positive integer $n = R_k(m; r)$ so that $n \longrightarrow (m)^k_r$.*

In 1935, Erdős and Szekeres [320] proved a case of Ramsey's theorem that they then used to solve a problem about points in convex position in the plane. At that time, they were unaware of Ramsey's theorem. For positive integers a, b, let $R(a, b)$ be the least integer n so that if the edges of K_n are coloured with red and blue, then there exists either a copy of K_a all of whose edges are red, or a copy of K_b all of whose edges are blue.

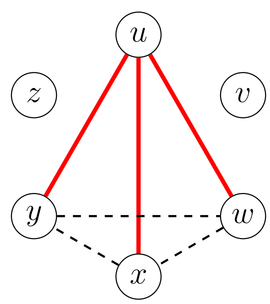
Exercise 208. *For $s \geq 2$, show that $R(2, s)$ exists and find its value.*

Theorem 7.1.3. $R(3, 3) = 6$.

Proof: There are two things to show: $R(3, 3) \leq 6$ and $R(3, 3) > 5$.

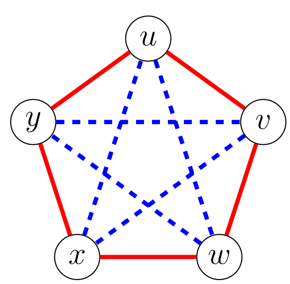
$\mathbf{R(3, 3) \leq 6}$: Let a copy of K_6 have vertices u, v, w, x, y, z, and let $c : E(K_6) \to \{\text{red,blue}\}$. Of the five edges incident with u, by the PHP, at least three are of the same colour, say

$$c(\{u, w\}) = c(\{u, x\}) = c(\{u, y\}) = \text{red}.$$



If any of $\{w, x\}$, $\{x, y\}$ or $\{w, y\}$ are also coloured red, then a monochromatic (red) triangle is found. If these three edges are all coloured blue, then a monochromatic (blue) triangle is found. Thus $R(3, 3) \leq 6$.

$\mathbf{R(3, 3) > 5}$: Let a copy of K_5 have vertices u, v, w, x, y. Colour the edges $\{u, v\}, \{v, w\}, \{w, x\}, \{x, y\}, \{y, u\}$ red, and the remaining five edges blue. Under this colouring, there is no monochromatic triangle, so $R(3, 3) > 5$. □



Theorem 7.1.3 is often stated as a "party problem": if six people are at a party, show that either there are 3 that either all know each other or 3 that are all strangers to one another (assuming that "know" is symmetric).

Theorem 7.1.3 appeared in 1935 in the seminal paper (that marked one of two beginnings of the study of Ramsey numbers) by Erdős and Szekeres [320]; this result again appeared 1955 in a paper by Greenwood and Gleason [390] (although they did not mention the first paper). The $R(3,3) \leq 6$ part of Theorem 7.1.3 also appeared as a question in the 1953 Putnam exam.

Using $k = 2$ and $m = \max\{a, b\}$, the existence of $R(a, b)$ is guaranteed by Ramsey's theorem. On the other hand, Erdős and Szekeres proved the existence of $R(a, b)$ by proving a recursion:

Theorem 7.1.4 (Erdős-Szekeres recursion, simple case, 1935)**.** *For integers $s, t \geq 2$,*

$$R(s+1, t+1) \leq R(s, t+1) + R(s+1, t).$$

Proof: Put $n_0 = R(s, t+1)$, $n_1 = R(s+1, t)$, and $n = n_0 + n_1$, and let $\Delta : E(K_n) \to \{0, 1\}$. Fix $x \in V(K_n)$ and for each $i \in \{0, 1\}$, set $X_i = \{y : \Delta(x, y) = i\}$. By the pigeonhole principle, for some $i \in \{0, 1\}$, $|X_i| \geq n_i$.

If $|X_0| \geq n_0$, then in the graph induced by X_0, there is either a 0-monochromatic K_s or a 1-monochromatic K_{t+1}; if there is a 0-monochromatic K_s, then together with x a 0-monochromatic K_{s+1} is formed.

The proof follows similarly when $|X_0| \geq n_1$. □

The following consequence of Theorem 7.1.4 gives an explicit upper bound on the numbers $R(k, \ell)$:

Corollary 7.1.5 (Erdős-Szekeres, 1935 [320])**.** $R(s,t) \leq \binom{s+t-2}{t-1}$.

Exercise 209. *Prove Corollary 7.1.5 by induction.*

Using Stirling's formula for approximating factorials, $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, Corollary 7.1.5 yields the following upper bound for diagonal Ramsey numbers.

Theorem 7.1.6. *There exists a universal constant c so that for every $k \geq 3$, $R(k, k) < \frac{c}{\sqrt{k}} 2^{2k}$.*

Proof:

$$\binom{2k-2}{k-1} = \frac{(2k-2)!}{((k-1)!)^2}$$

$$
\begin{aligned}
&= (1+o(1))\frac{\sqrt{2\pi(2k-2)}(2k-2)^{2k-2}}{2\pi(k-1)(k-1)^{2k-2}}\\
&= (1+o(1))\frac{2^{2k-2}}{\sqrt{\pi(k-1)}}\\
&< \frac{1}{6\sqrt{k}}2^{2k}.
\end{aligned}
$$

□

Upper bounds for $R(k,k)$ have been improved slightly over the years (see [203] for the best known bound and further references), but these bounds are all asymptotically "close" to 2^{2k}.

The above recursion idea can be used for $k \geq 3$ and for more colours. (see, e.g., [117] for more details), thereby giving another proof that showing that all (finite) Ramsey numbers exist.

Numbers of the form $R(k,k)$ are called *diagonal Ramsey numbers*; unfortunately, the only diagonal Ramsey numbers known are $R(3,3) = 6$ and $R(4,4) = 18$. Other known Ramsey numbers (for two colours) are: $R(3,4) = 9$, $R(3,5) = 14$, $R(3,6) = 18$, $R(3,7) = 23$, $R(3,8) = 28$, $R(3,9) = 36$), $R(4,5) = 25$. See the dynamic survey by Stanislaw Radziszowski [740] for other bounds on Ramsey numbers.

To prove a lower bound for some Ramsey number $R(k,\ell)$, one can either construct an example of a 2-colouring of the edges of some large K_n so that no red K_k or no blue K_ℓ is formed, or one can show that for some large n, such a "bad" 2-colouring exists. To construct such examples, Paley graphs (quadratic residue graphs) can be useful. On the other hand, Erdős gave a remarkable proof that (for appropriately chosen n) that such bad colourings exist by using only the simplest of ideas from probability.

The following result is not required later in this text, but due to its simplicity and novelty, it is worth including here. This new viewpoint has been said to be the beginning of a powerful technique in Ramsey theory called the "probabilistic method" (for example, see [24] or [317] or for an introduction.) It is not clear where this proof appeared first, but Erdős is credited with its discovery—since it is, essentially, a restatement of a counting proof he gave in 1947 (in 1957, Erdős [300] called his 1947 proof "combinatorial–probabilistic", but I have likely not found the earliest such source). Only a very basic understanding of probability is assumed in the proof below.

Theorem 7.1.7 (Erdős, 1947 [299]). *If k and n are integers so that*

$$\binom{n}{k}2^{1-\binom{k}{2}} < 1,$$

then $R(k,k) > n$.

Proof: Rather than consider red-blue colourings of a complete graph, let red stand for "edge" and let blue stand for "non-edge". Let G be a random graph on n vertices, where the probability of an edge is $p = 1/2$. (One can consider the sample space as the $2^{\binom{n}{2}}$ graphs on n labelled vertices.) The probability that a fixed set of k vertices induces a complete graph K_k or its complement $\overline{K_k}$ is $2^{1-\binom{k}{2}}$. So the probability that some set of k vertices does so is at most

$$\sum_{\text{all } k\text{-sets}} 2^{1-\binom{k}{2}} = \binom{n}{k}2^{1-\binom{k}{2}},$$

which is less than one. So there exists a graph on n vertices that does not contain any K_k or $\overline{K_k}$. □

A lower bound for diagonal Ramsey numbers now follows with some simple manipulation and Stirling's formula (like in the proof of Theorem 7.1.6)

Corollary 7.1.8 (Erdős). $(1+o(1))\frac{k}{e\sqrt{2}}2^{k/2} < R(k,k)$ *where $o(1)$ tends to zero as k increases.*

The above lower bound has since been improved only very slightly (see [24]).

Ramsey's theorem has applications in geometry, especially when the graphs in question are embedded in some Euclidean space. In the next section, the connection between Ramsey's theorem and convex sets is developed. In Chapter 18, Ramsey's theorem (and other "Ramsey-type" theorems) are discussed for configurations in Euclidean space.

7.2 Convex n-gons, Erdős and Szekeres

In the early 1930s, three young Hungarians, Paul Erdős, Esther Klein and George Szekeres, were examining a problem in geometry.

Esther noticed that whenever 5 points are in general position (no three in a line), then some four of these points are the vertices of a convex 4-gon.

Exercise 210. *Show that if if 5 points in a plane are in general position, some four of these points are the vertices of a convex 4-gon.*

The three Hungarians quickly generalized the problem from convex quadrilaterals. For $n \geq 3$, let $f(n)$ be the minimum number (if it exists) so that if $f(n)$ points are in the plane in general position there exist n of these points forming a convex n-gon.

So $f(3) = 3$ and $f(4) = 5$. Does $f(n)$ always exist? Esther Klein and George Szekeres had conjectured that it does. Together with Paul Erdős (in a paper by only Erdős and Szekeres), they showed that $f(n)$ exists.

Theorem 7.2.1 (Erdős–Szekeres, 1935 [320])**.** *For any $n \geq 3$, there exists a least integer $f(n)$ so that if $f(n)$ points in the plane are in general position, then some n of them form a convex n-gon.*

In fact, Erdős and Szekeres proved Theorem 7.2.1 by giving a simple upper bound on $f(n)$:

Theorem 7.2.2 (Erdős–Szekeres, 1935 [320])**.** *For each $n \geq 3$,*

$$f(n) \leq \binom{2n-4}{n-2} + 1.$$

The proof of Theorem 7.2.2 appears below. To prove Theorem 7.2.2, Erdős and Szekeres proved a stronger theorem that showed $\binom{2n-4}{n-2}+1$ points guaranteed a convex n-gon in one of two shapes.

Definition 7.2.3. For $m \geq 2$, Let $M = \{p_1, p_2, \ldots, p_m)$ be a set of points points in $\mathbb{R}^2$, written in order with strictly increasing x-coordinates. Then M forms an m-cup if and only if for $i = 1, 2, \ldots, m-1$, the slopes of the line segments $p_i p_{i+1}$ are increasing. Similarly, M forms an *m-cap* if and only if slopes between adjacent vertices are decreasing.

So an m-cup is a collection of points on a graph of a (strictly) convex function. In the original paper [320], cups were called "concave sets" and caps were called "convex sets". Probably to avoid any confusion of the meaning of "convex", modern write-ups use cups and caps.

Theorem 7.2.4 (Erdős–Szekeres [320])**.** *For each $k, \ell \geq 2$, if $\binom{k+\ell-4}{k-2} + 1$ points in the plane are in general position (no 3 collinear) with no two points having the same x-coordinate, either some k of these points forms a k-cup, or some ℓ of these points form an ℓ-cap.*

Proof: For $k, \ell \geq 2$, let $S(k, \ell)$ be the statement that if $\binom{k+\ell-4}{k-1} + 1$ points are in general position with no same x coordinates are chosen in $\mathbb{R}^2$, then some k of these points form a k-cup or some ℓ of these points form an ℓ-cap. The proof is by induction on $k + \ell$.

Base step: When $k = 2$ and $\ell = 2$, any two points (with different x-coordinates form both a 2-cup and a 2-cap, and since $\binom{2+2-4}{2-2} + 1 = 2$, the case $S(2, 2)$ is true. If no 3 points form a cap, then all points form a cup; similarly if no 3 points form a cup, then all points form a cap. Hence, for all $m \geq 2$, both $S(3, m)$ and $S(m, 3)$ hold.

Inductive step: Fix $s, t \geq 3$ and suppose that both $S(s-1, t)$ and $S(s, t-1)$ hold. It remains to show that $S(s, t)$ holds. Put

$$N = \binom{s+t-4}{s-2} + 1 = \binom{s+t-4}{t-2} + 1,$$

and let $P = \{p_1, \ldots, p_N\}$ be points listed in order with strictly increasing x-coordinates. In hopes of a contradiction, suppose that no subset of P forms either an s-cup or a t-cap. Since

$$N \geq \binom{s+t-5}{s-3} + 1$$

by $S(s-1, t)$, the set P contains $(s-1)$-cups. In fact, P contains *many* $(s-1)$-cups: Let R be the set of rightmost points of all $(s-1)$-cups. $P \setminus R$ is a set with no $(s-1)$-cups, so by $S(s-1, t)$,

$$|P \setminus R| \leq \binom{s+t-5}{s-3}.$$

Hence

$$\begin{aligned}
|R| &\geq N - \binom{s+t-5}{s-3} \\
&= \binom{s+t-4}{t-2} + 1 - \binom{s+t-5}{t-2} \\
&= \binom{s+t-5}{t-3} + 1 \qquad \text{(by Pascal's id.)}
\end{aligned}$$

and since R contains no s-cups, by $S(s, t-1)$, R contains at least one $(t-1)$-cap, say $\{q_1, \ldots, q_{t-1}\}$, given in left-right order, where each q_i is the rightmost point of some $(s-1)$-cup. Suppose that q_1 is the right endpoint of the $(s-1)$-cup on (after relabelling) points $p_1, \ldots, p_{s-2}, p_{s-1} = q_1$.

If the slope of $p_{s-2}p_{s-1}$ is greater than or equal to that of q_1q_2, then the points p_{s-2}, $p_{s-1} = q_1$, q_2, $\ldots$, q_{t-1} form a t-cap. If the slope of $p_{s-2}p_{s-1}$ is less than that of q_1q_2, then p_1, p_2, $\ldots$,$p_{s-2} = q_1$, q_2 forms an s-cup. Either position of q_2 contradicts the initial assumption, so $S(s,t)$ holds, thereby completing the inductive step.

By mathematical induction on $k+\ell$, for every $k, \ell \geq 2$, $S(k,\ell)$ holds. □

Since any k-cap or k-cup forms the vertices of a convex k-gon, Theorem 7.2.2 (and so also Theorem 7.2.1) is proved.

Exercise 211. *Show that the bound in Theorem 7.2.4 is tight when $k = \ell = 4$ by giving a configuration of six vertices (with no two x-coordinates the same) with no 4-cup and no 4-cap. In general, by studying the proof of Theorem 7.2.4, find a placement of $\binom{k+\ell-4}{k-2}$ points with no k-cup and no ℓ-cap.*

Theorem 7.2.4 gives an upper bound for $f(n)$, (using $k = \ell = n$), but for only special kinds of convex n-gons; for the k-cup ℓ-cap problem, the bound is tight, but is likely far larger than a value for any convex n-gon to appear. A simple form for $f(n)$ is conjectured.

Conjecture 7.2.5 (Erdős-Szekeres [320])**.** *For $n \geq 3$, $f(n) = 2^{n-2} + 1$.*

Apparently, the reason for the conjecture is that they had also discovered $f(n) \geq 2^{n-2} + 1$ by producing an example with 2^{n-2} points and no convex n-gon; however, that proof never made it to press until 25 years later!

Theorem 7.2.6 (Erdős–Szekeres, 1960 [321])**.** *For each $k, \ell \geq 2$, there exists a configuration of 2^{n-2} points that contains no convex n-gon.*

Exercise 212. *Find a configuration of eight points in general position, no five of which form a convex pentagon.*

The example given in Exercise 212 was (according to [320]) shown to be optimal (i.e., no such example works for 9 points) by E. Makai; a proof of this fact is given in [522] or [126]. For an explanation of how to construct a set of 2^{n-2} points satisfying Theorem 7.2.6, see [622, pp. 578–580]. So for $n = 5$, the conjecture that $f(n) = 2^{n-2} + 1$ is confirmed. For $n \geq 6$, $f(n)$ is (at the time of writing this) not yet known.

In the 1969 IMO in Bucharest, the following related question was given:

Exercise 213. *Given $n > 4$ points in a plane such that no three are collinear, prove that one can find at least $\binom{n-3}{2}$ convex quadrilaterals whose vertices are four of the given points.*

7.3 Other proofs of the n-gon theorem

There are two proofs of the n-gon theorem that use Ramsey's theorem; however, neither of these proofs give bounds as good as those derived from Theorem 7.2.4.

7.3.1 Second proof of the n-gon theorem

The second proof of Theorem 7.2.1 given by Erdős and Szekeres is based on Ramsey's theorem (which they proved by their now-famous "Erdős–Szekeres recursion", as given inTheorem 7.1.4, but the more general version for hypergraphs and more colours). To streamline this proof, an observation is used:

Lemma 7.3.1. *Let X be a set of $n \geq 4$ points in $\mathbb{R}^2$ in general position so that every four points in X induce a convex quadrilateral. Then X is the vertex set of a convex n-gon.*

Proof: One could use Carathéodory's theorem in the plane (Theorem 5.2.1), which says that any point in the convex hull of a set $X \subset \mathbb{R}^2$ is a convex combination of at most three points from X. Since any four points are convex, no point is the convex combination of three others. Thus the interior of the convex hull of X contains no points. $\square$

Second proof of Theorem 7.2.1: Let $N = R_4(n; 2) = R_4(n, n)$ be the Ramsey number for colouring 4-tuples with 2 colours (i.e., $N \longrightarrow (n)^4_2$). Let X be a set of N points in general position. Colour a 4-tuple $T \in [X]^4$ red if T induces a convex quadrilateral, and colour T blue otherwise. By the choice of N, there exists an n-tuple $Y \in [X]^n$ so that either all of its 4-tuples are red, or all of its 4-tuples are blue. If $n \geq 5$, this colour is not blue (since by the result in Exercise 210, any five points induce at least one convex quadrilateral), so the colour is red. By Lemma 7.3.1, X is the vertex set of a convex n-gon. $\square$

7.3.2 Another Ramsey proof of Theorem 7.2.1

During an exam in Haifa, a student named A. Tarsi found another proof of Theorem 7.2.1 using Ramsey numbers (see [410].) He invented this proof during the exam since he had missed the lecture(s) when other proofs of Theorem 7.2.1 were given.

Third proof of Theorem 7.2.1: Let $N = R_3(n;2) = R_3(n,n)$ be the Ramsey number for colouring triples with 2 colours (i.e., $N \longrightarrow (n)^3_2$). Let $X = \{x_1, x_2, \ldots, x_N\}$ be a set of N points in the plane in general position. For each $1 \leq i < j < k \leq N$, colour the triple $\{x_i, x_j, x_k\}$ red if the walk from x_i to x_j to x_k is a "righthand turn", and blue if the walk is a "lefthand turn". By the choice of N, there exists a (homogenous) set X with n points so that all of its triples are coloured the same. In this case, it is not difficult to verify that X is the vertex set of a convex n-gon. □

For yet one more proof of Theorem 7.2.1, similar to Tarsi's proof, see [508].

7.4 Bounds for the n-gon problem

For the next few results, let $f(n)$ denote the smallest number so that any set of $f(n)$ points in the plane in general position contains a convex n-gon. Erdős and Szekeres [320, 321] proved

$$2^{n-2} + 1 \leq f(n) \leq \binom{2n-4}{n-2} + 1.$$

For $n \leq 5$, the lower bound is correct, and it is conjectured to be the correct bound in general. Many improvements have occurred since, but still the lower bound remains. In a sequence of three papers all appearing in the same journal (all in 1998!) the upper bound was brought down. First, Chung and Graham [190] showed that $f(n) \leq \binom{2n-4}{n-2}$; then Kleitman and Pachter [556] showed $f(n) \leq \binom{2n-4}{n-2} + 7 - 2n$; Tóth and Valtr (see [873] and [874]) reduced this further (by about a factor of 2) to $f(n) \leq \binom{2n-5}{n-2} + 2$. A few years later, Tóth and Valtr [875] were able to trim off 1 more to give $f(n) \leq \binom{2n-5}{n-2} + 1$.

In 2016, Suk [840] proved that the lower bound $f(n) \geq 2^{n-1} + 1$ of Erdős and Szekeres is nearly correct by showing that $f(n) = 2^{n(1+o(1))}$.

Theorem 7.4.1 (Suk, 2016 [840]). *For each $n \geq 3$, let $f(n)$ be the minimum number of points in general position that guarantee n of these points form a convex n-gon. Then*

$$f(n) \leq 2^{n+6n^{2/3}\log_2 n}.$$

In the second version of his paper, Suk reports that after the first version, Gábor Tardos gave the following improvement:

$$f(n) = 2^n + O(\sqrt{n \log_2 n}).$$

7.5 Extending the n-gon problem

Here is a generalization of the Erdős-Szekeres theorem that appeared as an exercise in Grunbaum's *Convex Polytopes* [427, p. 22]

Theorem 7.5.1 (Grunbaum, 2003). *Given integers d and v with $2 \leq d < v$, there exists an integer $e(d, v)$ with the property that whenever $A \subset \mathbb{R}^d$ consists of at least $e(d, v)$ points in general position, there exists $B \subset A$ so that $|B| = v$ and $B =$ ext conv$B$.*

For the proof of Theorem 7.5.1, a hint in [427] is given, namely, apply Ramsey's theorem for 2-colouring $(d+2)$-tuples, guaranteeing either a red v-set or a blue $(d+3)$-set, where the $(d+2)$-tuples coloured are those C with $C =$ ext conv$C$. Grunbaum goes on to mention that $e(d, d+2) = d+3$, and $e(2, 5) = 9$ is shown in [320].

In 1994, Harborth and Möller [458] investigated the n-gon problem for the (real) projective plane. They first asked if there is a function f_1 so that for each $k \geq 3$, if $f_1(k)$ lines (not points) are given in the plane in general position, is there a convex k-gon whose sides use k of these lines. They observed that for $k \geq 5$, $f_1(k)$ fails to exist by exhibiting a family of lines so that any k of them determine only triangles and quadrilaterals. However, if one translates the problem to projective planes, one might use the duality between points and lines.

For each $k \geq 3$, Harborth and Möller say that k points determine a convex k-gon in the (real) projective plane if there is a mapping of the projective plane onto itself so that the k points form a convex k-gon in the Euclidean sense, that is, the convex k-gon has no point on the line at infinity. Define $p(k)$ to be the least number of points in the projective plane, no three collinear, so that some k of them determine a convex k-gon. Using duality, they showed that for $k = 3, 4, 5$, $p(k) = k$ and that $p(6) = 9$. Harborth and Möller also investigated the similar problem where not only points are replaced by lines, but by pseudolines.

Other extensions of the Erdős–Szekeres n-gon problem include those for higher dimensions (see [526], [528], or [839]) or other convex bodies (see [263] and [342]). See also [89] and [105]. The survey by Morris and Soltan [670] also includes many references for generalizations of the n-gon problem.

A striking example (see [645, Ex. 5.4.3, p. 99]) for higher dimensions guarantees a polytope combinatorially equivalent to a cyclic polytope (see Section 6.3). Also, a sequence of papers (see [56], [701], and [732]) appeared regarding a "positive fraction" version.

See [408] for another geometric analogue to Ramsey's theorem. For Ramsey-type theorems for metric spaces, see [566].

7.6 Empty convex n-gons

For $n \geq 3$, define $g(n)$ (if it exists) to be the minimum number of points in the plane in general position (no three collinear) so that some n points form an *empty* convex n-gon (an n-gon with no other points in its interior, also called a *hole*).

Harborth [456] proved that $g(n)$ exists for $n = 3, 4, 5$. Exercise 214 asks to show $g(5) \geq 10$.

Exercise 214. *Construct a set of 9 points in general position that has no 5-hole (five points forming a convex pentagon with no other points inside, a planar C_5 with empty interior).*

In 1983, Horton [488] showed that for $n \geq 7$, $g(n)$ fails to exist.

The existence of $g(6)$ was open for nearly 25 years, but was finally solved by two different authors in the same year.

Theorem 7.6.1 (Gerken, 2008 [385] and Nicolás, 2008 [683]). *For sufficiently large m, if m points are in the plane in general position, there exists a set of 6 points forming an empty convex hexagon.*

Much research has been done on these and similar problems; in particular, Pavel Valtr wrote an article [893] on 7-holes in special sets, and an extension of the problem to higher dimensions [894]. One research area is to bound the number of empty k-holes (when $k = 3, 4, 5$), or to find the number of empty triangles sharing two points; see, e.g., the recent paper by Valtr [896] for details. Valtr [896] mentions a few other problems, one of which is to find a simple proof that any set of 10 points in general position in the plane contains a 5-hole.

Chapter 8

Dissections

8.1 Dissection for two equal area polygons

8.1.1 Introduction

Given any two polygons of equal area, is there a dissection of one into finitely many pieces (using only straight cuts) so that these pieces can be rearranged to form the other? If one polygon can be dissected into pieces to be re-arranged into another polygon, what is the minimum number of pieces required?

It turns out (see below) that in fact the answer to the first question in the problem is "yes", and a solution is relatively simple; solutions were found by a few different authors in the 1800s (more detailed references appear below). The answer to the second question is considerably more complicated, and not too much is said about it here.

Theorem 8.1.1 (Wallace–Lowry–Bolyai–Gerwein, 1808, 1814, 1832, 1833)**.** *Given any two polygons with the same area, there is a dissection of one using only straight cuts so that the pieces can be re-arranged to form the other.*

A proof of Theorem 8.1.1 (given below) relies on a few simple dissections between a triangle and a rectangle with a given base.

The history of this problem and its solution seems complicated. In Kraitchik's famous book *Mathematical recreations* [574, 193–198] it says that Hilbert (1862–1943) solved the problem. In fact, Hilbert shows a solution in his 1899 book *Grundlagen der Geometrie* [468, Ch. 4], but makes no mention of previous solutions. (Hilbert's solution seems to show that two polygons with the same area can be decomposed into congruent *triangles*, not just any smaller polygons, but I can not be certain as the text is rather long and filled with different definitions, like "equicomplementable") On page 66 of the translation of Hilbert's book [471] he says "In §20 we have found that polygons having equal content have also equal measures of area. The converse of this is also true." (Here, "having equal content" involves more definitions that would take a longer explanation than is desired here.) Two pages later, this fact is recorded as Theorem 30. Hilbert's proof relies on cutting a polygon into triangles, as do earlier proofs, but he seems to have a different approach (using the affine Pappus theorem, Theorem 1.8.2 here) than what is described (see below) by earlier authors. In another translation of a revised (tenth) edition of *Grundlagen der Geometry* [473], Hilbert uses different definitions and in Supplement III (by Bernays) more discussion of the problem is given (in that volume, the only other previous work on this problem I could find a reference to [473, p. 200] was one by van der Waerden given in 1934–1936).

In Rouse Ball and Coxeter's book [770, pp. 89–93], Hilbert does not receive mention with regard to this dissection problem, which led me to wonder who actually solved this problem first. An article by Gunderbühler and Nüsken [492] seems to have a good collection of references, some of which are used here (I also highly recommend this article for many details regarding 3-dimensional analogues—see Section 8.9). It seems as if many popular websites also have completely wrong dates and names, with no bibliographic references to back them up.

Apparently, this problem was first proposed by Farkas Bolyai in the 1790s, and was solved by William Wallace circa 1808 (but his solution [910] was published only in 1831). A solution was published by Lowry (known only as "Mr. Lowry" in literature sources I could find) in 1814 [623], and again by Bolyai [121] in 1832 and Gerwein [386] in 1833. See Rouse Ball and Coxeter's book [770] for references to subsequent solutions (some for special cases) by Euzet in 1854, E. Guitel in 1895, E. Holst in 1896, A. Mineur (by 1931), and Michael Goldberg (by 1947). Greg Frederickson's *Dissections: Plane & fancy* [348] is also a resource for the history of dissection of polygons, only a few of which are examined here.

8.1.2 Steps for such a dissection

There are different ways to accomplish a dissection of a polygon so that the pieces can be rearranged into another polygon of equal area. One way is to take any polygon P with area p and show how to dissect it and to rearrange the pieces into a square with area p (or a rectangle with unit base and height p). This process is reversible, and so one can go between two different polygons with the same area by going through the square/rectangle as an intermediate step. Here are the steps, with details given below:

Step 1 Cut P into triangles.

Step 2 For each triangle from Step 1, dissect and rearrange into a parallelogram, with the same base as the triangle.

Step 3 For each parallelogram from Step 2, dissect and rearrange into a rectangle with the same base as the parallelogram.

Step 4 For each rectangle with from Step 3, dissect and rearrange into a rectangle with given base.

Step 5 Stack the rectangles from Step 4 to create a larger rectangle with unit base and height p (or a square with side length $\sqrt{p}$).

Step 1: A polygon into triangles

As shown in Exercise 46, any n-sided (simple) polygon can be dissected into $n-2$ triangles (the polygon need not be convex, but it must be simple, that is, no crossing edges).

Step 2: A triangle into a parallelogram

For a triangle $\triangle ABC$, let DE be the segment (parallel to AB) where D bisects AC and E bisects BC. Cut along DE, and rotate $\triangle DEC$ about E to form a parallelogram. See Figure 8.1 for examples when $\triangle ABC$ is acute or obtuse (in fact, the second is unnecessary, since AC could be used as a base and then the first diagram applies). In each case, it is trivial to check that indeed $ABFD$ is a parallelogram.

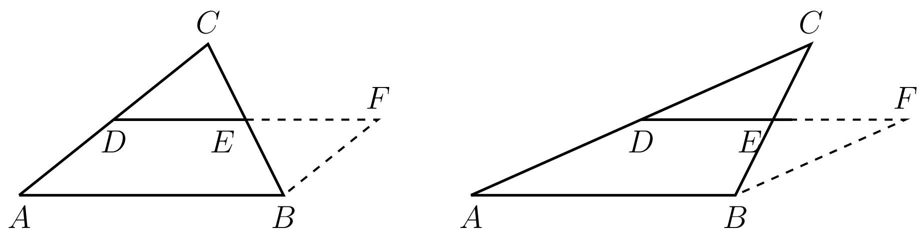


Figure 8.1: Triangle into a parallelogram

Step 3: A parallelogram into a rectangle

The area of a parallelogram is the length of its base times its altitude; one way to prove this is to chop off one end and move it to the other, as in Figure 8.2.

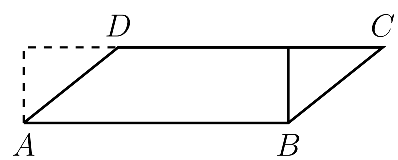


Figure 8.2: Parallelogram into rectangle

Step 4: A rectangle into a rectangle with a given base

Let R_1 be a $\ell \times h$ rectangle, and suppose that a dissection is desired to a rectangle R_2 with base b. (If the area of the original polygon is p, the only value of b needed here is $\sqrt{p}$ or 1.) Without loss of generality, let $b < \ell$. When $\ell < 2b$, follow the construction in Figure 8.3.

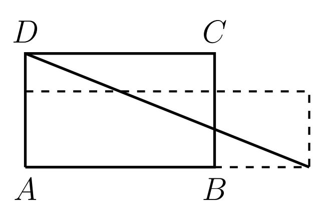


Figure 8.3: Rectangle into rectangle with given base

When $\ell \geq 2b$, cut one rectangle into pieces first so the above construction works. (For more details, see, *e.g.*, [770].)

8.2 Dissecting two squares into one

8.2.1 A pythagorean-square puzzle

In the November 1971 issue of *Scientific American*, Martin Gardner gave the following puzzle in his regular column "Mathematical games" [370].

Two squares are presented (see Figure 8.4) with the large one dissected into four pieces. According to Gardner, this puzzle had been produced in various plastic versions, the "the handsomest" was (in 1971) marketed as "Madagascar Madness". I do not know its origin.

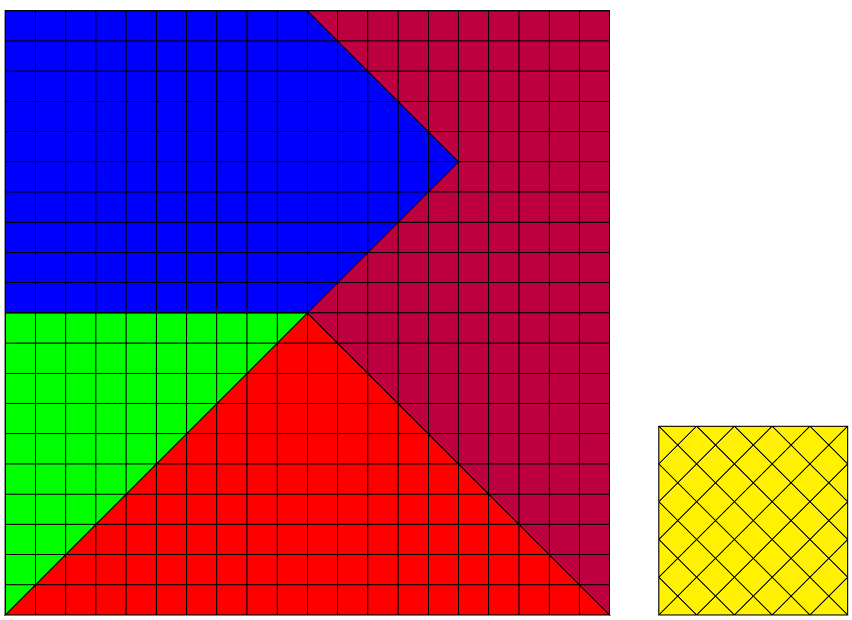


Figure 8.4: The Pythagorean-square puzzle: rearrange the five pieces into a square

The goal is to rearrange these five pieces into one larger square. The solution given in Figure 8.5 was published in the December 1971 issue [371].

This puzzle shows that $(20)^2 + (5\sqrt{2})^2 = 400 + 50 = 450 = (15\sqrt{2})^2$, which is of the form $a^2 + b^2 = c^2$; if a, b, c represent lengths found in a right

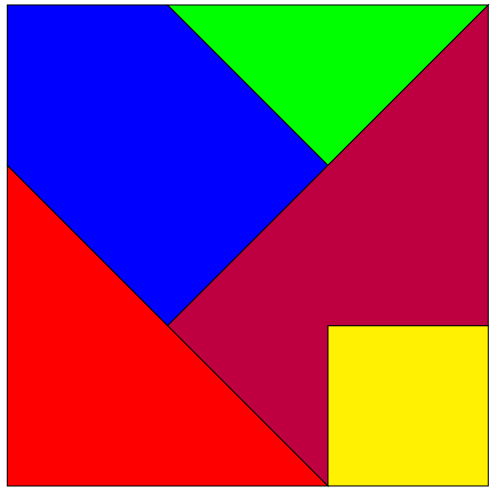


Figure 8.5: A Pythagorean-square puzzle solution

triangle, then this puzzle gives the Pythagoras equation for such a triangle.

There is no obvious way to extend this puzzle for two squares of arbitrary relative size. Dissection patterns that work for any two squares are given in Sections 8.2.2 and 8.2.3.

8.2.2 Perigal's dissection

Henry Perigal (1801–1898) gave a "dissection proof" of the Pythagoras theorem [717]. His paper was published in November 1872 (but appears in the 1873 bound volume, where additional comments were given by an editor).

Perigal was a stockbroker and amateur mathematician. A biography was given by Rogers in 1897 [760]; a website posting from Bill Casselman [178] is also recommended.

Perigal's dissection (see Figure 8.6) takes any two given squares, and with two cuts, chops one (the large one, if they are of different sizes) into four pieces so that these pieces together with the smaller square are arranged into a square whose area is the sum of the areas of the two given squares. This process is also reversible (take one large square, dissect it, and re-assemble the pieces to form two smaller squares).

Let $ABCD$ and $EFGH$ be squares with side lengths a and b respectively, where $a \geq b$. (In the drawing below, I used $a = 5, b = 2$). As in Figure 8.6, place these two squares adjacent to each other with bases AD and FH along a common line, where D and E occupy the same point. Let M be the midpoint of AH (so $|AM| = \frac{a+b}{2}$), and let N be a point on DC so that $|DN| = \frac{a+b}{2}$. (So N is above E.) Draw a pair of perpendicular lines through the center of the large square, one passing through M and the other passing through N, thereby cutting the large square into four identical quadrilaterals. One of these quadrilaterals is adjacent to the small square; leave it fixed. Move the other three to surround the small square, creating a (larger) square with area equal to the sum of the two areas.

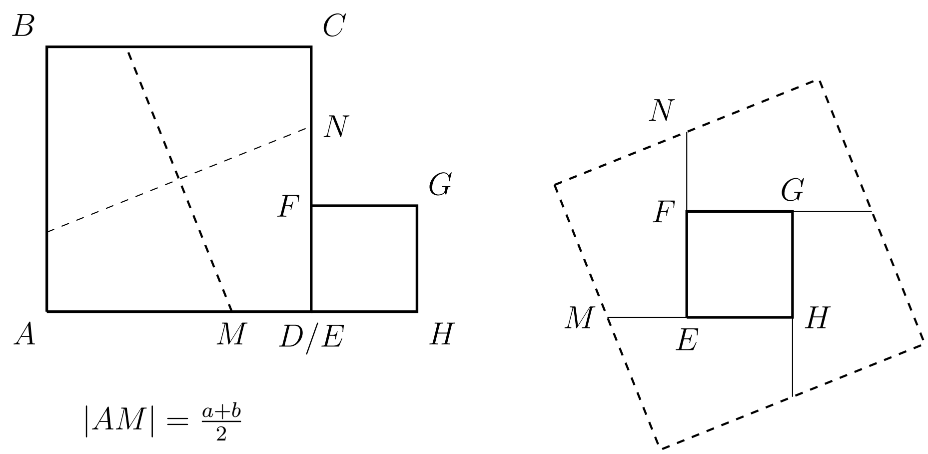


Figure 8.6: Perigal's dissection of two squares into one

Exercise 215. *For a Perigal dissection demonstrating* $3^2 + 4^2 = 5^2$*, compute the sizes and angles required for the four pieces.*

8.2.3 A two square dissection from Thābit

Henry Ernest Dudeney (1847–1930) wrote puzzle columns appearing in various periodicals (*The Strand Magazine*, *Cassell's Magazine*, *The Queen*, *Tit-Bits*, and *The Weekly Dispatch*). Readers submitted puzzles and solutions, and so it now seems impossible to trace origins of some puzzles or solutions. In 1917, Dudeney published *Amusements in Mathematics* [271], containing many puzzles that were discussed in these periodicals.

Two famous puzzles that are now attributed to Dudeney include a dissection of two squares into one (achieving the same goal as Perigal's dissection), and the hinged dissection of a square into an equilateral triangle (see Section 8.5). It seems that maybe neither dissection was invented by Dudeney.

For example, the following dissection was known long ago by the Baghdad scientist Thābit ibn Qurra (826–901 AD) (see [349, p. 34]; this dissection is credited to Dudeney in [770, p. 88]).

Let $ABCD$ and $EFGH$ be squares (possibly of different sizes). As in Figure 8.7, place the two squares side by side. Let X be the point on AD so that $|AX| = |DH|$, the length of the smaller square. Cut along BX and along XG.

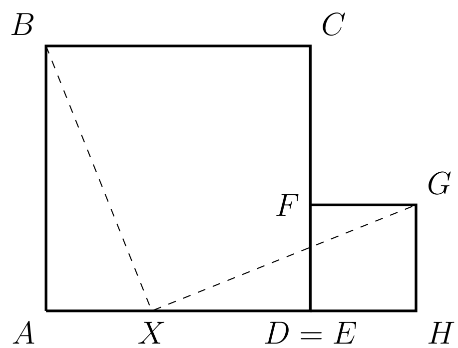


Figure 8.7: Thābit's dissection of two squares into one

It is left as an exercise to the reader to assemble the pieces of Thābit's dissection into one large square.

8.3 The minimum number of pieces required

As seen in Section 8.1.2, it is not difficult to dissect any triangle into 3 pieces that can be arranged into a rectangle (with the same base as the triangle).

The number of pieces generated by techniques in the proof of Theorem 8.1.1 are often much larger than the minimum number of pieces. In many cases, to prove that a dissection has the fewest pieces possible seems rather daunting, and so all I can recommend is to locate examples with fewest pieces from the common literature. As shown in Section 10.3.2, Zbigniew Moroń gave a dissection of a rectangle into 9 different squares in the early 1900s; it took a while to prove that 9 was minimal (and there are only two such examples!).

Exercise 216. *Find a dissection of a square (using straight cuts) into 5 pieces that can be re-arranged to form an equilateral triangle of the same area.*

In Section 8.5, a dissection of a square to an equilateral triangle is given with only 4 pieces; this was found only in the early 1900s.

If three equal squares are given, how many pieces must they be cut into so that the pieces can be rearranged to form one large square? Abul Wefa (940–998) found such a dissection using nine pieces. Perigal found one using only eight pieces, a record that stands today (as far as I know). See Moscovich's wonderfully illustrated *The puzzle universe: A history of mathematics in 315 puzzles* [671, p. 100] for drawings of each.

In 1933, Travers [876] gave a dissection of a regular octagon into 5 pieces that can be reassembled into a square. A. H. Wheeler divided a regular pentagon into 6 pieces that are assembled into an equilateral triangle. These are just a few of hundreds of examples of dissections; for further reference, one might begin with [348] or [613]. For a list of many known minimal polygon-to-polygon dissections, see the wonderfully illustrated article *Dissection* [866] on Wolfram's site by Gavin Theobald and Eric Weisstein. (there are also many references).

8.4 Miscellaneous dissection puzzles

Many (maybe hundreds of) books on recreational math contain problems of dissection (e.g., [770], [272], [570, Ch. IV], [826]). The book *Creative puzzles of the world* [139] contains many standard examples and photos of models. Some books are dedicated to dissection problems, notably those by Frederickson [348] and Lindgren [613].

Dissection puzzles can be addicting, and since I had only intended to give a quick survey of such puzzles with respect to "combinatorial geometry", I leave the interested reader with the above references to get started. More dissection puzzles occur later in this volume, and many of these might properly be called "combinatorial geometry puzzles", but there are, perhaps, thousands of other popular dissection puzzles.

Let me mention just one puzzle from a book that was an early favourite of mine, *536 Puzzles & curious problems* [272, Prob. 359]. I include it here (although it may not be considered by some to be combinatorial geometry) because of its simplicity as a "tiling of a square" problem with a somewhat surprising solution. The puzzle is called "The patchwork cushion", first proposed about someone taking 20 triangles of silk to make a cushion cover.

Exercise 217 (The patchwork cushion)**.** *Use twenty identical* $1 : 2 : \sqrt{5}$ *right triangles to tile a square.*

8.5 Hinged dissections and other variants

8.5.1 Introduction

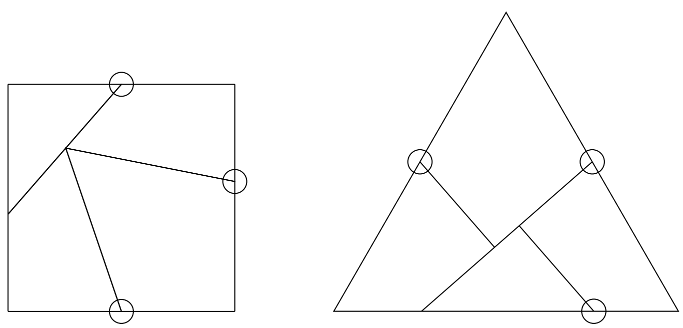
This section is just enough to introduce the reader to another kind of dissection problem: hinged dissections. The authoritative resource for this topic is Frederickson's book [349], with variations also in [351].

A (straight line) dissection of one polygon into another can often have all pieces held together with hinges on some vertices (the axes of the hinges are orthogonal to the plane, and so pieces are rotated in the plane about the vertices with hinges).

One classic hinged dissection is one of an equilateral triangle into a square using only four pieces. (Calculations of all angles and lengths are given in Section 8.5.2 below.) Here is a little history of this dissection. From 1896 to 1903, Dudeney (1897–1930) ran a puzzle column in the *Weekly Dispatch* called "Puzzles and Prizes" (see [349] for publication details). On 6 April 1902, Dudeney wrote "No. 440. - The Triangle and the Square. The puzzle is to cut any equilateral triangle (that is, a triangle whose three sides are of equal length) into as few pieces as possible that will fit together and form a perfect square." (See [349] for more details.)

On 4 May 1902, Dudeney gave a solution using four pieces, and credits Charles McElroy for providing a solution. In the following diagram, small circles represent hinges, something that Dudeney later noted could be done (and so in this dissection, no pieces need to be turned over).

Figure 8.8: Hinged dissection of an equilateral triangle to a square, by DSG



It is not clear whether or not Dudeney had discovered that solution, or if indeed McElroy should be credited with the solution (see [349, pp. 8–10, 90–92, 131–133] for discussion—it is not known if Dudeney was previously aware that only four pieces are needed). See Figure 8.8 for a version I made from mahogany and brass hinges.

On the internet, many animations of the hinged triangle-to-square dissection can be found. Note that in the construction of a wooden hinged model, hinge plates need to be recessed so that each hinge point is precisely where it needs to be and the model fits tight; tape is also a good hinge, provided it is durable but thin tape (fiber packing tape works, but upon Fredrickson's suggestion, I have since used 3M Polyester Film Tape 853, available in various widths).

Exercise 218. *Find all lengths and angles in the square-triangle dissection only based on the sketch above (also see Figure 8.8).*

[One solution to Exercise 218 is below in Section 8.5.2; I encourage the reader to try these calculations without peeking.]

In a series of articles called "The Canterbury puzzles" published in *London Magazine* (see [349] for details), Dudeney repeated this puzzle in 1903, calling it the "haberdasher's puzzle".

Dudeney made a hinged version of mahogany with brass hinges (perhaps similar to the one in Figure 8.8) and presented it to the Royal Society in 1905 (from what I remember reading, in a failed attempt to gain membership in the Society).

In 1907, Dudeney published the hinged solution in *The Canterbury puzzles and other curious problems* [270, No. 26]. (Perhaps this is why so some authors give 1907 as the date of this puzzle.) It also appeared in his 1917 *Amusements in mathematics* [271]. This puzzle has since been discussed in many classic books on recreational mathematics (*e.g.* [770, p. 92], [272], [368, p. 34],[826, pp. 3–4] [829, p. 169], [918, pp. 61–62]).

In class a few years ago, I asked students if they could find a dissection from a triangle to a pentagon. Nobody in the class could solve it, including myself. After a few days of playing with the problem, (around 2003) I wrote to Greg Frederickson, and he kindly responded with drawings and measurements for hinged dissections of triangle-to-pentagon, and triangle-to-hexagon. I then made models of these for myself and for Greg (in mahogany) and sent him a copy of each; I only hinged the pentagon-triangle puzzles, as the hexagon-triangle one had corners too narrow to accept screws (both Greg and I used tape for hinges on that one). My versions appear in Figure 8.9.

Greg posted pictures of his copy on his website [350] (expanded to show which pieces are connected). Both constructions can be found in Frederickson's book [349, pp. 108–109].

It is not difficult to verify that (as also noted by Frederickson [349, pp. 33–34]) both the Perigal and Thābit dissections of two squares into one (see Figures 8.6 and 8.7) are also hingeable. Until recently (see Frederickson [350]) many of the polygon-to-polygon dissections were shown to be hingeable, but no general proof was available that showed all were. Note that as a

Figure 8.9: Hinged dissections, pentagon and hexagon to equilateral triangle

consequence of Theorem 8.1.1 and its proof, for each dissection, there is one that does not require pieces to be turned over.

In 2012, Abbott, Abel, Charlton, E. Domaine, M. Demaine, and Kominers [1] published a proof showing that any two polygons of equal area have a hinged dissection. Their paper also gives a history of the problem, with many detailed pictures. (Perhaps early drafts of this proof became available as early as 2007.)

Frederickson [350, 351] also examined other kinds of hinged dissections, where some hinges twist (with axis orthogonal to an edge) through 3D (as in the classic two-piece dissection of an ellipse into a heart), or models with two layers that use piano hinges along edges (Greg gave me two such models, finely crafted with a laser).

8.5.2 Dimensions for the Dudeney–McElroy dissection

There are different ways to find the dimensions and angles for the pieces of this dissection. One way is to start with the square, and derive some relations. Another way is to begin with the equilateral triangle arrangement, and then derive. The second way is, in a sense, easier, since all four right angles are in the center of the diagram, thereby allowing one to project points from the perimeter onto lines inside. Details for this second method can also be found on the website by Alfinio Flores [338]; however, here all dimensions and angles are made explicit.

Suppose the square has side length 1; then an equilateral triangle with area 1 has sidelength $s = \frac{2}{3^{1/4}}$. Label the vertices and angles as in Figure

8.10, where hinges are at vertices F, D, B.

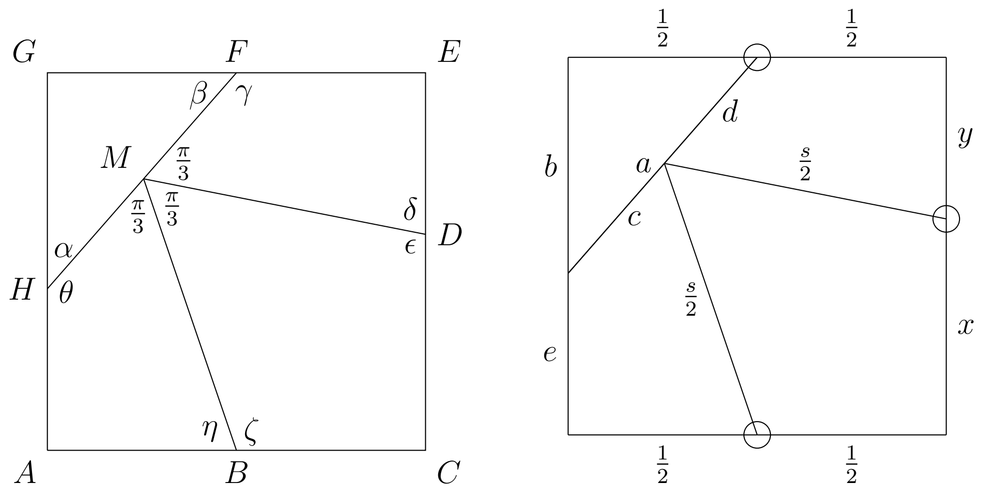


Figure 8.10: Labelling for square-to-triangle dissection

To verify some of the labelled angles and lengths given in Figure 8.10, because the way the puzzle folds, $|GF| = |FE| = \frac{1}{2}$ and $|AB| = |BC| = \frac{1}{2}$. Also the sides MD and MB fold out to a side of the triangle, so they both have length $s/2$. Since $a+c+d = s$ (looking at one edge of the triangle) and $a = c + d$, it follows that $a = |HF| = \frac{s}{2}$ (but M does not necessarily bisect HF). The three small angles at M are all $\frac{\pi}{3}$, because these form the corners of the triangle. (The four corners at A, C, E, G are right angles.)

Computing angles, there are four relations found from each of the four pieces (angles of a quadrilateral sum to 2π), and there are four pairs of supplementary angles (along the sides of the square), the following eight equations are obtained:

$$\alpha + \beta = \frac{\pi}{2};$$
$$\beta + \gamma = \pi;$$
$$\gamma + \delta = \frac{7\pi}{6};$$
$$\delta + \epsilon = \pi;$$
$$\epsilon + \zeta = \frac{7\pi}{6};$$
$$\zeta + \eta = \pi;$$

$$\eta + \theta = \frac{7\pi}{6};$$
$$\theta + \alpha = \pi.$$

Summing all angles around the square makes the last equation superfluous, and the solution has a parameter (say, θ):

$$\alpha = \pi - \theta;$$
$$\beta = \theta - \frac{\pi}{2};$$
$$\gamma = \frac{3\pi}{2} - \theta;$$
$$\delta = \theta - \frac{\pi}{3};$$
$$\epsilon = \frac{4\pi}{3} - \theta;$$
$$\zeta = \theta - \frac{\pi}{6};$$
$$\eta = \frac{7\pi}{6} - \theta;$$
$$\theta = \theta.$$

The angle θ is most easily found by first finding α, and this can be done by looking at $\triangle FGH$. In that triangle, $|HF| = a = \frac{s}{2} = \frac{1}{3^{1/4}}$, and since $|GF| = \frac{1}{2}$, $\sin(\alpha) = \frac{1}{2 \cdot 3^{1/4}}$, giving the answer in degrees (to 2 decimal places)

$$\alpha = \sin^{-1}\left(\frac{1}{2 \cdot 3^{1/4}}\right) \sim 41.15^\circ,$$

and so,

$$\theta = 180 - \alpha \sim 138.85^\circ.$$

Using this value in the previous solution to the system of equations for angles, all the angles can be computed (again, to two decimal places):

$$\alpha \sim 41.15^\circ;$$
$$\beta \sim 48.85^\circ;$$
$$\gamma \sim 131.15^\circ;$$
$$\delta \sim 78.85^\circ;$$
$$\epsilon \sim 101.15^\circ;$$

$$\zeta \sim 108.85^\circ;$$
$$\eta \sim 71.15^\circ;$$
$$\theta \sim 138.85^\circ.$$

To calculate the lengths of each segment, it was already observed that $a = \frac{1}{3^{1/4}} \sim .7598357$, and so by Pythagoras,

$$b = \sqrt{\frac{1}{\sqrt{3}} - \frac{1}{4}} \sim .5721453,$$

and thus

$$e = 1 - b \sim .4278547.$$

Knowing all the angles and three of the lengths of the quadrilateral $ABMH$, one can then compute the coordinates of M to be (u, v), where

$$u = \frac{1}{2} - \frac{s}{2}\sin\left(\frac{\pi}{2} - \eta\right) \sim .254503,$$

and

$$v = \frac{s}{2}\cos\left(\frac{\pi}{2} - \eta\right) \sim .71908.$$

[A different method of calculation gives the precise value of $u = \frac{3-\sqrt{4\sqrt{3}-3}}{4}$.]

Then

$$c = |HM| \sim \sqrt{(.254503)^2 + (.71908 - .42785)^2} \sim .38676,$$

and so

$$d = a - c \sim .37307.$$

Using the coordinates for M, one can then similarly find $x \sim .57214$ and $y \sim .42786$. □

8.6 Sperner's lemma for colouring vertices of a triangulation

8.6.1 Introduction

Emanuel Sperner is a name famous for his 1928 theorem [814] that says if $\mathcal{F}$ is a family of sets on n elements so that no one set in $\mathcal{F}$ contains any other, then $|\mathcal{F}| \leq \binom{n}{\lfloor n/2 \rfloor}$. This result is now commonly known as "Sperner's theorem", which is a central theorem in combinatorics, partially ordered sets, and extremal graph theory. In the same year, Sperner announced another theorem (now called "Sperner's lemma" to differentiate it from his more famous theorem) about colouring vertices of a "triangulation".

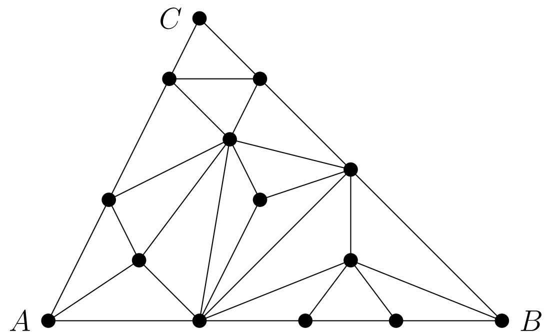
The 2-dimensional version of Sperner's lemma is given and proved in Section 8.6.3, along with how the lemma implies the famous Brouwer' fixed point theorem for the plane. For a little history and the same two proofs, see [13]. The general n-dimensional version is given in Section 8.6.4. To state Sperner's lemma, some terminology is needed.

8.6.2 Triangulations

Recall from Exercise 46 that if P is a polygon (with no crossing edges), then P can be "triangulated" by adding chords so that the interior of P is divided into triangles. (See Figure 1.5 for the ways to triangulate a pentagon.) In Exercise 48, it was shown that the vertices of a triangulated polygon can be coloured with three colours so that *every* triangle receives all three colours.

It is also said that a planar graph G is "triangulated" if any plane drawing of G divides the plane into regions (including the outer region), that are all triangles; in other words, the plane is triangulated by G. (A pentagon can be triangulated by adding two chords, but the plane is then not triangulated by the resulting graph since the outer region is not a triangle.)

For the statement of Sperner's lemma, a "triangulation of a triangle" is used. Let T be a triangle on vertices A, B, C, and let $\triangle ABC$ be dissected into triangles. Let V be the set of vertices of all triangles in the triangulation, and let E be the set of edges formed by these triangles. The planar graph $G = (V, E)$ is then called a *triangulation of* T. For example, here is a triangulation of a triangle $\triangle ABC$:

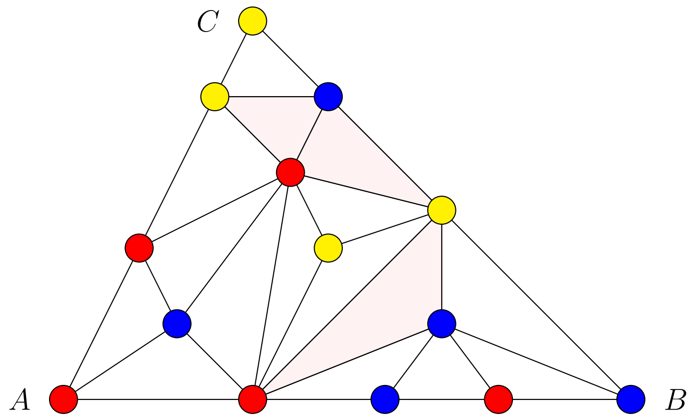


8.6.3 Sperner's lemma for two dimensions

Here is the 2-dimensional version of Sperner's lemma (there is a version for every dimension, but the 2-dimensional version is a good starting point for the general theorem which colours simplices in higher dimensions).

Theorem 8.6.1 (Sperner's lemma, 1928 [815]). *Let $T = \triangle ABC$ and let $G = (V, E)$ be a triangulation of T. Assume that $c : V \to \{0, 1, 2\}$ is a 3-colouring that satisfies $c(A) = 0$, $c(B) = 1$, $c(C) = 2$ and the vertices along any edge of $\triangle ABC$ are coloured only using one of the colours at either end (for example, vertices on the edge AB are coloured either 0 or 1), and interior vertices are coloured arbitrarily. (Such a colouring is called a Sperner colouring.) Then at least one of the triangles in G has its vertices all coloured differently.*

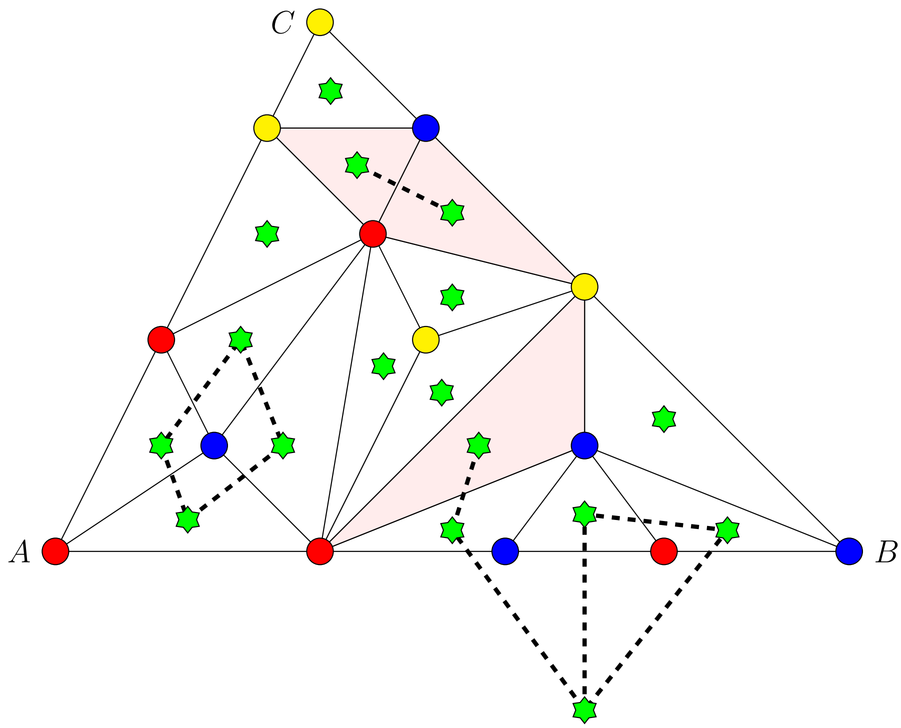
Before giving the proof of the 2-dimensional case of Sperner's lemma, here is an example of such a 3-coloured triangulation for the graph above, and "rainbow" triangles thereby determined (where red=0, blue=1, yellow=2):



Exercise 219. *Formulate and prove a 1-dimensional version of Sperner's lemma.*

Proof of Sperner's lemma for 2 dimensions: The proof given here is really an elegant double-counting proof but given in terms of simple graph theory, and can be found in many sources (e.g., see *Proofs from the book* [13, pp. 147–149]).

Let V^* be the vertices of the dual of G, that is, to each region of G, assign a vertex in V^*. Define a partial dual G^* of G on vertex set V^* by putting an edge between regions (including the outer region) sharing an edge coloured with 0 and 1. Using the above example, the graph G^* is indicated by (green) stars as vertices and dashed lines for edges):



Every vertex in G^* representing an inner triangle of G has degree at most 2, since no triangle in G has all three edges coloured 0 and 1. The vertex in G^* corresponding to the outer region of G has odd degree because on the path from A to B, the vertices in G change colour an odd number of times (since they start with 0 and end with 1).

By the handshaking lemma (Lemma 2.1.2), the number of odd degree vertices in G^* is even. Since the "outside" vertex in G^* has odd degree, the number of vertices from interior regions having odd degree is odd—and so is at least one. Since 1 is the only odd number from 0 to 2, there exists an interior vertex in G^* with degree exactly 1. Such a vertex corresponds precisely to an inner triangle with all vertices coloured differently. $\square$

Corollary to proof of Sperner's lemma (Theorem 8.6.1). For any Sperner colouring, the number of multicoloured triangles is odd.

For a function f on some set X, a element $x \in X$ is called a *fixed point* (of f) if and only if $f(x) = x$. Luitzen Brouwer's fixed point theorem [155] from 1912 says that for any $n \in \mathbb{Z}^+$ and any compact convex set $S \subseteq \mathbb{R}^n$, any continuous function $f : S \to S$ has a fixed point. Often Brouwer's fixed point theorem is only stated for the case when S is an n-dimensional ball.

When $n = 1$, and S is a closed interval, then Brouwer's fixed point theorem is true by the Intermediate Value Theorem.

Theorem 8.6.2. *Sperner's lemma for 2 dimensions implies Brouwer's fixed point theorem for* $n = 2$.

Proof: The proof given here is standard, found in, for example, [13]. Since any compact convex set in $\mathbb{R}^2$ can be continuously deformed into a triangle, it suffices to prove the theorem when S is a triangle. For convenience, choose the triangle $S = \triangle ABC$ in $\mathbb{R}^3$ with vertices $A = (1,0,0) = \mathbf{e}_1$, $B = (0,1,0) = \mathbf{e}_2$ and $C = (0,0,1) = \mathbf{e}_3$ (containing all points inside and on the border), and let $f : \triangle ABC \to \triangle ABC$.

Let $G = (V, E)$ be a triangulation of $\triangle ABC$. Define the *mesh length* of G to be

$$\|G\| = \max\{\|\mathbf{u} - \mathbf{v}\| : \{\mathbf{u}, \mathbf{v}\} \in E\},$$

the length of the longest edge in the triangulation G. Construct a sequence of triangulations $G = G_0, G_1, G_2, G_3, \ldots$ of $\triangle ABC$, each a refinement of the previous, so that $\|G\| \to 0$ (this is possible recursively by, at each step, adding the centers of triangles from the previous triangulation).

For any $\mathbf{p} \in \mathbb{R}^3$, write $\mathbf{p} = (p_1, p_2, p_3)$ and write $(\mathbf{p})_i = p_i$ to denote the ith coordinate of $\mathbf{p}$ with respect to the standard basis for $\mathbb{R}^3$.

Assume, for the moment, that f has no fixed point. Writing each $G_j = (V_j, E_j)$, then for any vertex $\mathbf{v} \in V = \cup_j V_j$, $f(\mathbf{v}) \neq \mathbf{v}$, and since each such vertex lies on the plane $x + y + z = 1$, at least one of the three coordinates $(f(\mathbf{v}) - \mathbf{v})_i$ is negative and another positive. For each $j = 0, 1, 2, \ldots$, define a 3-colouring $c_j : V_j \to \{1, 2, 3\}$ as follows: for $\mathbf{v} \in V_j$, let $c_j(\mathbf{v})$ be least $i \in \{1, 2, 3\}$ so that $(f(\mathbf{v}) - \mathbf{v})_i$ is negative.

Claim: Each colouring c_j is a "Sperner colouring"—it satisfies the colouring condition in Sperner's lemma, but with $1, 2, 3$ playing the role of $0, 1, 2$.

Proof of claim: Pick $j \in \{0, 1, 2, 3, 4, \ldots\}$. The coordinates of any $\mathbf{p} \in S = \triangle ABC$ are non-negative, and so the coordinates of any $f(\mathbf{p})$ are also. Hence, for $A = \mathbf{e}_1 = (1, 0, 0)$, the only coordinate of $f(A) - A$ that is negative is the first one. So $c(A) = 1$. Similarly, $c(B) = 2$ and $c(C) = 3$. Any point on $\mathbf{q} \in \overline{BC}$ is of the form $(0, y, z)$, and so the first coordinate of $f(\mathbf{q}) - \mathbf{q}$ is not negative; hence $c(q) \in \{2, 3\}$. Similarly, for any point

$\mathbf{r} \in \overline{AB}$, $c(\mathbf{r}) \in \{1, 2\}$ and for any point $\mathbf{s} \in \overline{AC}$, $c(\mathbf{s}) \in \{1, 3\}$. So c is indeed a Sperner colouring.

By Sperner's lemma, in each G_j, there is a triangle with corners $\mathbf{v}_1^j, \mathbf{v}_2^j, \mathbf{v}_3^j$ so that for each $i = 1, 2, 3$, $c_j(\mathbf{v}_i^j) = i$. Looking at the first point of each Sperner triangle, the sequence $(\mathbf{v}_1^j)_{j \geq 0}$ is an infinite sequence in a compact space, and so has a convergent subsequence, converging to a point $\mathbf{v}_1$. Similarly, the sequences $(\mathbf{v}_2^j)_{j \geq 0}$ and $(\mathbf{v}_3^j)_{j \geq 0}$ contain subsequences that converge to respective points $\mathbf{v}_2$ and $\mathbf{v}_3$. However, since the mesh goes to 0, these three points are the same.

For each j, the first coordinate of $f(\mathbf{v}_1^j)$ is smaller than the first coordinate $\mathbf{v}_1^j$, and so by continuity, the first coordinate of $f(\mathbf{v}_1)$ is at most the first coordinate of $\mathbf{v}_1$. The same is true for second and third coordinates. Then no coordinate of $f(\mathbf{v}) - \mathbf{v}$ is positive, contradicting the assumption that for any $\mathbf{v}$, $f(\mathbf{v}) \neq \mathbf{v}$. □

In 1997, Su [838] showed how (with a direct construction) Brouwer's fix point theorem follows from the Borsuk–Ulam theorem (Theorem 17.5.1).

8.6.4 Higher dimensional Sperner's lemma

Recall that a simplex in n-dimensional Euclidean space is a polytope on $n+1$ vertices. So a 2-dimensional simplex is a set of 3 affinely independent points together with their interior (convex hull), namely, a triangle. A 3-dimensional simplex is a tetrahedron.

Theorem 8.6.3 (Sperner's lemma, n-dimensional version)**.** *Let $n \in \mathbb{Z}^+$, and let $A_1, A_2, \ldots, A_{n+1}$ be a set of affinely independent points in $\mathbb{R}^n$, forming a simplex T_n. Let T_n be partitioned into a finite number of smaller simplices, $S_1, S_2, \ldots$, and let V be the set of vertices thereby determined. If a colouring $c : V \to \{1, 2, \ldots, n+1\}$ satisfies*

- *for each i, $c(A_i) = i$, and*
- *for any $k \leq n$ and a k-dimensional subspace of T_n given by points $A_{i_1}, \ldots, A_{i_{k+1}}$, any point from V on this subspace is coloured with one of $i_1, i_2, \ldots, i_{k+1}$,*

then there is a small n-dimensional simplex so that all of its vertices are coloured differently, and the number of such "rainbow" simplices is odd.

For early work on Sperner's lemma and related results, see [199] or [816]. For computational aspects of Sperner's lemma (and references for application, including game theory) see [50], and for computational aspects of fixed points, see [869]. For a type of Sperner's lemma applied to cubes instead of simplices, (and its relation to another combinatorial theorem called Tucker's Lemma, not covered in these notes) see [578]. For a generalization of Sperner's lemma to polytopes, see [247]. For more on triangulations and algorithms, see [248].

8.7 Dissecting polygons into equal area triangles

Trivially, one can dissect a square into two (or four) identical triangles. In Exercise 217, a square is dissected into 20 congruent (right) triangles.

Exercise 220. *Show that for any $n \in \mathbb{Z}^+$ a square can be divided into $2n$ congruent right triangles.*

Exercise 221. *Show that there is no dissection of a square into three triangles of equal area.*

Theorem 8.7.1 (Monsky, 1970 [663])**.** *A square can not be dissected into an odd number of triangles so that all triangles have equal area.*

Monsky's proof of Theorem 8.7.1 used p-adic analysis and Sperner's lemma (see Section 8.6). For more on p-adic analysis, see, e.g., [16]. For more on Monsky's theorem, see [13] or [552]. The result in Theorem 8.7.1 was first given as a problem in 1967 by Richman and Thomas [749] and the following year, Thomas [867] gave a solution when points of the triangles in a unit square are limited to rational points.

Another result for "balanced" polygons similar to Monsky's theorem was conjectured by Stein in 2000 [820] (see also [821]) and a proof was given by Rudenko in 2014 [772] for the case where all points of the polygons are rational vertices. (The case for polyominoes had also been discussed by Stein.)

8.8 Heilbronn's problem

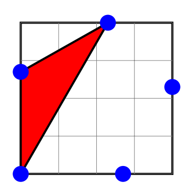
A set S of n points in some Euclidean geometry determine $\binom{n}{3}$ triangles, some of which may be degenerate (when 3 are collinear). Heilbronn (prior to 1950, see [763]) asked if n points are placed in the unit square, what is the maximum area of the smallest triangle determined? As found in Exercise 49, n points in the *interior* of a square can determine $2n+2$ triangles using these points and the four corners, but in Heilbronn's problem, points may be chosen on the border of the square as well.

Heilbronn's question can be asked for other compact shapes, not just the unit square. Only a few basic facts about the Heilbronn problem are given here, concentrating mainly on the problem for the square. Early work on the Heilbronn problem includes the five papers by Roth [763, 765, 766, 767, 768], and a paper by Schmidt [778]. It seems that this problem first appeared in print in Roth's first paper on the subject in 1951. The problem seemed to remain dormant until the 1970s, and since then, this problem continues to attract researchers.

For a compact bounded set $K \subset \mathbb{E}^2$, and an integer $n \geq 3$, let $H_n(K)$ denote the maximum, taken over all placements of n points in K, of the area of a smallest triangle determined by such points. If K is the unit square, the K is omitted and only H_n is used (in some articles, e.g., [763], the notation H_n is replaced by Δ_n). In 1972, configurations for each $n \leq 16$, were given by Goldberg [393], some of which were later proved to be optimal.

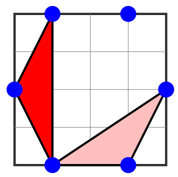
Exercise 222. *What is the largest area of a triangle contained in the unit square In other words, what is H_3? Also, find H_4.*

Goldberg gave a placement of five points given by the vertices of a regular pentagon with an affine transformation (so that it fits in a square), showing $H_5 \geq \frac{3-\sqrt{5}}{4} \sim .191$. The value $H_5 = \frac{\sqrt{3}}{9} = .192...$ was shown optimal by Yang and Zeng, 1992, unpublished (see [268]). This is achieved by:

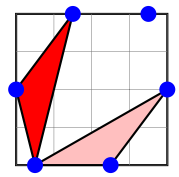


$H_5 = \frac{\sqrt{3}}{9}$

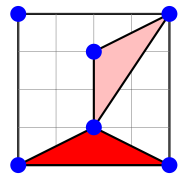
There are infinitely many optimal configurations for H_6; an affine deformation of a regular hexagon gives an optimal configuration. Shifting points slightly or putting two in the interior also works. The value $H_6 = \frac{1}{8} = .125$ was shown optimal by Dress, Yang, and Zeng, 1995 [268].



$H_6 = \frac{1}{8}$

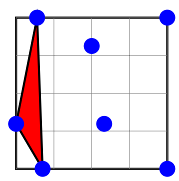


$H_6 = \frac{1}{8}$



$H_6 = \frac{1}{8}$

For seven points, if the points are all placed on the border of the square, the maximum is not reached, but with two points on the interior, the optimal bound is reached. The following configuration was given by Comellas and Yebra [202] in 2002 and proved to be optimal by Chen and Zeng [186] in a paper in 2011.



$H_7 = .0838...$, a root of $152x^3 + 12x^2 - 14x + 1$.

For more results along with wonderful diagrams and references, see Friedman's website [353] or *Wolfram Mathworld* [915].

Points $P_1, \ldots, P_n$ in the unit square, no three collinear, determine $n-2$ or $n-1$ internally disjoint triangles containing P_1, so by the pigeonhole principle, the minimum area is at most $O(1/n)$. Heilbronn conjectured that the area of the smallest triangle is $O(\frac{1}{n^2})$. Various authors have found configurations giving a lower bound on H_n (see [393], [353], or [915] for references); some of this data is in Figure 8.11.

If the Heilbronn conjecture $H_n = O(\frac{1}{n^2})$ is correct, it is best possible, as shown by a delightfully simple argument by Erdős. The following theorem was communicated by Erdős to Roth, who put it in his 1951 paper [763].

Theorem 8.8.1 (Erdős, $\leq$1951, see [763])**.** *As $n \to \infty$,*

$$H_n \neq o\left(\frac{1}{n^2}\right).$$

Proof: Let $n \geq 3$. Let p be the largest prime below n. For $i = 0, 1, \ldots, p-1$, let $P_i = (x_i, y_i)$ be the point defined by $x_i = \frac{i}{p}$, and $y_i = \frac{i^2 \pmod{p}}{p}$.

n	H_n	$\frac{4}{n^2}$
3	.5	.444...
4	.5	.25
5	$\frac{\sqrt{3}}{9} = .192...$	.16
6	.125	.111...
7	.0838...	.0816...
8	$\geq \frac{\sqrt{13}-1}{36} = .0723...$	.0625
9	$\geq \frac{9\sqrt{65}-55}{320} = .0548...$	.04938...
10	$\geq .046537...$	.04
11	$\geq \frac{1}{27} = .037037...$	.03305...
12	$\geq .032599...$	.02777...
13	$\geq .02669743...$	.02366..
14	$\geq .024304...$	.020408...
15	$\geq \frac{29}{1395} = .020789...$	.01777...
16	$\geq \frac{7}{341} = .0205...$	.015625...

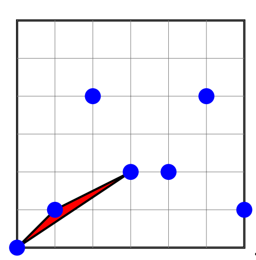
Figure 8.11: Some known values of H_n, compared to $\frac{4}{n^2}$

For $i < j < k$, the points P_i, P_j, P_k are not collinear since (using a form of Corollary 1.5.3 and the well-known determinant of a Vandermonde matrix):

$$\begin{vmatrix} 1 & px_i & py_i \\ 1 & px_j & py_j \\ 1 & px_k & py_k \end{vmatrix} = \begin{vmatrix} 1 & i & i^2 \\ 1 & j & j^2 \\ 1 & k & k^2 \end{vmatrix} = (j-i)(k-i)(k-j) \not\equiv 0 \pmod{p}.$$

By Pick's theorem (Theorem 1.11.8), for example, if a $p \times p$ lattice is formed, the area of any non-trivial triangle with vertices as lattice points is at least $\frac{1}{2}$, so in the construction above, each triangle has area at least $\frac{1}{2n^2} \neq o\left(\frac{1}{n^2}\right)$. $\square$

When $n = p = 7$, the Erdős construction gives the following configuration:


.

Only two of the points in the 7-point construction by Erdős are on the border of the unit square, perhaps suggesting that this construction can be improved.

A nontrivial upper bound for H_n was given in 1951 by Roth [763]:

$$H_n = O\left(\frac{1}{n\sqrt{\ln\ln n}}\right).$$

In 1972, Schmidt [778] proved that there exists a constant c so that

$$H_n < c\frac{500}{n\sqrt{\ln n}}.$$

For small values of n, Goldberg [393] observed that H_n seems to be loosely approximated by $\frac{4}{n^2}$ (see Figure 8.11), but this observation did not lead to a proof for general n. For large n, bounds of the form (where c_1, c_2 are constants)

$$c_1\frac{\ln n}{n^2} < H_n = O\left(\frac{2^{c_2\sqrt{\ln n}}}{n^{8/7}}\right) \tag{8.1}$$

were given in 1982 and 1981, respectively, by Komlós, Pintz, and Szemerédi [568, 567]; the lower bound (obtained probabilistically) disproved Heilbronn's conjecture. In 2000, an algorithm was given by Bertram-Kretzberg, Hofmeister, and Lefmann [77] that gives all triangles with area $\Omega(\ln n/n^2)$, giving another proof of the lower bound in (8.1).

In 2002, Jiang, Li, and Vitányi [504] proved that if n points are chosen uniformly at random (and independently of one another) in $[0,1]^2$, the expected value of the area of a minimal triangle is $\Theta(1/n^3)$.

In 2002, Comellas and Yebra [202] used a technique called "simulated annealing" to give a few new bounds for the square problem (see their paper for many more references).

If an optimal configuration is attained (for points in a square), how many minimal-area triangles occur? One might instead ask how many differently shaped minimal triangles occur. For $n = 3$ points, there is only one triangle. For $n = 4$, all four triangles in the optimal configuration are minimal. For $n = 5$, there are two types of minimal triangle (see diagram above). For $n = 6$, in the first configuration above, there $\binom{6}{3} = 20$ triangles, six of which have minimal area, and these six consist of two of one type and four of another. (For $n = 7$, the configuration given above appears to determine at least 8 minimal triangles, many of which are different.) From the *Wolfram MathWorld* website [915], "As can be seen, the solutions have a great deal of symmetry, with a large number of maximum minimum triangles sharing the same area."

The following heuristic (showed to me by Lefmann [601]) might give a lower bound on the possible number of minimal triangles found in an optimal arrangement of points in a square. If there is a point not used in a minimal area triangle, then move it slightly; not all triangles containing that point get larger (if the point is an interior point), so it can be moved to create at least one additional minimal triangle. From this heuristic, all vertices occur in at least one minimal triangle, and so there are at least $n/3$ minimal triangles. In known configurations, it seems that vertices often occur in more than one minimal triangle, so perhaps this bound of $n/3$ can be increased significantly. It seems unlikely that for most values of n the number of minimal triangles exceeds $O(n)$, but I have no real evidence for such a claim.

Goldberg [393, p.141] also mentions that there are many minimal area triangles by the following reasoning. If there are only a few triangles of (the same) minimal area, then one might be able to move points in the minimal triangles (at the expense of larger triangles) to make them even larger.

Question 8.8.2. *If n points in a square form an optimal configuration for the maximum of the minimum area triangles, how many minimum area triangles are there or can there be? How many of the minimal area triangles can be congruent?*

Extensions of the Heilbronn problem include replacing the square with a convex body (see, e.g., [268, 269] for 6 points in any convex body); in particular, the cases where the square is replaced by either an equilateral triangle or a circle have been well studied (see [915]).

Another way to generalize the Heilbronn problem is to ask for the maximum minimum area of not just a triangle, but of the convex hull of any $k \geq 3$ points. In 2008, Lefmann [600] used results on the independence number of linear hypergraphs (namely, hypergraphs whose hyperedges intersect in at most one vertex) to give the following strong result:

Theorem 8.8.3 (Lefmann, 2008 [600])**.** *For integers $3 \leq k \leq n$, there exists a placement of n points in the unit square $[0,1]^2$ so that for every $j = 3, \ldots, k$,* *the area of the convex hull of any j of these n points is at least*

$$\Omega\left(\frac{(\log n)^{\frac{1}{j-2}}}{n^{\frac{j-1}{j-2}}}\right).$$

Such a placement is found by a deterministic algorithm that runs in $o(n^{6k-4})$ time.

In Theorem 8.8.3, when $k = j = 3$ the bound agrees with the lower bound in (8.1), and improves a bound (for convex k-gons) found earlier by Bertram-Kretzberg, Hofmeister, and Lefmann [77]. One configuration gives (good?) bounds for all of the different j-sets (at once).

Heilbronn's problem is also generalized to higher dimensions. See [597] for an algorithm in three dimensions. In 2001, Barequet [58] gave a lower bound for the Heilbronn problem in d dimensions (where simplices play the role of triangles), and a few years later [59] gave an on-line algorithm for such a bound. In 2003, Lefmann [598] improved on Barequet's bounds by showing that in d dimensions, there exists a placement of n points in $[0,1]^d$ so that every $d+1$ points produces a simplex with volume on the order of $\frac{\log_2 n}{n^d}$. See Lefmann's 2008 paper [599] for more details and references for results for the Heilbronn problem in higher dimensions. See [599] (again, using independence numbers for linear hypergraphs), for a lower bound of volumes of simplices in higher dimensions given by a random polynomial in n time algorithm.

8.9 Dissecting polyhedra of equal volume

Two polyhedra (of equal volume) are said to be "scissor-congruent" if there is a dissection of one using straight cuts so that the pieces can be re-arranged to form the other. There are many examples throughout history showing the scissor congruence of various shapes.

In 1900, Hilbert [469] gave his famous list of 23 problems, some of which he presented at the ICM that year. (See also [470], and for an English translation of these problems, see [472].) His third problem (one that was not given to the ICM) was: Given any two polyhedra of equal volume, is it always possible to cut the first into finitely many polyhedral (actually, tetrahedral) pieces that can be reassembled to yield the second? In particular, are there two tetrahedra with same height and same base area that are not scissor congruent? Here is the text of the earliest English translation [472]:

> 3. THE EQUALITY OF THE VOLUMES OF TWO TETRAHEDRA OF EQUAL BASES AND EQUAL ALTITUDES.
> In two letters to Gerling, Gauss* expresses his regret that certain theorems of solid geometry depend upon the method of exhaustion, i. e., in modern phraseology, upon the axiom of continuity (or upon the axiom of Archimedes). Gauss mentions in particular the theorem of Euclid, that triangular pyramids of equal altitudes are to each other as their bases. Now the analogous problem in the plane has been solved.† Gerling also succeeded in proving the equality of volume of symmetrical polyhedra by dividing them into congruent parts. Nevertheless, it seems to me probable that a general proof of this kind for the theorem of Euclid just mentioned is impossible, and it should be our task to give a rigorous proof of its impossibility. This would be obtained, as soon as we succeeded in *specifying two tetrahedra of equal bases and equal altitudes which can in no way be split up into congruent tetrahedra, and which cannot be combined with congruent tetrahedra to form two polyhedra which themselves could be split up into congruent tetrahedra.*

Max Dehn (1878–1952), a student of Hilbert, solved the problem within a year [243] (see also [244]), showing that indeed such a dissection does not exist for certain pairs of tetrahedra. Dehn first proved that a tetrahedron could not be dissected (using straight cuts) into finitely many pieces to form any prism, including a cube, of the same volume. For more on Dehn's proof, Stillwell's description [831, pp. 159–170] might be helpful; also, see Boltianskii [119] So, there is no 3-dimensional analogue of Theorem 8.1.1.

Early work in dissections of polyhedra include work on a variation of the "Delian Problem" (doubling the cube); see [492] for more details. Stewart Coffin wrote a book [198] on polyhedral dissections from a puzzleist's point of view (rich with many types of puzzles, including interlocking and burr puzzles).

8.10 The Banach–Tarski paradox

The notation, definitions, and theorems given in this section can be found in the book *The Banach–Tarski paradox* by Tomkowicz and Wagon [870]; I highly recommend any interested reader to at least skim the first few chapters—the mathematics behind the Banach–Tarski paradox and other similar geometric paradoxes seems quite rich. This section is (however) only an introduction to the paradox, with no proofs. For proofs and many more related facts, see [870].

A popular version of a theorem called "the Banach–Tarski paradox" is the following:

Theorem 8.10.1 (Banach–Tarski, 1924 [48])**.** *Any ball of radius r in $\mathbb{R}^3$ can be decomposed into finitely many pieces so that these pieces can be rearranged into two balls, each with radius r.*

So out of one ball, it is possible to double its volume by making two copies out of one—and the two copies are seemingly "identical" to the original. The proof relies on the Axiom of Choice and the existence of "non-measurable sets". The same paradox is true if $\mathbb{R}^3$ is replaced with any higher dimensional $\mathbb{R}^d$, but not for $d = 1$ or $d = 2$. The Banach–Tarski paradox is a result that might make one question "what is volume?".

There is an old standard (among mathematicians?) riddle about the Banach–Tarski paradox—I do not remember its origin. To understand the riddle, recall that an anagram of a word is a re-arrangement (or permutation) of its letters. For example, an anagram of "microbotanist" is "combinatorist".

Riddle: What is an anagram of "BanachTarski"?

The word "batrachians" (meaning amphibians) only uses 11 of the 12 letters. The answer is given at the end of this section.

For a positive integer d, say that two sets $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^d$ are *equidecomposable* if and only if there is a partition of X into finitely many pieces so that the pieces can be rearranged to form a copy of Y. So the Wallace–Lowry–Bolyai–Gerwein theorem (Theorem 8.1.1) says that two polygons of equal area are equidecomposable, and the proof uses pieces that are themselves polygons. Dehn's theorem (Theorem 10.3.3) says that a tetrahedron and a cube of the same volume are not equidecomposable using pieces that are polyhedra.

A stronger form of the Banach–Tarski paradox says that for any two bounded subsets X and Y of $\mathbb{R}^3$ with non-empty interior, X can be decomposed into finitely many pieces so that these pieces can be rearranged to form Y (regardless of the "volumes" of X and Y).

Sometimes the Banach–Tarski paradox is stated in terms of the group G_3 of isometries of $\mathbb{R}^3$ and that $\mathbb{R}^3$ is "G_3-paradoxical". (For a group G, the expression "G-equidecomposable" was used by Banach and Tarski [48].) In Theorem 8.10.1, it is tacitly assumed that "rearranged" means that each piece is moved under a local isometry. Similar statements can be made (and conjectured) when the group G_3 is replaced by, say, the group of translations. (In Euclidean Ramsey theory—see Chapter 18—often a monochromatic "copy" of some set is desired, where "copy" usually means "congruent copy", but some theorems are stronger, guaranteeing a monochromatic translation—not just a congruent copy.) As in the Banach–Tarski paradox, can a ball be taken apart in pieces so that these pieces can be arranged by using translations only to form two new balls?

In 1925, Tarski [864] asked if a circle and a square of the same area are equidecomposable. Laczkovich answered this question 65 years later with an even stronger result.

Theorem 8.10.2 (Laczkovich, 1990 [584])**.** *A circle and a square of the same area are equidecomposable by translations alone.*

For a proof and more on Laczkovich's theorem and related work, see [870, Ch. 9] and [378].

In the Banach–Tarski theorem, it takes only a moment of thought to see that the pieces into which the ball is dissected are not polyhedra or any convex sets, since the volumes would not add up. The pieces are very strange. By the stronger form of Banach–Tarski, a tetrahedron can be dissected into finitely many pieces to form a cube of the same volume. For this simple case, can there be any understanding of the type of pieces used? This next theorem (for those who understand "measurable sets") says that for a tetrahedron-to-cube dissection, the pieces can be fairly well-behaved (but not polyhedral, as Dehn proved). According to [870, p. 34], the following was only recently proved (but not yet published):

Theorem 8.10.3 (Grabowski–Máthé–Pikhurko, 2016?)**.** *Any tetrahedron in* $\mathbb{R}^3$ *is equidecomposable with a cube using isometries and pieces that are Lebesgue measurable.*

See [870, p. 34 and Ch. 9] for more information on this and related work; that book also covers versions of the Banach–Tarski paradox in non-Euclidean spaces.

Answer to anagram riddle: BanachTarskiBanachTarski.

8.11 "Dissection" paradoxes

8.11.1 64 = 65?

The American puzzlist Samuel Loyd (31 January 1841 – 10 April 1911), is known for his many puzzles. (Note: his name is not Lloyd, as it sometimes appears.) The following puzzle (Figure 8.12) has been attributed to Sam Loyd, but Martin Gardner [364, pp. 133–134] attributes it to Sam Loyd's son (with the same name) but Greg Frederickson [348, pp. 273–274] says that it was invented by Walter Dexter in 1901.

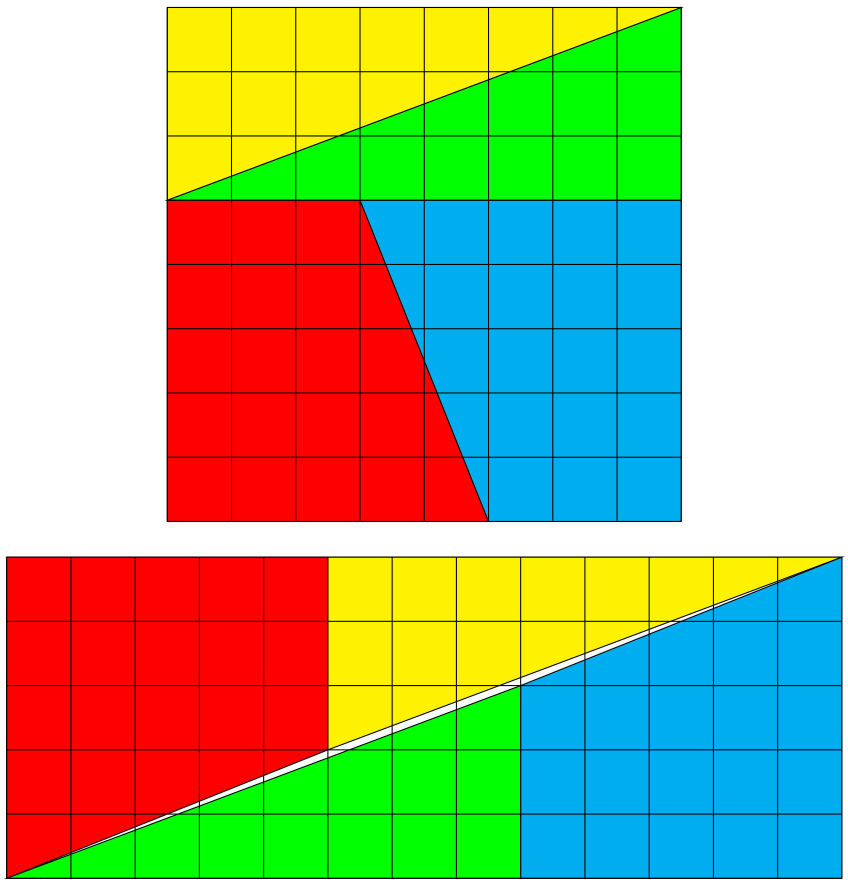


Figure 8.12: Top area is $8 \times 8 = 64$; bottom is $13 \times 5 = 65$?

8.11.2 Curry's triangle paradox

The following puzzle (see Figure 8.13) is by Paul Curry, an amateur magician, in 1953 (see [364, pp. 139–150] for reference). (There is another puzzle called "Curry's paradox" by Haskell Curry, a logician.) The form given here is an adaptation of the original made popular by Martin Gardner in 1956 (the original used pieces forming a larger isosceles triangle). In each of the images

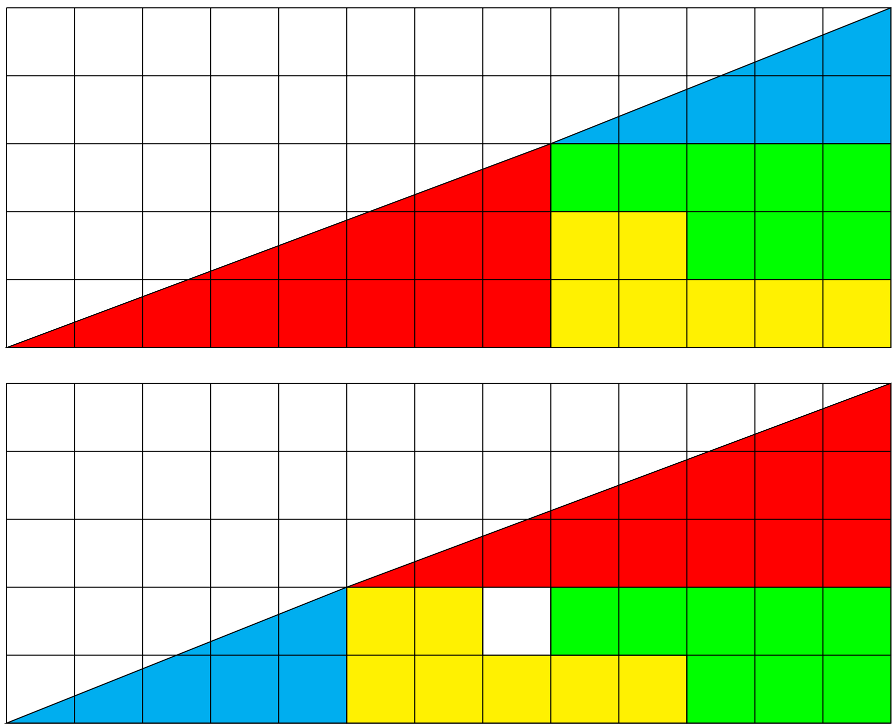


Figure 8.13: Curry's paradox: where did the hole come from?

in Figure 8.13, the pieces assemble to what appears to be a right triangle—and since the base and height of this "triangle" is the same in both cases, it seems as if the areas of the "triangles" are the same.

The solution to Curry's paradox can be explained in terms of slope. The slope of the red triangle is $\frac{3}{8} = 0.375$ and the slope of the blue triangle is $\frac{2}{5} = 0.4$. So what appears to be the hypotenuse of the large "triangle" is not a straight line (the top one has a slight dip downward and the bottom one has a slight bulge. The difference in area in the two "triangles" is made up by the crooked "hypotenuses".

8.11.3 Fibonacci numbers and dissection paradoxes

Recall from Definition 1.15.1 that the Fibonacci numbers $F_0, F_1, F_2, F_3, \ldots$, are defined recursively by $F_0 = 0$, $F_1 = 1$, and for $n \geq 2$, $F_n = F_{n-1} + F_{n-2}$. The first few are 0, 1, 1, 2, 3, 5, 8, 13, 21.

In the "64 = 65" dissection, the dimensions include the numbers 3, 5, 8, and 13, all Fibonacci numbers. In Curry's paradox, dimensions are 1, 2, 3, 5, 8, and 13, again Fibonacci numbers.

One key property of the Fibonacci numbers used here is that if F_k and F_{k+2} are Fibonacci numbers, their product is very close to F_{k+1}^2. For example, $F_5 \cdot F_7 = 5 \cdot 13 = 65$, which is one away from $F_6^2 = 8^2 = 64$. This pattern holds in general by what is now sometimes called "Cassini's identity":

Theorem 8.11.1 (Cassini's identity)**.** *For every positive integer n,*

$$F_{n-1}F_{n+1} = F_n^2 + (-1)^n.$$

Cassini's identity was presented by the Italian astronomer Giovanni Domenico Cassini (1625–1712) to the Royal Academy (Paris) in 1680, but these proceedings [179] were published only in 1733 (under the name Jean Dominique Cassini).

Cassini's identity is also often called "Simson's identity", since it was independently discovered in 1753 by the Scottish geometer Robert Simson (1687–1768) [796] (see also, *e.g.* [220, pp. 165–168], and [224, p. 41]). Perhaps the identity should be known as "Kepler's identity", as, according to Graham, Knuth, and Patashnik [407, p. 292], Johannes Kepler (1571–1630) knew of it in 1608. (They offer [543] as evidence of this.)

Exercise 223. *Prove Theorem 8.11.1 (Cassini's identity).*

The puzzle in Figure 8.12 confirms Cassini's identity

$$F_5 \cdot F_7 = 5 \cdot 13 = 65 = 64 + 1 = 8^2 + 1 = F_6^2 + 1.$$

The difference (of 1) in areas is made up by the small sliver in the second image).

Exercise 224. *The Sam Loyd "64=65" puzzle is based on the equation $8^2 = 5 \cdot 13 - 1$. Repeat the construction with 13 replacing 8 and the equation $13^2 = 8 \cdot 21 + 1$. What is the outcome? Repeat the same exercise when 13 is replaced with 21. Explain the pattern.*

Another key property of Fibonacci numbers helps the Curry paradox to "work". As already mentioned in Section 1.15, the ratio of consecutive Fibonacci numbers approaches the golden ratio. So the ratio of Fibonacci numbers two apart approaches the golden ratio squared. In Curry's paradox, the two slopes $\frac{2}{5} = \frac{F_3}{F_5}$ and $\frac{3}{8} = \frac{F_4}{F_6}$ are already close enough so that the eye is fooled. Using larger Fibonacci numbers may very well lead to puzzles where differences in slopes are even smaller.

Exercise 225. *Repeat the construction of the Curry paradox but with 13 being replaced with the next Fibonacci number 21. Instead of an extra square, there is a missing square. Show how Cassini's identity relates.*

8.11.4 Martin Gardner's vanishing tile puzzle

Martin Gardner gave the "vanishing cube puzzle" or "the mystery of the missing tile", in a 1961 Scientific American column. This puzzle is reprinted in chapter 11 (Mr. Apollinax visits New York) of his 1966 book *New mathematical diversions* (also in the revised 1995 edition [375, pp. 124–133]). In the 1995 edition, Gardner added that the puzzle was made in China and marketed in the 1990s in the United States under the name "Puzzle Mania" by Playtime Toys, Louisville, Kentucky (although the puzzle was produced with no mention of Gardner).

The paradox arises when 17 tiles are arranged in a square as in Figure 8.14, but 16 of these tiles can also make up the same square. (For convenience, I have drawn the two pictures slightly tilted; the original had both squares oriented horizontally, which might make the puzzle more challenging.)

Gardner says that he based his puzzle on Curry's triangle paradox (see Section 8.11.2). Gardner gives the following solution:

> "When all seventeen tiles are formed into a square, the sides of the square are not absolutely straight but convex by an imperceptible amount. When one cube is removed and the sixteen tiles re-formed into a square, the sides of the square are concave by the same imperceptible amount. This accounts for the apparent change in area."

The blue right triangles have slope 2/5 and the grey triangles have slope 3/8. Just as in the Curry paradox, these two slopes are sufficiently close to

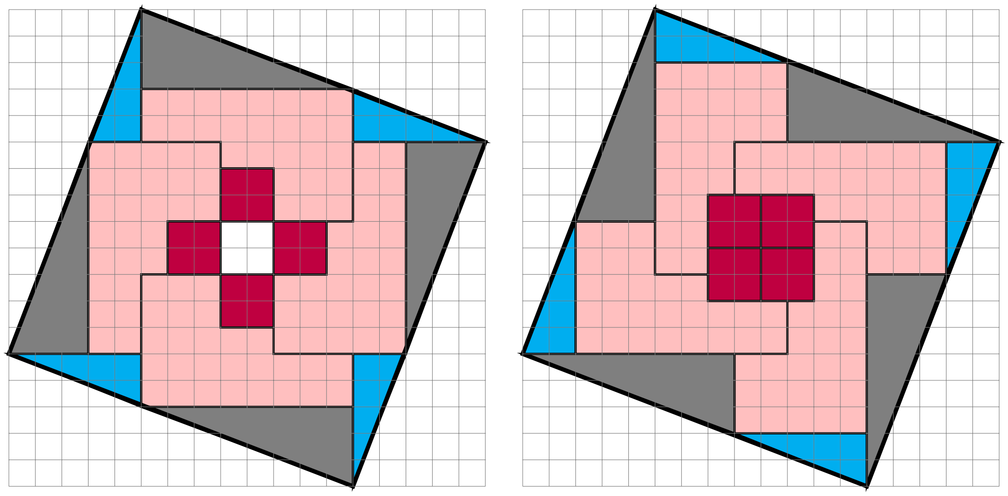


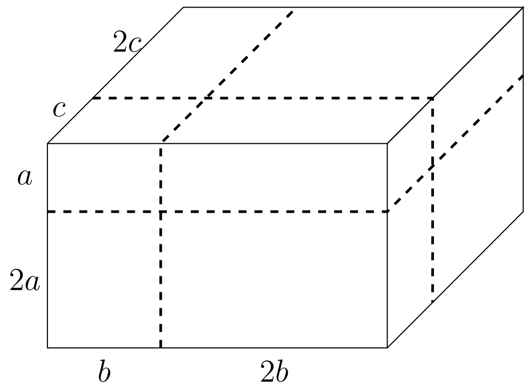
Figure 8.14: Gardner's missing tile puzzle: where did the white tile go?

one another that together they give the illusion of a straight line. The puzzle is made less obvious by drawing the outside border with a thicker line so as to hide the respective bulges or indentations in what appear to be two squares.

8.11.5 O'Beirne's melting box

Thomas H. O'Beirne was a Scottish mathematician who ran a puzzle column for *The New Scientist* during 1961–1962, but perhaps is more famous for his 1965 book *Puzzles and paradoxes* [688], a compilation of problems from his column. One of his most famous puzzles involving packing rectangular blocks, and is sometimes called *O'Beirne's melting box* (not to be confused with "O'Beirne's cube", another puzzle). This melting box puzzle was first shown to me by Richard Guy in Calgary in 1999.

The starting point is a brick cut into 8 pieces with three slices, each slice parallel to a wall of the brick, and each slice cutting the sides of the brick in a 1:2 ratio. If the smallest piece is a brick with dimensions a, b, c, the picture might look like:



So a box holding these 8 pieces has volume $27abc$, with the smallest piece having volume abc.

What is the puzzle? Suppose that an additional small abc piece is added, giving 9 pieces—can all 9 be repacked into the box?

In the model above, indeed the extra piece can be fitted inside the box! It can take a while to find the new arrangement of the 9 pieces inside the box, but it can be done.

What are the dimensions a, b, c so that such a trick is possible? The secret relies on the fact that instead of making the box with dimensions $3a \times 3b \times 3c$, make the box with one millimeter of spare room, say, with dimensions $(3a+1) \times (3b+1) \times (3c+1)$. In other words, for the trick to work, since the volume of the 9 pieces is $27abc + abc = 28abc$,

$$28abc = (3a+1)(3b+1)(3c+1). \tag{8.2}$$

To find values for a, b, c so that equation (8.2) holds, look at the factors of $28abc$. The only reasonable factorization of 28 is $2 \cdot 2 \cdot 7$, and since in the picture above, a is the smallest of the dimensions, put

$$\begin{aligned} 7a &= 3b+1 \\ 2b &= 3c+1 \\ 2c &= 3a+1. \end{aligned}$$

In matrix form,

$$\left[\begin{array}{rrr|r} 7 & -3 & 0 & 1 \\ 0 & 2 & -3 & 1 \\ -3 & 0 & 2 & 1 \end{array}\right],$$

which has the unique solution $a = 19$, $b = 44$, $c = 29$. So cut the pieces and hope that you can find an arrangement of all 9. (It might take a while!)

In 1999, Richard Guy challenged me with the following (I have not yet found an answer):

Problem 8.11.2 (Guy, 1999). *Find dimensions for a puzzle like O'Beirne's that uses a box that is closer to being a cube.*

For example, a solution to Problem 8.11.2 might be found by starting with the first brick that instead of being sliced in a 1:2 ratio, maybe try another ratio (like 3:5). Even if dimensions for the first eight pieces can be found that work algebraically, it might be an intractable question to find a packing of the resulting 9 pieces.

O'Beirne's puzzle is commercially available from, e.g., *Creative Crafthouse* (see `http://www.creativecrafthouse.com/index.php?main_page=product_info&products_id=712`) where the box has a lid.

Chapter 9

Lattice points

9.1 Simple lattice point problems

In discrete mathematics, the word "lattice" is used to denote a partial order with special properties (see Section 21.1). Here, the word "lattice" denotes something more visual.

A *lattice point* in some Euclidean space is one with integer coordinates (relative to some basis, usually the standard basis). For $d \geq 2$, the set of all lattice points $\mathbb{Z}^d$ is sometimes referred to as an *integer lattice*.

Exercise 226. *For any five integer lattice points in the plane, at least one pair has a midpoint that is also a lattice point.*

In fact, much more is true:

Exercise 227. *Prove that for $n \geq 1$, for any $2^n + 1$ lattice points in $\mathbb{Z}^n$ there exist two points whose midpoint is also a lattice point.*

Exercise 228. *Prove that for any nine lattice points in $\mathbb{Z}^3$, no three collinear, there is a midpoint which is a lattice point.*

Exercise 229. *For each $n \geq 0$, consider the set S_n of lattice points bounded by $x \geq 0$, $y \geq 0$, and $x + y \leq n$. Prove that S_n can be covered by no fewer than $n + 1$ lines.*

The next few results show that the square is the only regular n-gon that can have all vertices as lattice points.

Lemma 9.1.1. *No three points in the integer lattice $\mathbb{Z}^2$ form an equilateral triangle.*

Proof: Let T be an equilateral triangle with side length c, and suppose that the corners of T are lattice points. If two of these points have (integer) coordinates (x_1, y_1) and (x_2, y_2), then by the distance formula, $c^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$ is an integer. Hence the area $\frac{\sqrt{3}}{4}c^2$ is irrational. However, the area of any polygon with vertices on the integer lattice is rational (see Pick's theorem, Theorem 1.11.8, if necessary). $\square$

Thus, for each $k \in \mathbb{Z}^+$, no regular $3k$-gon can have all vertices as integer lattice points (because such a polygon has vertices that determine an equilateral triangle). For example, no regular hexagon can have all integer lattice points for vertices. Of course, it is easy to find a square with integer lattice points. The result in the next exercise might seem rather strong, but one proof is surprisingly simple.

Exercise 230. *Show that for each positive integer $n \geq 5$, no regular n-gon exists whose vertices are integer lattice points.*

9.2 Geometry of numbers

It might be said that most of what is now called "mathematics" had origins in both geometry and number theory. These two fields did not evolve independently. For example, the theorem of Pythagoras for right triangles immediately leads to the question "which integers are the sum of two squares?", a question in number theory. Today, the area of mathematics called "algebraic geometry" is largely motivated by questions in number theory.

In this section, some of the relations between geometry and number theory are studied; in particular, theorems about convex sets are related to vectors with integer coordinates. This connection is now sometimes called "the geometry of numbers", as coined by Hermann Minkowski (1864–1909), with a lecture with this title given in 1891 (see [941]) and his book *Geometrie der Zahlen* [658], published a few years later (some references say that his book was published posthumously in 1910, but others give 1896 as the publication date—I have not seen a copy). One of Minkowski's students was Albert Einstein, but two of his doctoral students are more worthy of mention here, namely, Constantin Carathéodory (see Theorem 5.2.1) and Dénes König (one of the founding fathers of graph theory).

Minkowski was known for many things, including his work in relativity, continued fractions, and analysis. In these notes, only his work on convex sets intersecting lattices is studied. For a more thorough introduction to geometry of numbers, see the book [420].

9.3 Minkowski's theorem for integer lattice points

For a positive integer d, the integer lattice in $\mathbb{R}^d$ is the set of points

$$\mathbb{Z}^d = \{(x_1, \ldots, x_d) : x_1 \in \mathbb{Z}, \ldots, x_d \in \mathbb{Z}\}.$$

The integer lattice $\mathbb{Z}^2$ can be viewed as a pattern generated by unit squares. Each unit square is called a "cell" or "fundamental region" generated by the two unit vectors $(0, 1)$ and $(1, 0)$.

Exercise 231. *(easy) Find a convex body in $\mathbb{R}^2$ that has area 4 and contains the origin, but no other integer lattice point.*

Exercise 232. *Find a convex body C that has infinite area but no lattice point (in $\mathbb{Z}^2$).*

A set $S \subseteq \mathbb{R}^d$ is called *centrally symmetric* if and only if $\mathbf{x} \in S$ implies $-\mathbf{x} \in S$.

Exercise 233. *For $d \geq 2$, show that if $C \subseteq \mathbb{R}^d$ is convex, centrally symmetric, and has finite non-zero area (and so contains the origin), then C is bounded.*

In 1910, the book *Geometrie der Zahlen* [658] by Minkowski was published posthumously. Apparently, one of the major theorems (regarding quadratic forms) in that book was proved in 1891 (but published in 1896); an underlying theorem for convex bodies and lattice points in that work is now called "Minkowski's theorem". The simplest form of that theorem is given below as Theorem 9.3.2, and a slightly more general form is given below as Theorem 9.4.2.

In 1914, Blichfeldt [108] observed a simple principle that implies Minkowski's theorem; modern day proofs of Minkowski's theorem (e.g., see [645]) for lattice points now implicitly use Blichfeldt's principle. According to [645], van der Corput also discovered this principle. Only a simple form of the principle is given here.

Theorem 9.3.1 (Blichfeldt's principle, 1914 [108]). *Let X be a bounded measurable set in $\mathbb{R}^d$ with vol$(X) > 1$. Then there are points $\mathbf{x}_1, \mathbf{x}_2 \in S$, $\mathbf{x}_1 \neq \mathbf{x}_2$, so that $\mathbf{x}_1 - \mathbf{x}_2 \in \mathbb{Z}^d$.*

Proof idea: If there are no such points in S, then integer vector translates of $\frac{1}{2}X$ are disjoint, each translate having volume 2^{-d}. Together, these translates do not cover some large cube. Further details are given in the proof of Theorem 9.3.2 below. $\square$

Theorem 9.3.2 (Minkowski, 1891 (see [658])). *Let $C \subset \mathbb{R}^d$ be a centrally symmetric convex set with vol$(C) > 2^d$. Then C contains at least one integer lattice point different from the origin $\mathbf{0}$.*

Proof: (The proof given here is from [645].) Let $C' = \frac{1}{2}C$. Then C' is also convex and centrally symmetric. To complete the proof, a claim is used.

Claim: There exists $\mathbf{v} \in \mathbb{Z}^d \backslash \{\mathbf{0}\}$ so that $C' \cap (C' + \mathbf{v}) \neq \emptyset$.

Proof of claim: Suppose not; assume that for any non-zero $\mathbf{v} \in \mathbb{Z}^d$, C' and $C' + \mathbf{v}$ are disjoint. This assumption in fact implies that any two integer translates are disjoint, because if some $\mathbf{u}, \mathbf{v} \in \mathbb{Z}^d$ and $\mathbf{x}_1, \mathbf{x}_2 \in C'$ satisfy

$$\mathbf{x}_1 + \mathbf{u} = \mathbf{x}_2, +\mathbf{v},$$

then $\mathbf{x}_1 - \mathbf{x}_2 \in \mathbb{Z}^d$, contrary to the assumption that $C' \cap (C + \mathbf{x}_1 - \mathbf{x}_2) = \emptyset$. So every two copies of C' are disjoint.

Let R be a large integer (how "large" depends upon how close vol(C') is to 1—see below). Consider the family $\mathcal{C}$ of translates of C' by integer lattice vectors in $[-R, R]^d \cap \mathbb{Z}^d$. If D is the diameter of C', all of $\mathcal{C}$ is found inside the cube $K = [-R - D, R + D]^d$. (Actually, D could be replaced by the radius of C'.) See Figure 9.1 for a depiction when $d = 2$ and $R = 3$ (based on the diagram in [645, p. 18]). Then

$$\begin{aligned}
\mathrm{vol}(K) &= (2R + 2D)^d \\
&\geq |\mathcal{C}|\mathrm{vol}(C') && \text{(because copies are disjoint)} \\
&= (2R + 1)^d \mathrm{vol}(C')
\end{aligned}$$

and so

$$\mathrm{vol}(C') \leq \left(\frac{2R + 2R}{2R + 1}\right)^d = \left(1 + \frac{2D - 1}{2R + 1}\right)^d,$$

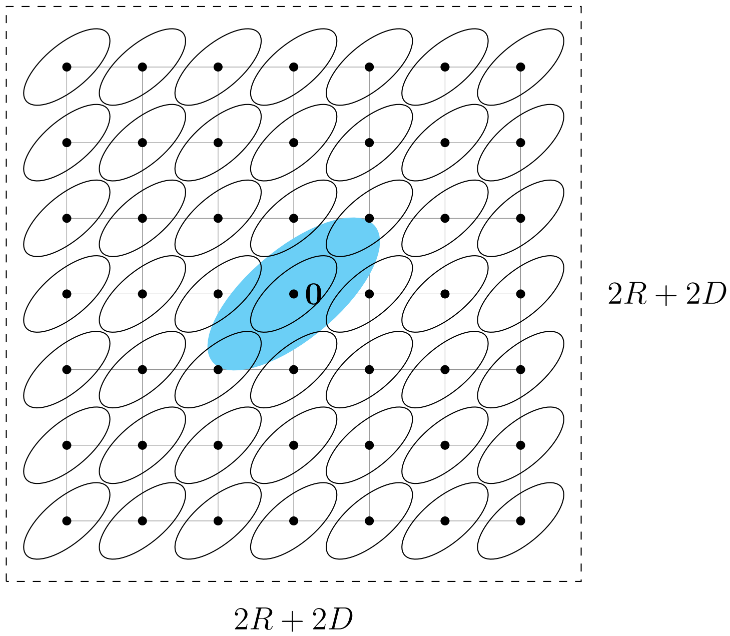


Figure 9.1: A step in the proof of Minkowski's theorem showing copies of C', with $R = 3$ and C in cyan.

which, for large enough R, can be made as close to 1 (but still greater than 1) as desired. However, since $\mathrm{vol}(C') = 2^{-d}\mathrm{vol}(C) > 1$, there exists $\epsilon > 0$ so that $\mathrm{vol}(C') > 1 + \epsilon$. This contradicts being able to get as close to 1 as desired, and this contradiction finishes the proof of the claim.

So by the claim, let vectors $\mathbf{v} \in \mathbb{Z}^d \backslash \{\mathbf{0}\}$ and $\mathbf{x}$ be so that $\mathbf{x} \in C' \cap (C' + \mathbf{v})$. Then $\mathbf{x} - \mathbf{v} \in C'$, and since C' is centrally symmetric, also $\mathbf{v} - \mathbf{x} \in C'$. Since C' is convex, the midpoint between $\mathbf{x}$ and $\mathbf{v} - \mathbf{x}$ is also in C'; in other words, $\frac{1}{2}\mathbf{v} \in C' \backslash \{\mathbf{0}\}$. Thus $\mathbf{v} \in C \backslash \{\mathbf{0}\}$ as desired. □

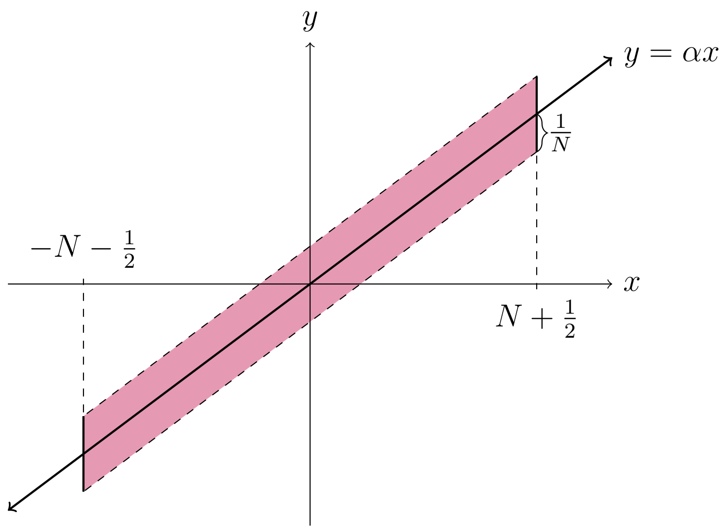
Minkowski's theorem can be used to prove an old result by Dirichlet:

Theorem 9.3.3 (Dirichlet, 1834 [261]). *Let $\alpha \in (0, 1)$ be a real number, and let $N \in \mathbb{Z}^+$. There are $m, n \in \mathbb{Z}^+$ with $n \leq N$ and*

$$\left|\alpha - \frac{m}{n}\right| < \frac{1}{nN}.$$

Proof: Let C be the parallelogram straddling the line $y = \alpha x$ defined by

$$C = \left\{ (x, y) : x \in \left[-N - \frac{1}{2}, N + \frac{1}{2} \right], |y - \alpha x| < \frac{1}{N} \right\}.$$



Then $\text{area}(C) = (2N+1)\frac{2}{N} > 4$. By Minkowski's theorem, there exists a non-zero integer lattice point $(n, m) \in C$. Observe that $\neq 0$, because there are no lattice points other than $(0, 0)$ inside of C on the y axis. Then $|m - \alpha n| < \frac{1}{N}$, and multiplication through by $\frac{1}{n}$ finishes the proof. $\square$

In the above proof, the value $\frac{1}{2}$ can be replaced by any $\epsilon > 0$; the extra $\frac{1}{2}$ does not add any more lattice points to the parallelogram, but makes its area just larger than the 4 required to apply Minkowski's theorem.

The original proof of Dirichlet's theorem used induction.

Proof: For any real number x, let $\{x\}$ denote the fractional part of x; for example, if $x = 3.12$, then $\{x\} = .12$. Suppose that α and N are given. (In fact, there is no need to assume that $\alpha \in (0, 1)$.)

If α is rational, then there is nothing to prove. So suppose that α is irrational and consider the $N + 1$ numbers

$$\{\alpha\}, \{2\alpha\}, \{3\alpha\}, \ldots, \{N\alpha\}, \{(N+1)\alpha\}.$$

Putting these numbers in the pigeonholes (open intervals)

$$\left(0, \frac{1}{N}\right), \left(\frac{1}{N}, \frac{2}{N}\right), \ldots, \left(\frac{N-1}{N}, 1\right),$$

there exist two in the same pigeonhole, say $\{a\alpha\}$ and $\{b\alpha\}$ with $a < b$, and so differ by at most $1/N$. Put $n = b - a$; since n is the difference between two different numbers in $1, 2, \ldots, N+1$, it follows that $1 \leq n \leq N$.

It remains only to see that there is an integer m so that $|n\alpha - m| < 1/N$, and then division by n finishes the theorem. □

9.4 Minkowski's theorem for general lattice points

The fundamental region (or cell) of the integer lattice $\mathbb{Z}^d$ is the unit square. Instead of unit squares, one could use cells that are, say, equilateral triangles, which are generated by the two vectors $(0, 1)$ and $(\frac{1}{2}, \frac{\sqrt{3}}{2})$. Note that these two vectors form a basis for $\mathbb{R}^2$.

Definition 9.4.1. Let d be a positive integer. A *lattice* in $\mathbb{R}^d$ is a set $\Lambda \subset \mathbb{R}^d$ so that there is a basis $\mathbf{v}_1, \ldots, \mathbf{v}_d$ for $\mathbb{R}^d$ with

$$\Lambda = \left\{ \sum_{i=1}^{d} z_i \mathbf{v_i} : (z_1, \ldots, z_d) \in \mathbb{Z}^d \right\}.$$

The d-dimensional parallelepiped determined by the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ is called a cell (or fundamental region), and the volume of a cell is denoted by $\det(\Lambda)$.

The reason for the notation "$\det(\Lambda)$" is that the volume of a parallelepiped is found by the absolute value of the determinant of the matrix formed by the generating vectors (see Section 3.4.2 for details).

Theorem 9.4.2 (Minkowski, general version,1896? [658]). *Let $d \geq 2$ be an integer. Let Λ be a lattice in $\mathbb{R}^d$, and let C be a centrally symmetric convex set in $\mathbb{R}^d$ with vol$(C) \geq 2^d \det(\Lambda)$. Then $C \cap \Lambda \neq \{\mathbf{0}\}$.*

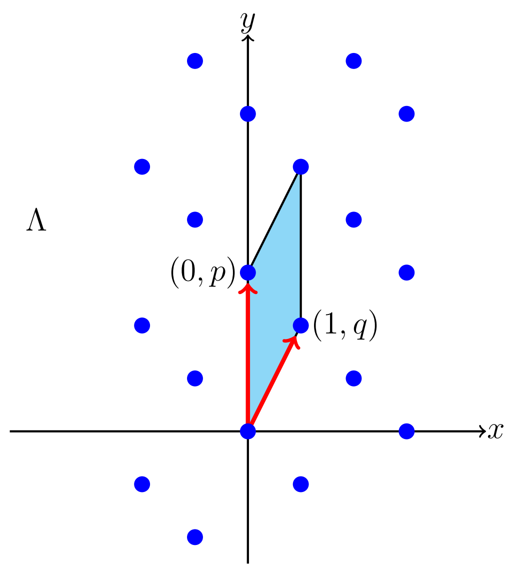
Exercise 234. *Using Theorem 9.3.2, prove Theorem 9.4.2.*

From the general version of Minkowski's theorem, one can prove major theorems in number theory (some proofs of these facts can be found in the appendix Chapter 22 on number theory). For the first such proof, a lemma from number theory is used (this lemma has an easy proof; see Chapter 22).

Lemma 9.4.3. *If p is a prime with $p \equiv 1 \pmod 4$, then -1 is a quadratic residue.*

Theorem 9.4.4 (Fermat and Euler)**.** *Let p be a prime with $p \equiv 1 \pmod 4$. Then p is the sum of two squares.*

Proof: By Lemma 9.4.3, let q be so that $q^2 \equiv -1 \pmod p$. To apply Minkowski's theorem, examine the lattice $\Lambda \subseteq \mathbb{R}^2$ defined by two vectors $(1, q)$ and $(0, p)$.



The area of a cell is p. Consider the region

$$C = \{(x, y) : x^2 + y^2 < 2p\}.$$

The volume (area) of C is $\pi \cdot 2p > 4p$, so by Minkowski's theorem, C contains a non-trivial lattice point $(a, b) \in \Lambda$. Let $i, j \in \mathbb{Z}$ be so that

$$(a, b) = i(1, q) + j(0, p) = (i, iq + jp).$$

Then

$$\begin{aligned} a^2 + b^2 &= i^2 + (iq + jp)^2 \\ &= i^2 + i^2q^2 + 2ijqp + j^2p^2 \\ &\equiv 2ijqp + j^2p^2 \pmod p \qquad\qquad (\text{since } q^2 \equiv -1) \end{aligned}$$

$$\equiv 0 \pmod{p}.$$

So $a^2 + b^2$ is a multiple of p; but $0 < a^2 + b^2 < 2p$, so $a^2 + b^2 = p$. □

Lagrange's theorem (Theorem 22.5.1) is stated and proved in Chapter 22, but for convenience, is repeated here:

Lagrange's theorem Every positive integer is the sum of at most four (integer) squares.

A proof of Lagrange's theorem that uses Minkowski's theorem appears in number theory texts, for example, by Erdős and Surányi [319] and by Montgomery, Niven and Zuckerman [664], and is outlined in a sequence of exercises in [645], which are attributed to Davenport.

Proof of Lagrange's theorem using Minkowski's theorem: By Lemma 22.5.2, it suffices to prove that every prime is the sum of at most 4 squares, so let p be a prime. If $p = 2$, the result is trivial, so suppose $p \geq 3$. By Lemma 22.5.3, let integers a, b be such that p divides $a^2 + b^2 + 1$.

Consider the lattice Λ in $\mathbb{R}^4$ generated by the vectors $(p, 0, 0, 0)$, $(0, p, 0, 0)$, $(a, b, 1, 0)$, and $(b, -a, 0, 1)$. Then $\det(\Lambda) = p^2$. Let C be the open ball in $\mathbb{R}^4$ with radius $\sqrt{2p}$. Then, by equation (3.5), $\mathrm{vol}(C) = \pi^2(2p)^2/2 = 2\pi^2p^2$. which is larger than 2^4p^2. By Minkowski's theorem (Theorem 9.4.2) there is a non-zero lattice point $(y_1, y_2, y_3, y_4) \in \Lambda \cap C$. Let $x_1, x_2, x_3, x_4 \in \mathbb{Z}$ be so that

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} p & 0 & a & b \\ 0 & p & b & -a \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}.$$

Then

$$\begin{aligned} y_1^2 + y_2^2 + y_3^2 + y_4^2 &= (px_1 + ax_3 + bx_4)^2 + (px_2 - bx_3 + ax_4)^2 + x_3^2 + x_4^2 \\ &\equiv (ax_3 + bx_4)^2 + (-bx_3 + ax_4)^2 + x_3^2 + x_4^2 \pmod{p} \\ &= (a^2 + b^2 + 1)(x_3^2 + x_4^2) \\ &\equiv 0 \pmod{p}. \end{aligned}$$

Since $0 < y_1^2 + y_2^2 + y_3^2 + y_4^2 < 2p$ and $y_1^2 + y_2^2 + y_3^2 + y_4^2 \equiv 0 \pmod{p}$, it follows that $y_1^2 + y_2^2 + y_3^2 + y_4^2 = p$, as desired. □

For other exercises that use a form of Minkowski's theorem, see [645] or [664, pp. 319–322].

Chapter 10

Grid problems

10.1 Rectangles with monochromatic corners

Lemma 10.1.1. *Let $k \geq 2$, and let $c : \mathbb{Z}^2 \to \{1, 2, \ldots, k\}$ be a k-colouring of the vertices of the integer lattice. Then there exists a rectangle (with vertical and horizontal sides) in $\mathbb{Z}^2$ whose vertices are monochromatic.*

Proof: Consider any $(k+1)\times(k^{k+1}+1)$ grid G. Since there are k^{k+1} different colouring patterns for columns of G, by the PHP, there are two such columns identically coloured. Each is coloured with k colours, so again by the PHP, there are two positions in these last columns that are the same. These two positions in the two columns are the corners of the desired rectangle. $\square$

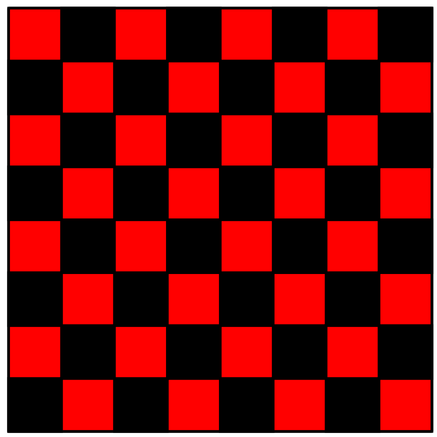
The next problem contains a basic result (which can be interpreted as a result in Euclidean Ramsey theory—see Chapter 18) with two slick proofs, one using the Cauchy–Schwarz inequality and the other using just the pigeonhole principle (PHP) and a shifting argument based on the convexity of the $\binom{x}{2}$ (virtually the same trick is often used in extremal graph theory):

Exercise 235. *Prove that if the squares of a 12×12 board are coloured with three colours, there exists a rectangle (oriented vertically/horizontally) with all four corner squares sharing the same colour.*

10.2 Checkerboards, chessboards, and polyominoes

10.2.1 Checkerboards and chessboards

Checkerboards have existed for centuries. The standard checkerboard is an 8×8 square, partitioned into 64 squares, and the squares are coloured using two shades, usually one light and one dark, with no two adjacent squares coloured the same.



The two colours are arbitrary, but usually one is a lighter shade and the other, very dark. In checkers, opponents sit on opposite sides, and the corner square closest to each player's right is red (a quarter turn of the board suffices). Checkers are played only on the black squares.

Chess is also played on a checkerboard (at which point, the board becomes a chessboard). In chess, the left-most square closest to each player is black (just like checkers, but stated differently!), and each queen begins on a square agreeing with her colour.

For many problems stated in terms of checkerboards or chessboards, the two colours chosen are usually arbitrary—it is not important that the lower left square is dark. Often the colours are not even considered (though many such problems are often given as grid problems). Some problems given below ask to colour squares in a particular manner or change the dimensions—which means that such problems are not really problems for checker/chessboards, but the terminology persists. For example, if $m, n \in \mathbb{Z}^+$, an $m \times n$ "checkerboard" has m rows and n columns (of 1×1 squares, also called "cells"). There are many so-called "chess puzzles" that require understanding of the game; only a few such simple puzzles occur here, chosen mainly because of the geometry behind them. About all one needs to know is that rooks (also called "castles") move to any cell in the same row or column, and queens can move along any row, column, or diagonal.

Exercise 236. *How many non-degenerate rectangles can be formed in an* 8×8 *checkerboard by using only lines in the board. For example, there are 64* 1×1 *squares, 49* 2×2 *squares, Is there a formula for an* $m \times n$ *board? Can you find a formula that gives the number of* $k \times \ell$ *rectangles in an* $m \times n$ *board?*

Exercise 237. *What is the maximum number of non-congruent integer-side rectangles that can be formed by cutting an* 8×8 *chessboard into pieces with all cuts parallel to the sides of the board.*

Exercise 238. *The numbers from 1 to 81 are written on the squares of a* 9×9 *chessboard. Show that there exist two neighbouring squares (adjacent in same row or same column) that have numbers differing by at least 6.*

The result in Exercise 238 is not optimal. By the next exercise, there are adjacent squares that differ by at least 9.

Exercise 239. *Let* $n \geq 2$ *and let the numbers* $1, 2, \ldots, n^2$ *be placed in the squares of an* $n \times n$ *chessboard (one number per square). Show that there are two adjacent squares (sharing an edge) whose numbers differ by at least* n.

Exercise 240. *At random, label the squares of an* 8×8 *chessboard with the numbers 1–64 (a unique number for each square). Define a* saddle square *to be one whose label is the maximum in its column but minimum in its row. What is the probability that a saddle square exists?*

Exercise 241. *Is it possible to paint the cells of a* 1990×1990 *chessboard with black or white so that the pattern is antisymmetric with respect to the center (a cell and its opposite have different colours) and each row/column has the same number of blacks?*

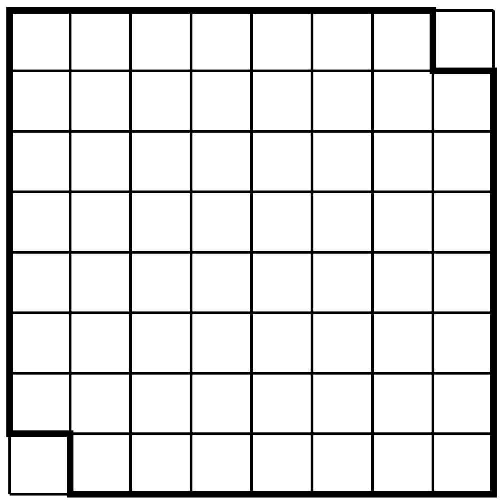
Exercise 242. *How many rooks can be placed on a* $10 \times 10 \times 10$ *3-dimensional chessboard so that no two can attack one another?*

10.2.2 Dominoes and checker/chessboards

In this section, consider a "domino" to be a 1×2 rectangle formed by two squares, each having the size of a square in a chess/checkerboard. (The pips ordinarily found on a domino are generally ignored in the following problems.)

Probably one of the most used domino problems in puzzle collections is in the following exercise, apparently first invented by the philosopher Max Black (1909–1988) in 1946 [106]. This puzzle is sometimes called the "mutilated chessboard puzzle", and has been made famous by Martin Gardner.

Exercise 243. *If two opposite corners of an* 8×8 *chessboard are removed, can the remaining shape be tiled with* 1×2 *dominoes?*



If so, show such a tiling; if not, prove why not.

Remark on Exercise 243: Ralph E. Gomory showed (see Honsberger's 1973 book *Mathematical Gems* [483]) that if any two squares of opposite colours are removed, the remaining board can be tiled with dominoes. For more on this domino problem, see an article [655] by Nathan S. Mendelsohn (1917–2006) (a former University of Manitoba professor and department head); this classic is also discussed in [684, pp. 126–9], for example, where it is called the "mutilated chessboard problem".

Exercise 244. *Show that for any tiling of a* 6×6 *chessboard with 18 dominoes, then one of the 10 lines forming an interior line of the board (5 horizontal, 5 vertical) can be sliced without cutting any dominoes.*

10.2.3 Trominoes and checkerboard/chessboards

One extension of the domino problem in Exercise 243 is to patterns called "trominoes". Trominoes are like dominoes, however are formed by three squares. There are two kinds of trominoes: three squares in a row, or three squares in the shape of an L. The following problem uses only the first kind:

Exercise 245. *If just one corner square of a chessboard is removed, can the remaining 63 squares be tiled with* 3×1 trominoes*?*

Note: the same trick used in the solution give for Exercise 245 is not available for the L-trominoes; does this mean that one can not use L-trominoes? In fact, L-trominoes can be used for this board and many more.

The following exercise is a strengthening of Exercise 245:

Exercise 246. *Using* 3×1 *trominoes, is it possible to tile the* 8×8 *chessboard with one square missing, and if so, in what positions can the uncovered tile be?*

Exercise 247. *Show that for any* m *and* n *that are multiples of 2 and 3 respectively, an* $m \times n$ *checkerboard (with no squares missing) can be covered with L-shaped trominoes.*

Exercise 248. *Show that for any* $n \geq 2$*, a* $6 \times n$ *checkerboard can be tiled with L-shaped trominoes.*

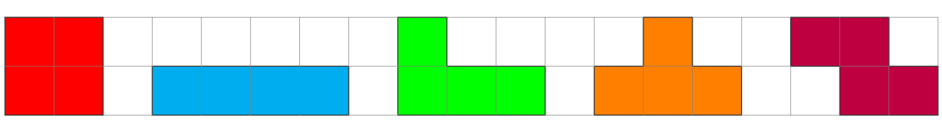
Exercise 249. *Show that for each* $n \in \mathbb{Z}^+$*, a* $2^n \times 2^n$ *chessboard with any one square removed can be covered with L-shaped trominoes.*

Exercise 250. *For any even number* $n \geq 14$ *that is not a multiple of 3, show that an* $n \times n$ *checkerboard with any one square removed can be covered by L-shaped trominoes.*

Exercise 251. *Show that for* $n > 5$*, every* $n \times n$ *checkerboard with a square missing can be covered with L-shaped trominoes if* n *is odd and 3 divides* $n^2 - 1$.

10.2.4 Tiling with tetrominoes

Tetrominoes were made famous by the game *Tetris*. There are five different tetrominoes:



Here is a problem involving only the fourth tetromino above.

Exercise 252. *Can a* 10×10 *chessboard be tiled with a T-shaped tetromino? For which* n *can an* $n \times n$ *board be done?*

10.2.5 Pentominoes

Pentominoes are flat shapes consisting of 5 squares, where squares are adjacent along entire sides. There are 12 different pentominoes. If one uses 5 cubes to construct each pentomino (some call these pentacubes), the total volume of the 12 pentominoes is 60. The number 60 can be factored in many ways, and for many of these factorizations there is a way to pack the pentominoes in the corresponding 3-dimensional rectangular parallelepiped (brick). For example, Figure 10.1 shows one of many ways how to fill the shape $6 \times 10 \times 1$.

Figure 10.1: Mahogany pentominoes in 6×10 rectangle

The following chart shows the bricks that can be formed by packing a set of 12 pentominoes (made from cubes) and the number of ways (again, ignoring reflections and rotations) each brick can be made.

dimensions	number of ways
$1 \times 3 \times 20$	2
$1 \times 4 \times 15$	368
$1 \times 5 \times 12$	1010
$1 \times 6 \times 10$	2339
$2 \times 3 \times 10$	12
$2 \times 5 \times 6$	264
$3 \times 4 \times 5$	3940

Computer searches were carried out for the numbers (in the above chart) of different packings with pentominoes by various authors. For example, the number of solutions for the $1 \times 6 \times 10$ case was found in 1960 [460]; in 1967, Bowkamp [141] finished the computer searches for all of the above packings and published his results in 1969 (in that paper, the twelve solutions to the $2 \times 3 \times 10$ case are given).

Here is kind of a "meta-question" inspired by the numbers above: if a problem has only a few solutions, does this make it harder to find one? It might be that if only a few solutions exist, each might be somehow "forced", making one easier to discover, or if after finding one, another is easy. On the other hand, if many solutions exist, a solution might be more easily found since a particular sequence in an attempt might be more likely to extend to a solution (and so a random start might have a better chance to be completed).

For pentominoes, the 3×20 solutions are rare, and so one might expect that either of the two solutions might arise from only a few "sub-patterns". Indeed, the two solutions for the 3×20 pentomino arrangement have a great deal in common, so finding one might help a great deal in finding the other.

Exercise 253. *Find the two* 3×20 *arrangements of the complete set of pentominoes.*

The interested reader may find many collections in popular literature (*e.g.*, [397] or [642]) for more problems on tiling with polyominoes.

10.3 Tiling rectangles with rectangles

10.3.1 Tiling with one rectangle

The old standard puzzle given in Exercise 243 asks to use 1×2 dominoes to tile a checkerboard (or chessboard) whose two black corners have been removed (it is impossible—each domino covers one square of each colour). Any beginning to an analysis of this sort of might first consider the problem of tiling of a rectangle with one copy of a smaller rectangle.

Much of what follows also has generalizations to higher dimensions (and rectangular parallelepipeds called "bricks"), but these topics are not stressed here. For an introduction to the higher dimensional versions, see [117].

A dissection of a large rectangle into smaller rectangles is also called a *tiling* of the larger one with smaller ones. Given a (small) rectangle R with integer dimensions $c \times d$, for which a and b can a (large) $a \times b$ rectangle be tiled with R?

Since small rectangles can be placed in a larger rectangle with two orientations, greater flexibility in arrangements is possible when using smaller lengths for tiles. Even with this flexibility, some surprises occur. For example, it might seem as if one could easily tile a 10×10 square with 1×4 rectangles—but in fact, it is impossible; a simple proof appears after the next theorem. In 1959, de Bruijn [164, Prob. 109] (in Hungarian) posed such a problem for $1 \times 2 \times 4$ bricks, with solutions given by de Bruijn, G. Hajós, G. Katona, and D. Szasz, and I. Thiry. Two years later, de Bruijn, gave a similar problem [165, Prob. 119] for the n-dimensional case (which was solved by de Bruijn, G. Hajós, G. Katona, and D. Szasz). In 1969, de Bruijn published these results in *The American Mathematical Monthly* [239].

The following theorem combines (the 2-dimensional versions of) two theorems from [239]:

Theorem 10.3.1 (de Bruijn, 1959, see [239])**.** *If a, b, c, d are integers so that an $a \times b$ rectangle is tiled with $c \times d$ rectangles, then both c divides one of a or b and d divides one of a or b.*

Since 4 does not divide 10, Theorem 10.3.1 shows that a tiling of a 10×10 square with 1×4 rectangles is impossible. Similarly, a 6×6 square can not be tiled with 1×4 tiles.

Theorem 10.3.2 (de Bruijn, 1959–61 see [239])**.** *If a rectangle R is tiled by rectangles, each of which has at least one integer side, then R has at least one integer side.*

Note that Theorem 10.3.2 generalizes Theorem 10.3.1 by dividing the width by c and/or the height by d. De Bruijn's proof used double counting and complex numbers.

Stan Wagon [909] collected fourteen proofs of Theorem 10.3.2, including other proofs by Rochberg, Ruzsa, Douady, Paterson, Robinson, Bishop, Seymour, Bachman & Yannakakis, Hochster, and Schmerl. Different proof techniques were used, including graph theory, induction, real and complex integrals, and checkerboards. The proof by Paul Seymour uses graphs, a minimal cut-set, and a planar dual.

10.3.2 Dissecting rectangles into squares and perfect squaring

A *squaring* of a rectangle R is a dissection of R into pieces, each of which is a square (naturally, non-trivial squarings use squares of different sizes).

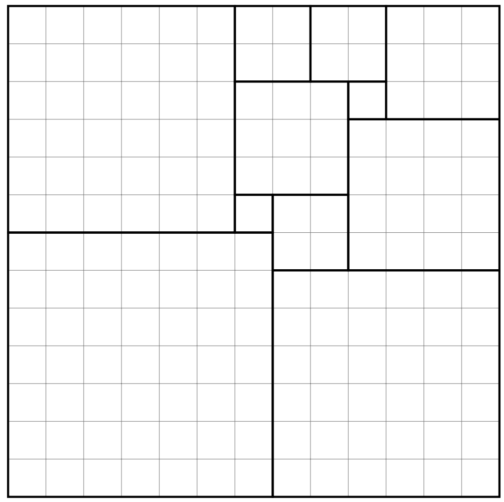
Theorem 10.3.3 (Dehn, 1903 [245])**.** *If a rectangle has a squaring, then the ratio of two neighbouring sides of the rectangle is rational, and the rectangle has infinitely many squarings.*

Note that, by Theorem 10.3.3, if a rectangle R has a squaring, then multiplication by a suitable integer produces another rectangle with integer length sides that also has a squaring. Similarly, one can assume that all squares used in a squaring of a rectangle also have integer lengths.

A dissection of a rectangle R into squares is called a *perfect squaring* of R if and only if all squares have different sizes. A rectangle R is called *perfect* if and only if R has a perfect squaring. A squared rectangle that contains a smaller squared rectangle is called *compound* and is called *simple* otherwise. The number of squares in a squaring is called the *order* of the squaring.

It seems that the earliest reference to finding different squares in a rectangle is the problem of "Lady Isabel's Casket", given by Dudeney in *The London Magazine*, 7 Jan. 1902 (and later appearing in *Canterbury Puzzles* [270, No. 40] in 1907). A 20 by 20 inch square was given and a strip of gold measuring $\frac{1}{4} \times 20$. The question was to tile the square using the strip and different squares. The answer was to use squares of sizes .25, 1, 2, 2.5, 2.75, 3, 5, 5.25, 7, 7.75, 8, and 12.

Another famous puzzle was called "The Patch Quilt" puzzle, later also called "Mrs. Perkins's quilt". Cut a 13×13 quilt (along grid lines) into smaller squares using as few squares as possible. One solution is possible with 11 squares (with sizes 1, 1, 2, 2, 2, 3, 3, 4, 6, 6, 7), and it is believed that the solution below is (up to various reflections) unique.



Using present vernacular, the solution above is neither simple nor perfect. The history of the Mrs. Perkins's quilt puzzle seems mixed, but apparently Sam Loyd advertised it some time between 1907 and 1914 (see [26] for details). Martin Gardner (see [272, Prob. 343]) says that Loyd published this problem in 1907 in the first issue of *Our Puzzle Magazine*, and mentions the possibility that Loyd got it from Dudeney, perhaps from one of his early magazine or newspaper articles. This puzzle was later given by Dudeney in [271, Prob. 173]. (Later Dudeney called a different problem "The patchwork quilt" [272, Prob. 340], where pieces are not necessarily square.) This problem is also discussed by J. H. Conway [208], G. B. Trustum [879], and given in Martin Gardner's column in *Scientific American* in September 1966 [369]. In general, it does not seem to be known for an $n \times n$ square how few smaller squares are required.

In 1925, Moroń [669] gave a perfect simple squaring of a 32×33 rectangle with 9 squares (see Figure 10.2, the first image in Figure 10.5, and Figure 10.6). Moroń also found a perfect squaring of order 10 for a 65×47 rectangle. In 1930, Kraitchik [573, p. 272] gave Moroń's order 9 squaring, and mentioned a conjecture by Lusin that no square can be divided into finitely many different "elements". In 1939, Rouse Ball mentioned Moroń's squaring of order 9 in his book *Mathematical recreations and essays* (see [770, p.93]) but did not give a picture. Moroń's order 9 dissection also appeared in Hugo Steinhaus' *Mathematical snapshots*, first appearing in 1950 (see [826, Prob. 11. p. 8] for a more recent edition).

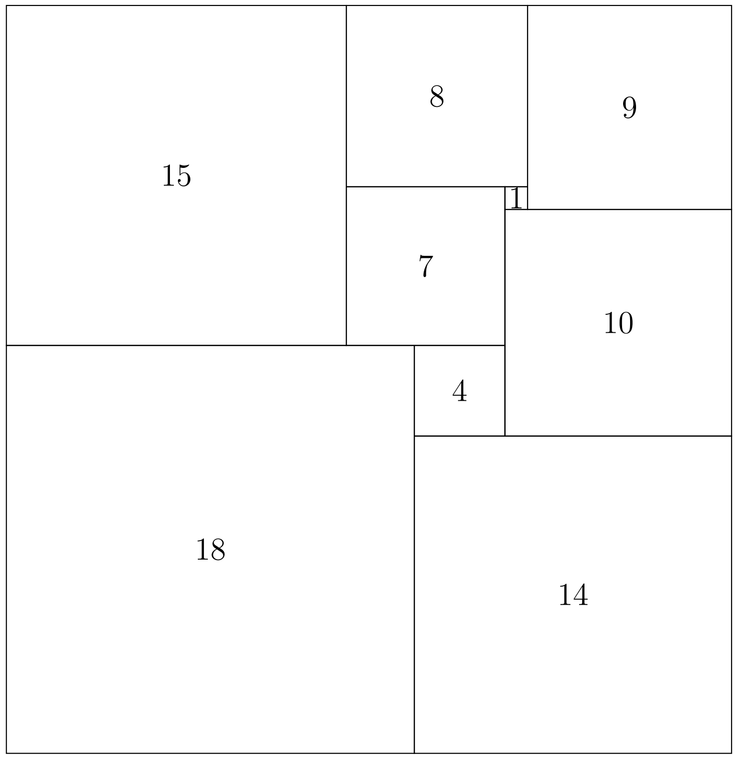


Figure 10.2: Moroń's order 9 perfect squaring of a 32×33 rectangle

Since perfect rectangles exist, it might be natural to ask a number of questions. For example, are there perfect squarings with order less than 9? Can the perfect rectangles be enumerated? Are there square perfect squarings? Does a rectangle exist that has two different squarings (where "different" might only mean a different arrangement of the same squares, or perhaps a different set of squares)?

In 1939, Sprague [817] was the first to give a perfect squaring of a square; his dissection used 55 different squares.

In the mid 1930s, these questions also caught the attention of Arthur Harold Stone (1916–2000), then a student at Trinity College, Cambridge. At first, Stone could not prove the assertion that perfect squarings of a square did not exist, but instead found a 176×177 rectangle cut into 11 unequal squares (with sizes 99, 78, 77, 57, 43, 34, 25, 21,16, 9). He began work with three other students, Rowan Leonard Brooks (1916–1993), Cedric Austin

Bardell Smith (1917–2002), and William Thomas Tutte (1917–2002). The four were later referred to as "the Trinity Four". (Brooks and Tutte became graph theorists; however, Smith became a statistician, and Stone became a well-known topologist.)

Smith had an idea to translate the problem to one of electrical flows, representing a circuit by a special 3-connected planar digraph, now called a *Smith diagram*. This idea was quite successful. (For their idea using Smith diagrams but explained in modern terms, see Bollobás' *Modern graph theory* [117], where many examples and exercises are also given.)

In 1940, Brooks, Smith, Stone, and Tutte [151] published their breakthrough paper. They first gave a different proof of Dehn's theorem (Theorem 10.3.3), and showed that every rectangle whose sides have a rational ratio (*i.e.*, have "commensurable" sides) have infinitely many squarings. This last result was also published by Sprague earlier the same year (Sprague's was noted in [151]).

Theorem 10.3.4 (Sprague, 1940 [818], Brooks *et al*, 1940 [151])**.** *A rectangle has a perfect squaring if and only if the ratio of two neighbouring sides is rational. Such rectangles have infinitely many perfect squarings.*

Brooks et al. [151] gave a few explicit examples of squared rectangles, and a compound perfect 608×608 square of order 26 (see Figure 10.3). Brooks et al. also found a perfect rectangle of order 9, different from Moroń's, with dimensions 61×69 (see Figure 10.4).

Brooks et al. [151] proved these were the only two perfect rectangles of order 9 (see Figure 10.5 for models of both I made in 2016). They proved that 9 is the minimum order of any perfect rectangle. The minimality of 9 was also shown by Reichard and Toepken [747] in the same year. (It might be interesting to note that these two results now appear as exercises in a recent textbook [117, Exercise 22, p. 61].)

In his 1942 book, Kraitchik [574] gave just one example of a (compound) perfect squaring of order 26 dissecting a 608×608 square using a slight rearrangement of the same squares given in Figure 10.3, the Brooks–Smith–Stone–Tutte [151, Fig. 9, p. 333] construction.

In 1950, Brooks independently found a simple perfect square of order 38, side length 4920 (see [885]).

In Figure 10.5 are pictures of such models I made in 2016: the two are the two different rectangles, roughly to scale, made from birch plywood and numbers burned in; in Figure 10.6 is a puzzle form of Moroń's dissection, but with finer woods (the woods are, in order of size, ebony, cocobolo(?), bubinga, chilean honey tree, pau amarello, padauk, zebrawood, ambrosia maple).

Duijvestijn [275] (using a DEC-10 computer) showed that the example in Figure 10.7 of order 21 uses the least number of squares that perfectly square a square, and that this squaring is unique (it is a 112×112 square).

Duijvestijn's proof used Smith diagrams, as well. Figure 10.7 is now the basis of the logo of the Trinity Mathematics Society (see [878]).

A perfect squaring of a slightly smaller square was found by the Bristol mathematician T. H. Willcocks but it uses one more square; it is a 110×110 square using 22 squares with side lengths

$$1, 2, 3, 4, 6, 8, 9, 12, 14, 16, 17, 18, 19, 21, 22, 23, 24, 26, 27, 28, 50, 60.$$

A compound squaring with 24 squares was found in 1948 [926] (see also [927]). In 1982, 24 was shown [276] to be the lowest order for a compound squaring, and Willcocks' example was shown to be unique. (In 1959, Willcocks also had the best known order for a perfect square, 37, with side length 1947.) There are at least eight perfect squarings of order 25 (see [275]).

The first copy I made of the 21 square model was given to Béla Bollobás in 2014. Since I only took a close-to-vertical shot of Béla's model before it was finished, the photo at least shows what the bare woods looked like:

It is made from (in order of smallest to largest) Macassar Ebony, Prima vera, Black Walnut, Holly, Osage Orange, kingwood, cocobolo, Shedua, Beefwood, Gabon Ebony, Tulipwood, Bloodwood, Pau Amerello, Lignum Vitae, Teak, American Mahogany, Ziricote, African Padauk, Tigerwood, and Zebrawood. Almost all wood samples above are 3/4 inch thick—the lignum vitae is slightly thinner. The frame is grey elm, *ulmus americana* (NE. North America); finish is clear lacquer; substrate is MDF.

I later made two more models (Figure 10.9) of the 21 squared square; woods are slightly different than the first.

All squares are 3/4 inch thick. The first frame is black walnut with an insert of grey elm; finish is clear lacquer; substrate is MDF. The second frame is lacewood, grey elm inner frame, clear lacquer, maple plywood base.

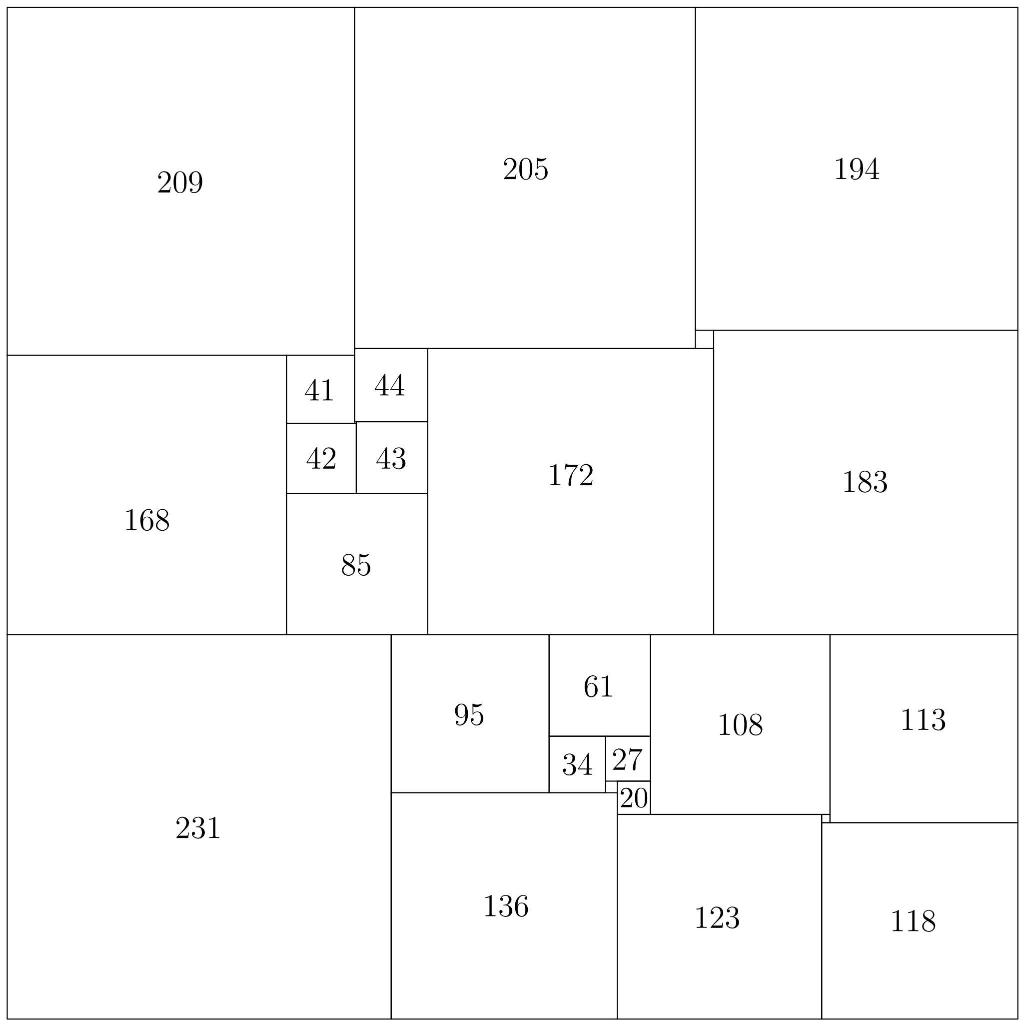


Figure 10.3: A compound perfect 608×608 square of order 26 by Brooks, Smith, Stone, and Tutte (3 squares unlabelled, including a unit square at the corners of the 41, 42, 43, and 44 squares, the other two of sizes 5 and 11)

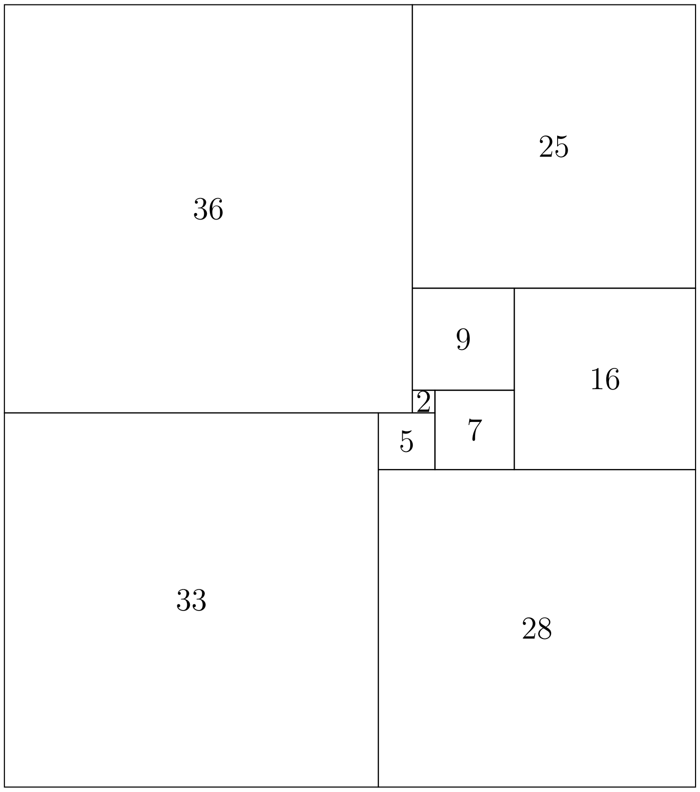


Figure 10.4: Order 9 perfect squaring by Brooks, Smith, Stone, and Tutte

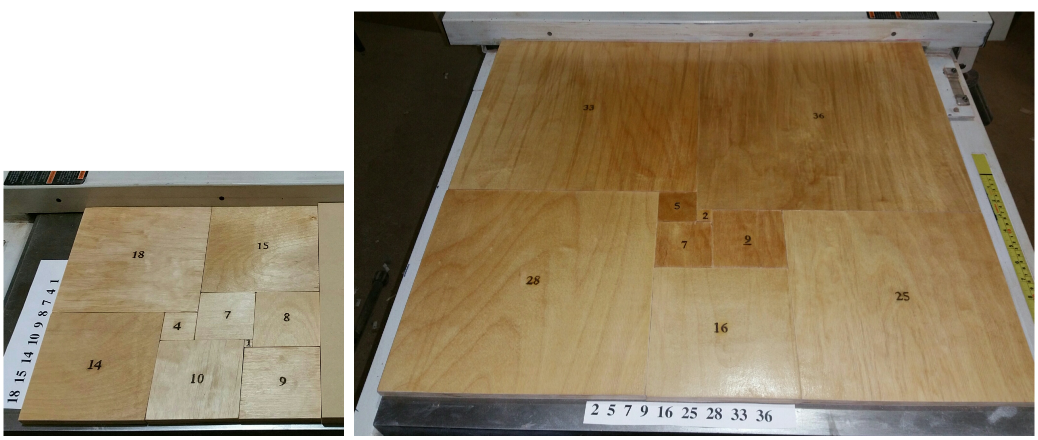


Figure 10.5: The two perfect squarings of (minimal) order 9 as puzzles.

Figure 10.6: Moroń's perfect squaring of order 9, a puzzle with exotic woods

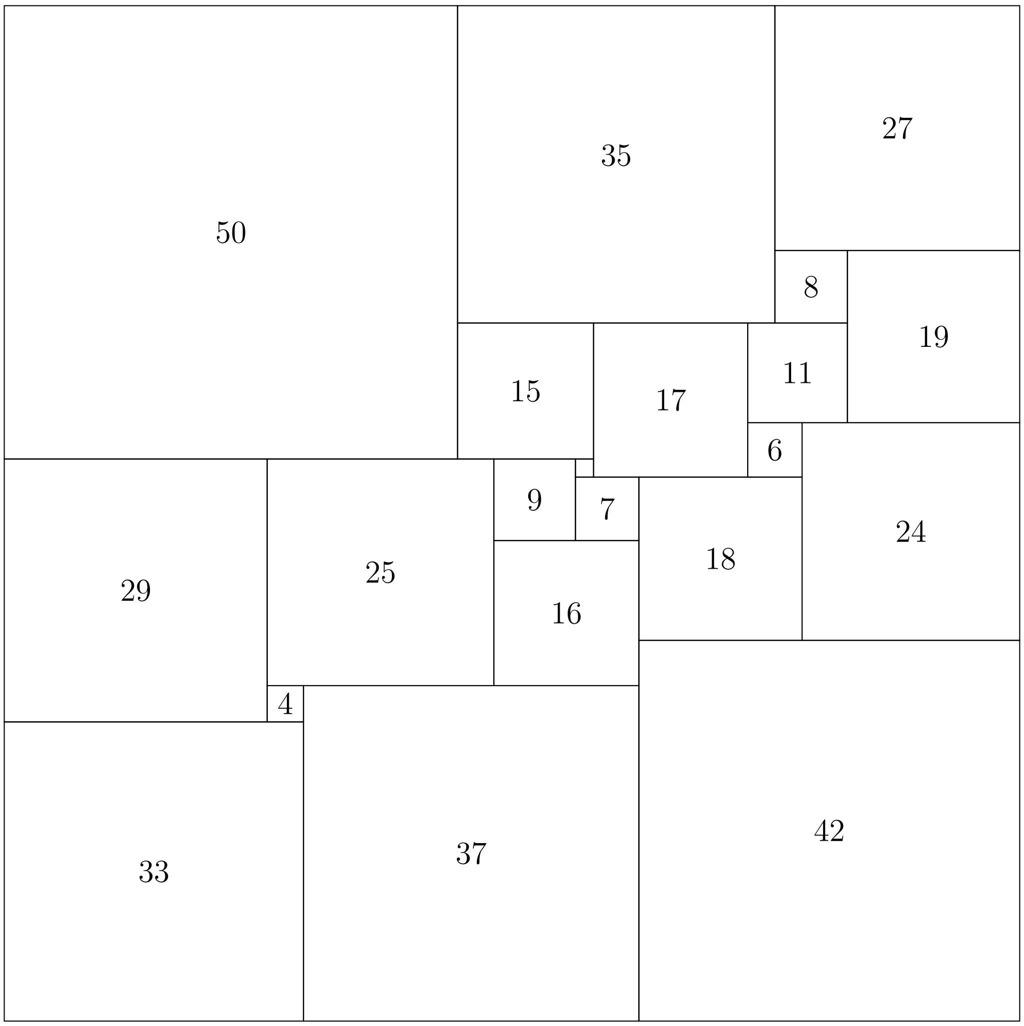


Figure 10.7: Duijvestijn's perfect squaring of 112×112 with 21 unequal squares, where the unlabelled square is 2×2

Figure 10.8: First model of Duijvestijn's squared 21 squares, before finishing

Figure 10.9: Two other models of Duijvestijn's squared 21 squares

10.3.3 Some references for squaring the square

See Tutte's article [886, pp. 186–209] for thoughts and progress made by the Cambridge four in the late 1930s. Tutte's article is also reprinted in Martin Gardner's *The 2nd Scientific American book of Mathematical Puzzles and Diversions* [368]. Tutte also wrote another article in 1965 [887] on finding perfect squares. Also see [885]. In Tutte's article is also a detailed description of how Smith diagrams work, including one for the order 13 squared 112×75 rectangle in Figure 10.10, one which Brooks' mother discovered had two essentially different arrangements, a key discovery for their method to work! (Can you find another arrangement?)

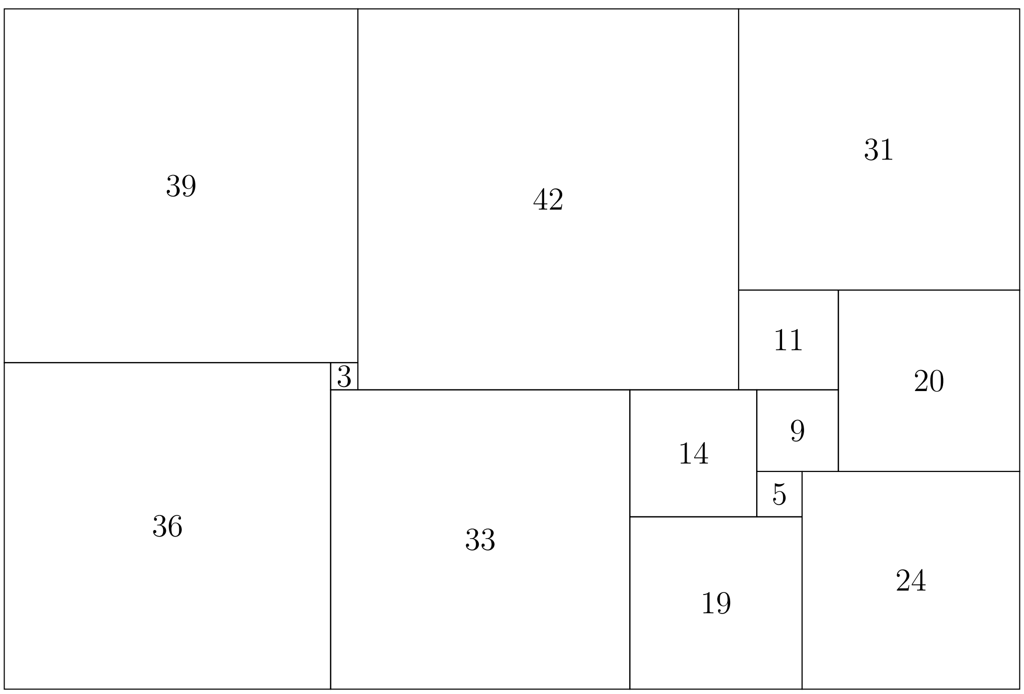


Figure 10.10: Mrs. Brooks' arrangement of a squared 112×75 rectangle

The website www.squaring.net by Stuart Anderson [26] contains a spectacular list of facts, references, links, and diagrams related to squaring of rectangle and squares. See also [799] for a survey on squaring the square, or [534] for some survey and results obtained by algorithms.

10.3.4 Dissecting a square into rectangles with proportion 1:3

Is there a square (with integer side lengths) that can be dissected into different sized rectangles, each with integer side lengths and a length to width ratio of 3 to 1? If so, what is the smallest number of such rectangles that work?

Such a dissection of a 96×96 square into 12 rectangles was given by Dick Hess [466, No. 15]. The rectangles had short sides 1, 3, 6, 7, 8, 9, 13, 15, 18, 19, 27, and 32. It does not seem to be known if 12 is minimal.

10.4 Dissecting an equilateral triangle into equiliateral triangles

In 1948, Tutte [884] similarly studied dissecting an equilateral triangle into equilateral triangles; Tutte found an example of order 15, later shown to be of minimum order (but it is not perfect; see [26] for details). Is there a perfect (equilateral) triangulation? Hmmm...

In a somewhat related recent development, Kupavskii, Pach, and Tardos (see [579]) proved that the plane cannot be tiled with pairwise noncongruent triangles of the same area and the same perimeter.

10.5 Higher dimensional versions of squaring the square

Cubes do not have a perfect "cubing":

Exercise 254 (Cubing the cube)**.** *Show that a cube cannot be dissected into finitely many differently sized cubes.*

The plot thickens. Note that for any $d \geq 4$, no d-hypercube can have a perfect cubing since a d-cube will have faces of dimension $d-1 \geq 3$, and so a simple induction down to 3 dimensions finishes the proof.

So, ..., squares are perfect, equilateral triangles might be, but cubes and hypercubes are certainly not!

Chapter 11

Finite geometries

11.1 Introduction

This chapter is only a brief introduction to finite geometries. Many books have been written on finite geometries; one notable book is by Dembowski [250], a very detailed study of various possible finite geometries (although due to notation used, it can be hard to read in places). A modern reference is the article by Peter Cameron [170]. Another article on extremal problems in finite geometries by Blokhuis [109] might be interesting. For finite geometries over finite fields, the volumes by Hirschfeld [477] and Hirschfeld and Thas [478] are standard references.

Finite geometries can be considered from different points of view. One perspective is that of having two classes of objects, points and lines, and an incidence relation between them. Another view is that a finite geometry consists of a set of points and subsets of points called lines. Both approaches are considered here.

Even though "geometry" is, as of yet, undefined, there are (at least) two ways to define a geometry. One is to begin with undefined terms, like "point" and "line" and then give some axioms (like "every line contains at least three points"). If one can create a "model" that satisfies all of the axioms, this model then is given as a witness to the existence of the particular geometry in mind. (There might even be non-isomorphic models that satisfy the same set of axioms on the same number of points.) Another way to define a geometry is to simply list the points, lines, and the relations among them; deriving properties of the resulting geometry is then often fairly easy.

Many finite geometries need not have a physical model. In many models for finite geometries, there is no concept of "distance", "in-betweenness", or "angle"—it is often convenient to consider a finite geometry as nothing more than a set system on a universe with finitely many elements (points). Such a structure is often referred to as a *hypergraph*.

11.2 Axioms and expressions

In many axiomatic systems, some terms are accepted without precise definition; "point" is such a term. In standard Euclidean geometry, a point is described only by its position, as is a line. In a finite geometry given by a set of axioms, both the expressions "point" and "line" are accepted without further description, even though both evoke a mental picture.

Another expression that appears in axioms for geometries is "is incident with". This expression usually gives a relationship between a point and a line, and is symmetric. Such an incidence relation is often denoted by "I". If a point P can be said to be incident with a line ℓ (in which case, the notations $PI\ell$ or $(P, \ell) \in I$ are sometimes used), then also ℓ is incident with P. Consistent with terminology for geometries where lines (and planes and ...) are considered as subsets of points, a point being incident with a line is often described by saying "the point is on the line" or "the line contains a point".

For two lines ℓ and m, if there is no point incident with both, then the lines ℓ and m are said to be *parallel*. If two lines are incident with a common point, the lines are said to *intersect*. If two or more lines are all incident with just one point, these lines are said to be *concurrent* (at that point). If two or more points are all incident with one line, these points are said to be *collinear*.

The *dual* of a geometry given by axioms is a geometry satisfying the axioms with the words "point" and "line" interchanged.

11.3 A definition of a finite geometry

A *space* or *configuration* $S = (X, \mathcal{L})$ is a collection X of points and a collection $\mathcal{L}$ of lines with an incidence relation I (or a set X of points together with a collection $\mathcal{L}$ of lines, where each line is a subset of X).

Some configurations are trivial, and are not intended to be called geometries. For example, is there a configuration defined by the following two axioms?

A1: each line is incident with 2 points.

A2: each line is incident with 3 points.

A trivial configuration with no lines indeed satisfies these two axioms. To eliminate this and other configurations for consideration as a geometry, some properties of a configuration are needed.

Definition 11.3.1. A configuration of points and lines is a *finite geometry* if and only if:

- There are finitely many points and lines.
- Each line is incident with the same number of points.
- Each point is incident with the same number of lines.
- For any pair of points, there exists at most one line incident with both. If two points A and B are incident with a line ℓ, then write $A \vee B = \ell$ (read "A join B").
- Two distinct lines are incident with at most one common point. If lines ℓ and m are incident with the point P, then write $\ell \wedge m = P$ (read "ℓ meet m").
- Not all points are incident with one line.
- There exists at least one line.

A configuration is called a *near linear space* if and only if every line contains (or is incident with) at least two points and any two points are on at most one line. If in addition, for any two points there exists exactly one line incident with both, the space is called a *linear space*.

A *projective geometry* is a geometry whose every pair of lines intersect. (See [80] for an extensive collection of facts regarding projective geometries; only a few passing comments are made here.) Axioms for a projective geometry vary. Here is one set:

PG1 There exist at least two lines.

PG2 Each line has at least three points.

PG3 Through any two points there exists a unique line.

PG4 Any two lines intersect.

Veblen and Young [901] used a different set of axioms for a projective geometry, where PG4 replaced by:

PG4a: Let A, B, C, and D be points such that the line AB intersects the line CD; then the line AC intersects the line BD.

Another formulation of this axiom is:

PG4b: If a line intersects two sides of a triangle, then the line also intersects the third .

The import of the fourth axiom is that parallel lines are forbidden. (See [80, 1.2] for discussion of these axioms, including a form due to Pasch.) A classical example of a projective geometry is formed by taking subspaces of a vector space. If $\mathbb{F}$ is field, and $d \geq 2$ is an integer, the d-dimensional projective geometry $\mathrm{PG}(d, \mathbb{F})$ is defined to be the set of all proper subspaces of the vector space $V = \mathbb{F}^{d+1}$; the points are the 1-dimensional subspaces (lines through the origin) of V, the lines are the 2-dimensional subspaces (planes), and so on. When $d = 2$, $\mathrm{PG}(2, \mathbb{F})$ is called a projective plane over $\mathbb{F}$. When $\mathbb{F}$ is finite, these are studied in Section 11.4.2. There is a more general definition for "projective plane" (see Definition 11.4.1) which includes planes over fields.

One notable fact is that for projective geometries, a d-dimensional projective geometry over a field always satisfies Desargues' theorem; there are other "projective planes" that need not.

An *affine geometry* is a geometry that satisfies Euclid's parallel postulate:

Euclid's parallel postulate: If ℓ is a line and P is a point not on ℓ, then there exists a line m containing P that is parallel to ℓ.

Affine planes can be derived from projective planes by deleting a line (see Section 11.5).

Exercise 255. *Let* $V = \{0, 1, 2, \ldots, 6\}$*, and let*

$$\mathcal{H} = \{(i, i+1, i+3) : i \in \{0, 1, \ldots, 6\}\},$$

where addition is modulo 7. Show that the set system $(V, \mathcal{H})$ *is a projective geometry.*

Some configurations that do not satisfy all the properties above are also sometimes called geometries in the literature; some may be called "partial geometries".

A *finite hyperbolic geometry* (see [51, p. 271]) is a configuration satisfying

- Any two points are incident with a unique line.
- Every line is incident with at least two points.
- There exist at least two lines.
- For any line ℓ and P not incident with ℓ, there exist at least two lines incident with P that do not intersect ℓ.

Compare the following example to Exercise 255.

Example 11.3.2. *A finite hyperbolic geometry on 13 points is given by points* $0, 1, \ldots, 12$*, and each line is either of the form* $(i, i+1, i+4)$ *or* $(i, i+2, i+7)$*, where* $i = 0, 1, \ldots, 12$ *and addition is modulo 13. So there are 26 lines.*

Exercise 256. *Show that the smallest hyperbolic geometry contains 5 points.*

A *weak projective plane* satisfies

- Any two points are incident with a unique line.
- There exist at least two lines.
- Every two lines intersect.

A *weak affine plane* satisfies

- Any two points are incident with a unique line.
- Every line is incident with at least two points.

- There exist at least two lines.
- For any line ℓ and P not incident with ℓ, there exists a unique line incident with P that does not intersect ℓ.

Does there exist a weak affine plane with 36 points? It can be shown (using a general theorem about mutually orthogonal latin squares and finite projective/affine planes—see, e.g., [898, Ch. 22]) that if a weak affine plane on 36 points exists, there is a pair of orthogonal latin squares of of order 6. In 1782, Euler [324] postulated the non-existence of such a pair (in his famous 36 officers problem, which asks for an arrangement of six ranks from six regiments in a 6×6 array so that in each row and each column, to ranks or regiments are repeated). In 1900 and 1901 [862, 863], Tarry proved by an exhaustive search that no such pair of latin squares exists.

11.4 Finite projective planes

For an introduction to general projective planes, see [451].

11.4.1 General definition

Definition 11.4.1. A projective plane $\mathcal{P}$ is collection of points, lines, and incidences between them that satisfy the following three axioms.

(PP1) For every pair of distinct points, there exists a unique line incident with both.

(PP2) For every pair of distinct lines, there exists a unique point incident with both.

(PP3) There exist four points, no three of which are incident with the same line.

Definition 11.4.2. A collection of four points satisfying (PP3) is called a *frame*, or *complete quadrangle*.

A frame or complete quadrangle can also be called a complete 4-point configuration (see Definition 15.4.7). See Figure 11.1 for two pictorial representations of a complete quadrangle; note that unmarked line crossings may indicate points, but such points are not part of the frame. In a projective plane, every pair of lines intersect, so the three opposite pairs of lines in a frame intersect; these three additional points are called *diagonal points of the frame.*

Exercise 257. *Show that for any frame in the Fano plane, the three diagonal points are collinear.*

For which geometries do the diagonal points of a frame lie on a line? If the first representation of a frame given in Figure 11.1 is considered in the real affine plane, then the diagonal points do not lie on a line. (Fano studied such geometries, and what is now known as Fano's axiom says that diagonal points are never collinear. Somehow, definitions have evolved so that now a Fano plane *fails* Fano's axiom.)

Definition 11.4.3. A complete quadrilateral is the dual of a complete quadrangle.

By definition, a complete quadrilateral is a collection of six points and four lines, no three lines concurrent; see Figure 11.1 for one pictorial representation.

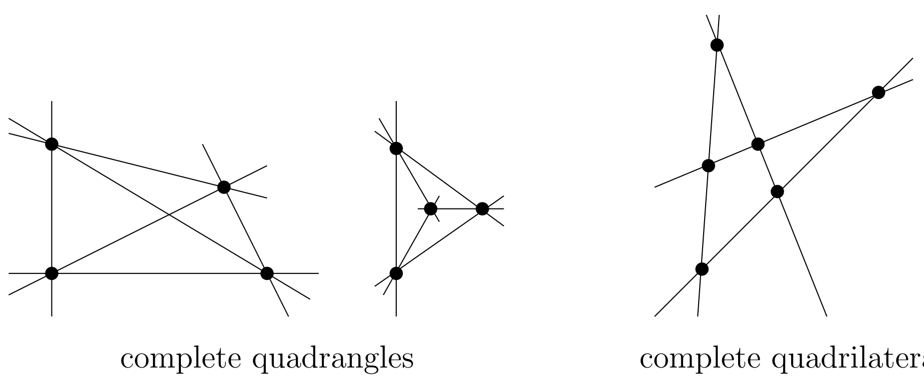


Figure 11.1: Two complete quadrangles (frames) and a complete quadrilateral

Lemma 11.4.4. *Any projective plane contains a complete quadrilateral.*

Proof: Let Π be a projective plane, and let A, B, C, D be points in a complete quadrangle. Opposite pairs of lines intersect in points (called diagonal points), say $E = (A \vee B) \wedge (C \vee D)$, $F = (B \vee C) \wedge (D \vee A)$, and

$G = (A \vee C) \wedge (D \vee B)$. Then A, B, C, D and any two of F, G, H together form a complete quadrilateral; for example, using just A, B, C, D, E, F, the four lines $\overleftrightarrow{ABE}$, $\overleftrightarrow{CDE}$, $\overleftrightarrow{BCF}$ and $\overleftrightarrow{DAF}$ form a complete quadrilateral. $\square$.

Theorem 11.4.5. *The dual of a projective plane is a finite projective plane.*

Proof: Due to the symmetry of axioms PP1 and PP2, the dual of a projective plane also satisfies PP1 and PP2. Also, by Lemma 11.4.4, since a projective plane contains a complete quadrilateral, the dual of a projective plane contains the dual of a complete quadrilateral, which is a complete quadrangle, and so the dual of a projective plane also satisfies PP3. $\square$

Lemma 11.4.6. *For any two lines in a projective plane, there is a point not incident with either line.*

Proof: Let ℓ and m be lines, and suppose that the lemma is false, that is, suppose that every point is on these two lines. By (PP3), let A, B, C, and D be a frame. By assumption, each of A, B, C, and D lie on either ℓ or m; since no three are collinear, two lie on each, say A and B on ℓ, and C and D on m. Let $X = (A \vee C) \wedge (B \vee D)$. If X is on $\ell = A \vee B$, then A, B, and C are collinear, a contradiction. A similar contradiction is obtained if X is on m. $\square$

Lemma 11.4.7. *There is a one-to-one correspondence between points on two lines in a projective plane.*

Proof: Let ℓ and m be lines in a projective plane. By Lemma 11.4.6, let P be a point not on lines ℓ and m. Define $\alpha : \ell \to m$ by $\alpha(X) = m \wedge (P \vee X)$. It remains to observe that α is indeed one-to-one. $\square$

Lemma 11.4.8. *For any line ℓ and point P in a projective plane $\mathcal{P}$, there is a one-to-one correspondence between the points on ℓ and the lines containing P.*

Proof: The proof is divided into two cases, depending on whether or not P lies on ℓ.

Case 1. Let $P \notin \ell$. For each $Q_i \in l$, by (PP1) there is a line $\ell_i = P \vee Q_i$ containing P. For each such line ℓ_i containing P, by (PP2) there is a point $R_i = \ell_i \wedge \ell$ on ℓ.

Case 2. Let $P \in \ell$. Since by (PP3) $\mathcal{P}$ contains a frame, there are points Q and R not on ℓ so that $(Q \vee R) \wedge \ell \neq P$. Let $m = Q \vee R$. By Case 1, there is a one-to-one correspondence between points on m and lines containing P. By Lemma 11.4.7 there is a one-to-one correspondence between points on ℓ and points on m and hence between points on ℓ and lines containing P. $\square$

The next theorem shows that if just one line in a projective plane has finitely many points, then every line does and the number of points in the plane is indeed finite.

Theorem 11.4.9. *If $\mathcal{P}$ is a projective plane with (at least) one line finite, then there exists a positive integer n so that each line is incident with exactly $n+1$ points, and each point is incident with exactly $n+1$ lines. Furthermore, $\mathcal{P}$ has n^2+n+1 points and n^2+n+1 lines.*

Proof: Let n be so that some line ℓ contains $n+1$ points and let Q be one of these points. By Lemma 11.4.8, there are $n+1$ lines containing Q, including l and say, n remaining lines $\ell_1, \ell_2, \ldots, \ell_n$. Since any two points (in this case, one of which is Q) lie on a unique line, every point in the plane is on one of these lines. By Lemma 11.4.7, each of the lines have $n+1$ points. By Lemma 11.4.8, each point has $n+1$ lines incident with it. So fixing a point, say Q, there are $n+1$ lines containing Q, each having n points other than Q, so in all, there are $(n+1)n+1$ points. The dual argument gives the result for the number of lines. $\square$

Definition 11.4.10. Any projective plane with $n+1$ points incident with each line (and hence $n+1$ lines through each point) is said to be a *finite projective plane (FPP) of order n.*

In the next section, for any prime power q, a FPP of order q is constructed. It is not known if any FPP exists with non-prime-power order. Some non-prime-power cases are eliminated by the following theorem, which appeared in the first volume of the *Canadian Journal of Mathematics*.

Theorem 11.4.11 (Bruck–Ryser, 1949 [158])**.** *If $n \equiv 1 \pmod 4$ or $n \equiv 2 \pmod 4$ and a projective plane exists of order n then n is the sum of (integer) squares.*

The proof of the Bruck–Ryser theorem is given in Section 12.3. The first values of n that the Bruck–Ryser theorem eliminates are $n = 6$ and $n = 14$.

Since 10 is indeed the sum of two squares, this case is not eliminated by the Bruck–Ryser theorem; however the case $n = 10$ was shown not to exist by Clement Lam *et al.* The proof was a computer search; first announcements of this result contained holes, with the final version published in 1991 (see [585], [589]).

For orders less than 20, FPPs of orders 2, 3, 4, 5, 7, 8, 9, 11, 13, 17, and 19 exist, and FPPs of order 6, 10, and 14 do not exist. The first few open cases are $n = 12, 15, 18$. The case $n = 6$ is eliminated by the Bruck–Ryser theorem, but this case was already proved not to exist by Tarry [863] in 1901.

It has been shown that for each of $n = 3, 4, 5, 7$, the finite projective projective plane of order n is unique. Marshall Hall settled the case $n = 7$ (see [452, 453]); this result was again shown by Kocay [558] using a different approach—looking at quadrangles that are not in some Fano subconfiguration. (By Theorem 11.4.20, if a finite projective plane over some field contains a Fano configuration, then the field has characteristic 2; and so the field plane of order 7 contains no Fano configurations at all.)

It is conjectured that for each prime p, there is only one FPP, namely the Desarguesian plane over the field $\mathbb{F}_p$ (this plane is denoted by $\mathrm{PG}(2, \mathbb{F}_p)$—see Section 11.4.2).

There are four FPPs of order 9: the Desarguesian plane (over the field of 9 elements), the "left nearfield plane" (or Hall plane), its dual (the right nearfield plane), and the (self-dual) Hughes plane of order 9 (for descriptions of each, complete with their automorphism groups, see, e.g., [828]). In 1964, Parker and Killgrove [712] analyzed planes of order 9 in terms of latin squares (they found 5 MOLS for each). See the same paper for a brief review of attempts to classify order 9 planes by groups. In [587] it was shown (by an exhaustive computer search) that this list of four planes of order 9 is complete. All three non-Desarguesian planes of order 9 were first discovered by Veblen and Wedderburn in 1907 [900] using near-fields.

In 1943, Hall generalized the near-field planes. In 1957 Hughes [489] confirmed that one of the Veblen–Wedderburn examples (the self-dual plane) is non-Desarguesian, and generalized the example for each odd prime p, and each $n \in \mathbb{Z}^+$, to a non-Desarguesian plane of order p^{2n}.

11.4.2 FPP over a finite field

It is well-known that if a field is finite, the number of elements in that field is a power of a prime, and for every such prime power q, there exists a unique field of order q, usually denoted by GF(q) (GF is short for "Galois field"), or simply by $\mathbb{F}_q$.

Definition 11.4.12. Let q be a power of a prime, and let $V = (\mathbb{F}_q)^3$ be the 3-dimensional vector space over the field $\mathbb{F}_q$. Define $\mathcal{P}(q) = \mathrm{PG}(2, q) = \mathrm{PG}(2, \mathbb{F}_q)$ to be the set system consisting of points and lines, where the points are the 1-dimensional (linear) subspaces of V and the lines are the 2-dimensional (linear) subspaces of V.

Below, it is shown that $\mathcal{P}(q)$ as defined in Definition 11.4.12 is indeed a geometry, namely, a finite projective plane, sometimes called the "classical finite projective plane of order q", also denoted by $\mathrm{PG}(2, q)$ or $\mathrm{PG}(2, \mathbb{F}_q)$. For $d \geq 2$, a higher dimensional analogue is also available, the *projective geometry* $PG(d, \mathbb{F}_q)$, where there are d kinds of substructure (points, lines, planes, ...) given by linear subspaces of the vector space $(\mathbb{F}_q)^{d+1}$, but these are not developed in this section. There are FPPs that are different from $\mathrm{PG}(2, q) = \mathcal{P}(q)$ (for example, of order 9, there are three more, one called the Hughes plane) but for higher dimensions, it is known that there is only $\mathrm{PG}(d, q)$.

Lemma 11.4.13. *Let q be a power of a prime. Then, as in Definition 11.4.12 $\mathcal{P}(q) = PG(2, \mathbb{F}_q)$ is a finite projective plane.*

Proof: Let $V = [GF(q)]^3$, and let $\mathrm{PG}(2, q)$ be as in Definition 11.4.12. To show that $\mathrm{PG}(2, q)$ is indeed a FPP, it suffices to check the three axioms of Definition 11.4.1.

The statement "any 2 points in $\mathrm{PG}(2, q)$ are contained in exactly one line" is equivalent to "any two lines through the origin in V are contained in precisely one plane", which is seen to be true by looking at the plane whose normal is a cross product of the direction vectors for the lines (or by using the definition of 2-dimensional subspace in V). So axiom PP1 holds.

Similarly, the statement "any two lines in $\mathrm{PG}(2, q)$ intersect in precisely one point" is equivalent to "any two planes in V through the origin intersect in a line in V", which is true by a simple algebraic deduction. Thus, axiom PP2 holds.

Let V have the standard basis $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. Each of the four vectors $(0, 0, 1)$, $(0, 1, 0)$, $(1, 0, 0)$ and $(1, 1, 1)$ in V generate a subspace of V (corresponding to a point in the projective plane), and since any three of these vectors are linearly independent, no three are contained in a subspace of

V of dimension 2 (a line in PG$(2, q)$). So PG$(2, q)$ contains a frame (complete quadrangle), verifying axiom PP3. $\square$

To find the number of points per line in PG$(2, q)$, count the number of lines in V through the origin contained in a plane (in V) containing the origin. Any one of $q^2 - 1$ points different from the origin exist on this plane, and $q-1$ of them yield the same line, so there are $(q^2-1)/(q-1) = q+1$ such lines. An analogous argument holds shows that the number of lines through a point is $q + 1$.

Rather than use Theorem 11.4.9, one can count the number of points and lines in PG$(2, q) = \mathcal{P}(q)$ directly. For any $\boldsymbol{v} \in V\backslash\{(0, 0, 0)\}$ and $k \neq 0$, both $\boldsymbol{v}$ and $k\boldsymbol{v}$ determine the same point in $\mathcal{P}(q)$, there are

$$\frac{q^3 - 1}{q - 1} = q^2 + q + 1$$

points in the plane $\mathcal{P}(q)$.

Each (hyper)plane (linear subspace of dimension 2) in $V = [GF(q)]^3$ contains q^2 points (including the origin), and is determined by two points (and the origin) not on a line through the origin, say $\mathbf{x}$ and $\mathbf{y}$. There are $q^3 - 1$ ways to pick $\mathbf{x}$, and having fixed $\mathbf{x}$, there are $q^3 - q$ remaining points not on a line through $\mathbf{x}$ and the origin. Hence there are $(q^3 - 1)(q^3 - q)$ possible ways to pick the ordered pair $(\mathbf{x}, \mathbf{y})$. There are many such pairs determining the same plane; on each plane containing the origin, there are $q^2 - 1$ choices for $\mathbf{x}$, and having chosen $\mathbf{x}$, there are $q^2 - q$ choices for $\mathbf{y}$ on the same plane. so there are

$$\frac{(q^3 - 1)(q^3 - q)}{(q^2 - 1)(q^2 - q)} = q^2 + q + 1$$

2-dimensional subspaces of V and hence as many lines in $\mathcal{P}(q)$.

11.4.3 Homogeneous coordinates

A convenient notation for points and lines in PG$(2, q)$ is called *homogeneous coordinates*. Any point in PG$(2, q)$ is a 1-dimensional subspace in $V = (\mathbb{F}_q)^3$, a line containing the origin, For each such line, there is a vector $(a, b, c) \in V$ so that the line is of the form $\{(ka, kb, kc) : k \in \mathbb{F}_q\}$. Choosing k to be the inverse of the first non-zero coordinate in (a, b, c), the line can be represented by a vector of the form $(1, b, c)$, $(0, 1, c)$ or $(0, 0, 1)$. Such a vector for a point in $\mathrm{PG}(2, q)$ is said to be the homogeneous coordinates for the point. As confirmed by Lemma 11.4.13, there are $q^2 + q + 1$ such vectors using homogeneous coordinates (q^2 of the form $(1, b, c)$, q of the form $(0, 1, c)$, and 1 of the form $(0, 0, 1)$).

Similarly, any line in $\mathrm{PG}(2, q)$ is a 2-dimensional subspace of V (a plane passing through the origin), and to each subspace, there is a normal vector. As in the case for points, there is a vector representing this plane (namely, a vector normal to the plane) written in homogeneous coordinates, but square brackets $[x, y, z]$ are used to indicate the corresponding line in $\mathrm{PG}(2, q)$. Just as for points, homogeneous coordinates for lines shows that there are q^2+q+1 lines in $\mathrm{PG}(2, q)$.

If (a, b, c) are homogeneous coordinates for a point in $\mathrm{PG}(2, q)$ and $[x, y, z]$ are homogeneous coordinates for a line in $\mathrm{PG}(2, q)$, then the point (a, b, c) is on the line $[x, y, z]$ if and only if in V, the vector (a, b, c) is orthogonal to the vector (x, y, z). Thus, in $\mathrm{PG}(2, q)$, a point (a, b, c) is on a line $[x, y, z]$ if and only if $ax + by + cz = 0$. In other words, a point (a, b, c) is on line $[x, y, z]$ if and only if their inner product (dot product) is zero (computed modulo q) .

Exercise 258. *In PG$(2, 5)$, compute* $(0, 1, 4) \vee (1, 2, 3)$.

Exercise 259. *In PG$(2, 3)$, compute* $[0, 1, 1] \wedge [1, 2, 0]$.

11.4.4 The Fano plane

The finite projective plane $\mathrm{PG}(2, 2) = \mathcal{P}(2)$ is called the Fano plane, named after Gino Fano (1871–1952). The $2^2 + 2 + 1 = 7$ points of the Fano plane, given in homogeneous coordinates, are $(0, 0, 1)$, $(0, 1, 0)$, $(0, 1, 1)$, $(1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 0)$ and $(1, 1, 1)$. Similarly, the lines are given by $[0, 0, 1]$, $[0, 1, 0]$, $[0, 1, 1]$, $[1, 0, 0]$, $[1, 0, 1]$, $[1, 1, 0]$ and $[1, 1, 1]$.

Exercise 260. *Show that the example given in Exercise 255 is isomorphic to the Fano plane.*

Recall that a point (a, b, c) is incident with a line $[x, y, z]$ iff $(a, b, c) \bullet (x, y, z) = 0$. Using this dot product, one can verify that the lines contain the following points:

$$[0, 0, 1] = \{(0, 1, 0), (1, 0, 0), (1, 1, 0)\}$$

$$[0,1,0] = \{(0,0,1),(1,0,0),(1,0,1)\}$$
$$[0,1,1] = \{(0,1,1),(1,0,0),(1,1,1)\}$$
$$[1,0,0] = \{(0,0,1),(0,1,0),(0,1,1)\}$$
$$[1,0,1] = \{(0,1,0),(1,0,1),(1,1,1)\}$$
$$[1,1,0] = \{(0,0,1),(1,1,0),(1,1,1)\}$$
$$[1,1,1] = \{(0,0,1),(0,1,0),(1,0,0)\}$$

See Figure 11.2 for a pictorial representation of the Fano plane. It is shown later (see Corollary 15.1.3) that by the Sylvester–Gallai theorem (Theorem 15.1.2), any drawing of a Fano plane cannot be given using only straight lines—at least one line needs to be curved.

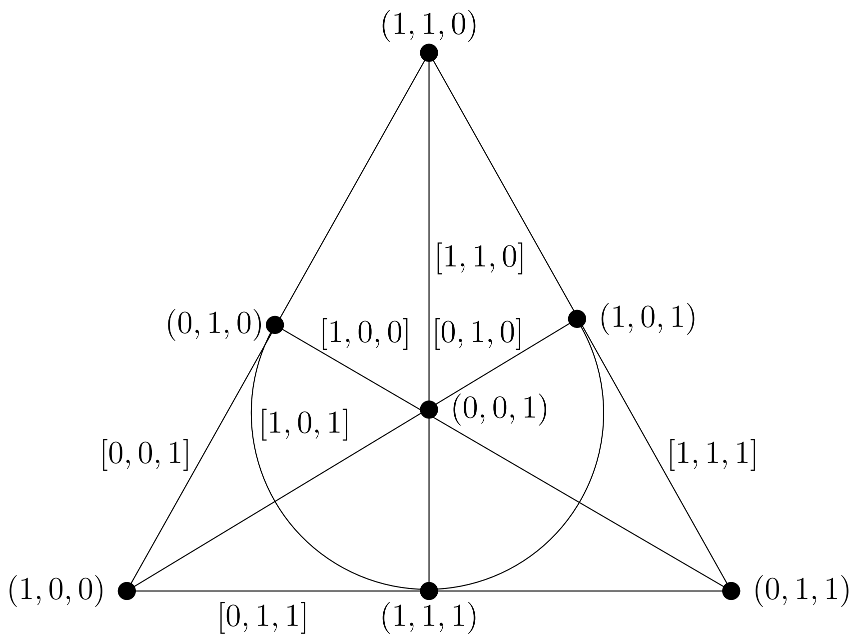


Figure 11.2: The Fano plane $\mathrm{PG}(2,2) = \mathcal{P}(2)$ with homogeneous coordinates

The Fano plane is also an example of what is called a “Steiner triple system” (STS). A STS on 7 points is a collection of 7 blocks, where each pair of points is contained in exactly one block. For example, if the seven points are labelled 1,2,3,4,5,6,7, then the blocks can be 123, 145, 167, 246, 257, 347, and 356. (The Fano plane is also a simple example of what is called a "design"—see Chapter 12.) For slightly more information on STSs, see Section 15.4.

The Fano plane is named after Gino Fano; however, it might be interesting to note that Fano was studying something now called "Fano's axiom", an axiom that the Fano plane *does not satisfy* (the three points of intersections of "diagonals" of a complete quadrilateral are not collinear).

It might be interesting to note that the standard drawing of the Fano plane can be used (with added arrows) to encode the multiplication rules for the octonians. The octonions are a non-associative algebra, and in the context of coordinatizing a FPP, octonions are also discussed (with a similar picture as alluded to above) in Kadison and Kromann's book [515, p. 181]. See also John Baez's website [41] for details. (Thanks to Ian Thompson for this reference.)

11.4.5 Characterizing FPPs

The following theorem characterizes the classical finite projective planes.

Theorem 11.4.14. *Let $\mathcal{P}$ be a finite projective plane. The following conditions are equivalent:*

(1) $\mathcal{P}$ arises from a finite field (i.e., for some prime power q, $\mathcal{P} = PG(2, q)$).

(2) $\mathcal{P}$ satisfies Desargues' theorem (see Theorem 1.8.3 for the statement in the real plane).

(3) $\mathcal{P}$ satisfies Pappus's theorem (see Theorem 1.8.1 for the statement in the real plane).

(4) The automorphism group of $\mathcal{P}$ is 2-transitive, that is, for any points A, B, C, D of $\mathcal{P}$, A and B distinct, and C and D distinct, there exists an automorphism of $\mathcal{P}$ that takes $\{A, B\}$ to $\{C, D\}$.

The implication (1)→(4) is shown below in Lemma 11.4.15. The implication (4)→(1) is due to T. G. Ostrom [696] in 1958 and together with A. Wagner [698] in 1959 (see also [697], [905] and [906]). The implications (1) $\leftrightarrow$ (2) can be found in many sources; for example, see [17, Thm. 5, p. 77] or [51, Thm. 13.2.3, pp. 267–268]. The implication (1) $\rightarrow$ (3) is given in many places, e.g., [51]. For a proof of (3) → (2), see, e.g., [17, Thm. 6, p. 78]. n 1956, Ostrom showed [695] that if n is odd and a non-square, and a FPP of order n has a doubly transitive collineation group, then it is Desarguesian (so showing in some cases, (4) → (2)).

Lemma 11.4.15. *Let $\mathbb{F}$ be a field, and let $\Pi = PG(2, \mathbb{F})$. Let A, B, D, E be points in Π, where $A \neq B$ and $D \neq E$. There exists an automorphism of Π that takes A to D and B to E.*

Proof: In homogeneous coordinates, let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{D}$ and $\mathbf{E}$ be vectors representing the points A, B, D, E, respectively. Since $A \neq B$, the vectors $\mathbf{A}$ and $\mathbf{B}$ are not multiples of one another, and so $\{\mathbf{A}, \mathbf{B}\}$ is an independent set in $V\mathbb{F}^3$. Similarly, $\{\mathbf{D}, \mathbf{E}\}$ is an independent set. Since any independent set in V can be extended to a basis, let $\mathbf{C}$ and $\mathbf{F}$ be so that $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ and $\{\mathbf{D}, \mathbf{E}, \mathbf{F}\}$ are bases for V. Define $\phi : V \to V$ to be the linear transformation satisfying $\phi(\mathbf{A}) = \mathbf{D}$, $\phi(\mathbf{B}) = \mathbf{E}$ and $\phi(\mathbf{C}) = \mathbf{F}$. (Essentially, the matrix for this linear transformation is a change of basis matrix, which is invertible.) Then ϕ naturally induces an automorphism of Π that takes A to D and B to E (and C to F. □

Since three non-collinear points in PG$(2, \mathbb{F})$ have (homogeneous) vectors that are linearly independent, the above proof also yields the following:

Corollary 11.4.16. *Let $\mathbb{F}$ be a field, and let $\Pi = PG(2, \mathbb{F})$. Then for two sets of non-collinear points $\{A, B, C\}$ and $\{D, E, F\}$, there is an automorphism that takes A to D, B to D, and C to F.*

Since Lemma 11.4.15 shows that the automorphism group of $PG(2, \mathbb{F})$ is 2-transitive, and Corollary 11.4.16 says that the group is "nearly" 3-transitive, a natural question might be to ask if 3-transitivity fails.

Exercise 261. *Give an example of $\mathbb{F}$ and two sets of three points in $PG(2, \mathbb{F})$ that shows that the automorphism group of $PG(2, \mathbb{F})$ is not 3-transitive.*

11.4.6 Subplanes of a FPP

A *subplane* of a finite projective plane $\mathcal{P}$ is a subset of points of $\mathcal{P}$ that form a finite projective plane with the same incidences as in $\mathcal{P}$.

Theorem 11.4.17 (Bruck, 1955 [157])**.** *Let $\mathcal{P}$ be a FPP of order n, and let $\mathcal{Q}$ be a proper subplane of order m. Then either $n = m^2$ or $n \geq m^2 + m$.*

Proof: Let P be the points of $\mathcal{P}$ and let $X \subset P$ be the points of $\mathcal{Q}$. Any line in $\mathcal{Q}$ extends to a line in $\mathcal{P}$. Let ℓ be a line in $\mathcal{Q}$ with $m+1$ points and let ℓ' be the line in $\mathcal{P}$ that extends ℓ. So ℓ' has $m+1$ points in X and $n+1-(m+1)=n-m$ points in $P \backslash X$.

Since any two lines in $\mathcal{Q}$ intersect, two lines in $\mathcal{P}$ that extend two lines in $\mathcal{Q}$ never intersect outside of X (so these sets of $n-m$ points are pairwise disjoint). Counting over the lines in $\mathcal{Q}$, $P \backslash X$ contains at least $(m^2+m+1)(n-m)$ points. Since the number of points in P is n^2+n+1,

$$n^2+n+1 \geq |X|+|P \backslash X| \geq m^2+m+1+(m^2+m+1)(n-m). \quad (11.1)$$

With a little simple algebra, (11.1) becomes

$$(m^2-n)(n-m) \leq 0.$$

Since $n>m$, then $m^2n-n \leq 0$ or $n \geq m^2$.

If $n=m^2$, equality in (11.1) is forced, showing that every point in P outside of X lies on some line extending from $\mathcal{Q}$.

Suppose that $n \neq m^2$. Then there exists a point A in P not incident with any line extending out of $\mathcal{Q}$. This means that any line containing A intersects X (the points of $\mathcal{Q}$) at most once. Since for each point B in X there is a line through A and B, there are at least $|X|=m^2+m+1$ lines through A. In $\mathcal{P}$, there are $n+1$ lines through A, and so

$$n+1 \geq m^2+m+1,$$

which gives the desired $n \geq m^2+m$. □

According to [490], until at least 1973, for planes of order n no subplanes had been found of order m with $n=m^2+m$.

Definition 11.4.18. Let $\mathcal{P}$ be a FPP, and let $\mathcal{Q}$ be a proper subplane of $\mathcal{P}$. Then $\mathcal{Q}$ is a Baer subplane if and only if every point on $\mathcal{P}$ is incident with some line extending a line in $\mathcal{Q}$.

The following is immediate from the proof of Theorem 11.4.17:

Corollary 11.4.19. *Let $\mathcal{P}$ be a FPP of order n and let $\mathcal{Q}$ be a subplane of order m. Then $\mathcal{Q}$ is a Baer subplane if and only if $n=m^2$.*

Exercise 262. *Give an example of n and m and a FPP of order n and subplane of order m so that $n>m^2+m$.*

I learned the following theorem from Padmanabhan [706], (who called it "Folklore") in 2016.

Theorem 11.4.20 (Folklore)**.** *Let $\mathbb{F}$ be any field (or division ring). Then any projective geometry over $\mathbb{F}$ contains a 7-point Fano subplane if and only if (in $\mathbb{F}$) $0 = 2$.*

Proof: Let $\mathcal{P}$ be a projective geometry over $\mathbb{R}$. Suppose that X is a set of 7 points forming a Fano configuration. Since $\mathcal{P}$ is over a field, without loss of generality, suppose that (using homogeneous coordinates) the points $A = (1, 0, 0)$, $B = (0, 1, 0)$, $C = (0, 0, 1)$ in X form the three points on the corners of a standard drawing of the Fano plane.

The line $A \vee B$ is given by $z = 0$, the line $B \vee C$ is given by $x = 0$, and the line AC is given by $y = 0$. For $b \in \mathbb{F}\backslash\{0\}$, let $E = (0, b, 1)$ be on $C \vee B$ and for some $a \in \mathbb{F}\backslash\{0\}$ let $F = (a, 0, 1)$ be a point in X on $A \vee C$. Let $D = (x, y, z)$ denote the "middle" point of X.

Since A, D, E are collinear,

$$\begin{vmatrix} 1 & 0 & 0 \\ x & y & z \\ 0 & b & 1 \end{vmatrix} = 0 \Rightarrow \begin{vmatrix} y & z \\ b & 1 \end{vmatrix} = 0 \Rightarrow y = bz.$$

Since B, D, F are collinear,

$$\begin{vmatrix} 0 & 1 & 0 \\ x & y & z \\ a & 0 & 1 \end{vmatrix} = 0 \Rightarrow \begin{vmatrix} x & z \\ a & 1 \end{vmatrix} = 0 \Rightarrow x = az.$$

Hence, the coordinates of the middle point D are $(az, bz, z) = (a, b, 1)$.

Finally, for $u, v \in \mathbb{F}\backslash\{0\}$ let $G = (u, v, 0)$ be the third point on $A \vee B$. Then C, D, G are collinear and so

$$\begin{vmatrix} 0 & 0 & 1 \\ a & b & 1 \\ u & v & 0 \end{vmatrix} = 0 \begin{vmatrix} a & b \\ u & v \end{vmatrix} = 0 \Rightarrow av = bu. \tag{11.2}$$

Also, since E, F, G are collinear,

$$\begin{vmatrix} 0 & b & 1 \\ a & 0 & 1 \\ u & v & 0 \end{vmatrix} = 0 \Rightarrow \begin{vmatrix} a & 0 \\ u & v \end{vmatrix} + \Rightarrow \begin{vmatrix} 0 & b \\ u & v \end{vmatrix} = 0 \Rightarrow av + bu = 0. \tag{11.3}$$

By equations (11.2) and (11.3), $av = -av$ or $2av = 0$. Since $a \neq 0$ and $v \neq 0$, and $\mathbb{F}$ has no zero-divisors, $2 = 0$ in $\mathbb{F}$. $\square$

Corollary 11.4.21. *Since the characteristic of $\mathbb{Q}$ and $\mathbb{R}$ are both 0, neither the projective plane (or affine plane) over $\mathbb{Q}$ or $\mathbb{R}$ contains a Fano plane as a subplane. Hence, the Fano plane cannot be drawn with straight lines in $\mathbb{R}^2$.*

11.5 Finite affine planes

For a positive integer n and a prime power q, the notation $\mathrm{AG}(n, q)$ denotes the affine geometry of dimension n over the finite field $\mathbb{F}_q$. Sometimes $\mathrm{AG}(2, q)$ is denoted by $\mathrm{AG}(2, \mathbb{F}_q)$ to emphasize the underlying structure of the plane. In this section, only 2-dimensional geometries (*i.e.*, planes) are discussed.

Definition 11.5.1. For a prime power q, the affine plane $\mathrm{AG}(2, q)$ is defined on point set $(\mathbb{F}_q)^2$, where lines are solution sets to linear equations (modulo q).

Exercise 263. *Show that for any prime power q, the affine plane $AG(2, q)$ has q^2 points and $q(q+1)$ lines, with q points per line.*

For example, the point set of $\mathrm{AG}(2, 3)$ is $\{(x, y) : x, y \in \{0, 1, 2\}\}$, and the lines are

$$\begin{aligned}
y = 0 &: \{(0,0),(1,0),(2,0)\};\\
y = 1 &: \{(0,1),(1,1),(2,1)\};\\
y = 2 &: \{(0,2),(1,2),(2,2)\};\\
x = 0 &: \{(0,0),(0,1),(0,1)\};\\
x = 1 &: \{(1,0),(1,1),(1,2)\};\\
x = 2 &: \{(2,0),(2,1),(2,2)\};\\
y = x &: \{(0,0),(1,1),(2,2)\};\\
y = x+1 &: \{(0,1),(1,2),(2,0)\};\\
y = x+2 &: \{(0,2),(1,0),(2,1)\};\\
y = 2x &: \{(0,0),(1,2),(2,1)\};\\
y = 2x+1 &: \{(0,1),(1,0),(2,2)\};
\end{aligned}$$

$$y = 2x + 2 : \{(0, 2), (1, 1), (2, 0)\}.$$

For a pictorial representation, see Figure 15.8. This plane is also called the "Hesse configuration" (see Section 15.4 to see why) or "Young's geometry".

It is not difficult to verify that an affine plane $\mathrm{AG}(2, q)$ is obtained from $\mathrm{PG}(2, q)$ by removing a line (the "line at infinity"), and to construct $\mathrm{PG}(2, q)$ from $\mathrm{AG}(2, q)$, add a line at infinity (each point of which extends all lines of the same "slope" in $\mathrm{AG}(2, q)$).

11.6 Incidence graphs for configurations or geometries

Suppose that $(X, \mathcal{L})$ is a configuration of points and lines (where each line $\ell \in \mathcal{L}$ is a collection of points in X). Then the *incidence graph* for this configuration is a bipartite graph $G = (V, E)$, where V is the (disjoint) union of X and $\mathcal{L}$, and a pair (x, ℓ) is an edge in G if and only f $x \in \ell$. Incidence graphs were studied by Levi, and so an incidence graph for a configuration is sometimes called the Levi graph for that configuration.

Incidence graphs for geometries are useful tools in combinatorics, since some geometries give rise to examples of graphs that are in some sense, usually "extremal" regular graphs.

For example, the incidence graph for the Fano plane, now called the Heawood graph (see Figure 11.3), has $7 + 7 = 14$ vertices, each of degree 3.

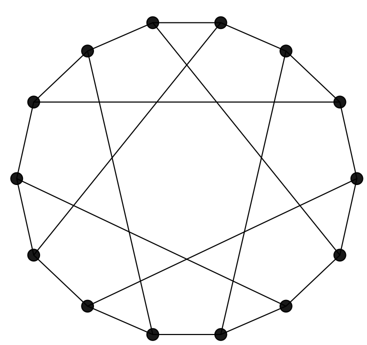


Figure 11.3: Heawood graph, 14 vertices; incidence graph for Fano plane

In a geometry, since no two points are on two different lines, the incidence graph for a geometry contains no 4-cycles, and since such a graph is bipartite, the smallest possible cycle in an incidence graph for a geometry is 6; in other words, if G is the incidence graph for a geometry, then $\text{girth}(G) \geq 6$. For example, the Heawood graph is the smallest 3-regular graph with girth 6 (and so it is called a $(3, 6)$-cage).

The incidence graph for a finite projective plane of order q has $2(q^2+q+1)$ vertices, $(q+1)(q^2+q+1)$ edges and contains no $C_4 \cong K_{2,2}$. Such a graph then shows:

Lemma 11.6.1 (Reiman, 1959 [748])**.** *If $n = q^2 + q + 1$ for some prime power q, then*

$$ex(n; K_{2,2}) > (1/2)(n^{3/2} - n^{1/2}).$$

There seems to be many results in extremal graph theory whose proofs use incidence graphs. Only two more are mentioned here. In 1966, Brown [156] proved that

$$\text{ex}(n; K_{3,3}) > (1 - o(1))\frac{1}{2}n^{5/3}.$$

Brown used, for an odd prime p, the 3-dimensional geometry over $\mathbb{F}_p$ (some number theory is also required for his proof). Ideas behind Brown's constructions, sometimes called "radius graphs" have since been extended (see, e.g., [564] or [23]).

Another notable example where the incidence graph for a FPP is used was provided by Erdős, Renyi, and Sós, who provided a novel proof of "the friendship theorem":

Theorem 11.6.2 (Erdős–Rényi–Sós, 1966 [315])**.** *Let G be a graph on $n \geq 3$ vertices with the property that between any pair of vertices there is a unique path of length 2. Then n is odd, say $n = 2k + 1$, and G is the graph F_k consisting of k triangles sharing a common vertex.*

Theorem 11.6.2 is often called "the friendship theorem", because the theorem implies that if there is a party where every pair of people have precisely one common friend, then there is someone who knows everyone else at the party. The resulting graph is called the *friendship graph*; see Figure 11.4 for an example.

Including the original, there are now (at least) three different types of proofs of Theorem 11.6.2. All three proof ideas begin in the same way—by showing that if there is no vertex adjacent to all others, then the graph is regular. In the original proof [315], incidence graphs and a result by Baer

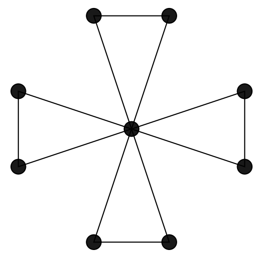


Figure 11.4: Friendship graph F_4

on finite projective planes shows that no such regular graphs exist (except trivial cases). This proof is presented in Section 13.4, after the notion of "polarity" is introduced.

A second proof (based on ideas of Wilf [924]) uses spectral theory (algebraic graph theory) to show that only one such graph exists, namely K_3. (There are many variants of this algebraic proof, but perhaps differing only in the way certain quantities are calculated.) A third proof [619] uses what one might call more straightforward counting.

For an example using the "anti-incidence" graph of a FPP, see Example 21.3.12.

11.7 Embedding graphs in geometries

If $G = (V, E)$ is a graph, for which geometries can G be "embedded" in the geometry; that is, there is an injection from V to points in the geometry, and each edge in G is a pair of points in some line. However, if the lines of a geometry are large enough, under this definition of an embedding, it is trivial that any graph on n vertices can be embedded into a geometry with one line having at least n points. So, the definition is made stronger.

Definition 11.7.1. Let $G = (V, E)$ be a (simple) graph and let $\mathcal{P} = (X, \mathcal{L})$ be a geometry. An embedding of G into $\mathcal{P}$ is an injective function $f : V \to X$ so that incidences from G are preserved in $\mathcal{P}$ and the induced function $g : E \to \mathcal{L}$ is also injective.

Of course, the above question is too broad to give any reasonable answer. In Exercise 256, it was established that K_5 is itself a hyperbolic geometry.

Perhaps of more interest, are embeddings of graphs into finite projective planes.

Since any FPP contains a complete quadrangle, any FPP embeds the complete graph K_4. If a FPP is large enough, then additional points can be found that embed a complete graph of any size.

Lemma 11.7.2. *Let $\mathcal{P}$ be a FPP of order q (a prime power), and let $n \geq 1$. If $q \geq \binom{n}{2}$, then K_n can be embedded into $\mathcal{P}$.*

Exercise 264. *Prove Lemma 11.7.2.*

Using homogeneous coordinates, a result much stronger than Lemma 11.7.2 is given in Exercise 279. See also Theorem 15.4.8 for embedding arbitrary graphs in finite projective planes.

Exercise 265. *Show that for any $n \geq 2$, the complete bipartite graph $K_{n,n}$ can be embedded into a FPP of order n.*

Embeddings of graphs in an FPP are sometimes looked at with respect to how many there are, what conditions suffice, and what properties do such embeddings have. Some results are related to structure of Baer subplanes. For more on graphs embedded into finite projective planes, see [653] (where references are given also for embedding cycles in FPPs). See also [904] for counting cycles in an FPP.

As suggested in [653], it might be an interesting project to classify (at least in part) the relationship between the automorphism group of a graph and the automorphism group of a plane that embeds the graph.

Chapter 12

Finite geometries as block designs

12.1 Introduction to block designs

In modern mathematics, the term "design" is often used as an abbreviation for "combinatorial block design", a finite set system with certain "uniformity" properties. Many authors use the word 'design" as an abbreviation for a more restricted class of set systems called "balanced incomplete block designs" (BIBDs) Some finite geometries are examples of BIBDs, where instead points are called "varieties" or "vertices" and lines are called "blocks".

The reason that these set systems are called "designs" is that they arose from the study of designing agricultural experiments. Many combinatorial questions and results from the 18th and 19th century are now considered to be central in design theory, although design theory *per se* has had its name only since the 20th century. An example of such a question is due to Kirkman [547]: 15 schoolgirls walk to school each of seven days. Can you arrange them into groups of three each day so that every pair of girls walk together precisely once? (The answer is "yes"; see, e.g., [770] for an introduction, and, e.g., [171] or [259] for more details.)

The material here is only a brief introduction to designs, in particular, as they relate to the combinatorics in finite geometries. (Much of this chapter is taken verbatim from my notes on finite projective planes [434].) For introductions to design theory in general, see, for example, [152], [171], [259], [454, Ch. 10], or [491].

A *set system* is an ordered pair $(X, \mathcal{S})$ where X is a set and $\mathcal{S}$ is a collection of subsets of X, where subsets may be repeated.

Definition 12.1.1. A *block design* is a set system $(X, \mathcal{B})$ so that all elements of $\mathcal{B}$ are of the same cardinality. Elements of X are called *vertices*, *points*, or *varieties*. Elements of $\mathcal{B}$ are called *blocks*.

A block design is called *simple* if no blocks are repeated; unless otherwise specified, all block designs are assumed to be simple. So, a block design is a simple set system where all blocks have the same number of vertices. (In graph theoretic terms, a design is a uniform hypergraph.)

The *complete design* on a set X is the collection of all subsets of X and the *empty design* contains no blocks; a *trivial design* is either complete, empty, or consists of a single block. In the literature, the term "incomplete block design" usually means only that the entire base set X does not appear as a block.

A finite geometry is a block design if all of its lines have the same number of points. By definition (see Definition 11.3.1), finite geometries studied here have this property.

Definition 12.1.2. For any positive integer t, a design on X is t-wise balanced if and only if every set of t vertices from X occurs in the same number of blocks. A t-wise balanced design is called a t-design.

A 1-design is called *pointwise balanced* since every vertex occurs in the same number of blocks, and, similarly, a 2-design is called *pairwise balanced*. [Note: Some authors define a tactical configuration to be a 1-design.]

In the definition of a finite geometry (Definition 11.3.1), every point is incident with the same number of lines, so a finite geometry is pointwise balanced (a 1-design). The condition that there are at most one line through any pair of points is not enough to say that a geometry is pairwise balanced, but some geometries studied here are pairwise balanced as well.

In a t-design, the number of times a t-subset occurs in a block is denoted by λ, and is called the *index* of the design. If $|X| \geq 2$, it is trivial that each vertex of X in a pairwise balanced block design on X appears in at least one block, since each vertex appears as a member of some pair.

Definition 12.1.3. A *balanced incomplete block design* (BIBD) is an incomplete block design that is both pointwise and pairwise balanced, that is, a BIBD is both a 1-design and a 2-design.

The next theorem shows that a pairwise balanced design is also pointwise balanced, and so an incomplete 2-design is a BIBD.

Theorem 12.1.4. *Let $X, \mathcal{B}$ be a 2-design, and where $|X| = v$, each block in $\mathcal{B}$ contains k elements of X and every pair of vertices from X is contained in precisely λ blocks. Then each vertex of X appears in the same number of blocks, namely, in*

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

blocks.

Proof: Fix $x \in X$ and suppose that x is contained in only the blocks $B_1, B_2, \ldots, B_r$. The proof is by counting the pairs

$$\{(y, B) : B \in \mathcal{B} \text{ and } \{x, y\} \subset B\}$$

in two ways.

In each of $B_1, \ldots, B_r$, there are $k-1$ pairs of the form $\{x, y\}$, so in all, there are $r(k-1)$ such pairs. On the other hand, there are $v-1$ pairs from X containing x, and each of these pairs occurs in λ blocks, giving $\lambda(v-1)$ such pairs. So

$$r(k-1) = \lambda(v-1). \tag{12.1}$$

Since x was arbitrary, the proof is complete. □

The parameter r is called the *replication number*. Is it true that if a design is both pairwise balanced and point-wise balanced then each block is of the same size? If this were to be true, then one could feel comfortable in defining a BIBD to have both pairwise and pointwise balance. The design that has as its blocks all subsets of say, two sizes, is incomplete, pairwise balanced and pointwise balanced, however, does not have uniform block size.

Theorem 12.1.5. *In a BIBD on X with $|X| = v$ where each block is of size k and each vertex appears in r blocks, there are $b = vr/k$ blocks (and so $bk = vr$).*

Proof: Let b be the number of blocks. The number of pairs $\{x, B\}$ where $x \in B$ is $bk = vr$. □

In view of Theorems 12.1.4 and 12.1.5, it suffices to know three of the parameters of a BIBD, from which the other two follow; however, it is convenient to list all five when discussing a particular BIBD. A BIBD on v vertices (or varieties), with b blocks, each block of size k, where each vertex occurs in r blocks and each pair of vertices occurs in λ blocks, is often referred to as as a (v,b,k,r,λ) design, and occasionally, to emphasize the pairwise balanced aspect, as a $(v,b,k,r,2,\lambda)$ design, or a (v,b,k,r,λ) 2-design, or a 2-(v,b,k,r,λ) design. In some literature, the pairwise balanced property of a BIBD is sometimes stressed by calling the design a PBIBD, however, this can lead to confusion since there are other classes of designs, including "partial designs" or "partially balanced" designs and the notation PBIBD has also been reserved for these—which are not discussed here. In some texts, the order of the k and r are interchanged, so a reader is advised to check on this detail.

A (v,b,k,r,λ) BIBD is called a *symmetric* BIBD (denoted SBIBD) if and only if $v=b$. By Theorem 12.1.5, if $v=b$, then $k=r$, in which case the notation (v,v,k,k,λ) BIBD is abbreviated by (v,k,λ) SBIBD. (Note that if one of $v=b$ or $k=r$ holds, then so does the other.)

A BIBD design is a geometry only when $\lambda=1$. A FPP of order q is a $(q^2+q+1,q+1,1)$ SBIBD.

12.2 Incidence matrices for designs

Definition 12.2.1. Let $(X,\mathcal{B})$ be a set system with vertex set $X=\{x_1,\ldots,x_v\}$ and blocks $B_1,\ldots,B_b$. The incidence matrix $A=(a_{ij})$ of the design is the $v\times b$ 0–1 matrix defined by $a_{ij}=1$ if and only if $x_i\in B_j$.

Note: An incidence matrix for any set system depends on the order (or labelling) of the vertices and blocks. Also, another way to see the identity $bk=vr$ from Theorem 12.1.5 is to simply count the 1's in the incidence matrix of the design in two ways, first by columns, then by rows.

Given a set system with incidence matrix A, the two matrices AA^T and A^TA reveal a great deal of information about the set system. The proof of the following lemma is left to the reader:

Lemma 12.2.2. *Let $(X,\mathcal{B})$ be a set system on vertices $x_1,\ldots,x_v$ with blocks $B_1,\ldots,B_b$. Let A be the $v\times b$ incidence matrix for $(X,\mathcal{B})$ for this particular ordering. Define the $v\times v$ matrix $B=(b_{ij})=AA^T$.*

(i) For each $i=1,\ldots v$, b_{ii} is the number of blocks containing x_i, and

(ii) For each $i\neq j$, b_{ij} is the number of blocks containing both x_i and x_j.

Define the $b \times b$ matrix $C = (c_{ij}) = A^T A$. Then
(iii) For $i = 1, \dots, b$, $c_{ii} = |B_i|$, and
(iv) For $i \neq j$, $c_{ij} = |B_i \cap B_j|$.

By the conclusions in Lemma 12.2.2, when A is an incidence matrix for a BIBD, both AA^T and $A^T A$ have very simple forms, but before describing these forms, some extra (temporary) notation may be helpful.

Let $(X, \mathcal{B})$ be an arbitrary set system with $X = \{x_1, \dots, x_v\}$, $\mathcal{B} = \{B_1, \dots, B_b\}$, and incidence matrix A. For each $i = 1, \dots, v$, let r_i denote the number of blocks containing x_i, and for $i \neq j$, let λ_{ij} be the number of blocks containing both x_i and x_j. With this notation, conclusions (i) and (ii) of Lemma 12.2.2 say

$$AA^T = \begin{bmatrix} r_1 & \lambda_{1,2} & \lambda_{1,3} & \cdots & \lambda_{1,v} \\ \lambda_{2,1} & r_2 & \lambda_{2,3} & \cdots & \lambda_{2,v} \\ \lambda_{3,1} & \lambda_{3,2} & r_3 & \cdots & \lambda_{3,v} \\ \vdots & \vdots & & \ddots & \vdots \\ \lambda_{v,1} & \lambda_{v,2} & \lambda_{v,3} & \cdots & r_v \end{bmatrix}. \tag{12.2}$$

For each $i = 1, \dots, b$, let $k_i = |B_i|$, and for $i \neq j$, let $\beta_{ij} = |B_i \cap B_j|$. Then conclusions (iii) and (iv) of Lemma 12.2.2 say

$$A^T A = \begin{bmatrix} k_1 & \beta_{1,2} & \beta_{1,3} & \cdots & \beta_{1,b} \\ \beta_{2,1} & k_2 & \beta_{2,3} & \cdots & \beta_{2,b} \\ \beta_{3,1} & \beta_{3,2} & k_3 & \cdots & \beta_{3,b} \\ \vdots & \vdots & & \ddots & \vdots \\ \beta b, 1 & \beta_{b,2} & \beta_{v,3} & \cdots & k_b \end{bmatrix}. \tag{12.3}$$

When $(X, \mathcal{B})$ is a BIBD, all of the r_i's are the same, and all of the λ_{ij}'s are the same, and so (12.2) becomes quite simple:

Corollary 12.2.3. *Let A be an incidence matrix for a (v, b, k, r, λ) BIBD. If J is the $v \times v$ matrix of all 1's, and I_v is the $v \times v$ identity matrix, then*

$$AA^T = \lambda J + (r - \lambda) I_v. \tag{12.4}$$

Furthermore,

$$det(AA^T) = (r + (v-1)\lambda)(r - \lambda)^{v-1}. \tag{12.5}$$

Proof: Equation (12.4) follows directly from the definition of a BIBD and equation (12.2). To see equation (12.5), subtract the first column of AA^T from each of the other columns, then add all other rows to the first row, which gives a triangular matrix whose determinant is evident. □

Lemma 12.2.4. *For any* (v, b, k, r, λ) *BIBD,* $r > \lambda$.

Proof: If $r \leq \lambda$, then by Theorem 12.1.4, $k \geq v$, contrary to the definition of an incomplete design. □

Lemma 12.2.5. *If* A *be an incidence matrix for any* (v, b, k, r, λ) *BIBD, then*

$$det(AA^T) = rk(r - \lambda)^{v-1} > 0. \tag{12.6}$$

Proof: By Theorem 12.1.4, $r(k - 1) = \lambda(v - 1)$ and so

$$r + (v - 1)\lambda = r + r(k - 1) = rk.$$

Thus by equation (12.5) of Corollary 12.2.3, $\det(AA^T) = rk(r-\lambda)^{v-1}$, which is positive by Lemma 12.2.4. □

The next result is known as "Fisher's inequality"; the proof given here is essentially due to Bose [138]. (The statement and proof given here can also be found in, *e.g.*, [753, pp. 386–388].)

Theorem 12.2.6 (Fisher's inequality [335])**.** *In a* (v, b, k, r, λ) *BIBD,* $v \leq b$.

Proof: Let A be an incidence matrix for a (v, b, k, r, λ) BIBD. Suppose, in the hope of contradiction, that $b < v$. Adjoining $v - b$ columns of zeros to A gives a square matrix B with determinant zero, and hence $\det(BB^T) = 0$. Yet $AA^T = BB^T$ and by Lemma 12.2.5

$$\det(BB^T) = \det(AA^T) > 0,$$

a contradiction. Hence $b \geq v$. □

Fisher's original proof [335] instead showed $r \geq k$, which, by Theorem 12.1.5, is equivalent to the form of Fisher's theorem presented above.

The conditions in Theorems 12.1.4 and 12.1.5 are not sufficient for the existence of a design with certain parameters. For example, no $(21, 14, 6, 4, 1)$

BIBD exists (since Fisher's inequality fails), yet the equalities $bk = vr$ and $r(k-1) = \lambda(v-1)$ hold.

The following theorem gives an easy condition to check for the existence of symmetric designs of certain types. The proof is relatively easy now that the preliminaries are out of the way. I don't know who first proved this; maybe it is due to Bruck and Ryser, since it commonly appears as part of the Bruck-Ryser-Chowla Theorem (Theorem 12.3.1).

Theorem 12.2.7. *In a symmetric design, if v is even, then $k-\lambda$ is a square (of an integer).*

Proof: Since $\det(AA^T) = (\det(A))^2$, and $r = k$, the relation (12.6) becomes

$$(\det(A))^2 = k^2(k-\lambda)^{v-1},$$

and so $(k-\lambda)^{v-1}$ is a square. Then since $v-1$ is odd and $\det(A)$ is an integer, $k-\lambda$ is also a square. $\square$

Using Theorem 12.2.7, the conditions of Theorems 12.1.4 and 12.1.5 together with Fisher's inequality are *not* sufficient for the existence of a balanced incomplete block design. For example, with $v = b = 22$, $r = k = 7$, and $\lambda = 2$, both $bk = vr$ and $r(k-1) = \lambda(v-1)$ hold, however v is even and $k - \lambda = 5$ which is not a square. So by Theorem 12.2.7, a $(22, 7, 2)$ SBIBD does not exist.

12.3 The Bruck–Ryser–Chowla theorem

One of the more powerful theorems regarding existence of designs is the Bruck–Ryser–Chowla Theorem.

Theorem 12.3.1 (Bruck–Ryser–Chowla, 1950 [158, 187])**.** *If a (v, k, λ)-SBIBD exists with v even, then $k-\lambda$ is a square (of an integer). If v is odd, then*

$$z^2 = (k-\lambda)x^2 + (-1)^{(v-1)/2}\lambda y^2$$

has a solution in integers x, y, z, not all zero.

The even case of Theorem 12.3.1 is merely Theorem 12.2.7, which has been proved. The case in Theorem 12.3.1 when $\lambda = 1$ and the design is a projective plane is the Bruck–Ryser Theorem (Theorem 11.4.11), first published in 1949; the remaining cases extending the result to symmetric designs were proved by Chowla and Ryser a year later.

To prove the Bruck–Ryser–Chowla theorem, some number theory is used, including Lagrange's theorem (Theorem 22.5.1), which says every positive integer is the sum of at most 4 squares (one proof of this theorem is by matrices and quadratic forms; another is available as a consequence of Minkowski's theorem for lattice points—see Section 9.4). The Bruck–Ryser theorem (Theorem 11.4.11) can be derived from the Bruck–Ryser–Chowla theorem.

To simplify the proof, of the Bruck–Ryser–Chowla theorem, a sequence of lemmas is given, the statement of the first of which is simple, but central to the proof given here of the main theorem.

Lemma 12.3.2. *For a (v, k, λ) SBIBD with incidence matrix $A = (a_{ij})$, the relationship*

$$B = AA^T = (r - \lambda)I + \lambda J \tag{12.7}$$

of Corollary 12.2.3 is equivalent to

$$\sum_{j=1}^{v}\left(\sum_{i=1}^{v} a_{ij}x_i\right)^2 = (k-\lambda)\left(\sum_{i=1}^{v} x_i^2\right) + \lambda\left(\sum_{i=1}^{v} x_i\right)^2, \tag{12.8}$$

where $x_1, x_2, \ldots, x_v$ are indeterminates.

Proof: Let $v_1, \ldots, v_v$ be the vertices of the design. (Please excuse the double meaning of the symbol v.) For each $i = 1, \ldots, v$ associate x_i with vertex v_i. For each $j = 1, \ldots, v$, to the j-th column of A (j-th block) associate the linear form

$$L_j = \sum_{i=1}^{v} a_{ij}x_i.$$

So AA^T is associated with the dot product

$$(L_1, \ldots, L_v) \bullet (L_1, \ldots, L_v) = \sum_{j=1}^{v} L_j^2,$$

which is the left-hand side of (12.8). To derive the right-hand side of (12.8), start with the left-hand side.

$$\sum_{j=1}^{v}\left(\sum_{i=1}^{v} a_{ij}x_i\right)^2 \quad = \quad \sum_{j=1}^{v}\left[\sum_{i'\neq i} a_{ij}a_{i'j}x_ix_{i'} + \sum_{i=1}^{v} a_{ij}^2x_i^2\right]$$

$$\begin{aligned}
&= \sum_{j=1}^{v}\left[\sum_{i'\neq i} a_{ij}a_{i'j}x_i x_{i'} + \sum_{i=1}^{v} a_{ij}x_i^2\right]\\
&= \sum_{j=1}^{v}\sum_{i'\neq i} a_{ij}a_{i'j}x_i x_{i'} + \sum_{j=1}^{v}\sum_{i=1}^{v} a_{ij}x_i^2\\
&= \lambda\left[\left(\sum_{i}^{v} x_i\right)^2 - \sum_{i}^{v} x_i^2\right] + k\sum_{i=1}^{v} x_i^2\\
&= (k-\lambda)\left(\sum_{i=1}^{v} x_i^2\right) + \lambda\left(\sum_{i=1}^{v} x_i\right)^2 . \square
\end{aligned}$$

The following 'reduction' type result also simplifies the proof of the main theorem.

Lemma 12.3.3. *Given a linear form*

$$L(y_1, y_2, \ldots, y_m) = c_1y_1 + c_2y_2 + \cdots + c_my_m$$

in independent indeterminates $y_1, \ldots, y_m$, with rational coefficients, there exists a linear form

$$M(y_2, \ldots, y_m) = \alpha_2y_2 + \cdots + \alpha_ny_n$$

such that

$$L\left(M(y_2, \ldots, y_m), y_2, \ldots, y_m\right)^2 = M(y_2, \ldots, y_m)^2.$$

Proof: If $c_1 = 1$, then let $M(y_2, \ldots, y_m) = -\frac{1}{2}(c_2y_2 + \cdots + c_my_m)$. In this case,

$$\begin{aligned}
&L\left(M(y_2, \ldots, y_m), y_2, \ldots, y_m\right)^2\\
&\quad= \left[-\frac{1}{2}(c_2y_2 + \cdots + c_my_m) + c_2y_2 + \cdots + c_my_m\right]^2\\
&\quad= \frac{1}{4}(c_2y_2 + \cdots + c_my_m)^2\\
&\quad= M(y_2, \ldots, y_m)^2.
\end{aligned}$$

If $c_1 \neq 1$, then let $M(y_2, \ldots, y_m) = \frac{1}{1-c_1}(c_2y_2 + \cdots + c_my_m)$. In this case,

$$L\left(M(y_2, \ldots, y_m), y_2, \ldots, y_m\right)^2$$

$$\begin{aligned}
&= \left[\frac{c_1}{1-c_1}(c_2y_2+\cdots+c_my_m)+c_2y_2+\cdots+c_my_m\right]^2\\
&= \left[\frac{1}{1-c_1}(c_2y_2+\cdots+c_my_m)\right]^2\\
&= M(y_2,\ldots,y_m)^2.
\end{aligned}$$

In either case, M is found, as desired. □

Perhaps one more observation with respect to linear forms is in order.

Lemma 12.3.4. *Let b_1, b_2, b_3, b_4 be integers with $n = b_1^2 + b_2^2 + b_3^2 + b_4^2 \neq 0$. If x_1, x_2, x_3, x_4 and $y_1.y_2, y_3, y_4$ are variables (indeterminates) satisfying*

$$(b_1^2+b_2^2+b_3^2+b_4^2)(x_1^2+x_2^2+x_3^2+x_4^2) = y_1^2+y_2^2+y_3^2+y_4^2, \tag{12.9}$$

then one can find a solution for each x_i as a rational linear combination of the y_i's. Indeed,

$$\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4\end{bmatrix} = \frac{1}{n}\begin{bmatrix} b_1y_1+b_2y_2+b_3y_3+b_4y_4\\ -b_2y_1+b_1y_2+b_4y_3-b_3y_4\\ -b_3y_1-b_4y_2+b_1y_3+b_2y_4\\ -b_4y_1+b_3y_2-b_2y_3+b_1y_4\end{bmatrix}.$$

Proof: The following assignment of y_i's (the other identity mentioned in Lemma 22.5.2) satisfies (12.9):

$$\begin{aligned}
y_1 &= b_1x_1-b_2x_2-b_3x_3-b_4x_4,\\
y_2 &= b_2x_1+b_1x_2-b_4x_3+b_3x_4,\\
y_3 &= b_3x_1+b_4x_2+b_1x_3-b_2x_4, \text{ and}\\
y_4 &= b_4x_1-b_3x_2+b_2x_3+b_1x_4.
\end{aligned}$$

This assignment is equivalent to the matrix equation

$$\begin{bmatrix} b_1 & -b_2 & -b_3 & -b_4\\ b_2 & b_1 & -b_4 & b_3\\ b_3 & b_4 & b_1 & -b_2\\ b_4 & -b_3 & b_2 & b_1\end{bmatrix}\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4\end{bmatrix} = \begin{bmatrix} y_1\\ y_2\\ y_3\\ y_4\end{bmatrix}.$$

The determinant of the above 4×4 matrix, call it (b_{ij}), can be calculated by

$$|(b_{ij})|^2 = |(b_{ij})(b_{ij})^T| = \begin{vmatrix} n & 0 & 0 & 0 \\ 0 & n & 0 & 0 \\ 0 & 0 & n & 0 \\ 0 & 0 & 0 & n \end{vmatrix} = n^4.$$

The determinant of (b_{ij}) is n^2, non-zero, and hence the system is invertible; that is, each x_i can be solved for in rational terms of the y_i's. Since

$$(b_{ij})(b_{ij})^T = nI_4,$$

then

$$(b_{ij})^{-1} = \frac{1}{n}(b_{ij})^T,$$

and hence

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \frac{1}{n} \begin{bmatrix} b_1 & b_2 & b_3 & b_4 \\ -b_2 & b_1 & b_4 & -b_3 \\ -b_3 & -b_4 & b_1 & b_2 \\ -b_4 & b_3 & -b_2 & b_1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix},$$

and multiplying out proves the lemma. □

Proof of the Bruck–Ryser–Chowla theorem (Theorem 12.3.1): Since the theorem has already been proven for v even (Theorem 12.2.7), suppose that v is odd. Let $n = k - \lambda$ (which is positive by Lemma 12.2.4) and by Lagrange's Theorem, let b_1, b_2, b_3, b_4 be integers with $n = b_1^2 + b_2^2 + b_3^2 + b_4^2$. It remains to show that the equation

$$z^2 = nx^2 + (-1)^{(v-1)/2}\lambda y^2 \tag{12.10}$$

has a solution in integers x, y, z, not all zero.

Rewriting (12.7) or (12.8) with $L_j = \sum_{i=1}^{v} a_{ij}x_i$, obtain

$$\sum_{j=1}^{v} L_j^2 = n\sum_{i=1}^{v} x_i^2 + \lambda\left(\sum_{i=1}^{v} x_i\right)^2. \tag{12.11}$$

Examine two cases, $v \equiv 1 \pmod 4$ and $v \equiv 3 \pmod 4$.

Case 1: $v \equiv 1 \pmod 4$.

In this case, one has to show the existence of integers x, y, z, not all zero, so that

$$z^2 = nx^2 + \lambda y^2. \tag{12.12}$$

As in (12.9), for each $i = 0, 1, \ldots, (v-5)/4$, write (by Lemma 22.5.2)

$$n(x_{4i+1}^2 + x_{4i+2}^2 + x_{4i+3}^2 + x_{4i+4}^2) = y_{4i+1}^2 + y_{4i+2}^2 + y_{4i+3}^2 + y_{4i+4}^2, \tag{12.13}$$

converting all but x_v four at a time, the x_i^2's in (12.11), which now reads

$$\sum_{j=1}^{v} L_j^2 = \sum_{i=1}^{v-1} y_i^2 + nx_v^2 + \lambda\left(\sum_{i=1}^{v} x_i\right)^2. \tag{12.14}$$

Set $x_v = y_v$ and $w = \sum_{i=1}^{v} x_i$; then (12.14) becomes

$$\sum_{j=1}^{v} L_j^2 = \sum_{i=1}^{v-1} y_i^2 + ny_v^2 + \lambda w^2. \tag{12.15}$$

By Lemma 12.3.4 (again applied to four at a time) express the x_i's in terms of y_i's so that each of the L_j's and w become rational linear forms in the y_i's (which are independent—recall that (b_{ij}) was non-singular). By Lemma 12.3.3, let $M_1(y_2, \ldots, y_m)$ be a linear form so that

$$L_1\left(M_1(y_2, \ldots, y_m), y_2, \ldots, y_m\right)^2 = M_1(y_2, \ldots, y_m)^2.$$

Since (12.15) holds for any y_1, it also holds for the special case when $y_1 = M_1$, and thus (12.15) reduces to

$$\sum_{j=2}^{v} L_j^2 = \sum_{i=2}^{v-1} y_i^2 + ny_v^2 + \lambda w^2. \tag{12.16}$$

where now each of the terms $L_2, \ldots, L_v$ and w become rational linear forms in $y_2, \ldots, y_v$ (throughout, y_1 is replaced with M_1, a rational linear form not containing y_1).

Again apply Lemma 12.3.3 replacing y_2 with M_2, some rational linear form in only the variables $y_3, \ldots, y_v$, and continue doing so to reduce (12.16) until

$$L_v^2 = ny_v^2 + \lambda w^2, \tag{12.17}$$

where L_v and w are rational linear forms in y_v. Let $L_v = \frac{\alpha}{\beta}y_v$ and $w = \frac{\gamma}{\delta}y_v$, where $\alpha, \beta, \gamma, \delta$ are integers, $\beta \neq 0$, and $\delta \neq 0$. Setting $y_v = \beta\delta$, equation (12.17) becomes

$$(\alpha\delta)^2 = n(\beta\delta)^2 + \lambda(\gamma\beta)^2,$$

the desired form in (12.12). Observe that since the y_i's were independent, L_v did not vanish in the process of specializing and reduction (nor did w), concluding the proof of the case $v \equiv 1 \pmod 4$.

Case 2: $v \equiv 3 \pmod 4$.

In this case, it remains to show the existence of integers x, y, z, not all zero, so that

$$z^2 = nx^2 - \lambda y^2. \tag{12.18}$$

Let x_{v+1} be a new indeterminate, and add nx_{v+1}^2 to both sides of (12.11) yielding

$$\sum_{j=1}^{v} L_j^2 + nx_{v+1}^2 = n\sum_{i=1}^{v+1} x_i^2 + \lambda\left(\sum_{i=1}^{v} x_i\right)^2, \tag{12.19}$$

where w and the L_i's are linear forms in $x_, \ldots, x_v$. Now apply (12.13), as before, to groups of x_i's taken four at a time; this time however, all x_i's get 'used', leaving

$$\sum_{j=1}^{v} L_j^2 + nx_{v+1}^2 = \sum_{i=1}^{v+1} y_i^2 + \lambda\left(\sum_{i=1}^{v} x_i\right)^2, \tag{12.20}$$

where now each L_i and x_i is a linear form in $y_1, \ldots, y_v, y_{v+1}$.

Similar to the first case, apply Lemma 12.3.3 successively to eliminate the linear forms $L_1, L_2, \ldots, L_v$ to arrive at

$$nx_{v+1}^2 = y_{v+1}^2 + \lambda w^2, \tag{12.21}$$

where x_{v+1} and w are now linear forms in y_{v+1}. If $x_{w+1} = \frac{\alpha}{\beta}y_{v+1}$ and $w = \frac{\gamma}{\delta}y_{v+1}$, then (12.21) simplifies to

$$(\beta\delta)^2 = n(\alpha\delta)^2 - \lambda(\gamma\beta)^2,$$

the desired form of (12.18), concluding the proof of the case $v \equiv 3 \pmod 4$, and hence the theorem. $\square$

Example 12.3.5. *A* $(43, 7, 1)$*-SBIBD does not exist.*

Proof: Perhaps it is worth noting first that with $v = b = 43$, $k = r = 7$, $\lambda(v-1) = k(k-1)$ and so one can not trivially discount the existence of such a design. However, by the Bruck–Ryser–Chowla theorem, one needs integers x, y, z not all zero, so that

$$x^2 = (k - \lambda)y^2 + (-1)^{(v-1)/2}\lambda z^2,$$

that is,

$$x^2 = 6y^2 + (-1)^{21}z^2 = 6y^2 - z^2.$$

Looking at this last equation modulo 3, one needs x, z so that

$$x^2 + z^2 \equiv 0 \pmod 3.$$

It is nearly trivial to check that there are no such x, z. □

Recall from Section 11.4, the Bruck–Ryser theorem says:

Theorem 11.4.11 [Bruck–Ryser] If $n \equiv 1 \pmod 4$ or $n \equiv 2 \pmod 4$ and a projective plane exists of order n then n is the sum of (integral) squares.

A proof of the Bruck-Ryser theorem is given in many popular texts (see [454] or [490], for example); in the interest of completeness of exposition, here is one proof based on the Bruck-Ryser-Chowla theorem (Theorem 12.3.1).

Proof of the Bruck-Ryser theorem: Using $v = n^2 + n + 1$, $k = n + 1$, and $\lambda = 1$, then $k - \lambda = n$ (conveniently). The number

$$n^2 + n + 1 = n(n + 1) + 1$$

is certainly odd, so the Bruck-Ryser-Chowla theorem (Theorem 12.3.1) guarantees the existence of integers x, y, and z, not all zero, so that

$$z^2 = nx^2 + (-1)^{(n^2+n)/2}y^2. \tag{12.22}$$

If $n = 4m + 1$, then

$$(n^2 + n)/2 = (16m^2 + 12m + 2)/2 = 8m^2 + 6m + 1$$

is odd; similarly, if $n = 4m + 2$,

$$(n^2 + n)/2 = (16m^2 + 20m + 6)/2 = 8m^2 + 10m + 3,$$

also an odd number. In either case, (12.22) becomes

$$z^2 = nx^2 - y^2.$$

However, $nx^2 = y^2 + z^2$ implies $n = (y/x)^2 + (z/x)^2$, a sum of two rational squares. So by Corollary 22.4.6, n is a sum of integral squares. $\square$

Chapter 13

Arcs, ovals, and polarities

13.1 Introduction

Arcs and complete arcs in projective planes over fields were studied by Segre [785, 786].

Definition 13.1.1. An *arc* in a FPP is a set of vertices no three of which are collinear. An arc consisting of k vertices is called a k*-arc*.

In higher dimensional projective spaces, such a set (with no three collinear) is called a *cap*. A quadrangle is a 4-arc. Each line in a FPP contains at most 2 points of any arc. If a line in $\mathcal{P}$ intersects an arc in two points, the line is said to be a *chord* (or *secant* to the arc); if the line intersects an arc in a single point, it is *tangent* to the arc; if the line does not intersect the arc, it is an *outside line*.

Lemma 13.1.2. *[Folklore?] If A is a k-arc in a FPP of order q, then $k \leq q+2$. Through each point in A there are exactly $q+2-k$ lines tangent to A.*

Proof: Let A be a k-arc in a FPP and fix a point $x \in A$. The $q+1$ lines containing x are either secants or tangents to A. Each of the $q+1$ lines containing x can contain at most one other point of A, so $k \leq q+2$. Since any two points in a FPP determine a unique line, $k-1$ of the lines containing x are secants to A; hence there are $(q+1)-(k-1) = q+2-k$ tangents to A containing x. $\square$

The special case of Lemma 13.1.2 with $k = q+1$ is worth noting separately.

Corollary 13.1.3. *Let A be a k-arc in a FPP of order q. Then $k = q+1$ if and only if through any point of A there is a unique tangent to the arc.*

For more on maximal arcs and smallest maximal arcs, see [545].

There have been different definitions of an *oval* in a FPP. In some texts (e.g., [250] p. 147) an oval is defined to be an arc whose every point has a unique tangent, and in others (e.g., [490] p. 241), an oval is simply defined to be a $(q+1)$-arc. By Corollary 13.1.3, the two definitions are equivalent.

A $(q+2)$-arc in a FPP is often called a *hyperoval*. Bose [137] (see [250]) showed that hyperovals can only occur in planes of even order.

Theorem 13.1.4 (Bose, 1947 [137]). *If A is a $(q+2)$-arc in a FPP of order q, then q is even.*

Proof: Let A be a $(q+2)$-arc in a FPP of order q. If $x \in A$, there are $q+1$ lines containing x, each of which intersects A in one additional point. It follows that for any fixed $y \notin A$, the lines containing y which intersect A partition the points of A into pairs. So $q+2$, and hence q must be even. □

Natural questions might arise: Can large arcs always be constructed? Can respective $q+1$ and $q+2$ bounds always be attained? That is, for any FPP does there exist an oval/hyperoval? Is it easy to preclude some cases? Does construction of an oval depend on a field, that is, do available constructions work only on FPPs coordinatized over a finite field? Is the situation different for other planes? The study of ovals has contributed to studying existence of certain FPPs (e.g., for the FPP of order 10, see [586]) and ovals are used in error-correcting codes (see, e.g., [400, p. 10ff]).

Loosely speaking, a collection of points in a FPP based on a finite field is called a *conic* if they 'satisfy' an irreducible polynomial equation over the field, this equation having degree 2. It seems that a conic is a special kind of oval? This definition is made more precise in what follows.

Definition 13.1.5. A (k,m)-arc in a finite projective plane is a set of k vertices, at most m of which lie on any line.

So a k-arc is, by definition, a $(k,2)$-arc.

Lemma 13.1.6. *If a (k,m)-arc exists in a FPP of order q, then*

$$k - 1 \leq (q+1)(m-1), \tag{13.1}$$

where equality holds if and only if every line which intersects the arc does so in exactly m vertices.

Proof: Let A be a (k, m)-arc in a FPP $\mathcal{P}$ of order q and fix a point $x \in A$. Consider the $q+1$ lines of $\mathcal{P}$ containing x; each of these contain at most $m-1$ other points of A. $\square$

A (k, m)-arc in a FPP of order q is said to be maximal if equality holds in (13.1), that is, the upper bound for k is realized. Maximal and complete agree for $(k, 2)$-arcs.

Lemma 13.1.7. *If there exists a maximal (k, m)-arc in a FPP of order q, then m divides q or $m = q + 1$.*

For $m = 1$, a $(k, 1)$-arc in a FPP exists for only $k = 1$, namely, a point. When $m = q$, a maximal (k, q)-arc could be formed by deleting all points from one line in the FPP (giving an affine plane on q^2 points), and when $m = q + 1$, a maximal arc would be the whole plane. Say that a maximal (k, m)-arc is non-trivial if $1 < m < q$.

For any $m \geq 2$, Ball, Blokhuis, and Mazzocca [46] showed that there are no maximal non-trivial (k, m)-arcs in planes of odd order. The proof is too detailed to go into here. Denniston [251] found maximal arcs in even planes.

13.2 Terminology for morphisms on planes

A *homomorphism* from a projective plane $\mathcal{P}$ to a projective plane $\mathcal{Q}$ is a map α from the vertices *and lines* of $\mathcal{P}$ to the vertices and lines of $\mathcal{Q}$ which takes points to points and lines to lines so that incidence is preserved, that is, if a point P lies on a line ℓ in $\mathcal{P}$, then $\alpha(P)$ lies on $\alpha(\ell)$ in $\mathcal{Q}$.

An *anti-homomorphism* is like a homomorphism, except points go to lines and lines to points; in this case, the incidence is reversed. An *isomorphism* is a one-to-one and onto homomorphism, and analogously, an *anti-isomorphism* is a bijective anti-homomorphism. Two planes $\mathcal{P}$ and $\mathcal{Q}$ are *isomorphic* if there exists an isomorphism from $\mathcal{P}$ to $\mathcal{Q}$.

An *automorphism* of a projective plane $\mathcal{P}$ is an isomorphism from $\mathcal{P}$ onto itself. A *collineation* of $\mathcal{P}$ is a permutation of the points that takes lines to lines. In [490], a collineation is defined to be an automorphism, however in [250], these are defined differently (as here), apparently since a collineation is only defined to be a map of the vertices that induces an automorphism. An automorphism need not, *a priori*, induce a collineation, however, special automorphisms do (see [250, pp. 9,31] for the subtle differences).

An *anti-automorphism* (also called a *correlation*) of a projective plane $\mathcal{P}$ is an anti-isomorphism from $\mathcal{P}$ to itself. Note that a projective plane has an anti-automorphism if the plane is self-dual. A *polarity* is an anti-automorphism of order 2. One more notion attracting a great deal of attention in the literature is that of a *perspectivity*—an automorphism of a projective plane that fixes a line pointwise, called the *axis* (of the perspectivity). It is interesting to note (see [490, Thm 4.9]) that for a perspectivity different from the identity, there is a point, called the *center* of the perspectivity, so that each line containing the center is also fixed (but not necessarily point-wise) are fixed, and these are the only structures that are fixed (not even the points on different fixed lines) by the perspectivity. If the center is on the axis, the perspectivity is called an *elation*, otherwise it is called a *homology*.

Recall that for a finite projective plane $\mathcal{P} = (P, \mathcal{L})$, an *anti-automorphism* (or *correlation*) is a bijection α from P onto $\mathcal{L}$ and $\mathcal{L}$ onto P so that for every point $A \in P$ and line $\ell \in \mathcal{L}$, $A \in \ell$ if and only if $\alpha(\ell) \in \alpha(A)$. A FPP admits a correlation if and only if only the FPP is self-dual (and such planes are not completely classified, as far as I know). Note that an affine plane does not admit a correlation, since if k and ℓ are parallel, the line $\alpha(k) \vee \alpha(\ell)$ has no preimage.

A *polarity* is an anti-automorphism of order 2, that is, $\alpha\alpha$ is the identity. Hence, a polarity is a bijection from points to lines and from lines to points that satisfies $A \in \ell$ if and only if $\alpha(\ell) \in \alpha(A)$. A polarity can also be seen as a bijection α between points and lines that satisfies $A \in \alpha(B)$ if and only if $B \in \alpha(A)$. It is not known whether or not a polarity exists for an arbitrary self-dual plane.

In the case of a finite projective plane over a field using homogeneous coordinates, the map defined by $\alpha((x, y, z)) = [x, y, z]$ and $\alpha([x, y, z]) = (x, y, z)$ is a polarity. In general, does any (not necessarily a field plane) FPP admit a polarity?

Given a polarity α on a FPP $\mathcal{P}$, a point A (or line ℓ) is called *absolute* if and only if $A \in \alpha(A)$ (or $\alpha(\ell) \in \ell$). The main result of the next section is a result due to Baer [40]: if a FPP of order q admits a polarity, then it contains at least $q + 1$ absolute points. Note that a projective plane over an infinite field with a polarity may have no absolute points (*e.g.*, consider the field $\mathbb{R}$ and the polarity that pairs up (x, y, z) and $[x, y, z]$; then $x^2 + y^2 + z^2 = 0$ has no non-trivial solutions). The case for FPPs over a finite field is considered in Section 13.5, but first the case for a general FPP is considered.

13.3 Polarities and absolute points in a general FPP

The following results are due to Baer [40], with some proofs taken from [490, pp. 239–241].

Lemma 13.3.1 (Baer, 1946 [40])**.** *Let α be a polarity of a projective plane $\mathcal{P}$. Then (i) every absolute point is on a unique absolute line, and (ii) every absolute line contains a unique absolute point.*

Proof: Only (i) is shown, since (ii) follows by duality. Let A be an absolute point; by definition, A is on $\alpha(A)$, and $\alpha(\alpha(A)) = A$ lies on $\alpha(A)$, so $\alpha(A)$ is an absolute line, and A lies on this absolute line.

Suppose that A lies on an absolute line m. Since m is absolute, $\alpha(m) \in m$. On the other hand, $A \in m$ implies that $\alpha(m) \in \alpha(A)$. So both A and $\alpha(m)$ lie on $\alpha(A)$. If $A \neq \alpha(m)$, then because two points uniquely determine a line, $m = \alpha(A)$. If $A = \alpha(m)$, then applying α, again it follows that $\alpha(A) = \alpha^2(m) = m$. □

By Lemma 13.3.1, in a FPP, if there exists an absolute point (or line), then some line contains at most one absolute point, and so not all points are absolute. The next goal is to show the opposite: at least one point is absolute. One way to show this is to use eigenvalues.

Recall that if M is any incidence matrix for a FPP with $v = n^2 + n + 1$ vertices, (where J_v denotes the $v \times v$ matrix of 1s, and I_v is the $v \times v$ identity matrix), then Corollary 12.2.3 (with $\lambda = 1$, and $r = n + 1$) says,

$$MM^T = J_v + nI_v.$$

Lemma 13.3.2. *For $v = n^2 + n + 1$, let $C = J_v + nI_v$. Then the eigenvalues of C are $v + n$ (occurring once) and n (with multiplicity $v - 1$).*

Proof: Let $\mathbf{1}$ be the $v \times 1$ vector of all 1s. Then $C\mathbf{1} = v\mathbf{1} + n\mathbf{1}$ shows that $v+n$ is an eigenvalue. It is not difficult to verify that the $v - 1$ vectors of the form $\mathbf{x}_i = (0, 0, ...0, 1, -1, 0, \ldots, 0)^T$ (where 1 appears in position $i = 1, \ldots, v - 1$) are eigenvectors associated with the eigenvector n. □

Theorem 13.3.3 (Baer, 1946 [40])**.** *Let $\mathcal{P}$ be a FPP of order q and let α be a polarity of $\mathcal{P}$. Then α has at least one absolute point. If q is not a square, then α has exactly $q + 1$ absolute points.*

Proof: Let $v = q^2 + q + 1$, and let $X_1, \ldots, X_v$ be an arbitrary labelling of the points. Label the lines $m_1, \ldots, m_v$ so that for each i, $\alpha(X_i) = m_i$. Since α is a polarity, $X_i \in m_j$ if and only if $X_j = \alpha(m_j) \in \alpha(X_i) = m_i$, and so the incidence matrix M determined by this labelling is symmetric. Note also that a diagonal entry m_{ii} of M is 1 if and only if X_i is absolute. Hence, the trace of M is the number of absolute points (with respect to α).

By the comment preceding Lemma 13.3.2 (with n replaced by q) and the fact that M is symmetric, $M^2 = J_v + qI_v$. So by Lemma 13.3.2, M^2 has eigenvalues $v + q = (q+1)^2$ (with multiplicity 1) and q (with multiplicity $v - 1 = q^2 + q$). If λ is an eigenvalue of M, then λ^2 is an eigenvalue of M^2, so the only choices for λ are $\pm(q+1)$ and $\pm\sqrt{q}$.

Since each row or column of M has $q+1$ non-zero entries, the $v \times 1$ vector $(1, 1, \ldots, 1)^T$ is an eigenvector of M associated with the eigenvalue $q+1$. [It is left to the reader to show that $-(q+1)$ is not an eigenvalue.] If $\sqrt{q}$ is an eigenvalue for M with multiplicity r and $-\sqrt{q}$ is an eigenvalue with multiplicity s, then $r + s = q^2 + q$. Since the trace of M is the sum of its eigenvalues,

$$\mathrm{tr}(M) = q + 1 + (r - s)\sqrt{q}. \tag{13.2}$$

If q is not a square, the only integer solution to (13.2) is when $r = s$, in which case $\mathcal{P}$ has exactly $q+1$ absolute points. If q is a square, say $q = k^2$, the right side of (13.2) becomes $k^2 + 1 + (r-s)k$. If this last expression is 0, then $1 = (s-r)k - k^2 = k(s-r-k)$, which is impossible, so when q is a square, $\mathrm{tr}(M) \neq 0$, and so α has at least one absolute point. $\square$

To prove Baer's main theorem (Theorem 13.3.6, below, which says that if a FPP of order q admits a polarity, then it contains at least $q+1$ absolute points), one more observation is needed.

Lemma 13.3.4 (Baer, 1946 [40])**.** *Let $\mathcal{P}$ be a FPP of order q and let α be a polarity of $\mathcal{P}$. For any non-absolute line m of α, the number of absolute points on m is congruent to $q+1$ modulo 2.*

Proof: Let m be a non-absolute line of α; then for any point X on m, $\alpha(X) \neq m$. For each X on m, define $\theta(X) = m \wedge \alpha(X)$. Note that $\theta(X) = X$ iff $X = m \wedge \alpha(X)$ if and only if $X \in \alpha(X)$; in other words, a fixed point of θ is the same as an absolute point of α.

Note that θ is a permutation of points on m and

$$\theta(\theta(X)) = m \wedge \alpha(\theta(X))$$

$$
\begin{aligned}
&= m \wedge \alpha(m \wedge \alpha(X)) \\
&= m \wedge (\alpha(m) \vee X) \\
&= X \qquad\qquad\qquad\qquad \text{(since } X \in m \text{ and } \alpha(m) \notin m\text{).}
\end{aligned}
$$

So θ has order 2, and so points not fixed by θ are partitioned into disjoint pairs $\{X, \theta(X)\}$. Thus, the number of fixed points of θ is the same parity as the number of points on m. In other words, the number of absolute points on m has the same parity as $q + 1$. □

Corollary 13.3.5 (Baer, 1946 [40])**.** *Let $\mathcal{P}$ be a FPP of even order and α a polarity on $\mathcal{P}$. Then each line contains at least one absolute point.*

Proof: If m is absolute, then $\alpha(m) \in m$, and $\alpha(m)$ is an absolute point. If m is non-absolute, m contains an odd number of points, and so by Lemma 13.3.4, m has at least one absolute point. □

Theorem 13.3.6 (Baer, 1946 [40])**.** *Let $\mathcal{P}$ be a FPP of order q, and let α be a polarity of $\mathcal{P}$. Then α has at least $q + 1$ absolute points.*

Proof: Let $a(\alpha)$ denote the number of absolute points of α. By Theorem 13.3.3, $a(\alpha) \geq 1$. Instead of dividing the proof into two cases where q is or is not a square, consider the cases where q is even or odd:

Case 1: Suppose that q is even. By Corollary 13.3.5, each line contains an absolute point. Let t be the number of flags (pairs (X, m) where X is a point on the line m) such that the point is absolute. Then $t = a(\alpha)(q + 1)$. Every line is in at least one such flag and so $t \geq q^2 + q + 1$. Thus

$$a(\alpha)(q + 1) \geq q^2 + q + 1 > q^2 + q,$$

and so $a(\alpha) > q$. Since $a(\alpha)$ is an integer, $a(\alpha) \geq q + 1$.

Case 2: Suppose that q is odd. Since $a(\alpha) \geq 1$, let X be an absolute point of α. By Lemma 13.3.1, X is incident with q non-absolute lines. Since $q + 1$ is even, by Lemma 13.3.4, the number of absolute points on each of these non-absolute lines is even, and so at least two. This means that each of these non-absolute lines contains an additional absolute point, giving at least $q + 1$ in all. □

Note that by Theorem 13.3.3, when q is not a square, the bound of $q + 1$ in Theorem 13.3.6 is best possible.

13.4 Polarities and the friendship theorem

Recall that Theorem 11.6.2 says that if G is a graph on at least 3 vertices so that between any two vertices, there is a unique path of length 2, then G is a friendship graph. Here is the original Erdős–Rényi–Sós [315] proof, which uses incidence graphs for FPPs and polarities:

Proof: Let G be a graph on n vertices so that between any two vertices, there exists a unique path of length 2. Then G contains no C_4 and each edge is contained in at most one triangle.

Case 1: Suppose that there exists a vertex $v \in V(G)$ with $\deg(v) = n-1$. In this case, since each edge is in at most one triangle, each vertex in $V(G)\backslash\{v\}$ has degree at most one; since between v and any other vertex there is a path of length 2, all vertices in $V(G)\backslash\{v\}$ have degree at least one. So $V(G)\backslash\{v\}$ induces a matching, and so $G = F_{\frac{n-1}{2}}$ as desired.

Case 2: Suppose that there is no vertex with degree $n-1$.

Claim: G is regular. Let $x \in V(G)$ be of maximum degree. Since no vertex has degree $n-1$, let y be a vertex not adjacent to x. Let $N(x) = \{v_1, \ldots, v_k\}$ and $N(y) = \{w_1, \ldots, w_\ell\}$. If the degree of any vertex is 1, then between that vertex and its neighbour there is no path of length 2, so $\delta(G) \geq 2$. Thus $k \geq 2$ and $\ell \geq 2$.

Since there exists a unique path of length 2 between x and y, $|N(x) \cap N(y)| = 1$ (if the intersection of neighbourhoods is larger, there is more than one x-y path of length 2). Suppose that $v_1 = w_1$ is the common neighbour of x and y. Since there is a unique 2-path from x to v_1, there exists a unique vertex, say v_2 adjacent to both. Similarly, there is a unique vertex, say w_2, adjacent to both $v_1 = w_1$ and y. Thus, $\deg(v_1) \geq 4$, and since x is of maximum degree, $k \geq 4$.

For each $i = 3, \ldots, k$, there exists a unique 2-path from v_i to y, and so there is a unique $w_j \in N(y)\backslash\{w_1, w_2\}$ so that $\{v_i, w_j\} \in E(G)$. (If v_i is adjacent to two neighbours of y, say w_j and w_m, then a 4-cycle is formed, contrary to the property mentioned above.) So $k \leq \ell$. By the same argument, but replacing v_i with some w_j, $\ell \leq k$. So $k = \ell$ and thus $\deg(x) = \deg(y)$. Hence, any two non-adjacent vertices have the same degree (in this case, the degree is k).

For each $i = 2, \ldots, k$, v_i is not adjacent to y, so $deg(v_i) = \deg(y)$ as well, and since $\deg(x) = \deg(y)$, it follows that $\deg(v_i) = \deg(x) = k$. The same argument applied to $w_2, \ldots, w_k$ shows that for each $i = 2, \ldots, k$, $\deg(w_i) = k$.

Observe that v_k is not adjacent to v_1 (for otherwise, $v_k v_1 v_2 x$ forms a 4-cycle), so $\deg(v_1) = \deg(v_k) = k$. So any two adjacent vertices also have the same degree.

Then G is regular of degree $k \geq 4$, proving the claim.

Let $V = V(G) = \{v_1, v_2, \ldots, v_n\}$, and for each $i = 1, \ldots, n$, let $N_i = N(v_i)$, the neighbourhood of v_i, and put $\mathcal{N} = \{N_1, \ldots, N_n\}$. Form the bipartite graph on sets V and $\mathcal{N}$, where $(v_i, N_j) \in F$ iff $v_i \in N_j$.

Let $\mathcal{P} = (V, \mathcal{N})$ be a "geometry" with point set V and line set $\mathcal{N}$. Then each line contains k points and since each vertex of G is in the neighbourhood of its neighbours, each point is contained in k lines.

Claim: $\mathcal{P}$ is a finite projective plane.

Proof of claim: Any two vertices in G have exactly one common neighbour, and so for all $i \neq j$, $|N_i \cap N_j| = 1$. Thus, any two lines in P contain a unique point. For some $i \neq j$, let v_i, v_j be vertices of G. Since v_i, v_j have a unique common neighbour, say, v_k, then as points in $\mathcal{P}$, both are on the line N_k and no others. So, in $\mathcal{P}$, two points are on a unique line. So $\mathcal{P}$ satisfies the first two axioms of a projective plane.

It remains to check that $\mathcal{P}$ contains a frame (four points, no three of which are collinear). (Erdős *et al.* [315] did not give a check for this third axiom, but Wilf [924] did.) To see that $\mathcal{P}$ contains a frame, some cases are required. Since $k \geq 3$, G contains at least four vertices, and so $\mathcal{P}$ contains at least four points. (Even when $k = 2$, four points exist because no vertex has degree $n - 1$.) The cases are left as an exercise for the reader.

So $\mathcal{P}$ is a FPP of (of order $k - 1 \geq 2$). The map $v_i \leftrightarrow N_i$ is a polarity, but no point lies on its own polar (since no vertex is an element of its neighbourhood), which contradicts (Baer's) Theorem 13.3.3). Thus the assumption that G has no vertex of degree $n - 1$ leads to a contradiction, and so Case 2 never holds. $\square$

13.5 Polarities and absolute points in field planes

The main result in this section is a consequence of Segre's theorem (whose proof can also be found in [490, pp. 250–252]). Before stating Segre's theorem, a definition of "conic" is reviewed.

Definition 13.5.1. Let q be a prime power and let $\mathrm{PG}(2, q)$ be the projective plane coordinatized by the field $\mathrm{GF}(q)$. A conic in PG(2, q) is a set of points whose homogeneous coordinates (x, y, z) satisfy a homogeneous quadratic equation in three variables.

The equation in the above definition can be assumed to be irreducible over $\mathrm{GF}(q)$. The discussion of conics in FPPs is not a goal of these notes, but one theorem might deserve mention.

Theorem 13.5.2 (Segre, 1955 [784])**.** *Let $\mathcal{P}$ be a Desarguesian plane of odd order. Then every oval in $\mathcal{P}$ is a conic.*

Corollary 13.5.3. *Let $\mathcal{P}(q)$ denote the Desarguesian FPP of order q over a finite field. For any oval $\mathcal{C}$, there is a unique polarity whose absolute points are the points of $\mathcal{C}$.*

The situation for non-Desarguesian planes is different. The Hughes plane (of order 9) contains ovals that are not the absolute points of a polarity (see [490, p. 243] for details).

Chapter 14

Colouring a finite projective plane

14.1 Introduction

Finite projective planes have many "highly regular" features, and in many ways, can be considered as a "highly balanced" set system. For example, if viewed as a hypergraph, a FPP is "uniform" (every hyperedge has the same number of vertices) and regular (every vertex has the same degree). Every hyperedge intersects every other in precisely a single vertex, and any two vertices are contained in precisely one hyperedge. In some sense, these facts are first-degree measures of "uniformity" or "balance" at local levels. In this section, the "balance" in a FPP with respect to vertex colourings is examined. In some sense, there are vertex-colourings of FPPs that confirm a sense of "balance", however there are vertex colourings that reveal FPPs are not as "balanced" as one might think.

14.2 Chromatic number of a FPP

For $k \in \mathbb{Z}^+$, denote the set $\{1, 2, \ldots, k\}$ by $[1, k]$ (or simply $[k]$). A k-colouring of a set X is a function $\phi : X \to [1, k]$. When elements of X are considered as vertices, then such a k-colouring of X may be called a vertex-colouring (as opposed to colourings of pairs, or subsets, for example). It is often convenient to replace the set $[1, k]$ by a set of colours, for example, $\{\text{red}, \text{blue}\}$, or by another set of numbers, for example $\{-1, 0, 1\}$.

For a set X and a family $\mathcal{F}$ of subsets of X, a k-colouring $\phi : X \to [1, k]$ of the vertices of the set system $(X, \mathcal{F})$ is defined to be a *good k-colouring* if for every $F \in \mathcal{F}$, $|\phi^{-1}(F)| \geq 2$. The *chromatic number* of $\mathcal{F}$, denoted $\chi(\mathcal{F})$, is the least integer k so that there exists a good k-colouring of $\mathcal{F}$. (In some texts, $\chi(\mathcal{F})$ is called the *weak chromatic number*, as opposed to the *strong chromatic number* $\chi_s(\mathcal{F})$, where each member $F \in \mathcal{F}$ not only must receive at least two colours, but must be totally multicoloured.) So the chromatic number of a projective plane is the least number of colours used to colour the points so that each line receives at least two colours.

Theorem 14.2.1. *The chromatic number of the Fano plane is 3, and for any finite projective plane $\mathcal{P}$ of order $n > 2$, $\chi(\mathcal{P}) = 2$.*

Exercise 266. *Prove Theorem 14.2.1.*

14.3 Blocking sets

For a given set system $(X, \mathcal{F})$ (where $\mathcal{F}$ is a collection of subsets of X) and $Y \subset X$, say that Y *blocks* $\mathcal{F}$ if for every $F \in \mathcal{F}$, $Y \cap F \neq \emptyset$ holds. If $Y \subset X$ blocks $\mathcal{F}$, then one might be inclined to call Y a "blocking set", however convention now dictates a slightly stronger definition.

Definition 14.3.1. For a set X and a family $\mathcal{F} \subseteq \mathcal{P}(\mathcal{X})$, a subset $B \subset X$ is called a blocking set for $\mathcal{F}$ if and only if for every $F \in \mathcal{F}$,

$$0 < |F \cap B| < |F|.$$

If B is a blocking set for $\mathcal{B}$ say that B blocks $\mathcal{F}$.

Thus, a blocking set in a projective plane is a set of points so that every line of the plane contains points in the set and points not in the set.

By Theorem 14.2.1, every finite projective plane except the Fano plane has a blocking set. Blocking sets in FPPs have created considerable interest (for example, see [160], [163], or [858] for many of the basic results and further references). In [74] all blocking sets (not just minimal) in the FPP of order four are characterized.

A set of points in a FPP that intersect lines in at most one point is called a *nucleus* and nuclei are well studied (*e.g.*, Wettl [920] gives results and references).

Recall from Definition 11.4.18 that a Baer subplane intersects all lines of the original plane, and so as noted in 1970 by Bruen [159], a Baer subplane (if one exists) is a blocking set.

Exercise 267. *Prove that in a finite projective plane of order $n > 2$, there exists a blocking set with $2n$ elements.*

A blocking set that contains no smaller blocking set is called *minimal*. As an example (mentioned by Szőnyi [855]), in a FPP of order q (not a square), let A, B, C be non-collinear points; the set of $3(q-1)$ points on the triangle ABC but not including A, B or C, forms a minimal blocking set.

In 1970, Bruen [159] gave a tight lower bound for the order of a blocking set when the order q of the plane is a square; he later gave a proof for general q.

Theorem 14.3.2 (Bruen, 1971 [160])**.** *Let B be a blocking set in a finite projective plane Π of order q. Then*

$$q + \sqrt{q} + 1 \leq |B|,$$

with equality holding if and only if B is the set of points in a Baer subplane of Π.

Proof: The proof given here is due to Bruen and Silverman [162]; this proof can also be found in [65] and [109].
Claim 1: $|B| \geq q + 1$.
Proof of Claim 1: Each point in B blocks only $q+1$ lines in Π, and so at most $|B|(q+1)$ lines in total. If $|B| \leq q$, then at most $q(q+1) < q^2 + q + 1$ lines are blocked, which proves Claim 1.

So let $\alpha \geq 1$ be such that $|B| = q + \alpha$.

Claim 2: Every line in Π contains at most α points of B.
Proof of Claim 2: In hope of a contradiction, suppose that some line ℓ in Π has at least $\alpha + 1$ points of B. Since B is a blocking set, some point x is on ℓ but not in B. Also since B is a blocking set, each of the other q lines through x contain a point of B. Thus the total of points in B is at least $\alpha + 1 + q > q + \alpha$, the desired contradiction, proving Claim 2.

Let k be the maximum number of points of B in any one line of Π, (so $k \leq q$) and for $i = 1, \ldots, k$, define ℓ_i to be the number of lines that intersect

B in i points. (Since B is a blocking set, $\ell_0 = 0$.) Since every line of Π is blocked,

$$\sum_{i=1}^{k} \ell_i = q^2 + q + 1. \tag{14.1}$$

Examining incidences between points in B and lines in Π,

$$\sum_{i=1}^{k} \ell_i \cdot i = |B|(q+1) = (q+\alpha)(q+1). \tag{14.2}$$

Similarly, by counting incidences of ordered pairs of points in B and lines in Π,

$$\sum_{i=1}^{k} \ell_i \cdot i(i-1) = |B|(|B|-1) = (q+\alpha)(q+\alpha-1). \tag{14.3}$$

For each $i \in \{1, \ldots, k\}$,

$$\ell_i(i-1)(\alpha - i) = -\alpha\ell_i + \alpha\ell_i \cdot i - \ell_i \cdot i(i-1),$$

and so by equations (14.1), (14.2) and (14.3),

$$\begin{aligned}\sum_{i=1}^{k} \ell_i(i-1)(\alpha-i) =& -\alpha\sum_{i=1}^{k}\ell_i + \alpha\sum_{i=1}^{k}\ell_i \cdot i - \sum_{i=1}^{k}\ell_i \cdot i(i-1)\\ =& -\alpha(q^2+q+1) + \alpha(q+\alpha)(q+1) - (q+\alpha)(q+\alpha-1)\\ =& \alpha^2 q - 2\alpha q - q^2 + q.\end{aligned}$$

Since each term of the left sum above is (by Claim 2) non-negative, so also is the right side. Hence, $\alpha^2 - 2\alpha - q + 1 \geq 0$, or equivalently, $(\alpha-1)^2 \geq q$. Thus, $\alpha \geq \sqrt{q} + 1$, finishing the proof of the lower bound for a blocking set.

Suppose equality in the last inequality holds. Then q is a perfect square, $\alpha = \sqrt{q}+1$ and $\sum_{i=1}^{k} \ell_i(i-1)(\alpha-i) = 0$. For each $i > 1$, $\ell_i(\alpha-i) = 0$, and so for $i \notin \{1, a\}$, $\ell_i = 0$. Thus any line in Π intersects B in either one point or in α points. If two lines, each with α points in B, intersect outside of B, say at a point x, then since each of the remaining $q-1$ lines through x intersect B this gives at least $2\alpha + q - 1 = 2(\sqrt{q}+1) + q - 1 > q + \sqrt{q} + 1$ points, a contradiction. So any two "lines" in B intersect in B. Since $\lceil\sqrt{q}\rceil \geq 2$, then $|B| \geq q+3$, and thus B contains a quadrangle. Hence B is the set of points in a Baer subplane. $\square$

The complement of a minimal blocking set is a maximal blocking set, so the size of any blocking set in a FPP of order n is bounded by

$$n + \sqrt{n} + 1 \leq |B| \leq n^2 - \sqrt{n}.$$

Theorem 14.3.3 (Bruen–Thas, 1977 [163]). *If B is a minimal blocking set in a finite projective plane of order q, then*

$$|B| \leq q\sqrt{q} + 1,$$

with equality holding only when q is a perfect square and the complement of B is a Baer subplane.

Some authors (e.g., Bruen [160] and Bierbrauer [93, 92]) studied blocking sets to gain information on the existence of a FPP of order 10.

Blokhuis and Brouwer [110] showed that in the Desarguesian FPPs of odd order $q \geq 7$, q not a square, $q \neq 27$, blocking sets have at least $q+\sqrt{2q}+1$ points. In 1992, Szőnyi [855] showed that for $q \equiv 1 \pmod 4$, in $\mathrm{PG}(2, q)$, there are minimal blocking sets with more than $q \log_2(q/2)$ points, and gave a similar result for $q \equiv 3 \pmod 4$.

The parity of a blocking set (of many set systems, including FPPs) was studied in [322]. Szőnyi [856] showed that if q is a power of a prime p and $\mathcal{P}$ is a Desarguesian plane of order q, then if B is a blocking set with $|B| < \frac{3(q+1)}{2}$, then B intersects every line of $\mathcal{P}$ in 1 modulo p points. (Szőnyi also shows that in the case $q = p^2$, a nontrivial blocking set either contains a Baer subplane or has at least $\frac{3(q+1)}{2}$ points. Other proofs of known theorems are also then given.)

In 1990, Berardi and Innamorati [75] found all minimal blocking sets in $\mathrm{PG}(2, 5)$; there are one each of orders 9, 11, 12, and six of order 10. In 1991, Batten [64] looked at collineation groups fixing a blocking set B that is transitive on flags. (A flag in a blocking set S is a point P in S and the restriction in S to a line containing P.) In 1993, it was shown [112] that planes of order $n \geq 25$ have no blocking set of size $n\sqrt{n}$.

In 2002, Polverino and Storme [729] looked at 'small' (i.e., $|B| \leq 3(p + 1)/2$) minimal blocking sets in $\mathrm{PG}(2, q^3)$. In the same year, Polito and Polverino [728] gave results for blocking sets in planes of the form $\mathrm{PG}(2, q^4)$. If $q = p^h$ where $p > 7$, small minimal blocking sets satisfy $|B| = q^3+q^2+cq+1$ where $c \in \{0, 1\}$. In 1994, Blokhuis found that $PG(2, p)$ has no small blocking sets, and in $\mathrm{PG}(2, p^2)$, blocking sets are Baer subplanes.

14.4 Sets that do not block all lines

In the paper by Blokhuis [109, Thm. 25], a theorem is given as folklore; its proof follows by an argument similar to that used for Theorem 14.3.2 that uses double counting (except that, as in the notation used for the Bruen–Silverman proof, ℓ_0 might be positive); the proof is omitted.

Theorem 14.4.1 (Folklore, see [109]). *Let X be a set of x points in a FPP Π of order q. Put $x - 1 = (q+1)a - b$, where $0 \le b \le q$. Then the number of lines in Π intersecting X is at least*

$$\frac{x}{a(a+1)}[2a(q+1) - x + 1].$$

Equality holds if and only if all lines intersect X in 0, a, or $a+1$ points.

The next result appeared in a paper by Biró, Hamburger, Pór, and Trotter [100] on dimensions of posets. These authors said that this result follows from ideas in [109], but did not include the details. I found a proof that follows from Theorem 14.4.1, (which is stated in [109]), and after I showed my proof to Biró, he sent me a much more elegant proof using only first principles that they had discovered but omitted from their paper. I first give the proof that Biró et al. found, and then I give a slight improvement that uses Theorem 14.4.1.

Theorem 14.4.2 (Biró–Hamburger–Pór–Trotter, 2016 [100]). *Let Π be a finite projective plane of order q. Let X be a set of points in Π and Y be a set of lines in Π with the property that no line of Y intersects X; that is, for any line $\ell \in Y$, $|X \cap \ell| = \emptyset$. Then $|X| \cdot |Y| \le q^3$.*

Proof: This proof is from [98]. Let $|X| = b$ and for each $i = 0, 1, \ldots, q+1$, let ℓ_i be the number of lines in Π with i points of X on them. Then (as in the proof of Theorem 14.3.2),

$$\sum_{i=1}^{q+1} \ell_i = q^2 + q + 1 - \ell_0,$$

and

$$\sum_{i=1}^{q+1} i\ell_i = b(q+1). \tag{14.4}$$

Also, counting ordered pairs of points in X,

$$\sum_{i=1}^{q+1} i(i-1)\ell_i = b(b-1).$$

Putting the above equations together,

$$\sum_{i=1}^{q+1} i^2\ell_i = b(b+q).$$

Squaring (14.4) and using the Cauchy–Schwarz inequality,

$$\begin{aligned}
b^2(q+1)^2 &= \left(\sum_{i=1}^{q+1} i\ell_i\right)^2 \\
&= \left\| \left(1\sqrt{\ell_1}, 2\sqrt{\ell_2}, \ldots, (q+1)\sqrt{\ell_{q+1}}\right) \bullet \left(\sqrt{\ell_1}, \sqrt{\ell_2}, \ldots, \sqrt{\ell_{q+1}}\right) \right\|^2 \\
&\leq \left\| \left(1\sqrt{\ell_1}, 2\sqrt{\ell_2}, \ldots, (q+1)\sqrt{\ell_{q+1}}\right) \right\|^2 \cdot \left\| \left(\sqrt{\ell_1}, \sqrt{\ell_2}, \ldots, \sqrt{\ell_{q+1}}\right) \right\|^2 \\
&= \left(1^2\ell_1 + 2^2\ell_2 + \cdots + (q+1)^2\ell_{q+1}\right)\left(\ell_1 + \cdots + \ell_{q+1}\right) \\
&= \left(\sum_{i=1}^{q+1} i^2\ell_i\right)\left(\sum_{i=1}^{q+1} \ell_i\right) \\
&= b(b+q)(q^2+q+1-\ell_0).
\end{aligned}$$

Simplifying,

$$bq \leq q(q^2+q+1) - b\ell_0 - q\ell_0,$$

and so

$$b\ell_0 \leq q(q^2+q+1-b-\ell_0).$$

Since one can safely assume that $b + \ell_0 \geq q+1$, it follows that $b\ell_0 \leq q^3$. $\square$

The assumption $b + \ell_0 \geq q+1$ in the last line of the above proof might be strengthened. In other words, the above proof might be strengthened somewhat. Another proof indeed strengthens the result, but it is not nearly so elegant:

Theorem 14.4.3. *Let Π be a finite projective plane of order q. Let X be a set of points in Π and Y be a set of lines in Π with the property that no line of Y intersects X. Then $|X| \cdot |Y| \leq q^3 - 2q$.*

Proof: Using the notation in Theorem 14.4.1, let a, b be non-negative integers with $b \leq q$ so that $|X| = x = (q+1)a - b + 1$. The case $a = 0$ forces $b = 0$ and X consisting of a single point, in which case $|X| \cdot |Y| = |Y| \leq q^2 < q^3$, so the theorem is true. If $a \geq q+1$, then $|X| \geq q^2 + 2q + 1 - b + 1 > q^2 + q + 1$, which is impossible, so $a \leq q$. Even if $a = q$, then $|X| = q^2 + q - b + 1$, so exactly $b \leq q$ points remain outside of X, not enough for a line, and so $|X| \cdot |Y| = 0$. So assume that $1 \leq a < q$. Also note that $qa < x - 1 \leq (q+1)a$.

By Theorem 14.4.1, the number of lines that intersect X is at least

$$\begin{aligned}
&\frac{x}{a(a+1)}[2a(q+1) - (x-1)] \\
&\quad \geq \frac{x}{a(a+1)}[2a(q+1) - (q+1)a] \qquad \text{(since } x - 1 \leq (q+1)a\text{)} \\
&\quad = \frac{x}{a(a+1)}[a(q+1)] \\
&\quad = \frac{x}{a+1}(q+1) \\
&\quad \geq \frac{qa+2}{a+1}(q+1) \qquad \text{(since } x - 1 > qa\text{)} \\
&\quad = (q - \frac{q-2}{a+1})(q+1) \\
&\quad = q^2 + q - \frac{(q-2)(q+1)}{a+1}.
\end{aligned}$$

Hence

$$|Y| \leq q^2 + q + 1 - \left[q^2 + q - \frac{(q-2)(q+1)}{a+1}\right] = 1 + \frac{(q-2)(q+1)}{a+1}.$$

Thus

$$\begin{aligned}
&|X| \cdot |Y| \\
&\leq [(q+1)a + 1] \cdot \left[1 + \frac{(q-2)(q+1)}{a+1}\right] \\
&= [(q+1)(a+1) - q] \cdot \left[1 + \frac{(q-2)(q+1)}{a+1}\right] \\
&= (q+1)(a+1) + (q+1)^2(q-2) - q - \frac{q(q-2)(q+1)}{a+1} \\
&\leq (q+1)(a+1) + (q+1)^2(q-2) - q - (q-2)(q+1) \qquad \text{(since } a + 1 \leq q\text{)}
\end{aligned}$$

$$\begin{aligned}&\leq (q+1)q + (q+1)^2(q-2) - q - (q-2)(q+1) && \text{(since } a+1 \leq q\text{)}\\&= q^2 + q + q^3 - 3q - 2 - q - q^2 + q + 2)\\&= q^3 - 2q.\end{aligned}$$

□

For an application of Theorem 14.4.2 to "standard examples" used in the theory of poset dimension, see the article by Biró, Hamburger, Pór, and Trotter [100]. (For dimension theory and standard examples in posets, see Section 21.3.1.)

14.5 Blocking sets with small number of points per line

Not only can one ask about the size of a blocking set, but ask in how many points does the blocking set intersect each line. Conditions on when a blocking set intersects each line in one of two numbers of points have been examined in [252], for example.

An old question of Erdős asks if there is some universal constant c so that for any FPP there exists a blocking set that intersects each line in at most c points. [Erdős asked me this question in April of 1991, but apparently this question was already nearly twenty years older.] The question is still open and some progress has been made in the area, however the results seem to indicate a negative answer. It is known that for a projective plane of order 5, there is no blocking set which intersects every line in fewer than 4 points. Indeed, using the Paige–Wexler matrix one can verify this simply with some exhaustive checking.

In 1983, Erdős, Silverman, and Stein [316] showed that for any constant $k \geq 2e \sim 5.44$ and for sufficiently large n, there is a blocking set in a FPP of order n that intersects each line in at most $k \ln n$ points. This was later improved by Abbott and Liu [2] to bounding the intersection size to $\frac{2}{\ln 2} \ln n$. For some planes derived from fields, one can improve the Abbott–Liu result considerably. First, Bruen and Fisher [161] showed that in a FPP over the field $\mathrm{GF}(3^d)$, a blocking set could be found that intersected each line in at most 4 points, regardless of what $d \geq 1$ is! This turned out to be a certain instance of a more general theorem:

Theorem 14.5.1 (Boros, 1987 [129]). *If p is an odd prime and d is any positive integer, there exists a set B of points in the projective plane $\mathcal{P}(p^d)$ (derived from the Galois field GF(p^d)) so that for every line ℓ in the plane,*

$$1 \leq |B \cap \ell| \leq p+1.$$

Proof idea: First pick a non-zero element $t \in \mathrm{GF}(p^d)$ so that $t^{(p^d-1)/(p-1)} \neq 1$ (which is possible since all such elements comprise $\mathrm{GF}(p)$—check it—and $p > 2$). Now using Hall's coordinatization, let B be all points of the forms (x, x^p), (x, tx^p) and (∞) (where x runs through the field $\mathrm{GF}(p^d)$). The reader is referred to the original article for the verification that B satisfies the theorem (it is non-trivial, yet brief and elegant). □

In what is called 'discrepancy theory' (see Section 14.7), one investigates blocking sets that intersect each line (in the case of geometries, that is) in as close as possible to half the points per line. Next are some conjectures by Aharoni [9] regarding blocking sets in finite projective planes. They seem to have been inspired by, at least in part, a theorem of Bollobás. Katona [530] came up with a proof of the Bollobás result, somewhat more elegant than the original (which was by induction). Since Katona's proof is so novel, I feel it worthy of inclusion here, even though I stray from the main objective for a moment. There have been many proofs, one using tensor products, if I recall correctly. See the Babai–Frankl book [39] for references; see also the book *Combinatorics* [116] by Bollobás for remaining references.

Theorem 14.5.2 (Bollobás, 1965 [115]). *Let X be a set. Let $A_1, \ldots, A_k \in [X]^a$ be distinct a-subsets of X and let $B_1, \ldots, B_k \in [X]^b$ be distinct b-element subsets of X. If for each $i = 1, \ldots, k$,*

$$A_i \cap B_i = \emptyset, \tag{14.5}$$

and

$$\text{for every } j \neq i, A_i \cap B_j \neq \emptyset, \tag{14.6}$$

holds, then $k \leq \binom{a+b}{a}$.

Proof: Let (14.5) and (14.6) hold, and without loss, let $X = [n]$ be the union of all the A_i's and B_i's. For each $i = 1, \ldots, k$, define classes of permutations on $[n]$,

$$\Pi_i = \{\pi \in \mathcal{S}_n : \pi(A_i) < \pi(B_i)\}.$$

Claim: Each permutation $\pi \in \mathcal{S}_n$ is in at most one class.
Suppose for the moment that this is not the case, say with $\pi \in \Pi_i \cap \Pi_j$, that is, both $\pi(A_i) < \pi(B_j)$ and $\pi(A_j) < \pi(B_j)$. If $\max \pi(A_i) \le \max \pi(A_j)$, say, then $\pi(A_i) < \pi(B_j)$; but $A_i \cap B_j \neq \emptyset$, a contradiction. So the classes are disjoint as claimed.

Thus

$$\begin{aligned} n! = |\mathcal{S}_n| &\ge \sum_{i=1}^{k} |\Pi_i| \\ &= \sum_{i=1}^{k} \binom{n}{a+b} a!b!(n-a-b)! \\ &= \sum_{i=1}^{k} \frac{n!}{\binom{a+b}{a}}, \end{aligned}$$

and so

$$1 \ge \sum_{i=1}^{k} \frac{1}{\binom{a+b}{a}}. \tag{14.7}$$

It now follows that $k \le \binom{a+b}{a}$. □

Equation (14.7) is a special case of the "LYM inequality"; see [116] or Lubell's proof [624] of Sperner's theorem for families of sets where no set is contained in any other.

Conjecture 14.5.3 (Aharoni, 1994 [9])**.** *Suppose $(X, \mathcal{S})$ is a set system so that for every S and T in $\mathcal{S}$, $S \neq T$, $S \cap T \le \ell - 1$ holds, and let $\{S_1, \ldots, S_k\} \subset \mathcal{S}$ and $X_1, \ldots, X_k \in [X]^p$ satisfy (14.5) and (14.6). Then $k \le \binom{\ell+p}{\ell}$.*

Couched in terms of projective planes, with $\ell = 2$, the truth of Conjecture 14.5.3 implies the validity of the following conjecture.

Conjecture 14.5.4 (Aharoni, 1994 [9])**.** *Let $\mathcal{P}$ be a finite projective plane with points P and lines $\mathcal{L}$. If $\ell_1, \ldots, \ell_k \in \mathcal{L}$ and $X_1, \ldots, X_k \in [X]^p$ are so that (14.5) and (14.6) hold, then $k \le \binom{p+2}{2}$.*

Conjecture 14.5.5 (Aharoni, 1994 [9])**.** *Let $\mathcal{P}$ be a finite projective plane with points P and lines $\mathcal{L}$, and fix $L \in \mathcal{L}$. If for every $M \subset L$ with $|M| = \binom{p+2}{2}$ there exists $B = B(M)$, $|B| = p$ so that for every line $l \in M$ with $B \cap l \neq \emptyset$ (i.e., B blocks M) then L is blocked by p points.*

Conjecture 14.5.4 implies Conjecture 14.5.5: Suppose that Conjecture 14.5.5 is false, that is, for every collection of $\binom{p+2}{2}$ lines of $L \subset \mathcal{L}$, there is a blocking p-set and yet L is not blocked by any p-set. Let L be minimally so, that is, for any $\ell \in L$, $L \backslash \ell$ is p-blocked. Then $|L| \geq \binom{p+2}{2} + 1$. But then for each $\ell_i \in \mathcal{L}$, there exists a set of points X_i, $|X_i| - p$, so that $\ell_i \cap X_i = \emptyset$ and yet for any $j \neq i$, $\ell_j \cap X_i \neq \emptyset$. But $|L| \geq \binom{p+2}{2}$, contradicting Conjecture 14.5.4. $\square$

A few years after Aharoni stated these three conjectures, he informed me that a 2001 result by Füredi [356] shows that when $\ell = 2$, Conjecture 14.5.3 is false (as is shown in Theorem 14.5.6 below). However, for the cases $\ell > 2$, Conjectures 14.5.3, 14.5.4, and 14.5.5 might all still be open. The following is a disproof of Conjecture 14.5.3 based on the aforementioned result by Füredi [356].

Theorem 14.5.6 (Aharoni, Füredi)**.** *When $\ell = 2$, Conjecture 14.5.3 is false.*

Proof: Let $q \geq 7$ be prime, let r be the greatest integer such that $\binom{r}{2} < q$, and put $m = \binom{r}{2}$. Let $\mathcal{A}_q = \mathrm{AG}(2,q)$ be the affine plane of order q on vertex set V_0 with parallel classes $\mathcal{L}_1, \ldots, \mathcal{L}_{q+1}$, $\mathcal{L}_i = \{L_i^1, \ldots, L_i^q\}$. Let $V_1, \ldots, V_{q+1}$ be disjoint vertex sets of $q+1$ complete graphs on r vertices. For $1 \leq i \leq q+1$ label the edges of V_i as $\{E_i^1, \ldots, E_i^m\}$. Let $V = \cup_{1 \leq i \leq q+1} V_i$ be the vertex set of $\mathcal{H}$. Define the edges of $\mathcal{H}$ by

$$\mathcal{E}(\mathcal{H}) = \{L_i^j \cup E_i^j | 1 \leq i \leq q+1, 1 \leq j \leq m\} \cup \{L_i^j | m \leq j \leq q\}.$$

It is shown in [356] that any two edges in $\mathcal{H}$ intersect in at most one vertex (such hypergraphs are called "linear"), and for $\tau = q + (r-2) + (q-m) = q + O(\sqrt{q})$, this graph is τ-critical (*i.e.*, the smallest blocking set of $\mathcal{H}$ has τ points, yet every proper subgraph of $\mathcal{H}$ can be blocked by less than τ points).

Using the notation of Conjecture 14.5.3, let $\ell = 2$, $p = \tau$ and $\mathcal{S} = \mathcal{E}(\mathcal{H})$. Hence $k = |\mathcal{E}(\mathcal{H})| = q^2 + q$. For $S_i \in \mathcal{S}$, let b_i be the size of a blocking set of $\mathcal{H} \setminus S_i$. Since $\mathcal{H}$ is τ-critical, $b_i < \tau$. Let X_i be a blocking set of $\mathcal{H} \setminus S_i$, along with $\tau - b_i$ other vertices in $V_0 \setminus S_i$. It follows that $X_1, \ldots, X_k$ along with $\mathcal{E}(\mathcal{H})$ satisfy the conditions of 14.5.3 with $\ell = 2$ and $p = \tau$.

The conclusion of Conjecture 14.5.3 then says

$$k < \binom{p+\ell}{\ell}$$

$$= \binom{\tau+2}{2}$$
$$= \binom{q+O(\sqrt{q})+2}{2}$$
$$< \frac{O(q^2)}{2}.$$

However, for large values of q, this contradicts $k = q^2 + q$. Hence, Conjecture 14.5.3 fails for $\ell = 2$. □

14.6 Maximal strong representative systems

Throughout this section, let Π be a FPP of order q.

A *flag* in Π is a pair (P, ℓ) with $P \in \ell$. A set of flags $F = \{(P_1, \ell_1), \ldots, (P_k, \ell_k)\}$ is called a *strong representative system* (SRS) if $P_i \in \ell_j \iff i = j$. A *maximal strong representative system* (MSRS) is an SRS F so that for any flag $f \notin F$, the set $F \cup \{f\}$ is not a SRS.

The upper bounds for the sizes of MSRSs and blocking sets are the same; this may be no surprise because if B is a minimal blocking set, then through each point $P_i \in B$, there is a tangent line ℓ_i (that is, $|B \cap \ell_i| = 1$). Then $\{(P_1, \ell_1), (P_2, \ell_2), \ldots, (P_b, \ell_b)\}$ is a MSRS.

Theorem 14.6.1 (Illés–Szőnyi–Wettl [494], 1991)**.** *If F is an MSRS in a FPP of order q, then*

$$q + 1 \leq |F| \leq q\sqrt{q} + 1.$$

The proof of the upper bound is nearly identical to a proof of the upper bound by Bruen and Thas, again, using fairly simple double counting.

14.7 Discrepancy of the projective plane

14.7.1 Discrepancy of a set system

In the colouring provided in Theorem 14.2.1, lines do not have 'balanced' colourings; *i.e.*, on each line, there are many more of one colour than the other. How balanced can a good colouring of a FPP be? To formalize this concept, the following definitions are used.

Let $\mathcal{A} = \{A_1, A_2, \ldots, A_t\}$ be a set system on vertex set $A = \cup_{i=1}^{t} A_i$. For any 2-colouring $\phi : A \to \{-1, 1\}$, define the discrepancy of each A_i with respect to ϕ by

$$\mathrm{disc}_\phi(A_i) = \left| |\phi^{-1}(-1) \cap A_i| - |\phi^{-1}(1) \cap A_i| \right|.$$

Define the discrepancy of the set system $\mathcal{A}$ with respect to ϕ by

$$\mathrm{disc}_\phi(\mathcal{A}) = \max\{\mathrm{disc}_\phi(A_i) : 1 \le i \le t\}.$$

Finally, define the *discrepancy of* $\mathcal{A}$ to be

$$\mathrm{disc}(\mathcal{A}) = \min_{\phi: A \longrightarrow \{-1,1\}} \{\mathrm{disc}_\phi(\mathcal{A})\}.$$

The primary goal in this section is to bound the discrepancy of a finite projective plane. Keeping this in mind, some major theorems about discrepancy are reviewed, and where appropriate, consequences of these theorems as they relate to the primary goal. For more details regarding these and others, see the article by Beck and Sós [70] or the book by Alon and Spencer [24]. For those interested in more details about discrepancy in general and the many connections between discrepancy and combinatorial/computational geometry, I highly recommend the book *The discrepancy method* [185] by Bernard Chazelle (for example, this book gives many more bounds, and connections to VC-dimension, a topic only briefly introduced in here in Section 19.2).

A closer look at discrepancy for FPPs in particular is given in Sections 14.7.2 (upper bounds) and 14.7.3 (lower bounds).

If a set system has discrepancy 0, then there is a colouring that is completely balanced on every set in the family. Graphs with chromatic number 2 (*e.g.*, trees, even cycles) have discrepancy 0. In fact, an r-uniform hypergraph $\mathcal{H}$ is 2-colourable if and only if $\mathrm{disc}(\mathcal{H}) < r$. (This implies that determining discrepancy is NP-hard; see [185].)

As noted in Theorem 14.2.1, the chromatic number of an FPP of order $n > 2$ is two; that is, there is a good 2-colouring of any such FPP. However, the 2-colouring given in the proof of Theorem 14.2.1 as a witness is highly imbalanced on most lines. Are there more balanced 2-colourings?

In any good 2-colouring of the points of a FPP of order n, say with red and blue, the red-blue split on each line is at worst n to 1, so the discrepancy of a FPP of order $n > 2$ is at most $n - 1$. On the other hand, in a FPP of order n, the number of points (n^2+n+1) is odd, so the total discrepancy is at least one. Summarizing, if $\mathcal{P}(n)$ is a projective plane of order $n > 2$, then

$$1 \le \mathrm{disc}(\mathcal{P}(n)) \le n-1. \tag{14.8}$$

Before trying to refine the results on discrepancy for FPPs, a few old results regarding discrepancy are first presented.

Definition 14.7.1. A matrix is *totally unimodular* if and only if every square submatrix has determinant -1, 0, or +1. A hypergraph $\mathcal{H}$ is *unimodular* if and only if its incidence matrix is totally unimodular.

One of the earliest theorems involving discrepancy was given in 1962 by Ghouila-Houri. If $\mathcal{H}$ is a set system on a vertex set X, and $Y \subseteq X$, the restriction of the family to Y is the family of sets $\mathcal{H}_Y = \{H \cap Y : H \in \mathcal{H}\}$.

Theorem 14.7.2 (Ghouila-Houri, 1962 [387])**.** *A hypergraph $\mathcal{H}$ on vertex set X is unimodular if and only if for every $Y \subseteq X$,*

$$disc(\mathcal{H}_Y) \le 1.$$

Roth [764] gave a result that says, essentially, long arithmetic progressions in $[1, n]$ have fairly large discrepancy.

Theorem 14.7.3 (Roth, 1964 [764])**.** *There exists a universal constant c so that for any n and $\phi : [1, n] \to \{-1, 1\}$, there exists an arithmetic progression $P \subseteq [1, n]$ so that*

$$\big||\phi^{-1}(-1)| - |\phi^{-1}(1)|\big| > cn^{1/4}.$$

Related is the following conjecture due to Erdős, who once offered \$100 for its solution (I don't know its present status):

Conjecture 14.7.4. *For fixed $0 < \epsilon < 1/2$ and any ℓ, there exists $n = n(\ell) \le 2^\ell$ so that for any 2-colouring of $[1, n]$ there exists an AP_ℓ with $(\frac{1}{2}+\epsilon)\ell$ terms monochromatic.*

14.7.2 Upper bounds on discrepancy of a FPP

A trivial upper bound on discrepancy of a set system is the size of the largest set:

$$\mathrm{disc}(\mathcal{F}) \le \max_{F \in \mathcal{F}} |F|.$$

So for FPPs of order n, the discrepancy is at most $n+1$, a result that is slightly too large by Theorem 14.2.1. In a set system, deletion of pairs that occur in the same set of edges does not affect the discrepancy. If $|\mathcal{F}| = m$, then one can reduce the ground set so that there are at most $2^m - 1$ vertices per set, so in this case, $\mathrm{disc}(\mathcal{F}) \leq 2^m - 1$. This bound can be improved significantly:

Theorem 14.7.5 (Olson–Spencer, 1978 [689])**.** *For* $|\mathcal{F}| = m$,

$$disc(\mathcal{F}) \leq c\log(m)\sqrt{m}.$$

For FPPs of order n, Theorem 14.7.5 says that the discrepancy is at most on the order $n\log(n)$, again no improvement over the trivial bound of $n-1$. For $x \in X$, and $\mathcal{F}$ a family of subsets of X, define the degree of x by $\deg_{\mathcal{F}}(x) = |\{F \in \mathcal{F} : x \in F\}|$. Put $\Delta(\mathcal{F}) = \max_{x\in X}\deg(x)$. It is rather surprising (to me, at least) that a simple bound on discrepancy can be found in terms of maximum degree only:

Theorem 14.7.6 (Beck–Fiala, 1981 [69])**.** *For any family* $\mathcal{F}$ *of subsets of a finite set,* $disc(\mathcal{F}) < 2\Delta(\mathcal{F})$.

For a simple proof of the Beck-Fiala theorem, see [185]. What Beck and Fiala really proved was the following:

Theorem 14.7.7 (Beck–Fiala, 1981 [69])**.** *Let* $X = [1, n]$ *and* $\mathcal{F}$ *be a family of subsets of* X*. For each* $i \in X$*, let* $p_i \in [-1, 1]$*. For each* i *there exists* $\epsilon_i \in \{-1, 1\}$ *so that*

$$\max_{F\in\mathcal{F}}\left|\sum_{i\in F}(\epsilon_i - p_i)\right| < 2\Delta(\mathcal{F}).$$

The Beck–Fiala theorem applied to FPPs of order n uses $\Delta = n+1$, and so does not contribute to the knowledge of the discrepancy of a FPP. If the following conjecture were to be true, it might yield more interesting upper bounds for the discrepancy of FPPs.

Conjecture 14.7.8 (Beck–Fiala, 1981 [69])**.** *There exists some universal constant* $c > 0$ *so that* $disc(\mathcal{F}) < c\sqrt{\Delta(\mathcal{F})}$.

As found in (for example) [24], a straightforward application of the probabilistic method gives a more general upper bound on the discrepancy of a set system:

Theorem 14.7.9. *Let $|S| = m$, and let $\mathcal{A}$ be a family of n subsets of S. Then*

$$disc(\mathcal{A}) \leq \sqrt{2m\ln(2n)}.$$

Proof outline: Let $\chi : S \longrightarrow \{-1, 1\}$ be random. Put $\alpha = \sqrt{2m\ln(2n)}$. Then for $A \in \mathcal{A}$,

$$\text{Prob}[|\chi(A)| > \alpha] < 2e^{-\alpha^2/2|A|} \leq 2e^{-\alpha^2/2m} = \frac{1}{n}.$$

Let X be the number of A's with $|\chi(A)| > \alpha$ and let X_A be the indicator random variable for $|\chi(A)| > \alpha$. Then the expected value of X is

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{A\in\mathcal{A}} X_A\right] = \sum_{A\in\mathcal{A}} \mathbb{E}[X_A] < |\mathcal{A}|\frac{1}{n} = 1.$$

So for some χ, then $X = 0$; *i.e.*, $\text{disc}_\chi(\mathcal{A}) \leq \alpha$. □

For $n = m$, Spencer showed that the factor $\sqrt{\ln(2n)}$ in Theorem 14.7.9 can be omitted.

Theorem 14.7.10 (Spencer, 1985 [811])**.** *If $\mathcal{F}$ is a family of at most m subsets of an m-set, then*

$$disc(\mathcal{F}) \leq 6\sqrt{m}.$$

Up to a constant, Spencer's bound is best possible (there are at least two proofs of this claim, one probabilistic, and one using Hadamard matrices). See Section 14.7.3 for details.

By Theorem 14.7.10, an upper bound for the discrepancy of the finite projective plane of order n is $\sqrt{n^2 + n + 1}$, still worse than the trivial bound. Spencer was indeed able to improve this for FPPs:

Theorem 14.7.11 (Spencer, 1988 [812])**.** *There exists a constant $k > 0$ so that for sufficiently large n, the discrepancy of a finite projective plane of order n is at most $k\sqrt{n}$.*

This result is best possible up to the constant factor (by Theorem 14.7.16, for one).

The value for k in Theorem 14.7.11 was not made explicit. Rödl was able to show that $k < 12$ when the plane is one derived from a field of odd order. Let $\mathcal{P}(q) = \mathrm{PG}(2, \mathbb{F}_q)$ denote the projective plane of order q derived from $\mathrm{GF}(q)$ (where q is some prime power). Let $\mathcal{A}(q) = \mathrm{AG}(2, q)$ be the corresponding affine plane formed by deletion of a line of $\mathcal{P}(q)$. There are $q^2 + q$ lines in $\mathcal{A}(q)$, each containing q points.

Let $A = \{(x, y) : x, y \in \mathrm{GF}(q)\}$ be the vertex set of the affine plane with lines $\mathcal{A} = \mathcal{A}(q) = \{A_i : 1 \leq i \leq q^2 + q\}$.

Theorem 14.7.12 (Rödl, 1994 [758])**.** *There exists a constant c_2 so that for every odd prime power q, $d(\mathcal{A}(q)) < c_2\sqrt{q}$.*

Before giving the proof, some preliminaries are established. First, recall a result about the number of perfect squares in a finite field: If q is odd, then $|\{x^2 : x \in \mathrm{GF}(q), x \neq 0\}| = (q-1)/2$ and for every $a \neq 0$, the equation $x^2 = a$ has two solutions.

Consider the mapping $f : A \longrightarrow GF(q)$ defined by $f(x, y) = x + y^2$. Use the notation $f(A_i) = \{f(x, y) : (x, y) \in A_i\}$ to indicate the image of a line under f. To prove Theorem 14.7.12, the following lemma regarding f is used. Essentially, the lemma says that the number of distinct images $f(A_i)$ of lines is at most $q + 1$, and for those lines on which f is not one-to-one, with the exception of at most one point per line, exactly two points per line are mapped to any one element of the field.

Lemma 14.7.13. *Let q be any odd prime power and define $f : A \to GF(q)$ by $f(x, y) = x + y^2$. Then*

$$|\{f(A_i) : A_i \in \mathcal{A}\}| \leq q + 1, \tag{14.9}$$

For any fixed $A_i \in \mathcal{A}$, either $f(A_i) = GF(q)$ or

$$|\{z \in f(A_i) : |f^{-1}(z)| = 2\}| = (q-1)/2, \tag{14.10}$$

and

$$|\{z \in f(A_i) : |f^{-1}(z)| = 1\}| = 1. \tag{14.11}$$

Proof: Let $A_i \in \mathcal{A}$ be so that for some $b, m \in GF(q)$, $A_i = \{(0, b) + t(1, m) : t \in GF(q)\}$. In this case, $f(A_i) = \{t + (b + mt)^2 : t \in GF(q)\}$. When $m = 0$, $f(A_i) = \{t + b^2 : t \in GF(q)\} = GF(q)$. When $m \neq 0$, completing the square gives

$$t + (b + mt)^2 = \left(mt + \frac{2bm + 1}{2m}\right)^2 + b^2 - \left(\frac{2bm + 1}{2^2 m^2}\right)^2.$$

As t varies over $GF(q)$, so does the variable $w = mt + (2bm+1)/(2m)$. Therefore, with constant $k = b^2 - [(2bm+1)/(2^2m^2)]^2$,

$$f(A_i) = \{k + w^2 : w \in GF(q)\}. \tag{14.12}$$

The remaining form for a line is $A_i = \{(c,t) : t \in GF(q)\}$, where $f(A_i) = \{c + t^2 : t \in GF(q)\}$, precisely the form in (14.12). So in any case, either $f(A_i) = GF(q)$, or $f(A_i)$ is of the form (14.12) and so (14.9) is proved.

Since the equation $x^2 = a$ has exactly two solutions when $a \neq 0$, whenever $f(A_i)$ is the form (14.12), $|f^{-1}(k)| = 1$ and for $z \neq k$, either $|f^{-1}(z)| = 2$ or $|f^{-1}(z)| = 0$. Using the fact that $|\{x^2 : x \in GF(q), x \neq 0\}| = (q-1)/2$ gives (14.10) and (14.11). □

Oddly enough, no such proof seems available for even q. No polynomial seems to be a good choice for f. For f as defined above, a property similar to (14.10) still holds; however, (14.9) does not—there are many more than q images. No other similar technique seems to work for the even case. Now return to finish the proof of Theorem 14.7.12.

Proof of Theorem 14.7.12: Consider the set system

$$\mathcal{S} = \{S_1, S_2, \ldots, S_t\} = \{f(A_i) : A_i \in \mathcal{A}\},$$

where $\cup_{i=1}^{t} S_i = GF(q)$. By the first statement in Lemma 14.7.13, $t \leq q+1$. By Theorem 14.7.10 let $\phi : GF(q) \to \{-1, 1\}$ be given so that $d_\phi(\mathcal{S}) < c_1\sqrt{q+1}$. Let $\psi : A \to GF(q)$ be defined by $\psi(x,y) = \phi(x + y^2)$. By Lemma 14.7.13, $d_\psi(\mathcal{A}) \leq 2 \cdot d_\phi(\mathcal{S})$, and so $d_\psi(\mathcal{A}) < 2c_1\sqrt{q+1} < c_2\sqrt{q}$. □

Spencer [811] found an explicit bound of $c_1 < 5.32$ for sufficiently large n. Using this bound, obtain $c_2 < 11$ (also for sufficiently large odd q). One might improve this bound considerably (by a fraction—probably around 1/2 or 3/4), since in the proof of Lemma 14.7.13, the number of images of lines is smaller than $q+1$ (some are repeated, and it was not determined exactly how many were really different). However, lowering this bound would mean going through Spencer's proof to get a bound on the discrepancy of a set system with the number of sets being a fraction of the number of vertices. Observe that Theorem 14.7.12 implies a similar result for finite projective planes derived from fields. In the following, Theorem 14.7.12 gives an easy bound.

Theorem 14.7.14. *There exists a constant c_3 so that for every odd prime power q, $d(\mathcal{P}(q)) < c_3\sqrt{q}$.*

Proof: Let P be the vertex set of $\mathcal{P}(q)$. Coordinatize the vertices (for example, by Hall's coordinatization) so that vertices receive the same coordinates as for the affine plane, except for $q+1$ vertices on a line at infinity, say l_∞. Lines of the projective plane consist of l_∞ and lines formed by affixing one more point (at infinity) to a line from the affine plane $\mathcal{A}(q)$.

By Theorem 14.7.12, let $\phi : A \to \{-1, 1\}$ be given so that $d_\phi(\mathcal{A}) < c_2\sqrt{q}$. Extend ϕ to $\phi' : P \to \{-1, 1\}$ by 2-colouring the points on l_∞ as evenly as possible. The discrepancy of this new colouring ϕ' can increase at most by one, since the discrepancy in colouring any one line in $\mathcal{P}(q)$ is at most one more than the induced colouring of the corresponding line in $\mathcal{A}(q)$. Hence $d(\mathcal{P}(q)) < c_2\sqrt{q} + 1 < c_3\sqrt{q}$. □

Observe that for q sufficiently large, $c_3 < 11$.

14.7.3 Lower bounds for discrepancy of a FPP

In general, the lower bound for discrepancy of a set system is zero. One can immediately dream up some fairly simple examples with discrepancy zero or one; graphs with chromatic number two trivially have discrepancy zero. More complicated structures have been found with small discrepancy, for example, hypergraphs that are *unimodular* (every square submatrix of its incidence matrix has determinant -1, 0, or 1) have discrepancy zero or one (see, *e.g.*, the survey by Beck and Sös [70] for discussion).

On the other end of the spectrum, set systems having a Hadamard matrix as its incidence matrix have been found to have discrepancy larger than $\sqrt{n}/2$.

Theorem 14.7.15. *If there exists a Hadamard matrix of order n then there exists a family $\mathcal{A}$ of n subsets of an n-element set with*

$$disc(\mathcal{A}) \geq \frac{\sqrt{n}}{2}.$$

Proof: The proof given here is adapted from [24]; for another, more general proof, see [70].

Let H be a Hadamard matrix of order n; that is, $H = (h_{ij})$ is an $n \times n$ matrix with each $h_{ij} \in \{-1, 1\}$ so that rows (and hence columns) are orthogonal. Furthermore, let H be normalized, that is, each entry in the first row and first column is a 1. (This can be done by simply multiplying rows and columns by -1; orthogonality between rows is preserved.) Let H^* be the $n \times n$ matrix formed by changing all -1's to 0's. Then $H^* = \frac{1}{2}(H + J)$. View H^* as an incidence matrix for a set system $(X, \mathcal{H})$ consisting of n subsets of an n-element set, say $X = \{1, 2, \ldots, n\}$, each row corresponding to the characteristic function for each set (one set has n elements, and the remaining $n - 1$ sets each have $n/2$ sets). To be specific, let $\mathcal{H} = \{H_1, \ldots, H_n\}$, where $j \in H_i$ if and only if $h_{ij} = 1$.

Claim: $\mathrm{disc}(\mathcal{H}) \geq \frac{\sqrt{n}}{2}$. Let $\phi : X \to \{-1, 1\}$ be given and put $\mathbf{v} = (\phi(1), \ldots, \phi(n))$. Putting $H^*\mathbf{v}^T = \mathbf{d}$, where $\mathbf{d} = (d_1, \ldots, d_n)^T$, then $\mathrm{disc}_\phi(H_i) = d_i$, and so $\mathrm{disc}_\phi(\mathcal{H}) = \min\{|d_i|\}$. That is, to compute $\mathrm{disc}_\phi(\mathcal{H})$, compute $\frac{1}{2}(H + J)\mathbf{v}^T$ and find the coordinate with the smallest magnitude. If this smallest magnitude can be bounded by a value independent of the colouring, then a lower bound is found for the discrepancy of the system. This is done in essentially two stages: first looking at $H\mathbf{v}^T$, then by examining $(H + J)\mathbf{v}^T$.

For each $i = 1, \ldots, n$, let $\mathbf{r}_i$ denote the i-th row of H, and $\mathbf{c}_i$ denote the i-th column of H. Then for each $i \neq j$, both $\mathbf{r}_i \bullet \mathbf{r}_j = 0$ and $\mathbf{c}_i \bullet \mathbf{c}_j = 0$. For each $i = 1, \ldots, n$, put $L_i = \sum_{j=1}^n h_{ij}\phi(j)$. Then

$$H\mathbf{v} = \phi(1)\mathbf{c_1} + \cdots + \phi(n)\mathbf{c}_n = (L_1, \ldots, L_n)^T.$$

Computing some bounds for both $\sum_{i=1}^n L_i$ and $\sum_{i=1}^n L_i^2$,

$$\begin{aligned}\sum_{i=1}^n L_i^2 &= (H\mathbf{v}^T) \bullet (H\mathbf{v}^T)\\ &= \sum_{i=1}^n \phi(i)^2 \|\mathbf{c}_i\|^2 \qquad\qquad (\text{for } i \neq j,\ \mathbf{c}_i \perp \mathbf{c}_j)\\ &= n \cdot n = n^2;\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^n L_i &= \sum_{i=1}^n \sum_{j=1}^n h_{ij} v_j\\ &= \sum_{j=1}^n \sum_{i=1}^n h_{ij} v_j\\ &= \sum_{j=1}^n v_j \sum_{i=1}^n h_{ij}\end{aligned}$$

$$
\begin{aligned}
&= v_1 \sum_{i=1}^{n} h_{i1} + \sum_{j=2} v_j \sum_{i=1}^{n} h_{ij} \\
&= v_1 \cdot n + \sum_{j=2}^{n} 0 \\
&= v_1 \cdot n \\
&= \pm n.
\end{aligned}
$$

Put $\lambda = \sum_{i=1}^{n} \phi(i)$. Then

$$
(H+J)\mathbf{v} = \begin{bmatrix} L_1 + \lambda \\ L_2 + \lambda \\ \vdots \\ L_n + \lambda \end{bmatrix}.
$$

Using the notation at the beginning of the proof, for each i, $d_i = \frac{1}{2}(L_i + \lambda)$, so it remains only to give a lower bound on $\max_i |L_i + \lambda|$. A common technique is used—examine the sum of the squares and use averages:

$$
\begin{aligned}
\|(H+J)\mathbf{v}\|^2 &= \sum_{i=1}^{n} (L_i + \lambda)^2 \\
&= \sum_{i=1}^{n} (L_i^2 + 2L_i\lambda + \lambda^2) \\
&= n^2 + 2(\pm n)\lambda + n\lambda^2 \\
&= n(n \pm 2\lambda + \lambda^2).
\end{aligned}
$$

In this last expression, the quadratic in λ (by simple calculus) attains a minimum at $\lambda = \pm 1$. Assume that $n > 1$ and n is even (since Hadamard matrices only exist for $n = 1$ and n even); so λ is also even (if m values of $\phi(i)$ are -1, then $\lambda = -m + (n - m) = n - 2m$, an even number). Subject to this constraint of λ being an even integer, the quadratic above attains a minimum at either $\lambda = 0$ or at $\lambda = \pm 2$. In any case,

$$
\sum_{i=1}^{n} (L_i + \lambda)^2 \geq n^2
$$

and so for some i, $(L_i + \lambda)^2 \geq n$, or $|L_i + \lambda| \geq \sqrt{n}$. Thus, for some i, $|d_i| \geq \frac{\sqrt{n}}{2}$ as claimed. $\square$

Before returning to examine a lower bound for discrepancy of a FPP, first note that by a probabilistic argument (see [24, pp. 208–209]) there is a constant λ so that a family $\mathcal{A}$ of n subsets of an n-set exists with $\mathrm{disc}(\mathcal{A}) \geq \lambda\frac{\sqrt{n}}{2}$.

The following theorem is well-known, and the version of its proof presented here appears in [812] (a slightly different format, apparently due to Lovász and Sós , can be found in [70]).

Theorem 14.7.16. *Let $\mathcal{P}$ be any projective plane of order n. Then $d(\mathcal{P}) \geq \sqrt{n}$.*

Proof: Let $\mathcal{P}$ be a projective plane of order n on points $P = \{P_i : 1 \leq i \leq n^2+n+1\}$ and lines $\mathcal{L} = \{l_j : 1 \leq j \leq n^2+n+1\}$. Let $A = (a_{ij})$ be an incidence matrix for the plane $\mathcal{P}$, where $a_{ij} = 1$ if and only if $P_j \in l_i$. Letting I be the $(n^2+n+1) \times (n^2+n+1)$ identity matrix, and let J be the matrix of the same size consisting entirely of 1's. Note that $A^TA = AA^T = nI + J$. Let $\chi : P \to \{-1, 1\}$ be a 2-colouring of the points of the plane, and let

$$\boldsymbol{v} = (\chi(P_1), \ldots, \chi(P_{n^2+n+1})).$$

Letting

$$A\boldsymbol{v}^T = (d_1, \ldots, d_{n^2+n+1})^T,$$

see that

$$[d_\chi(\mathcal{P})]^2 = \max\{(d_i)^2 : 1 \leq i \leq n^2+n+1\}.$$

Instead of looking at the maximum of the $(d_i)^2$'s, add them all up and find the average:

$$\begin{aligned}
\sum_{i=1}^{n^2+n+1} (d_i)^2 &= (A\boldsymbol{v}^T)^T(A\boldsymbol{v}^T) \\
&= \boldsymbol{v}A^TA\boldsymbol{v}^T \\
&= \boldsymbol{v}(nI+J)\boldsymbol{v}^T \\
&= n\boldsymbol{v}\boldsymbol{v}^T + \boldsymbol{v}J\boldsymbol{v}^T \\
&= n(n^2+n+1) + \sum_{i,j} \chi(P_i)\chi(P_j) \\
&= n(n^2+n+1) + \left(\sum_i \chi(P_i)\right)^2
\end{aligned}$$

$$\geq \quad n(n^2+n+1)$$

Hence there is a d_i with $(d_i)^2 \geq n$, that is, $d_\chi(\mathcal{P}) \geq |d_i| > \sqrt{n}$. □

14.8 Legitimate colourings of FPPs

Let $\mathcal{H}$ be a hypergraph. A *legitimate colouring* of $\mathcal{H}$ is a colouring of its vertices with finitely many colours so that any two distinct hyperedges receive different multisets of colours. Trivially, giving each vertex of a hypergraph a unique colour gives a legitimate colouring.

In 1989, Alon and Füredi [22] showed that if a finite projective plane $\mathcal{P}$ has order at least 5, then the fewest number of colours needed to find a legitimate colouring of $\mathcal{P}$ (viewed as a hypergraph) is in the interval $[5, 8]$. This result may be surprising since the number of colours does not depend on the order of the FPP. According to [22], Alex Rosa showed that three colours is not enough for planes of order $n \geq 5$. Alon and Füredi conjecture that for n sufficiently large, the minimum number of colours for a legitimate colouring is either 6 or 7.

Legitimate colourings of hypergraphs are related to "harmonious colourings" of graphs, not necessarily proper vertex colourings so that each edge receives a different pair of colours. The concept of harmonious (or "line-distinguishing") colouring was introduced by Frank, Harary, and Plantholt in 1982 [343]. See [503] for references.

Chapter 15

Other finite configurations of points and lines or curves

15.1 The Sylvester–Gallai theorem

In this section, lines in the real plane (or the real projective plane) and their intersection points are considered. Consider the following warm-up exercise:

Exercise 268. *Let $S \subseteq \mathbb{R}^2$ be an infinite set of points. Show that if any two points have integer distance, then all points of S are collinear.*

J. J. Sylvester posed the following:

Problem 15.1.1 (Sylvester, 1893, [846])**.** *Is it true that if n points on the plane are not all collinear, then there exists at least one line with only two points?*

Erdős "rediscovered" (see [303] and the preface by Erdős in [143, p. VIII]) Sylvester's problem in 1933 while reading the 1932 book *Anschauliche Geometrie* by Hilbert and Cohn-Vossen (later translated and published as *Geometry and the imagination* [474]), and Tibor Gallai (1912–1992) found a proof the same year. [Gallai's name was previously Tibor Grünwald, not to be confused with another colleague, Géza Grünwald (1910–1943), a holocaust victim, who studied more in analysis.]

In 1943, Erdős posed the problem in the *American Mathematical Monthly* [296], with the "ingenious proof" [304, p. 208] due to Gallai (together with a proof by R. Steinberg) following a year later [318]. Steinberg's proof was in the real projective plane (also see [223, pp. 29–30] for Steinberg's proof). See [136], [240], [304], [538], [675] for more history and generalizations of this problem.

The following is now known as the "Sylvester–Gallai theorem".

Theorem 15.1.2 (Sylvester 1893, Gallai 1933)**.** *Let $\mathcal{P}$ be a finite set of points in the plane, not all on a line. Then there is a line containing exactly two points from $\mathcal{P}$. [Stated another way: if every line contains at least three points, then all points are collinear.]*

Gallai's proof (which can also be found in [240]) uses angles. Two other proofs are given here, the first of which seems to be the simplest of all known proofs.

First proof: The proof given here is attributed to L. M. Kelly (1914–2002), given in 1947 and appearing in an article by Coxeter [217] the following year.

Let $\mathcal{L}$ be the set of lines determined by pairs of points from $\mathcal{P}$. Let $\ell \in \mathcal{L}$ and let $P \in \mathcal{P}$ be a point not on ℓ so that the distance from P to ℓ is minimal. Let X be the foot of the perpendicular dropped from P onto ℓ, and let A and B be two points of $\mathcal{P}$ on ℓ. Both A and B cannot be on the same side of X, since if A, B, X occur in order on ℓ (with $B = X$ allowed), then B is closer to the line $\overleftrightarrow{PA}$ than P is from ℓ (see Figure 15.1), contradicting that P and ℓ have minimal distance. Thus, each of the two closed rays of ℓ determined by X contains at most one point of $\mathcal{P}$. Hence, ℓ contains precisely two points of $\mathcal{P}$. $\square$

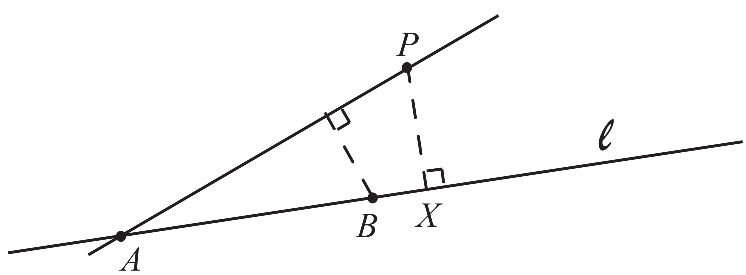


Figure 15.1: Kelly's proof

Second proof of Theorem 15.1.2; This proof is by Lin, 1988 [612], and is very similar to the Gallai proof given in [240].

Let S be a set of points in the plane, not all on a line. For points $A, B, C \in S$, let $\angle BAC$ denote the angle (less than 180 degrees) between the rays $\overrightarrow{AB}$ and $\overrightarrow{AC}$. Say that an angle $\angle BAC$ is "admissible" if and only if there exists a fourth point on $\overrightarrow{AB} \cup \overrightarrow{AC}$. Admissible angles exist (for example, when A is on the boundary of the convex hull of S). Suppose, in hope of contradiction, that every line through two points of S contains a third.

Let $\angle BAC$ be a largest admissible angle, and without loss of generality, suppose that $D \in S$ is a point on AC so that C is between A and D. Since each line contains three points, the line $\overleftrightarrow{BC}$ contains contains an additional point E. Since $\angle BCD$ is larger than $\angle BAC$, the angle $\angle BCD$ is not admissible, and hence E does not lie on $\overrightarrow{CB}$; thus E lies on $\overleftrightarrow{BC}$ on the opposite side of C that B lies (see Figure 15.2).

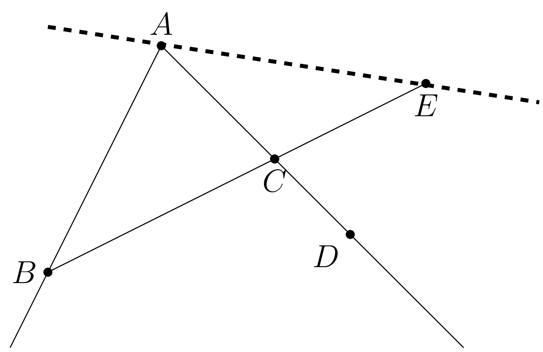


Figure 15.2: Lin's proof of the Sylvester–Gallai theorem

The line $\overleftrightarrow{AE}$ contains an additional point F; if F lies on $\overrightarrow{AE}$, then the admissible $\angle BAE$ is larger than $\angle BAC$, a contradiction; if F lies on the other side of A, then the admissible $\angle FAC$ is larger than $\angle BAC$, another contradiction. In either case, a contradiction is reached, and so S is collinear. □

For more references surrounding the Sylvester–Gallai theorem (and proofs), see, *e.g.*, Bollobás [118, Prob. 33] or [685].

Exercise 269. *Formulate and prove a dual theorem to Theorem 15.1.2, and define "ordinary points".*

The Sylvester–Gallai theorem (Theorem 15.1.2) shows that many finite geometries (those with three or more points on every line) can not be drawn in the plane with straight lines. One of the simplest such examples is worth noting.

Corollary 15.1.3. *The Fano plane (7 points, 7 lines, each line containing 3 points) can not be drawn in the Euclidean plane with straight lines.*

For a little more on configurations that can or cannot be drawn in the plane, see Section 15.4.4.

The next theorem (which is related to the Sylvester–Gallai theorem) is a theorem only about sets (with no mention of geometry or lines), but can be applied to geometric configurations by interpreting sets as lines.

Theorem 15.1.4 (de Bruijn–Erdős, 1948 [240])**.** , *Let $X = \{x_1, \dots, x_n\}$ be a set and let $\mathcal{L} = \{\ell_1, \dots, \ell_m\}$ be be a family of proper subsets of X so that every pair $\{x_i, x_j\}$ is contained in precisely one ℓ_k, then $n \leq m$, where equality $n = m$ holds if and only if the set system $(X, \mathcal{L})$ is one of the two following forms:*

(i) $\ell_1 = \{x_1, x_n\}$, $\ell_2 = \{x_2, x_n\}$, $\dots$, $\ell_{n-1} = \{x_{n-1}, x_n\}$, *and*

$$\ell_n = \{x_1, x_2, \dots, x_{n-1}\}.$$

(Such a configuration is called a "near pencil".)

(ii) There exists k so that $n = k(k-1)+1$, for each $i = 1, \dots, m$, $|L_i| = k$, and each x_i is contained in k members of $\mathcal{L}$.

Proof: Call each x_i a point and each L_j a line. For each point $x_i \in X$, define the degree of x_i, denoted $\deg(x_i)$, to be the number of lines containing x_i. By double counting point-line incident pairs,

$$\sum_{i=1}^{n} \deg(x_i) = \sum_{j=1}^{m} |L_j|. \tag{15.1}$$

Assume that each line contains at least two points, for if a line contains only a single point, delete that line (for if the theorem is true with no such lines, then it is also true with them). One may also assume that not all points are on one line, since lines are proper subsets of X. So $m \geq 2$.

If for some x_i and ℓ_j, if $x \notin \ell_j$, then for each point on ℓ_j, there is a unique line containing that point and x_j, and so

$$[x_i \notin \ell_j] \Rightarrow [\deg(x_i) \geq |\ell_j|]. \tag{15.2}$$

Suppose that x_n is a point with smallest degree, say $\deg(x_n) = d$, and without loss of generality, let $\ell_1, \ldots, \ell_d$ be the lines containing x_n. Since every two points are in some line, and not all points are on ℓ_1, it follows that $d \geq 2$. Since each line has at least two points, for $i = 1, \ldots, d$, let x_i be a point (other than x_n) on ℓ_i.

By (15.2), $\deg(x_1) \geq |\ell_2|$, $\deg(x_2) \geq |\ell_3|$, $\ldots$, $\deg(x_{d-1}) \leq |\ell_d|$ and $\deg(x_d) \geq |\ell_1|$, and adding all of these inequalities gives

$$\left(\sum_{i=1}^{d} \deg(x_i)\right) \geq \sum_{i=1}^{d} |\ell_j| = n - d + 1. \tag{15.3}$$

Subtracting (15.3) from (15.1) gives

$$\sum_{i=d+1}^{n} \deg(x_i) \leq \sum_{j=d+1}^{m} |\ell_j|. \tag{15.4}$$

Because the minimum degree of vertices is d, the left side of (15.4) is at least $(n-d)d$. Since for each $j = d+1, \ldots, m$, x_n is not on ℓ_j, by (15.2), $|\ell_j| \geq \deg(x_n) = d$, and so the right side of (15.4) is bounded above by $(m-d)d$. So (15.4) implies

$$(n-d)d \leq (m-d)d,$$

which gives $n \leq m$, proving the first part of the theorem.

Suppose now that $n = m$. Then all inequalities in (15.2) become equalities, and so $\deg(x_1) = |\ell_2|$, $\deg(x_2) = |\ell_3|$, $\ldots$, $\deg(x_{n-1}) = |\ell_n|$, $\deg(x_d) = |\ell_1|$, and for all $j = d+1, \ldots, m$, $|\ell_j| = d$. By relabelling points (or lines), suppose that

$$\deg(x_1) = |\ell_1| \geq \deg(x_2) = |\ell_2| \geq \cdots \geq \deg(x_d) = |\ell_d| > 1.$$

Consider two cases:

(a) Suppose that $\deg(x_1) > \deg(x_2)$. Then for $i = 2, 3, \ldots, n$, $\deg(x_1) > |\ell_i|$, and so by (15.2), each of $x_2, \ldots, x_n$ lie on ℓ_1; since $x_1 \notin \ell_1$, case (i) is achieved.

(b) Suppose that $\deg(x_1) = \deg(x_2)$.

Claim: All degrees are equal.

Proof of claim: Let $j \geq 3$ and suppose that $\deg(x_j) < \deg(x_1) = \deg(x_2)$. Then by equation (15.2), x_j lies on both ℓ_1 and ℓ_2. If there were two such js, two points would occur on two lines (namely ℓ_1 and ℓ_2), so there is at most one such j with $\deg(x_j) < \deg(x_1)$, namely $j = n$. Hence,

$$\deg(x_1) = \deg(x_2) = \cdots = \deg(x_{n-1}) > \deg(x_n) = d \geq 2.$$

Since $n - 1 \geq |\ell_1| = \deg(x_1) > \deg(x_n)$, it follows that $\deg(x_n) \leq n - 2$, and so there are two lines not containing x_n, say ℓ_i and ℓ_j; All that is needed is one that is not ℓ_n, say $\ell_i \neq \ell_n$. Then x_n is not on ℓ_i, so by (15.2), $\deg(x_n) >\geq |\ell_i| = \deg(x_i)$, contradicting the fact that for all $j = 1, \ldots, n-1$, $\deg(x_n) < \deg(x_j)$. Thus the claim is proved, and all degrees are equal (to d).

Then each line has d points. In this case, $(X, \mathcal{L})$ is a $(n, d, 1)$-SBIBD. Since the d lines through a single point cover all points, $n = d(d-1) + 1$, finishing the proof of this case. Since any SBIBD has a dual that is also a SBIBD with the same parameters (see any book on designs), it follows also that every pair of lines intersect in a unique point. □

In the statement of Theorem 15.1.4, if the ℓ_is are considered as lines in a geometry, the configuration (i) is called a "near-pencil". A finite projective plane (of order $d - 1$) satisfies condition (ii). (According to [240], Levi [608] produced another "non-projective" example with $d = 9$. Surely, what de Bruijn and Erdős meant was "non-Desarguesian", since a SBIBD with parameters from above is a FPP. Indeed, in Levi's book, a recipe is given for creating a non-Desarguesian FPP, and there seems to be no mention of a "non-projective" example with degree 9.)

Recall that Fisher's inequality (Theorem 12.2.6) says that in a (v, b, k, r, λ)-BIBD, $v \leq b$. A BIBD with $\lambda = 1$ satisfies the conditions of Theorem 15.1.4, and so Theorem 15.1.4 generalizes the case of Fisher's inequality for $\lambda = 1$.

Corollary 15.1.5 (de Bruijn–Erdős, 1948 [240]). *If $n \geq 3$ points in the Euclidean plane do not all lie on a line, then at least n of the lines joining them are different.*

De Bruijn and Erdős noted that Corollary 15.1.5 also follows easily from the Sylvester–Gallai theorem (Theorem 15.1.2); showing this is a popular exercise (see, *e.g.*, [294, 8.35, p. 209]); Corollary 15.1.5 is sometimes attributed to only Erdős (e.g, by Coxeter [223, p. 30]), so Corollary 15.1.5 might have been known prior to 1948.

Exercise 270 (Another proof of Corollary 15.1.5)**.** *Prove Corollary 15.1.5 by applying the Sylvester–Gallai theorem (Theorem 15.1.2) and mathematical induction.*

Exercise 271. *Use Theorem 15.1.2 to show that for any $n \in \mathbb{Z}^+$, n lines in the plane, not all through one point or not all parallel, determine at least $n-1$ intersection points.*

In a point-line configuration, a line that contains only two points is sometimes called an *ordinary line* (or a *Gallai line*). Let $g(n)$ be the minimum number of ordinary lines for n points in $\mathbb{R}^2$. In 1941, Melchior [652] published that $g(n) \geq 3$ (this was again proved in 1948 by de Bruijn and Erdős [240]). Melchior's proof was given for the dual form: every collection of $n \geq 3$ non-concurrent lines (in the real projective plane) has at least 3 ordinary points (see [685] for a proof based on Euler's formula for planar graphs).

Exercise 272. *For $k \in Z^+$, show that $g(2k) \leq k$.*

The cases where n is odd proved to be more challenging. In 1951, Motzkin [675] showed that $g(n) \geq 2\sqrt{n} - 2$.

It was conjectured by Dirac [260] (in 1951) that $g(n) \geq n/2$, however this is false; the Kelly–Moser configuration (given in 1958 [539]) in Figure 15.3 shows that $g(7) \leq 3$.

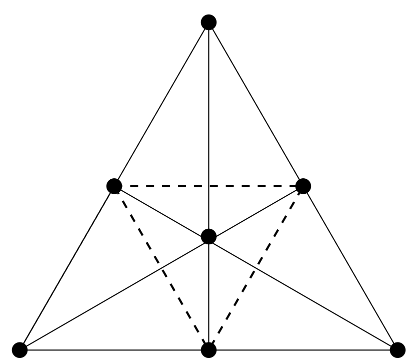


Figure 15.3: The Kelly–Moser configuration with Gallai lines dashed

Exercise 273. *Find an example that proves $g(13) \leq 6$.*

Much work in the study of configurations was done in terms of "pseudoline arrangements", which were introduced by F. W. Levi (1888–1966) in 1926 [607]. Very briefly, a pseudoline arrangement is a collection of closed curves (in the real or the real projective plane) so that any two intersect in at most one point. Pseudoline arrangements are not studied here; for more information see [399].

Remark 15.1.6. *The full name of F. W. Levi is Freidrich Wilhelm Daniel Levi, not to be confused with other geometers named Levy. Levi was a (German) Jesuit who went to India to spread the word of the bible (and maybe to avoid the war and the Nazis), but seems to have had success in teaching mathematics to Indian statisticians; I seem to remember someone telling me that he started out studying group theory, yet gave a sequence of six lectures on finite geometries [608] given in Calcutta in February 1940. It has been said that Levi's manuscript is the first book on finite geometries.*

On the other hand, there are at least three authors named Levy, all studying some kind of geometry. Harry Levy wrote a book Projective and related geometries *[609]. Lawrence S. Levy is from Wisconsin, known for his popular 1970 book* Geometry: Modern mathematics via the Euclidean plane *[610]. Finally, Silvio Levy edited a book called* Flavors of geometry *[611] published in 1997 (the first chapter of which is the treatise on convexity by Keith Ball).*

In 1958, Leroy Kelly and Willy Moser [539] showed that $g(n) \geq \frac{3n}{7}$, and by the Kelly–Moser configuration in Figure 15.3, this bound is tight for $n = 7$. In 1972, Kelly and Rottenberg [540] gave the same lower bound of $\frac{3n}{7}$ for pseudoline arrangements.

Constructions by Böröcky (and "partially rediscovered by Mckee"—see [228]) showed that if n is odd:

$$g(n) \leq \begin{cases} \frac{3(n-1)}{4} & \text{if } n \equiv 1 \pmod 4; \\ \frac{3(n-3)}{4} & \text{if } n \equiv 3 \pmod 4. \end{cases}$$

Until very recently, the best known lower bound was:

Theorem 15.1.7 (Csima–Sawyer, 1993 [230])**.** *If n points are given in the plane, at least $\frac{6n}{13}$ lines are ordinary. In other words, $g(n) \geq \frac{6n}{13}$.*

[Joe Csima was a combinatorics professor at McMaster until the mid-to-late 90s.]

By August 2012, Green and Tao had proved a version of Dirac's conjecture (also called "the Dirac–Motzkin conjecture"):

Theorem 15.1.8 (Green–Tao, 2013 [415]). *For n sufficiently large, $g(n) \geq \lfloor n/2 \rfloor$. Furthermore, if n is odd, $g(n) \geq 3\lfloor n/4 \rfloor$.*

One main tool in the proof of Theorem 15.1.8 was the use of cubic curves (and Melchior's original proof of the Sylvester-Gallai theorem that used projections and Euler's formula).

There are many other results on ordinary lines—too many to list here. For example, Motzkin [676] and M. O. Rabin (see [303]) independently answered a question of Ron Graham, showing that ordinary lines have, in a sense, a Ramsey property (for every 2-colouring of the points, there exists a monochromatic ordinary line). See the survey by Erdős and Purdy [313] for a proof or see the article by Peter Borwein and M. Edelstein [135]. The article by Nilakantan [685] is also valuable for history, proofs, and references for problems related to the Sylvester–Gallai theorem and the Dirac–Motzkin conjecture.

Motzkin [677] also studied ordinary lines (and "near ordinary" lines, those with only 3 points) in the projective plane. Other variants of Sylvester's problem are numerous; for example, Peter Borwein [134] looked at three extensions to cases where sets of points are compact or countable or [133] in Haar spaces. Dvir [284] has also written extensively on Sylvester–Gallai type problems.

15.2 The Sylvester–Gallai problem for circles

There are different kinds of questions one can ask about intersection patterns of circles. For example, see Section 16.4 for regions determined by intersecting circles. This section contains only analogues of questions of the Sylvester problem, but for circles instead of lines.

In Exercise 271, it was asked to show that if n lines are not all parallel or do not all pass through one point, then the number of intersection points is at least $n - 1$. A corresponding result for circles was given by K. Bezdek and R. Connelly:

Theorem 15.2.1 (Bezdek–Connelly, 1988 [87]). *Let $\mathcal{C}$ be a family of n unit circles in the plane so each circle intersects at least one other, and no two circles are tangent. Then the number of intersection points is at least n.*

Proof: (This proof is reproduced from [82].) Let v be the number of intersection points. Assign to each intersection point a "charge" of 1. (So the total charge in the arrangement is v.) For each intersection point $\mathbf{p}$, distribute its charge equally among all circles passing through $\mathbf{p}$. (So if $\mathbf{p}$ has 3 circles passing through it, each of these circles receives a charge of 1/3.) The total charge on a circle is the sum of charges contributed from each of its vertices.

Count the charge from the perspective of each circle. Let $C \in \mathcal{C}$, and let $\mathbf{x}$ be a vertex on C that contributes the least charge, say $1/k$. Then there are k circles passing through $\mathbf{x}$, one of which is C, and $k-1$ more, say $C_1, \ldots, C_{k-1}$. Since there are no circles tangent, each of $C_1, \ldots, C_{k-1}$ intersect C again and in a point contributing at least $1/k$ to the charge of C. So the total charge on C is at least 1. Hence the total charge for the configuration is at least n, and so $v \geq n$. □

Comment: Since the idea in the above proof of Theorem 15.2.1 is a simple double counting argument, is there a similar graph-theoretic proof? For each intersection point, the number of circles through that point is half of the degree of that vertex. For any given circle, the sum of the degrees of vertices on that circle is twice the number of other circles intersecting the given circle. The proof above uses the fact that the number of circles intersecting a given circle is at least one less than half of the highest degree on that circle. Is there another proof, perhaps one that yields a stronger bound?

The dual form of the Sylvester–Gallai theorem says that if n lines do not all contain the same point (that is, they do not form a pencil) then there exists a point with only two lines containing it. If lines are replaced with circles, this is not true by the following example on four unit circles: Three distinct circles, each meeting in two points, pass through one point P and a fourth circle passes through three the three remaining intersection points. Such a configuration is called "exceptional" (see [725] for a diagram).

In 1999, the following conjectures were given by Andras Bezdek [84].

Conjecture 15.2.2. *Let $\mathcal{C}$ be a finite family of unit circles in the plane.*
(a) If at least two circles from $\mathcal{C}$ intersect, then there is an intersection point that is incident to at most three circles.

(b) If every pair of circles from $\mathcal{C}$ intersect, then there is an intersection point that is incident to at most three circles. If $\mathcal{C}$ is not exceptional (defined above), then there is an intersection point incident to exactly two circles.

In 2002, Ron Pinchasi [725] proved (b).

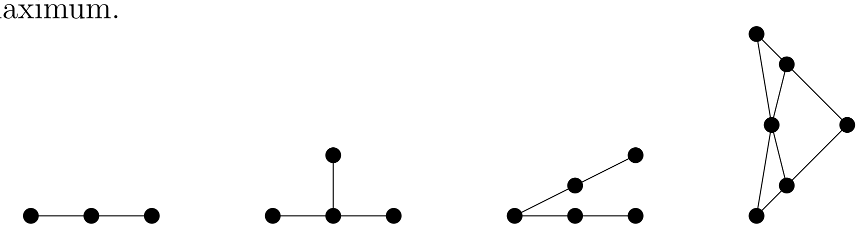
A related problem asks if a set of n points in the plane have diameter at most 2, is there a unit circle containing exactly two of these points? (This was conjectured by Bezdek, too.) See Pinchasi's paper [725] for more details and references, which include [43], [83], [84].

Also see [21], [85], and [86] for more Sylvester–Gallai type results for circles. See [928] for Sylvester's problem for conic sections, or [913] for spreads of curves.

15.3 Orchard tree planting

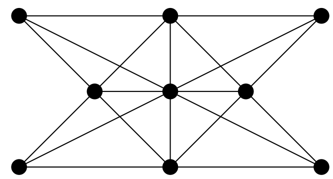
For this section, all geometry is in the Euclidean plane unless otherwise specified.

For each positive integer $n \geq 3$, what is the maximum number $f(n)$ of 3-point lines that can be found among n points in the plane? When $n = 3$, one line is possible. When $n = 4$, still only one such line is possible. However, with $n = 5$, two lines sharing a point can be constructed. With $n = 6$ points, four lines are possible, and with a little experimenting one sees that four is maximum.



So $f(3) = f(4) = 1$, $f(5) = 2$, and $f(6) = 4$. Other values are known: $f(7) = 6$, $f(8) = 7$, $f(9) = 10$, $f(10) = 12$, $f(11) = 16$, $f(12) = 19$, $f(13) = 22$, and $f(14) = 26$.

Here is a configuration that shows $f(9) \geq 10$:



Looking at the values for small n, it might be hard to guess that $f(n)$ is quadratic. By simple counting,

$$f(n) \leq \left\lfloor \frac{\binom{n}{2}}{\binom{3}{2}} \right\rfloor = \left\lfloor \frac{n^2}{6} - \frac{n}{6} \right\rfloor .$$

Using Theorem 15.1.7

$$f(n) \leq \left\lfloor \frac{\binom{n}{2} - \frac{6n}{13}}{3} \right\rfloor = \frac{n^2}{6} - \frac{25n}{13}.$$

The known lower bounds are also quadratic. For a lower bound, it suffices to give a construction. Sylvester showed that $f(n) \geq \frac{1}{8}n^2$ by using points on the curve $y = x^3$. In 1974, Sylvester's bound was improved to

$$f(n) \geq \left\lfloor \frac{n^2}{6} - \frac{n}{2} \right\rfloor + 1 \tag{15.5}$$

by Burr, Grünbaum and Sloane [168] (they used Weierstrass's elliptic functions); ten years later, Füredi and Palásti [359] obtained the same bound with a simpler construction using "hypocycloids". Finally, in 2013, Ben Green and Terence Tao [415] showed that the lower bound in (15.5) is indeed tight.

The orchard planting problem can also be posed in a more general form, where "3 trees per row" is replaced by "k trees per row". According to Erdős and Purdy [313], Croft and Erdős showed that the lower bound is proportional to n^2 but inversely proportional to k^3.

In *Canterbury puzzles* [270, pp. 19–20], Puzzle 21 is called "The ploughman's puzzle"; it is asked to find an arrangement of 16 trees into rows so that each row has four trees, and there are a maximum number of rows. An example is given with 12 rows in Figure 15.4

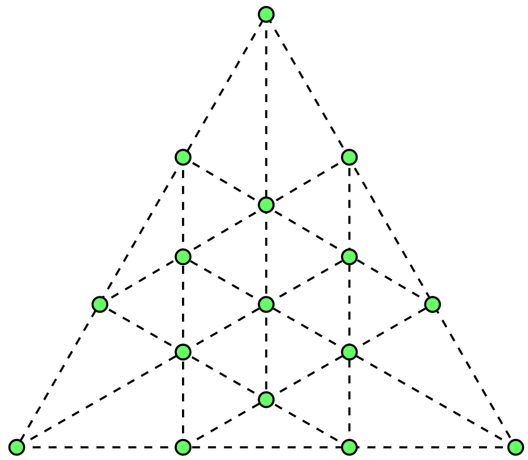


Figure 15.4: A 12 row example for the Ploughman's Puzzle

In his solutions, Dudeney [270, p. 140] gave the 15 row example shown in Figure 15.5, which he believed was maximal.

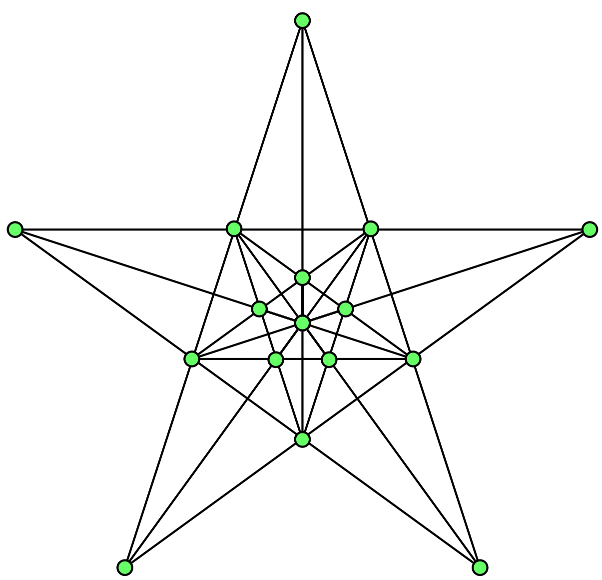


Figure 15.5: A 15 row solution to the ploughman's puzzle, given by Dudeney [270, p.140]

Dudeney [270] writes:

> This is in excess of what was for a long time believed to be the maximum number of rows possible, and though with our present knowledge I cannot rigorously demonstrate that fifteen rows cannot be beaten, I have a strong "pious opinion" that it is the highest number of rows possible.

15.4 Configurations of points and lines

15.4.1 Introduction

The word "configuration" has been used (here and elsewhere) to tacitly mean a collection of points and lines, usually in some plane. The Sylvester–Gallai theorem (Theorem 15.1.2) and the orchard problems (see Section 15.3) are about such configurations in the plane.

The word "configuration" can also be used to indicate a system of points and "lines", where "lines" are just sets—not necessarily straight lines in the plane, so that any two lines intersect in at most one point. If X is a set and $\mathcal{F} \subseteq \mathcal{P}(\mathcal{X})$ is a collection of subsets of X, say that $(X, \mathcal{F})$ is a "combinatorial" configuration The de Bruin–Erdős theorem (Theorem 15.1.4) is about (combinatorial) configurations—geometry or the Euclidean plane is not mentioned.

In a combinatorial configuration, it is usually assumed that "lines" behave as lines in that any two intersect in at most one point (such configurations have been called projective [560].) Another tacit assumption is that the configuration (or its underlying graph) is connected. Popular books on configurations are Grünbaum's *Configurations of points and lines* [429] and Pisanski and Seratius's *Configurations from a graphical viewpoint* [726].

If a configuration can be drawn in the Euclidean plane with straight lines, the configuration is called *geometric*, or *realizable in the Euclidean plane*. A configuration is often called *coordinatizable* over $\mathbb{R}$, $\mathbb{Q}$ or $\mathbb{C}$ if the configuration is realizable with points having respectively, real, rational, or complex coordinates.

Oddly enough, if a configuration is a finite geometry where every two points are in some line, and each line contains at least 3 points, then by the Sylvester–Gallai theorem (Theorem 15.1.2), the configuration is not geometric! As made explicit in Corollary 15.1.3, the Fano configuration is not geometric (but it is a geometry!). So one might take the expression "geometric configuration" with a grain of salt. If points of a configuration are given in the plane so that each line of the configuration can be drawn with a pseudoline, the configuration is called "topological"; however, topological configurations are not covered here.

Is there a classification of those combinatorial configurations that are realizable in the real plane? Some cases are considered where every point is on the same number of lines and every line has the same number of points. The Fano configuration is such an example (and as mentioned already, the Fano configuration is not geometric).

For example, a Steiner triple system is a set system $(X, \mathcal{S})$ such that X is a finite non-empty set and $\mathcal{S} \subset [X]^3$ is a set of triples so that for any two points in X, there exists precisely one triple containing these two points. If $|X| = n$, then $|\mathcal{S}| = \frac{1}{3}\binom{n}{2}$ and each point is contained in $\frac{n-1}{2}$ triples. An example of a STS for $n = 7$ was given in Section 11.4.4, which is equivalent to the Fano plane.

In 1847, Kirkman [547] proved that a Steiner triple system exists on n points if and only if n is congruent to either 1 or 3 modulo 6. In 1850, Kirkman [548] posed the following problem:

> Fifteen young ladies in a school walk out three abreast for seven days in succession: it is required to arrange them daily, so that no two will walk twice abreast.

Later in 1850, Cayley [182] published a solution to the "Kirkman schoolgirl problem"; Kirkman also published his solution [549] the same year. Stinson [832] gives an extensive survey of the Kirkman schoolgirl problem, its generalizations, and many references.

If a set X has n points, a Kirkman triple system (KTS) on X is a set $\mathcal{T}$ of triples from X so that $(X, \mathcal{T})$ is a STS and there exists a partition of $\mathcal{T}$ into "parallel classes", each parallel class being a set of disjoint triples that partition X (in Kirkman's original problem, the days of the week correspond to the parallel classes). If a KTS exists on n points, then n is divisible by 3; since a KTS is also a STS, it follows that $n \equiv 3 \pmod 6$. The general problem of those n for which there exists a KTS was solved in 1971 by Ray-Chaudhuri and Wilson [746], thereby showing that a KTS on n points exists if and only if $n \equiv 3 \pmod 6$.

The Kelly–Moser configuration (see Figure 15.3) is a collection of 7 points and 9 lines, each of six lines incident with 3 points and three lines with

2 points (the drawing given shows that the Kelly–Moser configuration is geometric).

Some configurations (like those outlined in the solutions to Exercises 272 and 273) can be considered as configurations in the real projective plane.

One of the reasons to study configurations is to identify properties of a geometry; for example, it has been shown that the Pappus and Desargues configurations are possible in the projective plane over a field, but that Pappus may fail in the plane over the quaternions. It has also been shown (e.g., in Corollary 11.4.21) that the Fano configuration is not realizable in the real projective plane.

Definition 15.4.1 (Extended complete quadrangle)**.** In a quadrangle on points A, B, C, D, there are three pairs of *opposite* lines, namely $(A \vee B, C \vee D)$, $(A \vee C, B \vee D)$, and $(A \vee D, B \vee C)$. The intersection points of these pairs are called *diagonal* points of the quadrangle. The configuration formed by the four points A, B, C, D and the three diagonal points is called an *extended complete quadrangle* (sometimes called the completed quadrangle).

If the three diagonal points of an extended complete quadrangle are collinear, then the extended complete quadrangle is a copy of the Fano plane $\mathrm{PG}(2, \mathbb{F}_2)$, also called a Fano configuration or Fano subplane (see Section 11.4.6 for more on subplanes).

Theorem 15.4.2 (Gleason, [389] 1956)**.** *If every complete quadrangle (four points, no three collinear, and the six lines thereby determined) in a projective plane can be extended to a Fano configuration, the plane is Desarguesian. In other words, if the diagonal points of every complete quadrangle are collinear, then the plane is Desarguesian.*

As one example demonstrating how configurations in a plane can say something about the plane, consider what Killgrove [544] called the "k-configuration" (some authors call it the "K-configuration"), as depicted in Figure 15.6.

Killgrove observed that if a plane contains a k-configuration, then it is not Desarguesian. One of the non-Desarguesian (finite) projective planes of order 9, called the Hughes plane, indeed contains a k-configuration. The projective Moulton plane also contains k-configurations, and, the affine Moulton plane (see [679]) is not Desarguesian.

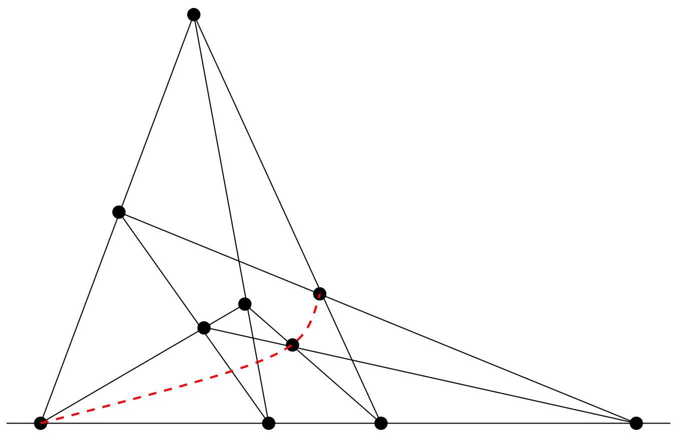


Figure 15.6: Killgrove's k-configuration

15.4.2 Tactical configurations

Some authors (e.g., Longyear [618]) call a collection of points and "lines" a *tactical configuration* if each line contains the same number of points. The term "tactical configuration" was used by E. H. Moore [665] in 1896 to describe a larger class of configurations (where, e.g., "lines" can intersect in more than just a single point). The word "tactical" comes from Cayley's [183, p. 294] division of algebra into "Tactic" and "Logistic"; here is Cayley's description:

> Algebra is an Art and a Science; *quá* Art, it defines and prescribes operations which are either tactical or else logistical; viz. a tactical operation is one relating to the arrangement in any manner of a set of things; a logistical operation (I prefer to use the new expression instead of arithmetical) is the actual performance, so as to obtain for the result a number, of any arithmetic operations (of course, given operations) finite in number, since these alone can be actually performed, upon given numbers.

Cayley then continues with the example of summing the numbers $1, 2, \ldots, n$ in two different ways giving a tactical proof: "…it is always tactic which determines what logistical operations are to be performed."

Tactical configurations include balanced incomplete block designs (which include finite projective planes). If further, two points are incident with at most one line and two lines are incident with at most one point, a tactical configuration is called an *incidence geometry* (see Chapter 11).

15.4.3 "Regular" configurations

Definition 15.4.3. For positive integers m, n, a, b, an (m_a, n_b) *configuration* is a (combinatorial) configuration with m points, and n lines, such that each point is contained in a lines, each line contains b points, and every pair of points is in at most one line.

Since combinatorial configurations are set systems, one can find an incidence matrix. In an (m_a, n_b) configuration, if points are labelled $P_1, P_2, \ldots, P_m$ and lines are labelled $\ell_1, \ell_2, \ldots, \ell_n$, then the configuration can be represented by an $n \times m$ 0-1 matrix $M = (m_{ij})$, where $m_{ij} = 1$ if and only if ℓ_i contains P_j. Note that an incidence matrix can be given for any ordering of the points and lines, and so is not unique (but any incidence matrix can be transformed into another one by permuting rows and columns). Each 1 in the incidence matrix represents what is sometimes called a "flag", a pair (P, ℓ) so that P lies on ℓ.

Double counting flags (or 1s in an incidence matrix) gives the following useful lemma.

Lemma 15.4.4. *For positive integers m, n, a, b, if an (m_a, n_b) exists, then $ma = nb$.*

By Lemma 15.4.4, a $(6_3, 4_4)$ configuration does not exist. An open problem is to classify all 4-tuples m, n, a, b so that an (m_a, n_b) configuration exists, and for such 4-tuples, how many configurations have these parameters.

For examples of configurations with small parameters, see Figure 15.7.

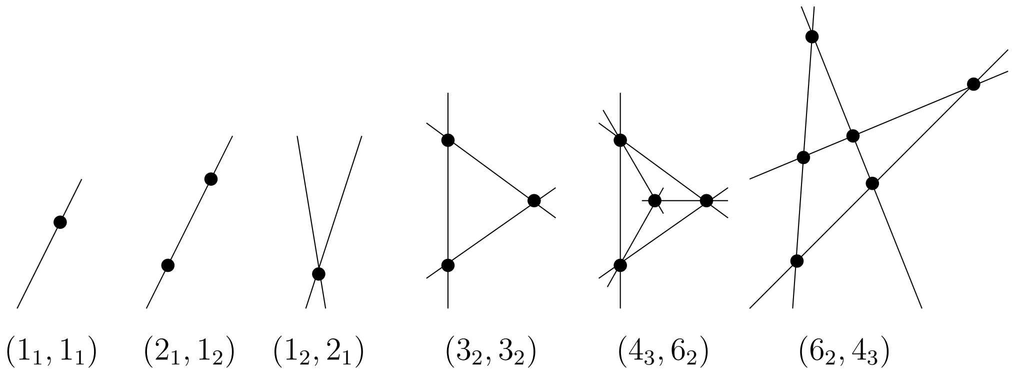


Figure 15.7: Some small configurations

The configuration given in Figure 15.4 for the Ploughman's Puzzle is a $(16_3, 12_4)$ configuration.

For example, what is sometimes called the "Hesse configuration" or "Young's geometry" (see Figure 15.8) the affine plane AG$(2, 3)$ is a $(9_4, 12_3)$ configuration.

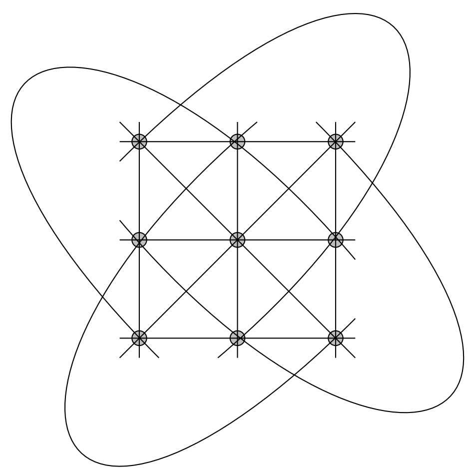


Figure 15.8: The Hesse $(9_4, 12_3)$ configuration, AG$(2, 3)$

According to a survey by Artebani and Dolgachev [36], the Hesse configuration was introduced by Colin Mclaurin (1698–1746) (a 1910 reference

[713, p. 384] is given as evidence). The Hesse configuration was made popular in 1844 when its structure was studied by Hesse [467], hence its present name. Hesse saw the configuration in terms of inflection points of plane cubic curves (and orbits in the "projective group" of the plane, a discussion not developed here—see [36] for details). The Hesse configuration can be seen as the affine plane AG$(2,3)$, the 2-dimensional affine geometry over the finite field $\mathbb{F}_3$, where lines are solutions to $ax + by \equiv c$ (modulo 3). There are 9 points and 12 lines, each point incident with 4 lines, and each line contains 3 points, and so the Hesse configuration is a $(9_4, 12_3)$ configuration.

Exercise 274. *Let q be a power of a prime. Show that the affine plane $AG(2, q)$ is a $(q^2_{q+1}, (q^2 + q)_q)$ configuration.*

If C is a (m_a, n_b) configuration with incidence matrix M, its dual (interchanging points and lines) is an (n_b, m_a) configuration with incidence matrix M^T.

When $m = n$ (and hence $a = b$) an (m_a, m_a) configuration is called *symmetric*, in which case the notation (n_a) is shorthand for (n_a, n_a). Hence, a finite projective plane of order q is a $((q^2 + q + 1)_{q+1})$ configuration.

Although the dual of a symmetric configuration is again a symmetric configuration with the same parameters, it is not clear when the dual of a configuration is isomorphic to its dual. For example, the two non-Desarguesian Hall planes of order 9 are duals of each other, but they are not isomorphic.

15.4.4 (n_3)-configurations

This section is a short survey inspired by the guest lecture Dr. Bill Kocay gave to my Combinatorial Geometry class on Wednesday 6 April 2016. Branko Grunbaum's 2009 book, *Configurations of points and lines* [429], gives an extensive history regarding the classification of (n_3) configurations. (Grunbaum was born 2 October 1929 and passed away 14 September 2018.)

As already mentioned, the study of (n_3) configurations seems to have originated in the mid-1800s with the work of Thomas Kirkman. Other authors that have laid the foundations for such configurations include Von Sternick (1860s), Schröter (1900s), Klein, Möbius (1820s), Steinitz, Kantor (late 1800s), and more recently, Grunbaum, Kocay and Pisanski.

The Fano plane is the only (7_3) configuration, but not a geometric one (see Corollary 15.1.3 or 11.4.21).

In 1828, Möbius asked [660] if there is a configuration consisting of two quadrilaterals, where vertices of each lie on the edges of the other. In 1882, Seligmann Kantor [524] produced a family of configurations, one of which answered the question of Möbius. What is now known as the Möbius–Kantor configuration, is the unique (8_3) configuration (which is not geometric in the real plane but is realizable in the complex plane), can be formed from the Hesse configuration by deleting a vertex. With the vertices numbered 1 through 8, the lines are given by 125, 148, 167, 236, 278, 347, 358, 456. There are two common drawings; one way is to put the vertices in a square as follows,

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

and then complete the drawing as in Figure 15.8 with the center vertex deleted (together with all lines containing that vertex). The Levi graph (incidence graph) of the Möbius–Kantor configuration (called the Möbius–Kantor graph) is well studied; for example, this graph is both vertex and edge-transitive, is a unit-distance graph (see Section 17.4.6) and is related to the Petersen graph. See [219] (or more recently, [634]) for more details and references.

The Pappus configuration (see Figure 1.2) is a (9_3) configuration and the Desargues configuration (see Figure 1.3) is a (10_3) configuration. Both the Pappus and the Desargues configurations are geometric. There are three (9_3) configurations and ten (10_3) configurations. All but one of the ten (10_3) configurations are geometric (see Figure 15.11 below).

A configuration is called a theorem in some plane if no matter how all but one of the lines are chosen (coordinatized) in the plane, the last line is also. For example, the Desargues configuration and the Pappus configuration are both theorems in the real projective plane. However in $\mathrm{PG}(2, \mathbb{H})$ (the projective plane over the quaternions), the Pappus configuration is not a theorem. Note that a Pappus configuration can be drawn by using lines in $\mathrm{PG}(2, \mathbb{H})$ if one chooses only real coordinates, but in general, Pappus' theorem fails in $PG(2, \mathbb{H})$.

One of the (9_3) configurations can be formed by nesting three equilateral triangles, as in Figure 15.9.

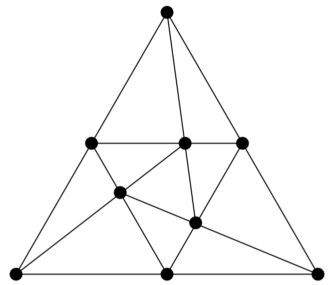


Figure 15.9: One of the two (9_3) configurations other than Pappus

Exercise 275. *Find a coordinatization for points in the* (9_3) *configuration given in Figure 15.9. You may assume that the points of the second triangle bisect edges of the first.*

Exercise 276. *Prove that the* (9_3) *configuration given in Figure 15.9 is not isomorphic to the Pappus configuration.*

A simple looking (12_3) configuration is also found by nesting three squares, as in Figure 15.10.

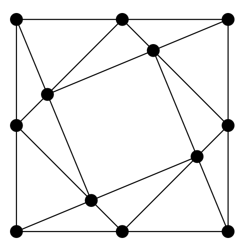


Figure 15.10: A (12_3) configuration

Exercise 277. *Find coordinates for the points in the* (12_3) *configuration in Figure 15.10. You may assume that the edges of the outer square are bisected by the points of the next square inside.*

For each $7 \leq n \leq 15$, number of (n_3) configurations are 1, 1, 3, 10, 31, 229, 2036, 21399, 245342. The ten (10_3) configurations appeared in a paper by Kantor [523] in 1881. Kantor gave drawings to show that all were geometric. However, Kantor made a mistake with one, which was shown by

Schroeter [782] in 1889 to be not geometric (but Schroeter confirmed that the remaining nine diagrams were correct—see also [781] for constructions, with no mention of Kantor).

So only one of the ten (10_3) configurations fails to be geometric (depicted in Figure 15.11), which has since been called the "anti-Pappian" (10_3) (by, e.g., Glynn [392]).

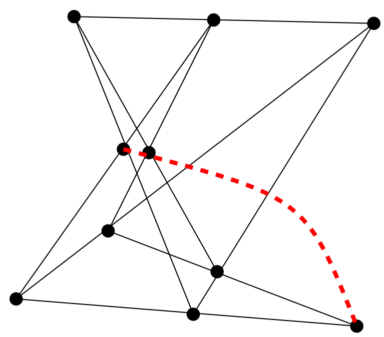


Figure 15.11: The only (10_3) configuration that is not geometric

Sturmfels and White [836, 837] showed that all (11_3) and (12_3) configurations are geometrically realizable in the *rational* plane.

Conjecture 15.4.5 (Grunbaum, 2009 [429, p. 151])**.** *Any (n_3) configuration that can be coordinatized over $\mathbb{R}$ can also be coordinatized over $\mathbb{Q}$.*

Grunbaum's conjecture was recently confirmed for (13_3)-configurations by Kocay.

Theorem 15.4.6 (Kocay, 2018, manuscript)**.** *All 2036 (13_3)-configurations are coordinatizable over the rational plane.*

For an (n_3) configuration C, its incidence graph (or Levi graph) is the equibipartite 3-regular (cubic) graph defined on $2n$ vertices (n vertices and n lines) where an edge between a point P and a line ℓ is present if and only if P lies on ℓ. (See Section 11.6 for more on incidence graphs.)

For example, the Heawood graph (see Figure 11.3) is the incidence graph of the Fano plane. Any equibipartite cubic graph also gives rise to an (n_3) configuration. It seems to be difficult to decide when such a graph gives rise to a geometric configuration. Insisting on additional properties of a bipartite cubic graph might help to analyze the associated configuration(s). For example, Grünbaum [428] studied the "Georges graph", a non-Hamiltonian bipartite cubic graph on 50 vertices (given by Georges in [382]), and found that the associated (25_3) configuration (now called the Georges configuration) can be coordinatized in the real plane. Kocay [557] later showed that the Georges configuration can be coordinatized over the rational plane, confirming yet one more case of Conjecture 15.4.5.

Kocay [559] also studied ways to extend a coordinatized configuration to one with one more point.

15.4.5 Complete n-point configurations

The simplest configuration consists of only a single point and no lines; such a configuration is of type $(1_0, 0_0)$. The $(2_1, 1_2)$ configuration given in In Figure 15.7 contains two points contained in one line. Also given in Figure 15.7 is a $(3_2, 3_2)$ configuration with three points, and through each pair of points, there is one line. As already mentioned, a quadrangle (four points, no three collinear, and the six lines thereby determined) is a $(4_3, 6_2)$ configuration. The dual of a complete quadrangle is called a complete quadrilateral, a $(6_2, 4_3)$ configuration, with four lines and six points, where for each pair of lines there is a unique intersection point.

Definition 15.4.7. For a positive integer n, a complete n-point is a configuration of n points, where each pair of points has a (unique) line containing them. A complete n-line is a configuration with n lines, so that each pair of lines has a different intersection point.

From the definition, a complete n point is a $(n_{n-1}, \binom{n}{2}_2)$ configuration. A frame (a complete quadrangle—see Definition 11.4.2) is a complete 4-point configuration.

By the axioms for a projective plane, every projective plane contains a complete 1-point, 2-point, 3-point, and 4-point. However, for $n \geq 5$, a complete 5-point is not always possible (see Exercise 278).

Exercise 278. *Show that in the Fano plane $PG(2, \mathbb{F}_2)$, there is no complete 5-point.*

In many larger projective planes, a complete 5-point indeed exists.

Exercise 279. *Let q be a power of a prime. Show that the projective plane $PG(2, \mathbb{F}_q))$ contains a complete $(q+1)$-point. Hint: Look at points with homogeneous coordinates of the form $(1, a, a^2)$ together with one more point.*

As pointed out in Section 11.7, it might be of interest to know that any (simple) graph can be embedded (see Definition 11.7.1) in any large finite projective plane:

Theorem 15.4.8 (Mellinger–Vaughn–Vega, 2015 [653])**.** *Let G be a simple graph on n vertices, and let Π be a finite projective plane of order q. If $q \geq \frac{n(n-3)}{2}$ then G can be embedded in Π.*

See [653] for the history of embedding graphs in finite projective planes. (Oddly, it seems that the authors of [653] were not aware of the example given in Exercise 279, since they showed, by induction, that if $q \geq n(n-3)/2$, then the plane contains an n-point, i.e., embeds a complete graph K_n.) The same authors also look at embeddings of graphs depending upon the existence of Baer subplanes (see Definition 11.4.18) and ovals (see Chapter 13).

15.4.6 Point-line incidences and the Szemerédi–Trotter theorem

For positive integers n and m, let $I(n, m)$ be the maximum number of incidences in any configuration of n points and m lines in the plane.

Theorem 15.4.9 (Szemerédi–Trotter, 1983 [851])**.** *For $m, n \geq 1$, $I(n, m) = O\left((nm)^{2/3} + n + m\right)$.*

Also in 1983, J. Beck [68, Theorem 1.5] proved a similar result.

The original proof of the upper bound in Theorem 15.4.9 used a complicated argument. In 1997, Székely [847] gave a simple proof of Theorem 15.4.9 that used "crossing numbers"; this proof is given in Section 17.3.8.

Erdős (see [305]) showed that the bound in the Szemerédi–Trotter theorem is tight by looking at "rich lines" formed by the $\sqrt{n} \times \sqrt{n}$ integer lattice. The Erdős proof relied on some number theory. Beck [68] also points out that this lattice shows the lower bound within a constant (Beck says " we leave the trivial proof to the reader."). Elekes found another example:

Example 15.4.10 (Elekes, 2002 [292]). *For some $N \in \mathbb{Z}^+$, consider the integer lattice $P = [1, N] \times [1, 2N^2]$. Define a set of lines on point set P by*

$$\mathcal{L} = \{y = ax + b : m \in [1, N], b \in [1, N^2]\}.$$

Then $|P| = 2N^3$ and $|\mathcal{L}| = N^3$. Since each line in $\mathcal{L}$ is incident with N points (using $x = 1, 2, \ldots, N$), the total number of incidences is N^4. Putting $m = 2N^3$ and $n = N^3$, the number of incidences is $N^4 = (N^3N^3)^{2/3} = (\frac{m}{2}n)^{2/3}$.

Csaba Tóth [871] proved a complex plane version of the Szemerédi–Trotter theorem and some subsequent corollaries. (Tóth's paper was submitted in 1999, but appeared only in 2015.) The Szemerédi–Trotter theorem was generalized to higher dimensions by Agarwal and Aronov [8]. For higher dimensions, a lower bound was given by Edelsbrunner [287]. Solymosi and Tao [806] gave a higher dimensional version (using the "polynomial ham sandwich theorem"). See Dvir's survey [284] for more on counting incidences.

In Exercise 280 is a result counting incidences between points and curves (its solution does not need the Szemerédi–Trotter theorem).

Exercise 280. *Let $V \subset \mathbb{R}^2$ be a set of n points and let $\mathcal{L}$ be a collection of n curves in the plane such that any two curves in $\mathcal{L}$ intersect in at most two points. Prove that the number of point-curve incidences is $O(n^{3/2})$.*

The bound in the Szemerédi–Trotter theorem does not hold for projective geometries, since in a FPP with $n = q^2 + q + 1$ points and n lines, there are on the order of $q^3 \sim n^{3/2}$ incidences.

Exercise 281. *Let P be a set of n points in the plane. Use the Szemerédi–Trotter theorem to prove that the number of triangles with area 1 (unit triangles) determined by points in P is $O(n^{7/3})$.*

Note that in the solution to Exercise 281, nowhere was it needed that the triangles have unit area—any positive area suffices.

When studying incidences between points and circles, the result in the following exercise might be interesting.

Exercise 282. *Show that if three circles of the same size intersect in a common point, then the remaining three points of intersection lie on a circle—of the same size as the first three! (Here, "intersect" does not include being tangent.)*

In 1984, Elekes [291] gave a construction of n points so that $O(n^{3/2})$ circles are thereby determined (each circle having at least three of the n points). The Elekes result also follows from Exercise 282 (start with a point and many circles through that point).

The following exercise is somewhat related to point-line incidence problems.

Exercise 283. *For each positive integer n, show that a straight line can cut at most $2n-1$ squares of an $n \times n$ chessboard.*

According to Bárány and Frankl [55], the result in Exercise 283 is folklore. (They also give a proof.) See also [55] for a proof that a single hyperplane cuts at most $\frac{9}{4}n^2 + 2n + 1$ cubes in a three dimensional $n \times n \times n$ board, and references for higher dimensional versions.

15.5 Graphs from intersecting lines and circles

Recall that a graph G is a pair (V, E) where V is a non-empty set and E is a set of 2-element subsets of V. Elements of V are called vertices and elements of E are called edges. For a graph $G = (V, E)$, a k-colouring $f : V \to [k]$ is called *proper* if and only if for every $\{x, y\} \in E$, $f(x) \neq f(y)$.

15.5.1 Graphs from lines in the plane

Given a collection of lines drawn in the plane, one can form the (planar) graph whose vertices are points of intersection between lines, and edges are pairs of points that are consecutive on a line. Such a graph is planar, and so (by the Four Colour Theorem) is 4-colourable (that is, there is a 4-colouring of the vertices so that any adjacent vertices are coloured differently). Call such a graph an *underlying graph*. Note that if no three lines are concurrent (intersect at a common point), then the resulting graph is 4-regular.

Theorem 15.5.1 (Folklore)**.** *For any finite collection of lines, no three of which are concurrent, the underlying graph is 3-colourable.*

Proof: By slightly rotating the picture, one can assume that there are no two intersection points one above the other. Colour points greedily from left to right. Upon colouring any point, there are at most two neighbours (to the left) that have already been coloured, leaving at least one colour available for the present point. □

There are many large classes of planar graphs that are 3-colourable. For example, in 1959, Grötzsch [419] showed that planar graphs without triangles are 3-colourable. Grötzsch also presented a triangle-free (non-planar) graph that is not 3-colourable; this graph is now called the Grötzsch graph (see Figure 15.12), which is also the second of the "Mycielski graphs" (see, e.g., [125, pp. 129–130]).

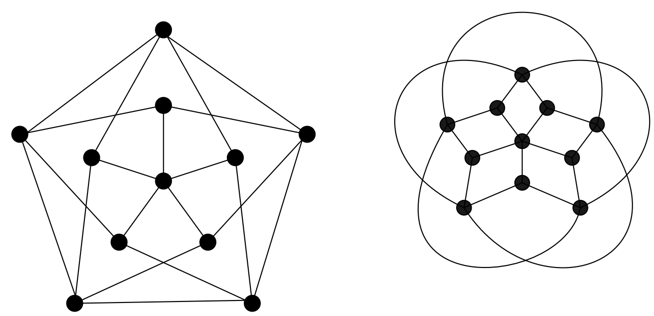


Figure 15.12: Drawings of the Grötzsch graph, 4-chromatic, girth 4, 11 vertices

In 1974, Chvátal [193] noted that the Grötzsch graph is the smallest 4-chromatic triangle-free graph, and gave a slightly larger example that is 4-regular (see Figure 15.13).

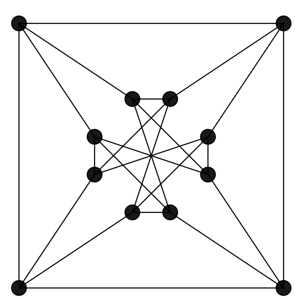


Figure 15.13: Chvátal's Graph, 4-chromatic, 4-regular, girth 4, 12 vertices

A few years later, Grünbaum [425] published the claim that if a planar graph has at most three triangles, then it is 3-colourable. (Gallai found an error in Grünbaum's proof, but it has been since corrected by Aksenov [15] in 1974. A recent proof of "Grünbaum's theorem" is by Borodin [127].) For much more on the 3-colourability of planar graphs, see Steinberg's survey [823].

15.5.2 Graphs from circles in the plane

If one forms the "underlying graph" using arcs or circles instead of lines (as in Theorem 15.5.1), the graph is not necessarily 3-colourable. One such example (see Figure 15.14) is due to Koester, [561, 562] given with 5 circles, no three of which share a common point and each two intersecting twice. Koester's graph has 20 vertices, is 4-regular, and has chromatic number 4.

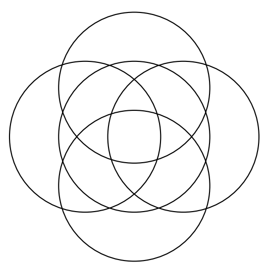


Figure 15.14: Koester's 4-chromatic graph given by 5 circles

In the early 1950s, Dirac introduced the notion of "critically k-chromatic" graphs, or in short, k-critical graphs.

Definition 15.5.2. For a positive integer k, a graph G is k-critical if and only if $\chi(G) = k$ but for any proper subgraph H of G, $\chi(H) < k$.

For example, the complete graph K_ℓ is ℓ-critical, and odd cycles are 3-critical. See Gallai [363] for early work on critical graphs (that paper is an English version of results published in two papers in 1963). Also see Kostochka's slides [128] for bounds on the number of edges in a k-critical graph.

Koester [563] produced a 4-regular critically 4-chromatic planar graph (see Figure 15.15) by using seven circles. Two of these circles do not intersect, but all other pairs of circles intersect in two points. Hence there $2\binom{7}{2}-2=40$ vertices (and hence 80 edges).

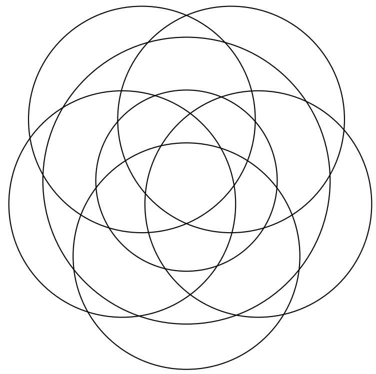


Figure 15.15: Koester's critical 4-chromatic graph given by 7 circles

Koester's example also destroyed conjectures of Gallai (planar 4-critical graphs have $|E| \leq 2|V|-2$, 1964) and Dirac (planar 4-critical graphs have a vertex of degree 3, 1957). See [264] for more on Koester graphs.

15.5.3 Great circle graphs

Another class of planar graphs generated by circles are called "great circle graphs". A *great circle* on a sphere S (in $\mathbb{E}^3$) is the intersection of S and a plane containing the center of S. For a collection of distinct great circles $C_1, C_2, \ldots, C_n$ on S, construct the graph G with vertex set $V(G) = \{x \in S : x \in C_i \cap C_j, 1 \leq i \neq j \leq n\}$ and edge set $E(G)$ consisting of all pairs x_1, x_2 so that x_1 and x_2 are adjacent points on some C_i. Such a graph is called a *great circle graph* (GCG). As the structure of a GCG does not depend on which sphere is chosen, without loss of generality, one can use $S = S^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\}$, the standard unit sphere centered at the origin. Two vertices $x, y \in V(G)$ are *antipodal* iff the associated vectors $\mathbf{x}, \mathbf{y} \in S^2$ satisfy $\mathbf{x} = -\mathbf{y}$.

The following conjecture was popularized by Stan Wagon; I do not know its origin, but the topic was mentioned in [332].

Conjecture 15.5.3. *If G is a great circle graph generated by circles no three of which are concurrent, then $\chi(G) \leq 3$.*

For example, the great circle graph formed by three circles (in general position on the sphere) is the graph of the octahedron, which is easily seen to be 3-colourable.

Chapter 16

Regions formed by geometric structures

16.1 Regions determined by lines

The next few results are of a type that one might call "partitioning space".

Lemma 16.1.1. *For an integer $n \geq 0$, let r_n denote the maximum number of regions determined by n lines. Then $r_{n+1} = r_n + n + 1$.*

Proof of Lemma 16.1.1: Consider a system S of n lines, and let ℓ be a line not in S and let $S' = S \cup \{\ell\}$. Without loss of generality, assume that ℓ is not vertical (for if it is, rotate the system slightly). In the most left part of the plane, there are two regions, one above and one below ℓ both comprising a single region in S. As one follows ℓ to the right, if ℓ is not parallel to any line in S, then ℓ crosses all lines of S. When ℓ intersects each line in S, it cuts a region of S into two new regions. So when ℓ is not parallel to any line in S, then S' has $n+1$ more regions than does S (if ℓ is parallel to any lines in S, then there are fewer). Hence, $r_{n+1} \leq r_n + n + 1$, with equality when ℓ is not parallel to any previous line. $\square$

A set of lines in the (Euclidean) plane are said to be in *general position* if no two are parallel and no three are concurrent.

Theorem 16.1.2. *For $n \geq 0$, n lines in general position in the plane partition the plane into $1 + \binom{n+1}{2}$ regions.*

The first published proof of Theorem 16.1.2 is due to the Swiss mathematician Jakob Steiner [824] in 1826. Three proofs of Theorem 16.1.2 are given here. The first is by induction (based on Lemma 16.1.1), and the second (rather elegant one, due to Moore [666]), is based on Euler's formula for planar graphs. The third proof separates regions into two classes, those with a lowest point and those without; this proof idea is mentioned in, for example, [294].

Inductive proof of Theorem 16.1.2: For $n \geq 0$, let r_n be the number of regions formed by n lines in general position (by Lemma 16.1.1, this number is unique), and let $P(n)$ be the proposition

$$r_n = 1 + \binom{n+1}{2}.$$

BASE STEP: When $n = 0$, there is only one region (the whole plane), and since $1 = 1 + \binom{0+1}{2}$, $P(0)$ is true.

INDUCTIVE STEP: Let $m \geq 0$ and suppose that $P(m)$ is true. Let S be a system of $m+1$ lines in general position. The number of regions in S is

$$\begin{aligned}
r_{m+1} &= r_m + (m+1) + 1 && \text{(by Lemma 16.1.1)} \\
&= 1 + \binom{m+1}{2} + m + 1 && \text{(by } P(m)\text{)} \\
&= 1 + \frac{m(m+1)}{2} + (m+1) \\
&= 1 + \frac{m(m+1)}{2} + \frac{2(m+1)}{2} \\
&= 1 + \frac{(m+2)(m+1)}{2} \\
&= 1 + \binom{m+2}{2},
\end{aligned}$$

showing that $P(m+1)$ holds, completing the inductive step $P(m) \rightarrow P(m+1)$.

By mathematical induction, for any number n of lines in general position in the plane, the number of regions determined is $1 + \binom{n+1}{2}$. $\square$

Second proof of Theorem 16.1.2: This proof is due to Moore [666]. Draw a circle around all the points of intersection of the n lines in general position.

Throw away the rays on the outside of this circle, and get a planar graph G. Then G has $\binom{n}{2}$ interior points, and since each line cuts the circle in two points, there are $2n$ exterior points. Since interior points have degree 4 and exterior points have degree 3, by the handshaking lemma, G has

$$\frac{1}{2}\left[4\binom{n}{2} + 3 \cdot 2n\right] = n(n-1) + 3n = n^2 + 2n$$

edges. (This number of edges is also found directly: each line gets cut $n-1$ times and so into n edges. The circle gets cut into $2n$ edges.) The number of interior regions of G is the same as the number of regions in the plane determined, so by Euler's formula $v + f = e + 2$, the number of regions (disregarding the external face of G) is

$$f - 1 = e + 1 - v = n^2 + 2n + 1 - \left[\binom{n}{2} + 2n\right] = \frac{n^2 + n + 2}{2} = 1 + \binom{n+1}{2}.$$

□

Third proof of Theorem 16.1.2: Let n lines in the plane be in general position. Assume that none of these lines are horizontal, for if one is, slightly rotate the entire picture.

Each intersection point is the lowest point of precisely one region, and thus there are $\binom{n}{2}$ regions with a lowest point. The number of regions without a lowest point is $n+1$, which can be seen by drawing a new horizontal line beneath all intersection points, and then observing that such a new line is cut by each of the n lines, and so into $n+1$ pieces.

So in all, there are $\binom{n}{2} + n + 1 = \binom{n+1}{2} + 1$ regions. □

See [921] for more history on the proofs of Theorem 16.1.2, including a "proof" by Roberts.

The number of regions determined by n lines in general position can be written in various ways. For example,

$$\binom{n+1}{2} + 1 = \binom{n}{2} + \binom{n}{1} + \binom{n}{0}.$$

Since the number of regions n points can cut a line into is $n+1 = \binom{n}{1} + \binom{n}{0}$, and the number of regions the plane can be cut into with n lines is $\binom{n}{2} + \binom{n}{1} + \binom{n}{0}$, is it possible that there is a pattern? In fact, there is a pattern for higher dimensions (see Theorem 16.7.2), and the third proof of Theorem 16.1.2 suggests the proof. The case for 3 dimensions is asked for in Exercise 312.

The next exercise is a special case of a result given in Exercise 287.

Exercise 284. *A rectangular sheet of veneer is cut into pieces using 25 straight cuts (each cut from one edge of the veneer to another edge). No three cuts go through the same point, cuts only intersect in the interior of the veneer, and there are 10 intersection points. How many pieces of veneer are formed? Is there a general formula for the number of pieces in terms of the number of cuts and number of intersections?*

Suppose the plane is divided into regions (by lines or curves). If one colours each region with one of k colours so that any two regions sharing a (non-trivial) common border receive two different colours, such a colouring is called a *proper k-colouring.*

Exercise 285. *For each $n \geq 0$ show that n lines in the plane in general position divide the plane into regions that can be properly 2-coloured.*

Exercise 286. *For $n \geq 3$, show that any n lines in general position determine $\binom{n}{2} - \binom{n}{1} + \binom{n}{0}$ bounded regions.*

Exercise 287. *If a convex region in the plane is crossed by ℓ lines with p interior points of intersection (no three intersecting at a point), then the number of disjoint regions created is $r = \ell + p + 1$. Give a proof of this statement by induction on ℓ. (Another solution is outlined in the solution to Exercise 284.)*

Exercise 288. *Suppose that n lines are in general position (no two parallel, no three concurrent). Show that it is possible to colour at least $\sqrt{n}$ of these lines blue so that no finite region has a completely blue boundary.*

Exercise 289. *Let $n = r + k$ lines be given in the plane with no three concurrent and exactly k of the lines are parallel (but no others). If $f(r, k)$ is the number of regions the plane is partitioned into, prove that*

$$f(r, k) = \frac{r^2 + r + 2}{2} + k(r + 1).$$

Exercise 290. *Given $N \geq 1$ lines in a plane in general position, prove that it is possible to assign a non-zero integer of absolute value at most N to each region of the plane determined by these lines such that the sum of the integers on either side of any of the lines is 0.*

Exercise 291 (Folklore). *Let $\ell_1, \ldots, \ell_n$ be lines in the plane. Show that some two of these lines form an angle of at most $\frac{\pi}{n}$.*

Exercise 292. *Consider the three points $P = (0,0)$, $Q = (1,0)$ and $R = (\pi, 0)$. Prove that for any other point X, at least one of the distances between X and P, Q, or R is irrational. Can you find just two points with this property?*

The next few exercises consider cutting an "area" into equal pieces; the underlying idea of using the Intermediate Value Theorem (IVT) is discussed in Courant and Robbins, [215, pp. 317–319]. For the following exercises, let "area" mean a bounded (measurable) area in the plane. For some exercises or results below, if you can not find a solution for a particular exercise below for general bounded sets, provide one for when the area is assumed to be the interior of a simple closed curve, or if necessary, when the area is convex,

Lemma 16.1.3. *Let region $B \subseteq \mathbb{R}^2$ be bounded by a simple closed curve. For any slope m, there is a unique line with slope m that bisects B (cuts B into two regions of equal area). Furthermore, there is a line containing the origin that bisects B as well (if B is convex and does not contain the origin, this line is unique).*

Proof outline: In the first case, let ℓ be a line with slope m so that the region B is entirely on one side of ℓ. Moving this line (in a parallel fashion) through B, the area on one side is a continuous (increasing) function, starting at zero and as the line passes through B, continuing until $|B|$. By the Intermediate Value Theorem, there is a line so that the area on each side is $|B|/2$. The analogous argument applies to a line containing the origin rotating and passing through B. $\square$

The result in the next exercise is sometimes called "the pancake theorem", since each region represents a (very flat) pancake in the plane. (A higher dimensional version also holds—see the "ham sandwich theorem", Theorem 16.2.1.)

Exercise 293. *Let A and B be two bounded regions in the plane separated by a line (so A is on one side of the line and B is on the other). Show that there exists a line that simultaneously bisects both A and B.*

Exercise 294. *Prove that a cake of uniform thickness but arbitrary (bounded) shape can be cut into four equal areas using a pair of vertical cuts along* *perpendicular lines. Give an example when the base of the cake is a 3-4-5 triangle.*

For more on cake cutting (including those of the type “I cut, you choose”), see [755]. Here are three such problems, which can be found in Dick Hess’s book [466, Prob. 22]. In these problems, there are two players who alternately cut a cake. It does not hurt to think of “the cake” as some standard convex shape and that cuts are made with straight lines, perhaps even vertical straight line cuts, but these assumptions are not necessary. Saying that someone “cuts a cake into two pieces” hides the assumption that cakes are connected (one piece). Even connectedness is not necessary if “pieces” is replaced by “portions”.

Exercise 295. *Alice cuts a cake into two pieces. Then Bob cuts one of these pieces into two pieces. Alice then gets both the largest and the smallest of the three pieces (or one of the largest or smallest, if two agree in size). How can Alice guarantee herself the largest portion?*

Exercise 296. *Alice cuts a cake into two pieces. Then Bob cuts one of these pieces into two pieces. Alice then gets both the largest and the smallest of the three pieces (or one of the largest or smallest, if two agree in size). Neither want too much cake (they are on a diet?). How much cake can Bob force Alice to take?*

Exercise 297. *Alice cuts a cake into two pieces. Bob then cuts one of these pieces into two pieces. Then Alice cuts one of the three pieces. Alice then gets a largest and a smallest piece. How can Alice be guaranteed as much as possible?*

Exercise 298. *Given any bounded plane region, show that there are three concurrent lines that cut the region into six pieces of equal area.*

See [368, p. 161] for discussion of how to cut a cake into n pieces so that n people think it is fair.

16.2 Ham sandwich theorem

The following theorem was proved by Banach (around 1938) for the case $n = 3$, and later by Stone and Tukey for general n.

Theorem 16.2.1 (Banach,1938 [825], Stone–Tukey, 1942 [833]). *Let $n \geq 3$. For any n sets in $\mathbb{E}^n$, each of finite outer Lebesgue measure, there exists a hyperplane (an $(n-1)$ dimensional affine subspace) that bisects all n sets, i.e., separates each of the given sets into two sets of equal n-dimensional volume.*

A 2-dimensional version of the ham sandwich theorem is sometimes called the "pancake theorem", given here as Exercise 293.

The 3D case was proposed by Hugo Steinhaus, who apparently (see [81]) thought of the question while hiding from the Germans in a Polish farmhouse; Steinhaus (and others) wrote a paper on the topic in 1938 [825] (see [81] for a translation and other historical facts), where a proof given for 3 dimensions was attributed to Banach (using the Borsuk–Ulam theorem, Theorem 17.5.1 here). The general case (for higher dimensions) was proved by Stone and Tukey [833] in 1942.

In popular literature, some say that the 3-dimensional version of Theorem 16.2.1 is called a "ham sandwich" theorem because it says that if a sandwich, consisting of two slices of bread and a piece of ham can be cut (with one slice) so that the three quantities (bread, ham, bread) can be bisected simultaneously, regardless of the size or positioning of the components. However, both slices of bread can be taken as one of the components, ham a second component, and, say, cheese as a third component. Nowhere in the theorem does it insist that each component be connected or even convex, just measurable, so such a sandwich can be sliced to have the bread, ham, and cheese on each side to have the same respective volumes. (Theoretically, the theorem also says that the sandwich could have components stacked of bread-cheese-ham-cheese–ham–bread, and another piece of ham sitting beside the original sandwich.)

The ham sandwich theorem was also popularized in Martin Gardner's 1961 book *The second Scientific American book of mathematical puzzles & diversions* [368, p. 147]. The 3D question appeared again in 1981 in *The Scottish book* [648] as problem 123.

In computational geometry, discrete versions of the ham sandwich theorem are of interest, where measurable sets are replaced by finite collections of points. For work on algorithms for finding such bisecting hyperplanes from the ham sandwich theorem, see, e.g., [288] and [616].

16.3 Lines or chords in circles

In a circle, a *chord* is a line segment between two points on the circle. . If a chord passes through the center of a circle, it is sometimes called "a diameter", however sometimes the word "diameter" is reserved for the maximum length across a shape (see Definition 1.14.1).

Exercise 299. *For $n \in \mathbb{Z}^+$, let $f(n)$ be the maximum number of pieces a circular cake can be cut into with precisely n vertical cuts. What is $f(n)$?*

Exercise 300. *Show that any n-element point set in the plane in general position can be partitioned by a line in exactly $\binom{n}{2} + 1$ ways. Dualize this result.*

The next problem is well-known (*e.g.*, see [442, Prob. 36], [441, Ex. 5], and sequence # 427 in [800]) and is yet another problem of partitioning space by using cuts. Mark n dots on the edge of a circle (like raisins on the edge of a cake) and then connect all dots with straight chords as in Figure 16.1; this cuts the circle (or cake) into various regions. With n dots, what is the maximum number of regions the circle can be cut into? For $n = 1$ dot, there are no chords, and hence only 1 region. The numbers of regions for the first five cases are 1, 2, 4, 8, 16 (see Figure 16.1).

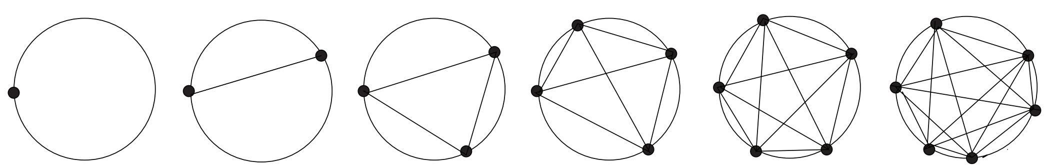


Figure 16.1: Cutting the cake

However, when $n = 6$, there are 31 regions, not the "expected" 32 one might first guess. (According to [917, pp. 119–120], the disruption in the pattern was discovered by the mathematician Leo Moser.)

Exercise 301. *Place n points on a circle and draw in all possible chords joining these points. If no three chords are concurrent, show that the number of regions created is $\binom{n}{4} + \binom{n}{2} + 1$.*

Exercise 302. *Mark n points around a circle and label them either red or blue. Prove that there are at most $\lfloor (3n + 2)/2 \rfloor$ chords that join differently labelled points and that do not intersect inside the circle.*

16.4 Regions formed by circles or other bent lines

The same idea as that in Exercise 285 occurs in the next exercise:

Exercise 303. *Prove that for $n \geq 0$ circles in the plane, the regions thereby determined can be properly 2-coloured.*

The next exercise has a solution similar to Moore's proof of Theorem 16.1.2 that uses Euler's formula, or it can be proved by induction.

Exercise 304. *Let n circles be in the plane so that any two circles intersect in two points, and no three intersect in a single point. Prove that these circles divide the plane into $n^2 - n + 2$ regions.*

A higher dimensional version of Exercise 304 to spheres occurs in Exercise 314.

It might be interesting to note that Exercise 304 has an unexpected consequence: since four circles divide the plane into at most 14 regions, there does not exist a Venn diagram using precisely 4 circles, no matter what their sizes, because a Venn diagram for four sets requires $2^4 = 16$ regions. Note that if three circles are concurrent, one gets fewer regions than given in Exercise 304.

Rather than use circles or lines to partition the plane, the next two exercises use "bent lines" and "zig-zag lines", as in Figure 16.2.

Let two rays originating from the same point be called a *bent line*. A continuous line made from one segment and two rays (and not self-intersecting) is called a *zig-zag* line. Following notation from [407], let Z_n be the maximum number of regions that n bent lines can partition the plane into. It is not too difficult to check that $Z_1 = 2$ and $Z_2 = 7$.

Exercise 305. *Prove that for each $n \geq 1$, $Z_n = 2n^2 - n + 1$.*

The similar problem for zig-zag lines is slightly more difficult.

Exercise 306. *Let ZZ_n be the maximum number regions determined by n zig-zag lines. For positive integers n, first prove the recursion $ZZ_n = ZZ_{n-1} + 9n - 8$, and conclude that*

$$ZZ_n = \frac{9n^2 - 7n + 2}{2}.$$

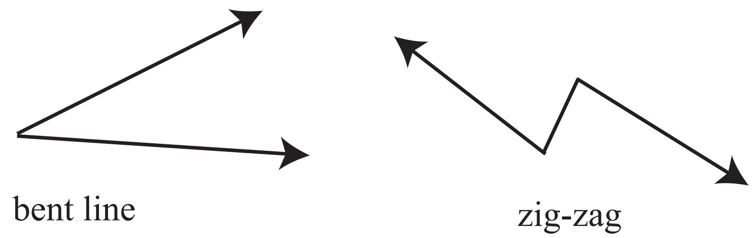


Figure 16.2: Partitioning the plane with other shapes

16.5 Regions determined by planes

The first exercise of this section was posed by Frank Hawthorne [461] in 1950; it has appeared in many puzzle collections since (e.g., [366, pp. 26, 33] and [542, C2]).

Exercise 307. *A $3 \times 3 \times 3$ cube can be cut into 27 $1 \times 1 \times 1$ cubes by using six cuts—two each in the direction of the three planes parallel to the faces. Can this be achieved by using fewer than six cuts if one allows the pieces from one cut to be rearranged (e.g., stacked) for a subsequent cut?*

Exercise 308. *Similar to Exercise 307, a $4 \times 4 \times 4$ cube can be cut into 64 unit cubes with 9 cuts (3 in each direction). If you are allowed to rearrange the pieces in between cuts, is it possible to make the 64 unit cubes with fewer than 9 cuts? What is the minimum number of cuts required?*

The general problem of cutting an $a \times b \times c$ block into abc unit cubes with the minimum number of cuts was posed in *Mathematical Magazine* (1951) by Leo Moser, with solutions announced in 1952 [673]. (See also Martin Gardner's description of the algorithm [372, p. 52].). Moser's answer is

$$3 + \lfloor \log_2(a-1) \rfloor + \lfloor \log_2(b-1) \rfloor + \lfloor \log_2(c-1) \rfloor.$$

For example, a $3 \times 4 \times 5$ block requires 7 cuts.

It might be interesting to note that the $3 \times 3 \times 3$ version appeared in Martin Gardner's column in *Scientific American* in 1957 [365] (again see [366]), and later that year, Ford and Fulkerson posed the $n \times n \times n$ problem in the *American Mathematical Monthly*, with solution appearing in 1958 [341]—along with a note from the editors that this was just a special case of the more general result of Moser from a few years earlier. The higher dimensional version has also been solved (Gardner gives a reference [737], but [341] gives the formula, too).

For use in the next exercise, a fairly easy lemma is provided:

Lemma 16.5.1. *Suppose that a line (for example, the real line) is covered by a finite collection of rays (half-lines). Then some two of these rays cover the entire line.*

Proof: Let p be the rightmost endpoint of all rays pointing to the left, and let q be the leftmost endpoint of all rays pointing to the right. By assumption, all rays cover the line, and so p is not to the left of q. So the two rays starting at p and q respectively cover the line. □

For the next exercise, a *half-plane* is a region of the plane on one side of a line. [It does not matter if one considers only open half-planes or only closed half-planes for this exercise.]

Exercise 309. *Show by induction on n that if a plane is covered with $n \geq 2$ half-planes, then there exist two or three half-planes that cover all of the plane.*

The result in Exercise 309 has a three-dimensional version, with a solution that follows the same idea; see [398, Ex. 35, pp. 120–121].

Exercise 310. *Prove that n planes, passing through one point in a way that no three pass through the same line, divide space into $n(n-2)+2$ parts.*

A *half-space* is a region on one side of a plane (or hyperplane).

Exercise 311. *Prove that if $n \geq 2$ half-spaces cover all of three-dimensional space, then there exist two, three, or four of them that cover the whole space.*

Planes in three dimensional space are said to be in *general position* if no three planes share a common line and no two planes are parallel.

Exercise 312. *Show that the maximum number of regions three dimensional space is divided into by n planes in general position is*

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \binom{n}{3},$$

and the number of infinite (unbounded) regions is

$$2\binom{n}{0} + 2\binom{n}{2}.$$

16.6 Regions determined by spheres

The first exercise in this section is not really about regions, but it is about spheres:

Exercise 313. *Three spheres intersect in a point P, but no line containing P is tangent to all three spheres. Prove that the spheres intersect in an additional common point.*

Exercise 314. *Show that $n \geq 2$ spheres, any two of which intersect, partition 3-space into at most*

$$\frac{n(n^2 - 3n + 8)}{3}$$

regions.

16.7 Regions formed by hyperplanes

For present purposes, a hyperplane in $\mathbb{R}^d$ is a $(d-1)$-dimensional affine subspace. Hyperplanes cut $\mathbb{R}^d$ into *cells*, just as lines cut the plane into regions.

Definition 16.7.1. An arrangement of hyperplanes in $\mathbb{R}^d$ is in *general position* if and only if

(i) for $2 \leq k \leq d+1$, the intersection of each k hyperplanes is $(d-k)$-dimensional;

(ii) for $k \geq d+1$, every d hyperplanes intersect in a single point and no $d+1$ have a common point.

The next few exercises can be found in [645, p. 129]:

Exercise 315. *Find the number of k-faces in an arrangement of n hyperplanes in general position in $\mathbb{R}^d$. (Begin by answering this when $k = 1, 2$.)*

Exercise 316. *Prove that for a fixed d, the number of unbounded cells determined by an arrangement of n hyperplanes in $\mathbb{R}^d$ is $O(n^{d-1})$.*

Exercise 317. *Show that an arrangement of d or fewer hyperplanes in $\mathbb{R}^d$ creates no bounded cell, and that $d+1$ hyperplanes in general position produces only one bounded cell.*

Theorem 16.7.2. *For integers $d \geq 1$ and $n \geq 0$, the number of cells determined by n hyperplanes in $\mathbb{E}^d$ in general position is*

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

There are, essentially, at least two proofs of Theorem 16.7.2, both inductive. Here is one:

Proof: For integers $d \geq 1$ and $n \geq 0$, put

$$\Phi_d(n) = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

For each n, d, let $S(d, n)$ be the statement that n hyperplanes (in general position) cut $\mathbb{R}^d$ into $\Phi_d(n)$ regions. The proof is by induction on d and n.

Base step [$\forall d, n,\ S(d, 0)$ and $S(1, n)$]: When $n = 0$, the entire space is the only cell, and so for every d, $S(d, 0)$ holds, since $\binom{n}{0} = 1$. Also, for every $n \geq 1$, $S(1, n)$ says that the number of cells is

$$\Phi_1(n) = \binom{n}{0} + \binom{n}{1} = n + 1,$$

but in $\mathbb{R}^1$, hyperplanes are only points, and n points on the real line cut the line into $n+1$ regions; so for all n, $S(1, n)$ holds.

Inductive step[$S(k, m-1) \wedge S(k-1, m-1) \to S(k, m)$]: Let $k, m \geq 2$ and suppose that $S(k-1, m-1)$ and $S(k, m-1)$ hold. Let an arrangement $\mathcal{A}$ of m hyperplanes in $\mathbb{R}^k$ be given. Fix one plane $P \in \mathcal{A}$, and consider $\mathcal{A}' = \mathcal{A} \backslash P$. Since P is $(k-1)$-dimensional, and each hyperplane in $\mathcal{A}'$ cuts P, by $S(k-1, n-1)$, the number of cells in P is $\Phi_{k-1}(m-1)$. Each ($(k-1)$-dimensional) cell in P partitions one cell from $\mathcal{A}'$ and so produces one more cell than what $\mathcal{A}'$ did. By $S(k, m-1)$, the $m-1$ planes in $\mathcal{A}'$ cut $\mathbb{R}^k$ into $\Phi_k(n-1)$ cells, and so the total number of cells in $\mathbb{R}^k$ formed by $\mathcal{A}$ is

$$\Phi_k(n-1) + \Phi_{k-1}(n-1).$$

With a simple application of Pascal's identity, this sum is equal to $\Phi_k(m)$, completing the proof of $S(k, m)$, and hence the inductive step.

By an alternate form of mathematical induction, for all $d, n \geq 1$, $S(d, n)$ holds. $\square$

Another proof of Theorem 16.7.2 is already mentioned in these notes just after the case for lines (Theorem 16.1.2), and is conceptually quite easy (this proof is well-known, perhaps for more than a century; see, e.g., [645, p. 128]): divide the cells into two types, those that have a minimum element, and those that don't. Suppose that $d \geq 3$, and consider n points in $\mathbb{R}^d$. Assume that no two points share a common coordinate (otherwise, tilt the space slightly). Count the cells according to whether or not they have a minimal point. To count the cells with a minimal point, each such point is determined by the intersection of d hyperplanes (and each selection of d hyperplanes determines a minimal point of some cell), so there are $\binom{n}{d}$ such cells. To count those cells without a minimal point, "project" these cells onto a copy of $\mathbb{R}^d$ placed below all intersection points. Inducting on d then finishes the proof.

Chapter 17

Geometric graphs

17.1 Introduction

Geometric graph theory is a field of modern mathematics that covers many topics, only a few of which are touched upon here.

Some authors refer to any graph arising from geometric concerns a "geometric graph". Other authors say that a geometric graph is a Euclidean graph, that is, a graph whose vertices are in some Euclidean space (usually $\mathbb{R}^2$ or $\mathbb{R}^3$ endowed with the usual metric) and edges are straight line segments joining vertices, each segment incident with only its two end vertices. (The term "topological graph" is often used for a graph embedded in some metric space where edges are simple Jordan curves, but these, too, can be called geometric graphs.) For the moment, consider geometric graphs that are Euclidean, drawn with straight lines.

If $a, b, c \in \mathbb{R}$ satisfy the triangle inequaility $a + b > c$, a triangle with these lengths exists, and so the graph $G = K_3$ with three edges can be drawn in the plane (or 3-space) realizing these three distances a, b, c. In general, what are the restrictions on a collection of distances so that a graph can be drawn in the plane (that is, there exists a geometric graph) realizing these distances?

If there are two geometric graphs realizing a collection of distances, how are they related? For example, one might ask if there is a rigid motion (distance preserving bijection) of the plane that takes one to the other.

Say that a geometric graph is *rigid* if there is no continuous motion of its vertices that produces another geometric graph with the same edge distances.

One might think of the fixed distance edges as stiff rods and the vertices as flexible connecting joints—in this case the geometric graph is rigid if and only if the rod-joint structure is architecturally stable and cannot be deformed. For example, any geometric K_3 is rigid, but the cycle C_4 drawn as a square is not (since a square can deformed into a rhombus). Similarly in 3 dimensions, the standard drawing of the cube-graph Q_3 is not rigid, but the regular tetrahedron graph K_4 is.

Determining which geometric graphs are rigid is an especially hard problem in three dimensions. The concept of rigidity has also been studied for polyhedra (not just their graphs). The difference is that in polyhedral rigidity, the faces are considered rigid (like plates of steel); under this assumption, is the structure rigid, or can the polyhedra be deformed (allowing bends at the an edge between two faces)?

It is known (by Cauchy [180], see [13]) that convex polyhedra with rigid faces are rigid, and certain triangulated geometric graphs are rigid [227]. One of the early results on non-rigid polyhedra is Connelly's 1978 result [206] that some (non-convex) polyhedra are not rigid (that is, they can be "flexed" somewhat). For more information on rigidity, see also two earlier articles by Connelly on rigidity [205], [207], or see the article by Goldberg [394] on unstable polyhedral structures (where the subtitle is "Unlike convex polyhedra, some non-convex forms may shake, snap, or flex").

Instead, the approach taken here looks at questions about geometric graphs regarding the number of times a particular distance can be realized, or the minimum number of different distances that may be realized by a geometric graph on n vertices. Having said this, I focus only on a few such problems, and the results given here are in no way indicative of the present state of affairs. Erdős loved problems of these types, and think it is a shame that I don't remember a large fraction of his results.

One tool that is often used in distance problems is extremal graph theory—the study of just how many edges in a graph forces the existence of a particular subgraph, or the existence of a more abstract property. Many types of results are used for counting distances in geometric graphs. Section 17.2 contains a few of the major results in extremal graph theory (where a "graph" is an ordinary graph, a simple reflexive binary relation on a vertex set), including bounds on extremal numbers $\mathrm{ex}(n; K_p)$, $\mathrm{ex}(n; K_{2,t})$ and $\mathrm{ex}(n; K_{3,3})$.

In Section 17.3 is a brief look at a result on crossing numbers. This might come as no surprise, since the crossing number of a graph is the minimum number of edge-crossings among all embeddings of the graph in the plane where edges can be curved; hence planar graphs have crossing number 0.

For Sections 17.2 and 17.3 it is assumed that the reader is familiar with basic graph theory (see Chapter 2 for a brief review of definitions in graph theory and notation used here). Also, in Section 22.4 a basic number theoretic result is reviewed: in how many ways can a positive integer be written as a sum of two squares. This result can also be used when counting distances in geometric graphs.

17.2 Extremal numbers and Turán's theorem

All graphs in this section are simple; in other words, a graph G is an ordered pair $(V, E) = (V(G), E(G))$, where V is a non-empty set and $E \subseteq [V]^2$. So a graph has no loops and no multiple edges.

For a graph G and positive integer n, define $\text{ex}(n; G)$ to be the maximum number of edges of a graph H on n vertices so that H contains no copy of G (say that H is G-free). Recall that K_m denotes the complete graph on m vertices.

Theorem 17.2.1 (Mantel, 1907 [633]). *For $n \geq 3$, $ex(n; K_3) = \lfloor n^2/4 \rfloor$.*

For many proofs of Mantel's theorem, see [13]. Mantel's theorem is a special case of a much stronger theorem proved by P. Turán (1910–1976). Turán was also known for his work in number theory and analysis. To state this theorem, a few definitions are given.

Definition 17.2.2. For positive integers n and k, let $T(n, k)$ be the complete k-partite graph on n vertices whose partite sets have sizes that are as nearly equal as possible. (The graph $T(n, k)$ is called "the Turán graph".) Denote the number of edges in $T(n, k)$ by $|E(T(n, k))| = t(n, k)$.

If $n = qk + r$, where q, r are non-negative integers with $0 \leq r < k$, then r of the partite sets in $T(n, k)$ have $q + 1 = \lceil n/k \rceil$ vertices, and the remaining $k - r$ have $q = \lfloor n/k \rfloor$ vertices. Hence

$$t(n, k) = \binom{r}{2}(q + 1)^2 + r(k - r)(q + 1)q + \binom{k - r}{2}q^2.$$

There are many ways to count the edges in $T(n, k)$; here is one convenient approximation:

Lemma 17.2.3. *For a fixed $k \geq 2$, as n increases,*

$$t(n,k) = (1+o(1))\frac{n^2}{2}\left(1-\frac{1}{k}\right).$$

Proof outline: Assume that k divides n (the remaining case is similar and is omitted). The number of "missing edges" in the Turán graph $T(n,k)$ is $k\binom{n/k}{2}$, so

$$\begin{aligned} t(n,k) &= \binom{n}{2} - k\binom{n/k}{2} \\ &= \frac{1}{2}\left[n(n-1) - k\frac{n}{k}\left(\frac{n}{k}-1\right)\right] \\ &= \frac{n}{2}\left[n-1-\frac{n}{k}+1\right] \\ &= \frac{n^2}{2}\left(1-\frac{1}{k}\right). \end{aligned}$$

□

In 1940, Turán was in a labour camp when he discovered his famous theorem, Theorem 17.2.4 below) (some say without the use of pencil and paper). A short account of the story can be found in [883], in the first issue of the *Journal of Graph Theory*. This result was also published in 1954 [882] in English.

Theorem 17.2.4 (Turán, 1941 [881, 882]). *For positive integers k and n,*

$$ex(n; K_{k+1}) = t(n,k).$$

Furthermore, $T(n,k)$ is the unique extremal K_{k+1}-free graph on n vertices.

There are at least six different proofs of the first sentence (or equation (17.1)) in the statement of the theorem (see, for example, [11], or [13, pp. 207–211] for five). The first proof given here (due to Turán) is by induction on n (this proof also yields the second statement), and is sometimes called a "chopping off" proof. A second proof provided here is due to Zykov [943], which uses "symmetrization". As noted in [795], another proof, due to Motzkin and Straus [678], can be seen as a variation of Zykov's proof.

Turán's Proof: Fix $k \geq 1$ and for each $n \geq 1$, let $S(n)$ be the statement that if G is a K_{k+1}-free graph on n vertices with $\text{ex}(n; K_{k+1})$ edges, then $G = T(n,k)$.

Base cases: For each $i = 0, 1, \ldots, k$, the graph with the most edges on i vertices is $K_i = T(i,k)$, so $S(i)$ holds.

Inductive step: Fix $m \geq k$ and suppose that $S(m-k)$ holds. Let G be a K_{k+1}-free graph on m vertices with $\text{ex}(m, K_{k+1})$ edges. As G is extremal for K_{k+1}, G contains a copy of K_k, call it H, on vertices $A = \{a_1, \ldots, a_k\}$. Put $B = V(G)\backslash A$, and let G^* be the graph induced on B.

Since G is K_{k+1}-free, each vertex in B is adjacent to at most $k-1$ vertices of A. Then

$$\begin{aligned}
|E(G)| &\leq \binom{k}{2} + |B|(k-1) + |E(G^*)| \\
&\leq \binom{k}{2} + (m-k)(k-1) + \text{ex}(m-k; K_{k+1}) && (G^* \text{ is } K_{k+1}\text{-free}) \\
&\leq \binom{k}{2} + (m-k)(k-1) + t(m-k,k) && (\text{by IH, } S(m-k)) \\
&= t(m,k) && (\text{structure of } T(m,k)).
\end{aligned}$$

Hence $|E(G)| \leq t(m,k)$. Also, since $T(m,k)$ is K_{k+1}-free and G has an extremal number of edges, $t(m,k) \leq |E(G)|$. Thus $|E(G)| = t(m,k)$, forcing equality in the equations above. Then each vertex in B is joined to exactly $k-1$ vertices of A.

For each $i = 1, \ldots, k$, put $W_i = \{x \in V(G) : \{x, a_i\} \notin E(G)\}$. Note that $a_i \in W_i$ and the W_i's partition $V(G)$ since every vertex in B is not adjacent to one of the a_i's. Each W_i is an independent set, since if some $x, y \in W_i$ were adjacent, x, y and $A\backslash\{a_i\}$ form a K_{k+1}. Hence, G is k-partite.

Since $T(m,k)$ is the unique k-partite graph with as many edges as possible, $G = T(m,k)$. This completes the inductive step $S(m-k,k) \rightarrow S(m,k)$.

By mathematical induction, for all $n \geq 0$, $S(n)$ is true. □

Zykov's proof: Let $G = (V, E)$ be a graph on n vertices, and suppose that G does not contain a copy of K_{k+1} as a subgraph. Let $v_1 \in V$ be a vertex with maximum degree. For each non-neighbour w of v_1, remove all edges incident with w and add all edges $\{\{w, v\} : \{v_1, v\} \in E\}$ to create a new graph G_1. Put $X_1 = \{v_1\} \cup \{w \in V : \{v_1, w\} \notin E$. Then each vertex in X_1 is now adjacent to all remaining vertices $V \backslash X_1$ and X_1 is an independent set. Also, G_1 has at least as many edges as does G, and more importantly, since G contains no K_{k+1}, neither does G_1.

Repeat this process of "symmetrization" as follows: pick $v_2 \in V \backslash X_1$ with maximum degree in G_1. For each non-neighbour w of v_2 in $V \backslash X_1$ remove all edges incident with w and add all edges $\{\{w, v\} : \{v_2, v\} \in E\}$ to create a new graph G_2. Put $X_2 = \{v_2\} \cup \{w \in V : \{v_2, w\} \notin E$. Again, X_2 is an independent set, with each vertex of X_2 adjacent to all vertices in $V \backslash X_2$, G_2 is K_{k+1}-free, and $|E(G_2)| \geq |E(G_1)| \geq |E(G)|$. Continue this process creating, for some r, the graph $G_r = T(n, r)$. Since G_r is K_{k+1}-free, $r \leq k$.

Finally, observe that if $|E(G)| = |E(T(n, k))|$, no edges were added, and so G was originally $T(n, k)$. □

Exercise 318. *Use Turán's theorem to show that if $r \geq 2$ and G is a K_r-free graph on n vertices, then*

$$|E(G)| \leq \left(1 - \frac{1}{r-1}\right)\frac{n^2}{2}. \tag{17.1}$$

The result in the next exercise has what might be seen as a surprising application counting pairs of points at unit distance in the plane (see Theorem 17.4.4 and Exercise 334).

Exercise 319. *Let G be a graph on n vertices, and let $t \geq 2$. Show that if*

$$\sum_{v \in V(G)} \binom{\deg(v)}{2} > (t-1)\binom{n}{2},$$

then G contains $K_{2,t}$. Deduce that if G contains more than $\frac{(t-1)^{1/2} n^{3/2}}{2} + \frac{n}{4}$ edges, then G contains $K_{2,t}$. (In other words, show an upper bound for $ex(n; (K_{2,t}))$.)

Here is (without proof) another result for bipartite graphs that is used later (see Theorem 17.4.6), also for unit distances but in 3-space.

Theorem 17.2.5 (Füredi, 1995 [355]). *As $n \to \infty$,*

$$ex(n; K_{3,3}) \leq \left(\frac{1}{2} + o(1)\right) n^{5/3}.$$

If a forbidden graph F is bipartite, one can ask for a *bipartite* graph G with the most number of edges that does not contain F oriented in G. The main theorem for such problems is the following (the number of copies was not in the original proof, but the added mention makes the proof easier), which was proved in response to a question by Zarankiewicz [937] about the number of 1s necessary in a 0-1 matrix for a certain submatrix to exist.

Theorem 17.2.6 (Kővári–Sós–Turán, 1954 [572]). *Let $G = (X_1, X_2, E)$ be a bipartite graph with $|X_1| = m$ and $|X_2| = n$, and let $s \le m$ and $t \le n$. If $|E| \ge t^{\frac{1}{s}} mn^{1-\frac{1}{s}}$, then G contains*

$$\binom{m}{s}\binom{\frac{n}{\binom{m}{s}}\binom{\frac{|E(G)|}{n}}{s}}{t}$$

copies of $K_{s,t}$ where the $s \le m$ vertices occur in X_1 and $t \le n$ vertices occur in X_2.

Proof: For any set S of vertices in a graph G, define the degree of S to be

$$\deg(S) = |\{y \in V(G) : \forall x \in S, \{x, y\} \in E(G)\}|.$$

The number of copies of $K_{s,t}$ in G (where the $s \le m$ vertices occur in X_1 and $t \le n$ vertices occur in X_2) is:

$$\begin{aligned}
\sum_{S \in [X_1]^s} \binom{\deg(S)}{t} &\ge \binom{m}{s}\binom{\text{avg } \deg(S)}{t} && \text{(by Jensen's ineq.)} \\
&= \binom{m}{s}\binom{\frac{\sum_{S\in[X_1]^s} \deg(S)}{\binom{m}{s}}}{t} \\
&= \binom{m}{s}\binom{\frac{1}{\binom{m}{s}} \sum_{x \in X_2} \binom{\deg(x)}{s}}{t} \\
&\ge \binom{m}{s}\binom{\frac{1}{\binom{m}{s}} n \binom{\text{avg}_{X_2} \deg(x)}{s}}{t} && \text{(by Jensen's ineq.)} \\
&= \binom{m}{s}\binom{\frac{n}{\binom{m}{s}}\binom{\frac{|E(G)|}{n}}{s}}{t}.
\end{aligned}$$

[Checking, if G is complete bipartite with mn edges, the number of oriented copies of $K_{s,t}$ works out to be $\binom{m}{s}\binom{n}{t}$, which is true.]

For the number $\binom{m}{s}\binom{\frac{n}{\binom{m}{s}}\binom{\frac{|E(G)|}{n}}{s}}{t}$ to be at least one, the following is needed:

$$\frac{n}{\binom{m}{s}}\frac{|E(G)|^s}{n^s s!} \geq t.$$

With a little basic algebra,

$$|E(G)| \geq t^{\frac{1}{s}} m n^{1-\frac{1}{s}}$$

suffices. With $m = n$, this becomes $|E(G)| \geq t^{\frac{1}{s}} n^{2-\frac{1}{s}}$. □

Extremal theory for geometric graphs has also been studied. For an ordinary graph G, let $\mathrm{ex}_{\mathrm{geom}}(n; G)$ be the extremal number of edges in a geometric graph with no copy of G. For example, if M_{k+1} denotes a matching with $k+1$ edges and no two edges crossing, in 1999, Tóth and Valtr [874] showed that for $k \leq \frac{n}{2}$,

$$\frac{3}{2}(k-1)n - 2k^2 \leq \mathrm{ex}_{\mathrm{geom}}(n; M_{k+1}) \leq k^3(n+1).$$

The upper bound was improved by Tóth in 2000 [872] to $2^9 k^2 n$. See Tóth's paper [872] for more references on extremal theory for geometric graphs, indicating that Erdős and Perles, Avital and Hanani, and Kupitz might have started this area with a few initial questions.

17.3 Crossing numbers

17.3.1 Definitions and examples

In this section, unless otherwise stated, all graphs are simple, that is, if $G = (V, E)$ is a graph on vertex set V, then the edge set is a subset of distinct pairs from V, written $E \subseteq [V]^2$. Thus, simple graphs are undirected, have no loops and no multiple edges.

Recall that a graph is called planar if and only if it can be drawn in the plane with no edges crossing (or touching) except at vertices of the graph. Such a drawing is often called a "plane drawing" of a planar graph or a "planar embedding". For the moment, let a "drawing" D of a graph $G = (V, E)$ in the plane be a set of $|V|$ points in $\mathbb{R}^2$ given by an injective embedding

$\phi : V \to \mathbb{R}^2$, together with, for each $\{x, y\} \in E$, a curve (or arc, or "edge") joining $\phi(x)$ and $\phi(y)$ but containing no other vertex $\phi(z)$.

Two edges in a drawing of a graph are said to *cross* if they share a common point other than an endpoint. An *edge crossing* is an instance of a cross, namely, a triple $(e_1, e_2, \mathbf{x})$, where e_1 and e_2 are distinct edges both containing the point $\mathbf{x}$, and a crossing point is a point $\mathbf{x}$ that witnesses a crossing. *A priori*, in a drawing of a graph, two edges may cross arbitrarily many times (even edges drawn from the same vertex), and a crossing point can be a witness to more than one crossing. If two edges in a drawing touch at a point but do not actually cross (when they are tangent), another drawing can be made where these two edges do not touch (at that point), so it is tacitly assumed that a pair of edges form a "crossing" if and only if one cuts through the other.

For a non-planar graphs G, one measure of the "non-planarity" is the minimum number of edge crossings required in any drawing of G.

Definition 17.3.1. The *crossing number* of a graph G, denoted by $\mathrm{cr}(G)$, is the minimum number of edge crossings among all drawings of G in the plane (or sphere).

The crossing number is often denoted in the literature by $\nu(G)$, but in graph theory, $\nu(G)$ is often used to denote the "matching number" of a graph (see, e.g., [117]), so here, the more descriptive notation $\mathrm{cr}(G)$ is used. For an introduction to crossing numbers, see the 1973 article by Erdős and Guy [311], or the blog post by Terence Tao [860]. For an extensive survey (including 409 references), see Marcus Schaefer's 2014 article [776]. There are numerous other surveys (e.g., [750]).

There are many variations on this "crossing number" (for example, see [704] or [776]). Only a very few variations are considered here. One might restrict drawings to straight lines (giving "rectilinear crossing numbers", defined below), or one might restrict drawings where edges can intersect only an odd number of times, or one might ask for only drawings where each edge is crossed by at most some number of times. If more than two edges are allowed to cross at a point, one might ask how few crossing points (not pairs of edges) are required (see "degenerate crossing numbers", defined below). One might also count the minimum number of crossings in a graph drawn on some surface S, in which case the notation cr_S (or $\nu_S(G)$) might be used. If no subscript is used, it is assumed that the surface is the plane (or equivalently, a sphere). There are also "spherical crossing numbers", where drawings are restricted to spheres but edges are geodesics. The survey by Schaefer [776] contains a rather extensive description of different kinds of crossing number, only a few of which are touched upon here.

To give an upper bound for the crossing number of a particular graph, it suffices to give a drawing with few crossings, but then to check that such a drawing is optimal can be non-trivial. In general, Garey and Johnston [379] showed that computing crossing numbers is NP-complete. (To see the complexity of other types of crossing numbers, see [776].) In 2013, Cabello and Mohar [169] showed that just among the class of planar graphs with one extra edge added, it is NP-hard to decide whether or not a graph has crossing number 1, even if one assumes an upper bound on the degree! This result might seem somewhat surprising, since determining planarity can be done in linear time, e.g., by the Hopcroft–Tarjan algorithm [486]. (Thanks to Stephane Durocher [281] for pointing out this reference.)

Definition 17.3.2. The *rectilinear crossing number* of a graph G, denoted by $\overline{\mathrm{cr}}(G)$ (or $\overline{\nu}(G)$ in some texts), is the minimum number of crossings for drawings of G in the plane that use only straight line segments.

Trivially, $\overline{\mathrm{cr}}(G) \geq \mathrm{cr}(G)$, and there are examples for which the rectilinear crossing number is larger than the crossing number. (For example, as noted below, $\mathrm{cr}(K_8) < \overline{\mathrm{cr}}(K_8)$.)

Since graphs can represent electrical circuits, those who study integrated circuits (both simple, or more complex VLSI circuits) are also interested in crossing numbers. In very large circuits, one might want to minimize the number of layers of integrated circuit boards necessary, or the minimum number of "jumper" connections needed on one board or between boards.

If a drawing D of a graph G has $\mathrm{cr}(G)$ crossings, then D is called an *optimal* drawing (as far as crossing numbers go). It was observed (e.g., in [311]) that optimal drawings satisfy a simple property.

Lemma 17.3.3. *Let G be a graph and D be an optimal drawing for $cr(G)$. Then in D, any two edges sharing a common vertex (in G) do not cross.*

So for the present discussion of crossing numbers, assume that all "good drawings" satisfy the condition that any two edges incident with the same vertex (sometimes called neighbouring edges) do not cross. This condition is not always mentioned explicitly in papers about crossing numbers. (When looking at only rectilinear drawings, this condition is automatic.) Also note that if two edges touch at a point other than at a vertex, these edges must actually cross, not just touch and then "bounce off" or be tangent.

17.3.2 Some examples

Since K_1, K_2, K_3 and K_4 are planar, for $1 \leq k \leq 4$, $\mathrm{cr}(K_k) = 0$, and by a standard drawing of K_4 with one vertex in the middle, $\overline{\mathrm{cr}}(K_4) = 0$. Other common planar graphs are trees, cycles, and the cube graph Q_3.

Recall that the graphs K_5 and $K_{3,3}$ are not planar; however, each can be drawn with precisely one pair of edges crossing.

Exercise 320. *Give drawings of K_5 and $K_{3,3}$ that each have exactly one pair of edges crossing, thereby showing $cr(K_5) = cr(K_{3,3}) = 1$. Show that such drawings exist using only straight line segments for edges, thereby showing that also $\overline{cr}(K_5) = \overline{cr}(K_{3,3}) = 1$.*

The drawing given in Figure 17.1 shows that $\mathrm{cr}(K_6) \leq 3$.

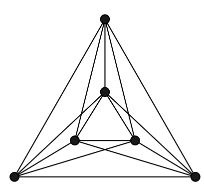


Figure 17.1: Drawing for $\mathrm{cr}(K_6) \leq 3$

Exercise 321. *Show that $cr(K_6) = 3$.*

A complete tripartite graph $K_{a,b,c}$ is a graph whose vertex set is a disjoint union $V = A \cup B \cup C$, where $|A| = a$, $|B| = b$, $|C| = c$, and all edges go between either A and B, A and C, or B and C. For example, the graph $K_{3,2,2}$ with 7 vertices and 16 edges is given in Figure 17.2.

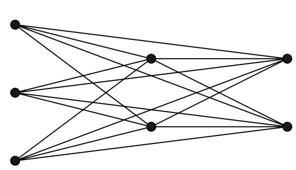


Figure 17.2: The graph $K_{3,2,2}$

Exercise 322. *Show that $cr(K_{3,2,2}) = 2$. Hint: Apply Euler's formula to an optimal drawing considered as a planar graph on the original 7 vertices plus crossing points.*

For more general results on tripartite graphs, see, e.g., [923, Thms 6-67, 6-68]. Also noted in [923] are results by Beineke and Ringeisen for crossing numbers of cartesian products of cycles and/or complete graphs.

A famous graph is the Petersen graph; it is often drawn with 5 crossings, as in Figure 17.3.

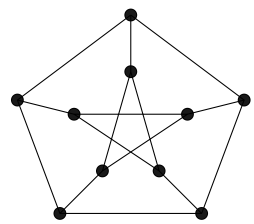


Figure 17.3: A standard drawing of the Petersen graph with 5 crossings

Exercise 323. *Find a drawing of the Petersen graph (see Figure 17.3 that shows its crossing number is at most 3 (such a drawing also exists with straight lines).*

Exercise 324. *Find a drawing of the Petersen graph with only 2 crossings. Can you find such a drawing that uses only straight line segments?*

Exercise 325. *Show that any optimal drawing of the Petersen graph has at least two crossings.*

The Heawood graph (see Figure 11.3) is the point-line incidence graph for the Fano plane, and so is a 3-regular bipartite graph on 14 vertices. The shortest cycle in the Heawood graph has length 6 (a 4-cycle would correspond to two points on two lines, contrary to the definition of a finite projective plane). The number of crossings in Figure 11.3 is 14, but the crossing number of the Heawood graph is 3; this bound is witnessed by different rectilinear drawings, one given in Figure 17.4.

Each of K_4, $K_{3,3}$, the Petersen graph, and the Heawood graph (see Figure 11.3) is 3-regular (called cubic). These are the smallest cubic graphs with respective crossing numbers 0,1,2,3. The smallest cubic graphs with crossing

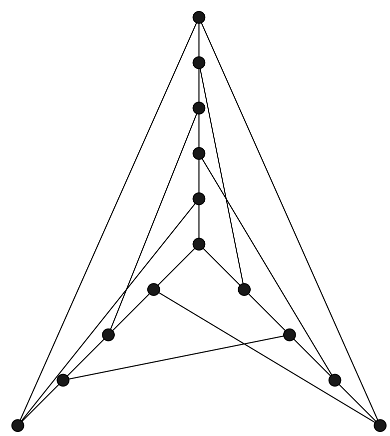


Figure 17.4: The Heawood graph with 3 crossings

numbers up to 8 have been found (see Sloane's sequence A110507). It was shown [475] that finding crossing numbers just for cubic graphs is still a hard problem.

Exercise 326. *Let Q_4 denote the 4-dimensional cube graph (where vertices are binary words of length 4 and edges are between words that differ in exactly one position). Give a drawing that shows $cr(Q_4) \leq 8$.*

Eggleton and Guy [290] once thought that

$$\mathrm{cr}(Q_n) \leq \frac{5}{32}4^n - \left\lfloor \frac{n^2+1}{2} \right\rfloor 2^{n-2},$$

but according to [311], the construction was found to have a gap; equality is still conjectured (also see [311]).

17.3.3 Crossing numbers for complete bipartite graphs

It is said that the study of crossing numbers began with Turán, who considered the number of crossings of tracks used in a brick factory (see [883] for Turán's description of the problem, which he thought about while in a work camp during WWII). The system of tracks formed a complete bipartite graph between kilns and storage sites. See [72] for a history of the brick factory problem. A similar problem arose in a 1944 paper in sociology [149] in representing relationships by graphs (sociograms), where the partite sets were boys and girls, and the aim was to draw the graph with fewest crossings so that information was more visibly apparent.

For any positive integer k, both $K_{1,k}$ and $K_{2,k}$ are planar, so $\text{cr}(K_{1,k}) = \text{cr}(K_{2,k}) = 0$. In 1953, Zarankiewicz [938] gave a general construction that gives an upper bound on the crossing numbers for complete bipartite graphs. For each $m, n \geq 3$, define a drawing of $K_{m,n}$ by putting the m vertices balanced on points $(\pm i, 0)$ on the x-axis, and similarly putting the other n vertices balanced on the y-axis, and draw all edges with straight lines. For example, see Figure 17.5 (which also appears in [311]) for such a drawing of $K_{7,7}$ showing $\text{cr}(K_{7,7}) \leq 81$.

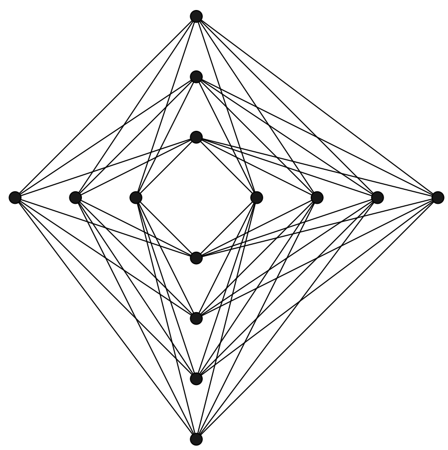


Figure 17.5: $\text{cr}(K_{7,7}) \leq 81$

With a bit of work, one can show that the Zarankiewicz construction gives

$$\begin{aligned}
\text{cr}(K_{2k,2\ell}) &\leq (k^2 - k)(\ell^2 - \ell), \\
\text{cr}(K_{2k,2\ell}) &\leq (k^2 - k)\ell^2, \\
\text{cr}(K_{2k,2\ell}) &\leq k^2\ell^2,
\end{aligned}$$

or equivalently (as observed by Urbanik, and Rényi and Turán, see [455]),

$$\text{cr}(K_{m,n}) \leq \frac{1}{4}\left\lfloor\frac{m}{2}\right\rfloor \cdot \left\lfloor\frac{m-1}{2}\right\rfloor \cdot \left\lfloor\frac{n}{2}\right\rfloor \cdot \left\lfloor\frac{n-1}{2}\right\rfloor. \tag{17.2}$$

In his 1953 paper, Zarankiewicz [938] announced that his construction was optimal, and he provided details in 1954 [939]. However, it was observed (see [438]) that Zarankiewiicz's proof was not complete; it now remains as a conjecture (e.g., see [311]) that the inequality in equation (17.2) is in fact equality.

Conjecture 17.3.4. *For all positive integers m and n,*

$$cr(K_{m,n}) = \frac{1}{4}\left\lfloor\frac{m}{2}\right\rfloor \cdot \left\lfloor\frac{m-1}{2}\right\rfloor \cdot \left\lfloor\frac{n}{2}\right\rfloor \cdot \left\lfloor\frac{n-1}{2}\right\rfloor.$$

Since the drawing proposed by Zarankiewicz uses only straight lines, the bound in (17.2) is also a bound for the corresponding rectilinear number $\overline{\mathrm{cr}}(K_{m,n})$.

In 1970, Kleitman [554] showed that for each $n \geq 3$,

$$\mathrm{cr}(K_{5,n}) = 4\left\lfloor\frac{n}{2}\right\rfloor \cdot \left\lfloor\frac{n-1}{2}\right\rfloor, \tag{17.3}$$

and

$$\mathrm{cr}(K_{6,n}) = 6\left\lfloor\frac{n}{2}\right\rfloor \cdot \left\lfloor\frac{n-1}{2}\right\rfloor,$$

confirming Conjecture 17.3.4 whenever $\min\{m, n\} \leq 6$. Kleitman [555] revised and clarified one part of his proof (a parity argument) a few years later. In 1993, Woodall [932] showed that the conjecture is true also when $m \in \{7, 8\}$ and $n \in \{7, 8, 9, 10\}$.

17.3.4 Crossing numbers for complete graphs

Since K_1, K_2, K_3 and K_4 are planar, for $1 \leq k \leq 4$, $\mathrm{cr}(K_k) = 0$, and by a standard drawing of K_4 with one vertex in the middle, $\overline{\mathrm{cr}}(K_4) = 0$. In Exercises 320 and 321, it was shown that $\mathrm{cr}(K_5) = 1$ and $\mathrm{cr}(K_6) = \overline{\mathrm{cr}}(K_6) = 3$.

One idea for a drawing of K_n with fewest crossings was given by Anthony Hill in 1959 (see [72, p. 45] for a brief story of how Hill's result was picked up by Richard Guy); see [3] for a brief description of Hill's construction. Also see [311], [437], or [455]. The basic idea is to put $\lfloor n/2 \rfloor$ vertices equally spaced around the base of a cylinder and $\lceil n/2 \rceil$ vertices around the top of the cylinder. For each circle, connect all vertices with straight lines, and for edges between the circles, use geodesics along the cylinder. According to Andrii Arman (personal communication, 2017), the similar construction on a sphere also gives the same bounds. ("Cylindrical crossing numbers" have since been well-studied—see [3] or [776].) (Such constructions were also given by Blažek and Koman [107] and by Guy and Jenkyns [443].) Such a "cylindrical" construction gives (with a bit of work) the following bound:

Theorem 17.3.5 (Hill 1959, see [437]).

$$cr(K_n) \leq \frac{1}{4} \left\lfloor \frac{n}{2} \right\rfloor \cdot \left\lfloor \frac{n-1}{2} \right\rfloor \cdot \left\lfloor \frac{n-2}{2} \right\rfloor \cdot \left\lfloor \frac{n-3}{2} \right\rfloor .$$

The drawing of K_6 in Figure 17.1 is based on a cylindrical construction.

Exercise 327. *Find a drawing of K_7 with 9 crossings. Hint: Use the Hill idea of putting 4 points in a square surrounded by 3 points in a triangle.*

In the Harary–Hill paper [455] are two drawings of K_8 that show $\mathrm{cr}(K_8) \leq 18$; the first of these drawings is formed with the 8 vertices on a circle; the second is based on the cylindrical method of Hill, and is given here in Figure 17.6.

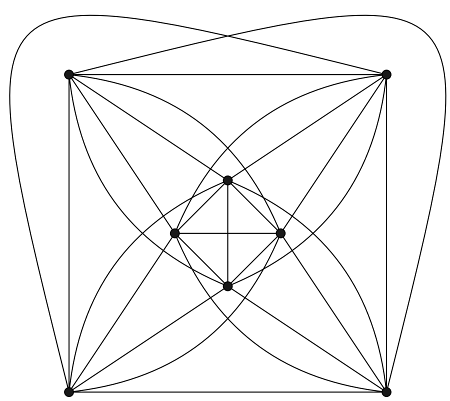


Figure 17.6: $\mathrm{cr}(K_8) \leq 18$

Crossing numbers for a few other small complete graphs have been computed (see [439] or [440]): $\mathrm{cr}(K_7) = 9$, $\mathrm{cr}(K_8) = 18$, $\mathrm{cr}(K_9) = 36$, and $\mathrm{cr}(K_{10}) = 60$. Based on these values, the following conjecture (probably first due to Hill; see [437]) became popular:

Conjecture 17.3.6 (see, e.g., [455]).

$$cr(K_n) = \frac{1}{4}\left\lfloor\frac{n}{2}\right\rfloor \cdot \left\lfloor\frac{n-1}{2}\right\rfloor \cdot \left\lfloor\frac{n-2}{2}\right\rfloor \cdot \left\lfloor\frac{n-3}{2}\right\rfloor.$$

According to [311], if Conjecture 17.3.6 is true for some odd n, then it is also true for $n+1$. Eggleton and Guy [290] showed that for n odd, $\mathrm{cr}(K_n)$ and $\binom{n}{4}$ have the same parity.

In 2007, Pan and Richter [711] confirmed that Conjecture 17.3.6 is true for $n = 11, 12$ as well, by showing that $\mathrm{cr}(K_{11}) = 100$, from which it follows that $\mathrm{cr}(K_{12}) = 150$.

To support Conjecture 17.3.6, a simple lower bound argument shows that $\mathrm{cr}(K_n)$ is of the order n^4. (In [72], this argument is credited to Guy.)

Lemma 17.3.7 (Guy, 1969 [438]). *For $n \geq 5$,*

$$cr(K_n) \geq \frac{1}{120}n(n-1)(n-2)(n-3).$$

Proof: Consider some drawing of K_n. For each of the $\binom{n}{5}$ copies of K_5 there is at least one pair of edges crossing. Each crossing uses 4 points, so each crossing is in at most $n-4$ copies of K_5. Thus $\mathrm{cr}(K_n) \geq \frac{\binom{n}{5}}{n-4}$, from which the result follows. $\square$

In 2017, Balogh, Lidický, and Salazar [47] used flag algebras to show that, asymptotically,

$$\mathrm{cr}(K_n) \geq 0.985 \cdot \frac{1}{4}\left\lfloor\frac{n}{2}\right\rfloor \cdot \left\lfloor\frac{n-1}{2}\right\rfloor \cdot \left\lfloor\frac{n-2}{2}\right\rfloor \cdot \left\lfloor\frac{n-3}{2}\right\rfloor.$$

The same authors [47] give a similar result for spherical crossing numbers, but with .996 replacing .985.

17.3.5 Rectilinear crossing numbers

Recall from Definition 17.3.2 that the rectilinear crossing number of G, $\overline{\mathrm{cr}}(G)$, is the least number of crossings when drawings are restricted to those whose edges are straight-line segments. Rectilinear linear crossing numbers appear to be introduced in the 1963 paper by Harary and Hill [455] (where they are denoted $\overline{c}(G)$).

The following fact is often called "Fáry's theorem", named after István Fáry, even though it was proved by Klaus Wagner [907] 12 years earlier, and later, independently by S. K. Stein [822] in 1951.

Theorem 17.3.8 (Wagner, 1936 [907]; Fáry, 1948 [330])**.** *Any planar graph has a plane drawing with all edges drawn as straight line segments. Hence for any planar graph G, $\overline{cr}(G) = cr(G) = 0$.*

One simple proof of Theorem 17.3.8 is by induction (see, e.g., [184, pp. 259–260]). So, by Fary's theorem, if G is planar, then $\overline{\mathrm{cr}}(G) = 0$.

Remark: In 1987, Harborth *et al.* [457] conjectured that any planar graph can be drawn with straight line segments having *integer lengths*. This conjecture has been confirmed for cubic (regular of degree 3) graphs [384].

By 2001, rectilinear crossing numbers for complete graphs of order $n \leq 10$ were found. As was shown above, $\overline{\mathrm{cr}}(K_4) = 0$, $\overline{\mathrm{cr}}(K_5) = 1$ (see Figure 24.9), and $\overline{\mathrm{cr}}(K_6) = 3$ (see Figure 17.1). Harary and Hill [455] conjectured that $\overline{\mathrm{cr}}(K_8) = 19$ (and gave a drawing like in Figure 17.7 witnessing the upper bound). Harary and Hill also showed $\overline{\mathrm{cr}}(K_9) = 36$ (they gave a drawing as in Figure 17.8, which shows that the rectilinear crossing number agrees with crossing number). Harary and Hill conjectured that $\overline{\mathrm{cr}}(K_{10}) = 63$; even though Harary and Hill did not accurately predict $\overline{\mathrm{cr}}(K_{10})$, their conjecture that for $n = 8$ and $n \geq 10$, $\overline{\mathrm{cr}}(K_n) > \mathrm{cr}(K_n)$ was indeed correct.

Exercise 328. *Find a drawing that shows $\overline{cr}(K_7) \leq 9$. Note that since $cr(K_7) = 9$, this inequality is equality. Hint: Does removing a vertex from the drawing in Figure 17.7 work? Can the drawing in Figure 24.14 be modified to give a rectilinear drawing?*

In 1971, David Singer [798] confirmed that $\overline{\mathrm{cr}}(K_8) = 19 > 18 = \mathrm{cr}(K_8)$ and

$$\overline{\mathrm{cr}}(K_{10}) \in \{61, 62\}.$$

Singer also showed that

$$\lim_{n\to\infty} \frac{\overline{\mathrm{cr}}(K_n)}{n^4} \leq \frac{5}{312}.$$

It wasn't until 2001 when Brodsky, Durocher (now a professor at U. Manitoba), and Gethner [147] showed that $\overline{\mathrm{cr}}(K_{10}) = 62$. See also [4], [5], or [148]

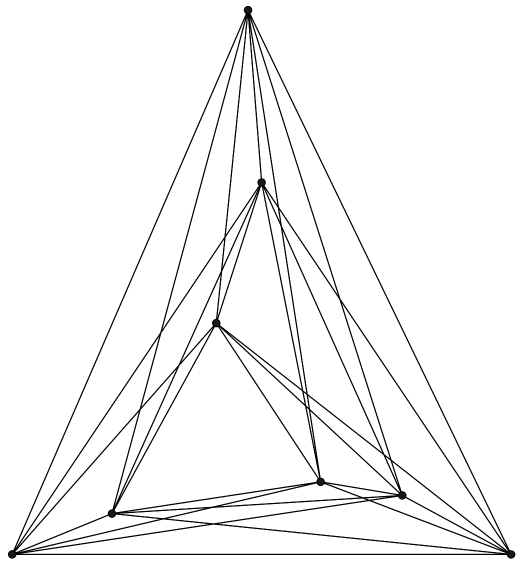


Figure 17.7: $\overline{\mathrm{cr}}(K_8) \leq 19$, drawing from [455, Fig. 6].

for more recent developments on evaluating $\overline{\mathrm{cr}}(K_n)$. In [4], it is claimed that

$$\overline{\mathrm{cr}}(K_n) \geq \frac{1}{4}\left\lfloor\frac{n}{2}\right\rfloor \cdot \left\lfloor\frac{n-1}{2}\right\rfloor \cdot \left\lfloor\frac{n-2}{2}\right\rfloor \cdot \left\lfloor\frac{n-3}{2}\right\rfloor.$$

According to [5], for all $n \leq 27$, the rectilinear numbers for K_n have been found. See [10] for a page with more recent results, including examples of optimal drawings, and bounds for $n \leq 100$.

In 1993, Bienstock and Dean [91] showed that if G is a simple graph with $\mathrm{cr}(G) \leq 3$, then $\overline{\mathrm{cr}}(G) = \mathrm{cr}(G)$. Bienstock and Dean also show that there are graphs with crossing number 4 but arbitrarily large rectilinear crossing number.

17.3.6 Degenerate crossing numbers

In calculating crossing numbers, it is required that only one pair of edges cross at any one point, or that any crossing point has precisely two edges through it. If a drawing is given with $k \geq 3$ lines all crossing at one common point, this one point is witness to $\binom{k}{2}$ (pairwise) crossings, and by perturbing the edges slightly, all such crossings can occur at $\binom{k}{2}$ distinct points. If one allows more than a pair of edges to cross at any one point, one might be able

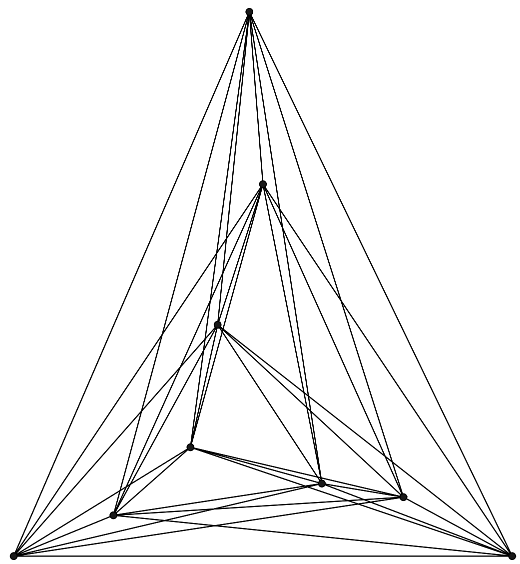


Figure 17.8: $\overline{\text{cr}}(K_9) \leq 36$, drawing from [455, Fig. 7].

to reduce the total number of "crossing points" by allowing more than two edges at a crossing point.

Definition 17.3.9. The *degenerate crossing number* of G, denoted $\text{cr}^*(G)$, is the minimum number of crossing *points* (where more than two edges may pass through the same point in $\mathbb{R}^2$).

By definition, for any G, $\text{cr}^*(G) \leq \text{cr}(G)$. There are examples of graphs that have a degenerate crossing number that is smaller than the crossing number; for example, by Kleitman's above result (17.3), $\text{cr}(K_{5,5}) = 16$; however, $\text{cr}^*(K_{5,5}) = 15$ (see [705]).

Exercise 329. *What is the degenerate crossing number for K_6?*

When limiting the definition of degenerate crossing numbers to drawings where edges intersect at most once, a lower bound is obtained.

Theorem 17.3.10 (Pach–Toth, 2009 [705])**.** *There exists a constant $c^* > \frac{1}{40^4}$ so that for every graph $G = (V(G), E(G))$ with $|E(G)| \geq 4|V(G)|$,*

$$cr^*(G) \geq c^* \frac{|E(G)|^4}{|V(G)|^4}.$$

17.3.7 General bounds for crossing numbers

Lemma 17.3.11. *Let D be a drawing of a graph on n vertices with m edges and $cr(D)$ crossings. Then $cr(D) \geq m - 3n + 6$.*

Proof: Create a planar drawing H (with no crossings) by removing one edge from each crossing in D; at most $\mathrm{cr}(D)$ edges need be removed (less may be required if there are edges in D with multiple crossings). So H has at least $m - \mathrm{cr}(D)$ edges and by Lemma 2.3.3) for planar graphs,

$$m - \mathrm{cr}(D) \leq 3n - 6,$$

from which the desired result follows. □

The following theorem (which was conjectured by Erdős and Guy [311, p. 56]) is sometimes (e.g., [645, p. 55]) called "the crossing number theorem", first published in 1982 by Ajtai, Chvatál, Newborn, and Szemerédi [14] (and independently a year later by Leighton [605] in work on VLSIs—see also [604] for the relations between crossing numbers and VLSIs).

Theorem 17.3.12 (Crossing number theorem)**.** *There exists a constant $0 < c < 1$ so that for every simple graph $G = (V, E)$ with $|E| \geq 4|V|$,*

$$cr(G) \geq c \cdot \frac{|E|^3}{|V|^2} - |V|.$$

Proof of Theorem 17.3.12: An original proof used $c = \frac{1}{100}$ and was complicated. The proof given here is a simple probabilistic one by Chazelle, Sharir, and Welzl (as found in [13] or [645]), that shows $c \geq \frac{1}{64}$.

Let $G = (V, E)$ have $n = |V|$ vertices and $m = |E|$ edges. Let D be a drawing of G with a minimal number of crossings. Let $p \in (0, 1]$ be a probability (to be identified later), and let H be a subgraph of D formed by selecting vertices of D independently and randomly with probability p (and deleting vertices of D with probability $1-p$). Then $\mathrm{cr}(H)$ is a random variable (as are $|E(H)|$ and $|V(H)|$). The probability that an edge of D remains in H is p^2, and since each crossing in D uses four vertices, the probability that a crossing in D survives in H is p^4. By linearity of expectation,

$$\begin{aligned} p^4\mathrm{cr}(G) &= \mathbb{E}[\mathrm{cr}(H)] \\ &\geq \mathbb{E}[|E(H)| - 3|V(H)| + 6] \qquad \text{(by Lemma 17.3.11)} \end{aligned}$$

$$= p^2 m - 3pn + 6,$$

and so

$$\text{cr}(G) \geq \frac{m}{p^2} - \frac{3n}{p^3}.$$

Using $p = \frac{4n}{m}$ (which is at most 1 because $m \geq 4n$),

$$\text{cr}(G) \geq \frac{1}{64}\left(\frac{4m^3}{n^2} - \frac{3m^3}{n^2}\right) = \frac{m^3}{64n^2}.$$

$\square$

For a short history of Theorem 17.3.12 and related facts, see [645, pp. 52–54]. Original bounds for c were apparently quite small (one review of the paper [14] gives $c = \frac{1}{100}$, although I have not seen the original). The value for c in the above proof is $\frac{1}{64} \sim .0156$. In 1997, Pach and Tóth [703] showed that $c \leq .09$, and for $m \geq 7.5n$, $c \geq \frac{1}{33.75} \sim .0296$. This last result was improved in 2006 for $m \geq \frac{103}{16}n$ by Pach, Radoičić, Tardos, and Tóth [700] to $c \geq \frac{1024}{31827} \sim .0321$.

In 2013, Ackerman [6] showed that if $m \geq 6.95n$, then a lower bound for c can be improved to $\frac{1}{29}$ if each edge is crossed at most 4 times.

The bound in Theorem 17.3.12 is asymptotically tight because for $5n \leq m \leq \binom{n}{2}$ there exists a graph with n vertices, m edges, and crossing number $O\left(\frac{m^3}{n^2}\right)$ (see [645, Ex. 1, p. 58]).

A multigraph version of Theorem 17.3.12 has also been shown by a number of authors. (One proof follows the above one. See, for example, [847].)

Theorem 17.3.13 (Various authors)**.** *Let $G = (V, E)$ be a multigraph on n vertices with edge-multiplicity at most k. Then for $|E| \geq 4k|V|$,*

$$cr(G) \geq \frac{1}{64}\frac{|E|^3}{|V|^2 k^3}.$$

For degenerate crossing numbers, Pach and Tóth [705] showed that $\text{cr}^*(G) = \Omega(\frac{m^4}{n^4})$. Ackerman and Pinchasi [7] showed that for a graph G with $m \geq 4n$, $\text{cr}^*(G) = \Omega(\frac{m^3}{n^2})$.

17.3.8 Applications of the crossing number theorem

In the 1997 paper by Székely [847] some combinatorial geometry consequences of the crossing number theorem (Theorem 17.3.12) are given (in some cases, the results had only previously been proved using far more complicated machinery). A few are mentioned in these notes. A third application of crossing numbers is shown regarding Sylvester's four point problem.

The Szemerédi–Trotter theorem

The upper bound part of the Szemerédi–Trotter theorem (Theorem 15.4.9) states that if $I(p,q)$ is the maximum number of incidences between p points and q lines in $\mathbb{R}^2$, then $I(p,q) = O((pq)^{2/3} + p + q)$. This upper bound is a relatively easy consequence of the crossing number theorem.

Proof of upper bound in Theorem 15.4.9: Suppose that a configuration of p points and q lines are given in the plane. Define a graph G whose vertices are points and whose edges are consecutive points on a line. Since any two lines cross at most once, $\mathrm{cr}(G) \le \binom{q}{2}$. In each line, the number of vertices is one more than the number of edges. So, counting over all lines, the number of incidences is at most $|E| + q$. Thus, $I(p,q) \le |E| + q$, and so $|E| \ge I(p,q) - q$. Then either $|E| \le 4|V|$ (in which case, $I(p,q) \le 4p + q$, satisfying the theorem), or $|E| \ge 4|V|$ and

$$\begin{aligned}
\binom{q}{2} &\ge \mathrm{cr}(G) \\
&\ge \frac{1}{64}\frac{|E|^3}{|V|^2} && \text{(by Thm. 17.3.12)} \\
&\ge \frac{1}{64}\frac{(I(p,q)-q)^3}{p^2},
\end{aligned}$$

and so

$$(I(p,q)-q)^3 \le 64\binom{q}{2}p^2.$$

Hence, there is a constant c so that $I(p,q) \le c(pq)^{2/3} + q$. The two cases give the desired result. $\square$

Distances between n points

As observed by Székely [847, 848], the crossing number theorem can be used to give an easy proof of a lower bound on the number of distinct distances between n points in the plane. See Section 17.4.2 for details (see also [645, Section 4.4]). Also observed by Székely is an easy proof for the number of times a unit distance can occur in a geometric graph in the plane (see Section 17.4.3).

Rectilinear crossing numbers and the Sylvester four point problem

In Section 4.2.3, Sylvester's famous four point problem was given, which is to find the probability that four "random" points are in convex position. In 1994, Scheinerman and Wilf [777] showed how rectilinear crossing numbers are related to Sylvester's four point problem (see also [925]). To describe this relation, a folklore result (found in [777, Thm. 2]) that extends Lemma 17.3.7 is first given.

Theorem 17.3.14 (Folklore)**.** *There exists a positive finite constant L so that*

$$L = \lim_{n\to\infty} \frac{\overline{cr}(K_n)}{\binom{n}{4}} = \sup_n \frac{\overline{cr}(K_n)}{\binom{n}{4}}.$$

Proof: Let $4 \le m < n$, and let D be an optimal rectilinear drawing of K_n (with $\overline{\mathrm{cr}}(K_n)$ crossings) on an n-element set V of vertices. Each crossing C in D is determined by 4 vertices, and so is contained in $\binom{n-4}{m-4}$ copies of K_m in D. For any subset A of m vertices, let $\mathrm{cr}(A)$ denote the number of crossings induced in D of the complete graph on A. Counting pairs (crossing C, copy of K_m containing C) in two ways,

$$\begin{aligned}\overline{\mathrm{cr}}(K_n)\binom{n-4}{m-4} &= \sum_{A\subset V(D);|A|=m} \mathrm{cr}(A)\\ &\ge \binom{n}{m}\overline{\mathrm{cr}}(K_m),\end{aligned}$$

where the inequality above is given in case any of the copies of K_m have more than the necessary number of crossings. Rearranging,

$$\frac{\overline{\mathrm{cr}}(K_n)}{\binom{n}{m}} \ge \frac{\overline{\mathrm{cr}}(K_m)}{\binom{n-4}{m-4}}.$$

Since $\frac{\binom{n}{m}}{\binom{n-4}{m-4}} = \frac{\binom{n}{4}}{\binom{m}{4}}$,

$$\frac{\overline{\text{cr}}(K_n)}{\binom{n}{4}} \geq \frac{\overline{\text{cr}}(K_m)}{\binom{m}{4}}.$$

Thus the sequence $\{\overline{\text{cr}}(K_n)/\binom{n}{4}\}$ is increasing, bounded above by 1 and below by $\overline{\text{cr}}(K_5)/\binom{5}{4} = \frac{1}{5}$. □

Since $\overline{\text{cr}}(K_{10}) = 62$, $L \geq \frac{31}{105}$. By a result of Singer [798], when n is a power of 3, $\overline{\text{cr}}(K_n) \leq \frac{5n^4-39n^3+91n^2-57n}{312}$. So $L \leq \frac{5}{312} \cdot 24 = \frac{5}{12}$, and thus

$$\frac{31}{105} \leq L \leq \frac{5}{12}.$$

□

Let q be the probability that in a convex open set R with finite area, if four points are chosen at random, they form a convex quadrilateral (see Sylvester's four point problem, Section 4.2.3). Scheinerman and Wilf [777] proved that $q = L$. (The proof is short, but omitted here; essentially, the proof relies on the correspondence between four points in convex position and a crossing.)

17.3.9 Crossing numbers on other surfaces

This section contains only a few facts and references on crossing numbers for surfaces other than the plane or sphere.

The complete graph K_7 is not planar and $\text{cr}(K_7) = 9$. However, K_7 can be drawn on a torus with no crossings; that is, if $\text{cr}_T(G)$ denotes the "toroidal crossing number" of a graph G drawn on a torus, then $\text{cr}_T(K_7) = 0$. See Figure 17.9 for a torus with K_7 embedded with no crossings. See [444] for early work on the toroidal crossing numbers for the complete graph, and [443] for early work on toroidal crossing numbers for $K_{m,n}$. Much more work has since been done on embedding graphs on tori; for example, see [360]. For early work on crossing numbers for projective planes or Klein bottles, see Koman's 1969 work [565]. Crossing numbers for a "torus" with g holes (and many other problems) were considered by Pach, Spencer, and Tóth in [702].

For more on embedding graphs on surfaces, see one of the standard references for topological graph theory, e.g., [416] (this book appears in Figure

Figure 17.9: K_7 embedded on torus, by K. Gunderson, circa 2008

17.9), [661] (where crossing numbers are not considered), or [923] (where crossing numbers are briefly introduced on pp. 76ff, including crossing numbers for graphs on surfaces with higher genus).

17.4 Geometric graphs and distances

17.4.1 Distances in a set with a given diameter

There are many questions one can ask regarding possible distances in a geometric graph. Such questions can often be solved by considering only vertex sets with some fixed diameter, usually 1. If V is a set of n points in the plane with diameter 1, how many times can the distance 1 be realized? How many times can any smaller distance be realized? How many or how few different distances are possible? How many distances can be found in any particular range of $(0, 1]$?

For example, when $n = 6$ vertices occur as vertices of a regular hexagon (with diameter 1), one can check that three pairs have distance 1, six pairs have distance 1/2, and six pairs have distance $\sqrt{3}/2$. Thus nine pairs have distance greater than $1/\sqrt{2}$. However, if six vertices are formed by splitting the three vertices of an equilateral triangle (where each split pair remain very close), then all pairs except the split pairs have distance greater than $1/\sqrt{2}$, giving 12 such pairs. Turán's theorem shows that such a result is optimal:

Theorem 17.4.1 (see [125, pp. 114–115]). *If $\{v_1, \ldots, v_n\}$ is a set of points in the plane with Euclidean diameter 1, then the maximum number of pairs at distance greater than $1/\sqrt{2}$ is $\lfloor n^2/3 \rfloor$; furthermore, there exists such a placement of the n points so that exactly $\lfloor n^2/3 \rfloor$ pairs have distance greater than $1/\sqrt{2}$.*

Proof outline: For any points u and v in the plane, let $d(u, v)$ denote their Euclidean distance. Consider the graph G on $V(G) = \{v_1, \ldots, v_n\}$, where v_i and v_j are adjacent iff $d(v_i, v_j) > 1/\sqrt{2}$.

Claim: G contains no K_4.
Proof of claim: Any four points in the plane determine at least one angle with measure at least $\pi/2$. (This can be seen since any four points have a convex hull that is either a line, a triangle, or a quadrilateral, and in each case, three points can be found that have a large angle.) Consider four points, say, v_1, v_2, v_3, v_4, and suppose that $v_1v_2v_3$ form a large angle. At least one of the distances $d(v_1, v_2)$, $d(v_2, v_3)$, is at most $1/\sqrt{2}$ since v_1v_3 has length at least as long as the hypotenuse of a $1/\sqrt{2} : 1/\sqrt{2} : 1$ right triangle, and $d(v_1, v_3) \leq 1$. Thus any 4-tuple of points has at least one pair that is not adjacent in G, proving the claim.

From Turán's theorem, $|E(G)| \leq t(n; 3) = \lfloor n^2/3 \rfloor$, which shows the first part of this theorem.

To see the second statement, form the arrangement of n vertices by separating them into three groups with sizes as close as possible. For each group, put the vertices in a very small circle, and arrange the three groups at the vertices of an equilateral triangle. □

Exercise 330. *(a) Show that if three points are in the plane, where distance between any two is at most one, then there is a disc of radius $\frac{1}{\sqrt{3}}$ containing them.*

(b) Show that if $X \subseteq \mathbb{R}^2$ is any finite set of diameter at most 1, then there is is a disc of radius $\frac{1}{\sqrt{3}}$ containing X.

The next two exercises can be solved by applying Turán's theorem in the same way as in the proof of Theorem 17.4.1.

Exercise 331. *Suppose a flat circular city has radius 6 kilometers, and is patrolled by 18 police cars, each with a two-way radio. If the range of a radio is 9 kilometers, show that at all times there are at least two cars that can communicate with at least five other cars.*

Exercise 332. *A flat circular city with radius 4 miles has 18 cell-towers, each tower with transmission range of 6 miles. Prove that no matter the placement of the towers in the city, at least two will each be able to transmit to at least 5 others.*

Exercise 333. *Show that the greatest distance arising from a set of n points in the plane is realized by at most n different pairs of points. Hint: try induction. Furthermore, for each n, find a configuration of n points so that exactly n pairs have greatest distance.*

Is there a higher-dimensional analog to the result in Exercise 333?

Conjecture 17.4.2 (Vázsonyi, given by Erdős in 1946 [298]). *The greatest distance determined by n points in $\mathbb{E}^3$ occurs at most $2n-2$ times.*

If P is a finite set of points in a Euclidean space, the *diameter* of P is the maximum distance between any two points in P. Letting P a set of vertices in a geometric graph, the *diameter graph* for P is formed by putting an edge between points at maximum distance. Vázsonyi's conjecture says that a diameter graph on n vertices in $\mathbb{F}^3$ can have at most $2n-2$ edges. An obvious example is the regular tetrahedron on $n=4$ vertices with 6 edges.

If V is a set of points in some Euclidean space, the "ball polytope" $B(V)$ is the intersection of all unit balls with centers in V. Vászonyi's conjecture was proved using ball polytopes, first by Grünbaum [424] in 1956, and by both Heppes [464] and Straszewicz [834] in 1957. (For more on ball polytopes and Vászonyi's conjecture, see also [582].) In 2008, Swanepoel [843] gave another kind of proof, and showed how extremal examples can be embedded in the projective plane. In 2009, Swanepoel [844] showed that maximal diameter graphs all arise from a specific construction, thereby answering Vászonyi's conjecture in a strong fashion.

17.4.2 The number of different distances between n points

In a 1946 paper [298], Erdős asked (and partially answered) two now famous questions in combinatorial geometry, in particular, in the area of geometric graphs. The first asks, among all possible sets of n points in the plane, what is the minimum number $f(n)$ of different distances realized by pairs of points? The second asks, among n points, what is the minimum number

$g(n)$ of times that any one particular distance can occur? Without loss of generality, one can ask this question for unit-distances, because scaling does the rest. In other words, how many edges can a unit-distance graph have when points are in the plane?

In this section, results concerning $f(n)$ are examined; in Section 17.4.3 the question about unit distances is considered. Modern extremal graph theory, crossing numbers, and the Szemerédi–Trotter theorem have had significant impact on the early questions regarding distances in geometric graphs. In this and the following section, log denotes $\log_2$, but most occurrences of log here can be replaced by ln.

Theorem 17.4.3 (Erdős, 1946 [298]). *Let $f(n)$ denote the minimum number of distances determined by n points in the plane. Then there exists a constant $c > 0$ (independent of n) so that*

$$\sqrt{n - \frac{3}{4}} - \frac{1}{2} \leq f(n) \leq \frac{cn}{\sqrt{\log n}}. \tag{17.4}$$

Proof: For the lower bound, let $P_1, \ldots, P_n$ be points in the plane. Let C be the least polygon containing all points, and suppose that P_1 is on C. Then the remaining points $P_2, \ldots, P_n$ lie on or on one side of a line through P_1. It is shown that either P_1 has many points at different distances or there is a circle centered at P_1 among which many distances are realized.

Let k be the number of different distances occurring among $P_1P_2, \ldots P_1P_n$; then trivially, $k \leq f(n)$. Let M be the maximum number of times any one distance occurs from P_1, and so $kM \geq n-1$, or equivalently, $k \geq \frac{n-1}{M}$ and so $f(n) \geq \frac{n-1}{M}$.

Suppose that r is a distance from P_1 occurring M times; then there are points $Q_1, \ldots, Q_M$ on a semicircle of radius r centered at P_1. Suppose that the points $Q_1, \ldots, Q_M$ lie in order along the semicircle. Then the distances from Q_1 to each of $Q_2, \ldots, Q_M$ are increasing, producing $M-1$ different distances. Thus, $f(n) \geq M - 1$.

Together, these above points show that $f(n) \geq \max\{\frac{n-1}{M}, M-1\}$. The maximum is minimized when $\frac{n-1}{M} = M - 1$ or when $M^2 - M - n + 1 = 0$, in which case, by the quadratic formula, $M = \frac{1 \pm \sqrt{1+4(n-1)}}{2}$, and since $M \geq 1$, $M - 1 = \sqrt{n - \frac{3}{4}} - \frac{1}{2}$, concluding the first inequality in equation (17.4).

To see the second inequality in equation (17.4), suppose that n is a perfect square and consider the example of points arranged in a $\sqrt{n} \times \sqrt{n}$ rectangular grid. Since distances between such points are of the form $\sqrt{a^2+b^2}$, where a, b are integers, the number of distances is given by the number of solutions to $m = a^2 + b^2$ where (using opposite corners of the grid), $m \le 2n$. It is known (see Theorem 22.4.9) that if x is a positive integer and $S(x)$ denotes the number of integers $y \le x$ that are a sum of two squares, then there is a constant $c_0 \sim 0.76422$ (called the Landau–Ramanujan constant) so that

$$\lim_{x\to\infty} \frac{S(x)}{x/\sqrt{\ln x}} = c_0.$$

The upper bound now follows by using $x = 2n$. □

Since the lower bound in (17.4) is achieved by looking at only one point and a circle containing some of its neighbours, one might think that this lower bound can be improved. Erdős conjectured that indeed the upper bound was closer to the truth. This conjecture was confirmed by a sequence of results that gave larger and larger lower bounds. In 1952, Leo Moser [672] showed that $f(n) > \frac{n^{2/3}}{2\cdot 9^{1/3}}$. In 1984, Fan Chung [189] gave a lower bound on the order of $n^{5/7}$. In 1990, Clarkson, Edelsbrunner, Guibas, Sharir, and Welzl [197] showed that there is always one point that shares in $O(n^{3/4})$ distances.

In 1992, Chung, Szemerédi, and Trotter [192] showed $f(n) \ge O(n^{4/5}/\log n)$. In 1997, using crossing numbers, Székely [847] gave a remarkably simple proof for a lower bound on the order of $n^{4/5}$ (and showed that there is a vertex that realizes this number of distances).

In 2001, Solymosi and Tóth [807] improved the lower bound for $f(n)$ to $O(n^{6/7})$, and in 2003, the lower bound was further improved to $O(n^{\frac{4e}{5e-1}-\epsilon})$ (so the exponent is roughly .8635) by Tardos [861]. A year later, Katz and Tardos [531] used entropy to improve the exponent to $\frac{48-14e}{55-16e)-\epsilon} \sim .8641$. Finally, in 2015, Guth and Katz [436] improved the lower bound to $O(\frac{n}{\log n})$, which is of the form $n^{1-o(1)}$, achieving Erdős's original conjecture.

For $k \ge 3$, if $f_k(n)$ denotes the minimum number of distinct distances between n points in $\mathbb{E}^k$, Erdős [298] noted that by using the same proof ideas as used in Theorem 17.4.3, there are constants $c_1, c_2 > 0$ so that

$$c_1 n^{1/k} \le f_k(n) \le c_2 n^{2/k}.$$

Again, it was thought that the upper bound was closer to the actual value. In 2008, Solymosi and Vu [808] showed that $f_k(n) \ge c n^{\frac{2}{k} - \frac{2}{d(d+2)}}$.

Erdős [298] conjectured that if n points are arranged in a convex polygon, then the number of distinct distances is at least $\lfloor n/2 \rfloor$, with equality holding only for the regular polygons. In 1952, Leo Moser [672] showed that this number is at least $\lfloor \frac{n+2}{3} \rfloor$. [I do not know the current state of this conjecture.]

17.4.3 The maximum number of unit distances

The amount of work done in this area is tremendous; here is only a simple introduction.

Theorem 17.4.4 (Erdős, 1946 [298]). *For $n \geq 2$, let $g(n)$ denote the maximum number of unit distances between n points in the plane. There exists a constant c_1 so that for any n,*

$$n^{1+c_1/\log\log n} \leq g(n) \leq O(n^{3/2}).$$

Erdős thought (see also [302]) that the lower bound was closer to being correct.

Note: The function $g(n)$ from Theorem 17.4.4 is also denoted by $U_2(n)$ in the literature, since then the notation is in place for higher dimensional questions. Erdős actually used a more general notation for any distance r, namely $g(n, r)$; in this context, the notation $g(n)$ above is really Erdős's $g(n, 1)$.

To prove the lower bound in Theorem 17.4.4, Erdős again gave the example of a $\sqrt{n} \times \sqrt{n}$ integer lattice point set, He then said "...we easily obtain (using well-known theorems about the number of solutions to $u^2 + v^2 = m$)" the desired result. For complete details of this proof, see, e.g., [645]; in these notes, see Theorems 22.1.11, 22.4.8, and 22.4.9. The upper bound in Theorem 17.4.4 was proved by fairly simple counting and an extremal graph result.

Exercise 334. *Use the result in Exercise 319 to show that if n points are placed in the plane, the number of pairs of points with distance exactly 1 is at most*

$$\frac{n^{3/2}}{\sqrt{2}} + \frac{n}{4}.$$

The upper bound in Theorem 17.4.4 has since been reduced by different authors. For example, in 1984, Spencer, Szemerédi, and Trotter [813] gave an upper bound of $g(n) \leq O(n^{4/3})$. In 1997, Székeley [847], again using crossing numbers (see Section 17.3), gave a simple proof of the Spencer–Szemerédi–Trotter theorem.

Theorem 17.4.5 (Spencer–Szemerédi–Trotter, 1984 [813])**.** *As $n \to \infty$, the maximum number of unit-distances in a geometric graph in the plane on n vertices is at most $O(n^{4/3})$.*

Proof by Székely [847]: Let X be a set of n points in the plane. Around each point in X, draw a unit circle. One of these unit circles may be incident with other points. Delete those circles that pass through less than two points. If a circle centered at some point P contains $m \geq 2$ points, then the circle is divided into m arcs, and these m points determine m unit distances. Since two circles intersect in at most 2 points, there may be at most two arcs between a pair of points that are adjacent on some circle.

Form the geometric multigraph H with vertex set X and the set of edges are the arcs drawn between consecutive vertices on a circle. If two vertices are connected by two arcs, arbitrarily delete one of these arcs thereby creating a simple graph G (still with vertex set X).

If N is the number of unit distances between points of X, then N is the number of point-circle incidences, and removing the circles with only one such incidence gives deletes at most n incidences, and since at most half of the arcs have been deleted, it follows that $|E(G)| \geq \frac{N-n}{2}$. For this value to be positive, one can assume that, say, $N > 5n$ (for if not, the bound on N is proved).

Now count the number of crossings in two ways. By the crossing number theorem (Theorem 17.3.12)

$$\mathrm{cr}(G) \geq \frac{|E(G)|^3}{64n^2} \geq \frac{(N-n)^3}{8 \cdot 64n^3}. \tag{17.5}$$

On the other hand, since any two circles intersect at most twice, there are at most $2\binom{n}{2} < n^2$ crossings. Thus, by (17.5)

$$n^2 > \frac{(N-n)^3}{512n^2},$$

form which it follows that $(N-n)^3 < 512n^4$, and so $N < 8n^{4/3}+n$, completing the proof. $\square$

Exercise 335. *Find a finite set X of points in the plane so that each point $\mathbf{x} \in X$ has at least 2018 other points in X at unit distance from $\mathbf{x}$.*

Let $U_3(n)$ denote the maximum number of unit distances among n points in $\mathbb{R}^3$.

Theorem 17.4.6 (Erdős, 1960 [301])**.**

$$\Omega(n^{4/3} \log\log n) \leq U_3(n) \leq \left(\frac{1}{2} + o(1)\right) n^{5/3}.$$

Proof of upper bound: Let P be a set of n points in $\mathbb{E}^3$. Form the unit-distance graph $G = (V, E)$ on vertex set $V = P$ (where two points in P form an edge if and only if they are at unit distance). Consider three vertices, say x, y, z. The intersection of their neighbourhoods lie on the intersections of 3 spheres. Since 3 unit spheres can intersect in at most 2 points, G contains no copy of $K_{3,3}$. The upper bound for $\text{ex}(n; K_{3,3})$ given in Theorem 17.2.5 finishes the proof. □

Erdős [305] asked similar questions about the maximum number of unit distances in spaces with different norms (e.g., a taxicab norm); for further information and references, see [142].

17.4.4 s-distance sets

Definition 17.4.7. For positive integers s and n, an s-distance set in $\mathbb{E}^n$ is a collection of points with precisely s different distances between pairs of points.

The three vertices of an equilateral triangle form a 1-distance set (in $\mathbb{E}^2$ or in higher dimensional spaces), and an isosceles triangle (that is not equilateral) yields a 2-distance set. The vertices of a regular pentagon also form a 2-distance set.

Exercise 336. *Let $n \geq 3$. In $\mathbb{E}^2$, how many (different) distances are determined by the vertices of a regular n-gon?*

For any $s, n \in \mathbb{Z}^+$, what is the maximum number of points in $\mathbb{E}^n$ forming an s-distance set? For a synopsis of work and open problems in this area see [120, pp. 389–391], [226, §F3], and [418, Chapter 17]. In 1947, Kelly [537] showed that the maximum number of vertices in a 2-distance set in $\mathbb{E}^2$ is 5, realized by the vertices of a regular pentagon. In 1962, Croft [225] showed that the maximum number of vertices in a 2-distance set in $\mathbb{E}^3$ is 6, realized by the vertices of a regular octahedron.

Exercise 337. *For each $n \geq 3$, give an example of 2-distance set in $\mathbb{E}^n$ with $\binom{n}{2}$ elements.*

Any 2-distance set in $\mathbb{R}^n$ can not have too many more than $\binom{n}{2}$ points, as is observed below (in Theorem 17.4.9). For the proof of this upper bound, a lemma is used.

Lemma 17.4.8. *For each $i = 1, \ldots, m$, and a set X, let $f_i : X \to \mathbb{F}$ be functions and $v_1, \ldots, v_m \in X$ so that (i) for each $i \in [1, m]$ $f_i(v_i) \neq 0$, and (ii) for each $i \neq j$, $f_i(v_j) = 0$. Then $f_1, \ldots, f_m$ are linearly independent in the space $f : X \to \mathbb{F}$.*

Exercise 338. *Prove Lemma 17.4.8*

Theorem 17.4.9 (Larman–Rogers–Seidel, 1977 [592]). *Any 2-distance set in $\mathbb{E}^n$ has at most $\binom{n}{2} + 3n + 2$ points.*

Proof: Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ be a 2-distance set in $\mathbb{E}^n$, with distances d_1 and d_2 (where $d_1 \neq d_2$ and each is non-zero). For each $i = 1, \ldots, m$, define the polynomial $f_i : \mathbb{E}^n \to \mathbb{R}$ by

$$f_i(\mathbf{x}) = \left(\|\mathbf{x} - \mathbf{a}_i\|^2 - d_1^2\right)\left(\|\mathbf{x} - \mathbf{a}_i\|^2 - d_2^2\right).$$

For each i, then $f_i(\mathbf{a}_i) = d_1^2 d_2^2 \neq 0$, and for $j \neq i$, $f_i(\mathbf{a}_j) = 0$. By Lemma 17.4.8, the f_is are linearly independent. Writing $\mathbf{x} = (x_1, \ldots, x_n)$, these f_is are in the span of the five types of polynomials of the forms (where $i, j \in [n]$)

$$\left(\sum_{i=1}^n x_i^2\right)^2, \quad \left(\sum_{i=1}^n x_i^2\right) x_j, \quad x_i x_j, \quad x_i, \ 1.$$

The numbers of respective types are $1,\ n,\ \binom{n}{2} + n,\ n,\ 1$, giving $\binom{n}{2} + 3n + 2$ in all. So the (linearly independent) polynomials live in a vector space of dimension $\binom{n}{2} + 3n + 2$. Thus $m \leq \binom{n}{2} + 3n + 2$. □

Exercise 339. *For $n \geq 2$, show that the midpoints of a regular simplex in $\mathbb{R}^n$ form a 2-distance set with $\binom{n+1}{2} = \binom{n}{2} + n$ points.*

For dimensions $n = 2, 6, 22$, Delsarte, Goethals, and Seidel [249] give examples of 2-distance sets in $\mathbb{E}^n$ consisting of $\binom{n}{2} + 2n$ points (in fact, these sets are spherical).

Exercise 340. *Find a configuration of 6 points in the plane so that any 3 form an isosceles triangle. Then prove that among any 7 points in the plane, there exist 3 points that do not form an isosceles triangle.*

The question in Exercise 340 was posed by Erdős in [395], and mentioned again [302, p. 244] in 1961, where he said that similar thresholds for graphs in higher dimensions are not known, even for graphs in $\mathbb{E}^3$.

17.4.5 Distance graphs

A graph is a pair $G = (V, E)$, where V is a non-empty set and E is a collection of pairs from V; elements of V are called vertices and elements of E are called edges. Of particular interest here (and in sections following) are graphs whose vertex set V is a collection of points in some Euclidean space, and pairs of points are adjacent in the graph if they are at some specified distance(s).

A proper (vertex) k-colouring of a graph G is a function $c : V(G) \to [k]$ so that for each $\{x, y\} \in E(G)$, $c(x) \neq c(y)$. Recall that the *chromatic number* of a graph G is the least number $k = \chi(G)$ so that there exists a proper k-colouring of $V(G)$. A graph G is called k-colourable if and only if there exists a proper k-colouring of $V(G)$.

Theorem 17.4.10 (Hadwiger, 1944 [445])**.** *For $n \geq 1$, if $n+1$ closed sets cover $\mathbb{E}^n$, then one of the sets realizes all possible distances.*

The proof of Theorem 17.4.10 given in [450, Pr. 59] relies on two previous lemmas about closed sets. Theorem 17.4.10 is improved upon in Theorem 18.1.2, where the condition of "closed" is dropped.

Later, Hadwiger [447] showed that if the plane is covered with 5 congruent closed sets, then at least one of the classes contains two points at unit distance.

Exercise 341. *Using a hexagonal packing of the plane, each regular hexagon with side length 1, show that there is a covering of the plane with 7 congruent closed sets so that the distances satisfying* $2 < d < \sqrt{7}$ *are not realized in any one set.*

For early work on distances realizable in coverings of spheres, see [590].

17.4.6 Unit distance graphs

The *unit distance graph* for a set $S \subseteq \mathbb{E}^n$ is the graph G whose vertex set is S and any pair of vertices $\{x, y\} \subset V(G)$, is an edge in G if and only if the distance between x and y is 1.

For an integer $k \geq 2$, say that a graph G is a *unit distance graph* in $\mathbb{E}^k$ if there exists a set $S \subseteq \mathbb{E}^k$ so that G is isomorphic to the unit distance graph for S.

Unit distance graphs (in the plane) include all cycles and stars. The wheel graph with six spokes is also easily seen to be a unit distance graph. Other notable unit distance graphs include the Moser spindle (see Figure 17.10), hypercube graphs Q_n, the Petersen graph (see Figure 17.3), the Möbius–Kantor graph (see Section 15.4.4), and (see [383]) the Heawood graph (see Figure 11.3).

Exercise 342. *Find a drawing of the Petersen graph in* $\mathbb{E}^2$ *that shows the Petersen graph is a unit-distance graph in the plane.*

17.4.7 Colouring the unit distance graph of the plane

As described in the last section, the *unit distance graph of the plane* is the graph G whose vertices are the points in $\mathbb{E}^2$, and two points are adjacent in G if and only if the two points are distance 1 apart.

If G is the unit distance graph of the plane, what is $\chi(G)$? The notation $\chi(G)$ below is replaced by the more descriptive $\chi(\mathbb{E}^2)$. Again, what is $\chi(\mathbb{E}^2)$?

Exercise 343. *Show that if all the points in the plane are coloured with two colours, there are (at least) two points at distance 1 with the same colour.*

So the result in Exercise 343 says that $\chi(\mathbb{E}^2) > 2$.

The quest for $\chi(\mathbb{E}^2)$ was made popular by Martin Gardner in 1960 [367], but was proposed much earlier (see [803] for more history). The problem of determining $\chi(\mathbb{E}^2)$ is often referred to as the "Hadwiger–Nelson problem" (named after Hugo Hadwiger and Edward Nelson), known since the 1950s (maybe earlier). Finding the chromatic number of the plane is essentially a finite problem:

Theorem 17.4.11 (de Bruijn–Erdős, 1951 [241]). *Let k be a positive integer and let G be an infinite graph. If every finite subgraph of G is k-colourable, then G is k-colourable.*

(Trivially, if G is k-colourable, then every finite subgraph is, too.)

Until recently, only the following bounds were known:

Theorem 17.4.12.

$$4 \leq \chi(\mathbb{E}^2) \leq 7.$$

Proof: To see that the chromatic number is at least four, a unit-distance graph called the Moser graph (named after the Mosers, who published it in 1961 [674]) or "Moser spindle" (see Figure 17.10) is not (properly) 3-colourable.

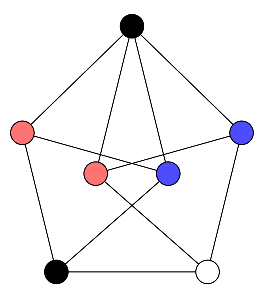


Figure 17.10: Moser's unit-distance graph, not 3-colourable

To show that at most 7 colours are required, John Isbell (for details, see [803]) tiled the plane with regular hexagons with diameter slightly less than 1, and coloured the hexagons in an obvious manner—see Figure 17.11. In order to have a proper 7-colouring, points on the borders of the hexagons are assigned to just one hexagon. (Using a 3×3 patch of squares, the similar idea shows an upper bound of 9.)

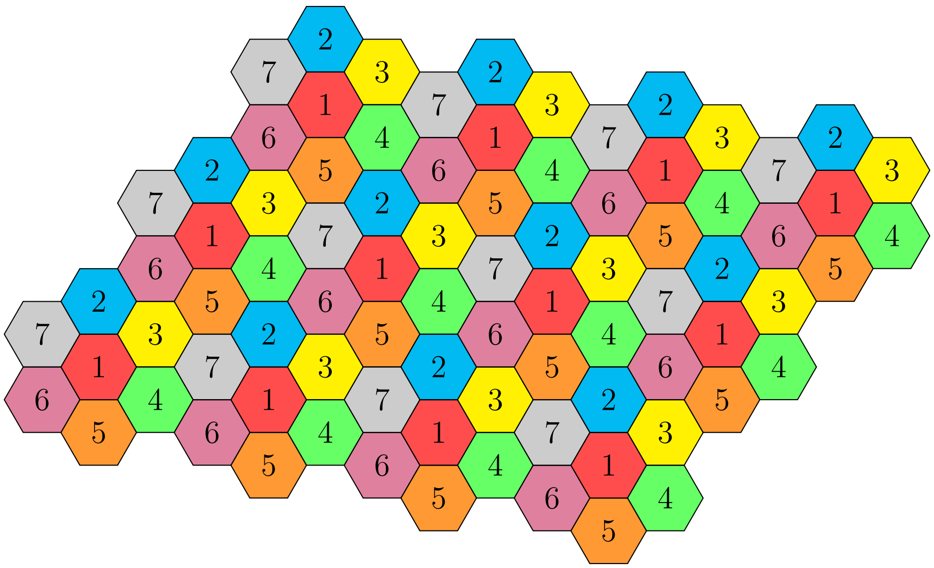


Figure 17.11: Isbell's 7-colouring of the plane

□

In 1998, Pritikin [735] showed that if a 7-chromatic unit-distance graph exists, then it has at least 6198 vertices (so all unit-distance graphs on at most 6197 vertices are 6-colourable).

There was no progress on the bounds for $\chi(E^2)$ for over half a century. Only in 2018 was a preprint posted by Aubrey de Grey that showed $\chi(\mathbb{E}^2) \geq 5$.

Theorem 17.4.13 (de Grey, 2018 [242])**.** *The unit distance graph of the plane is not 4-colourable.*

De Grey gave examples of unit distance graphs, the smallest of which has 1581 vertices, that are not 4-colourable. (This number was originally 1567 in de Grey's first draft. See Gil Kalai's blog [520] for more information.) The construction given starts by adding two vertices to the Moser spindle, one at unit distance from each of the two blue vertices in Figure 17.10, and one at unit distance to the two red vertices—these two new vertices are green in Figure 17.12), then by using this 9 point graph to consecutively create larger and larger graphs with more restrictions on any 4-colourings.

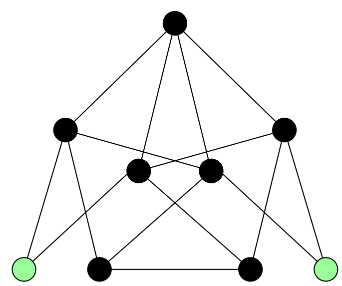


Figure 17.12: The 9 point unit distance graph used by de Grey

After de Grey's announcement of his construction, a *Polymath16* (see [659]). project reduced the number, as of 3 August 2019, to 510 vertices.

Exercise 344. *Show that the 9 point unit distance graph in Figure 17.12 has the bottom 4 vertices collinear and that the triangle formed by the three outer corners is equilateral.*

Shkredov [792] showed that if colour classes are restricted to measurable sets, then at least 5 colours are required. In 1981, Falconer [329] showed that if the extra axiom "all subsets of $\mathbb{E}^n$ are measurable" is assumed, then again, $\chi(\mathbb{E}^2) \geq 5$. See [803] for references and other related facts. A model with this extra axiom is called the *Solovay model*, given by Robert M. Solovay [805] in 1970.

17.4.8 Colouring unit distance graphs of $\mathbb{E}^n$

For three dimensions, Raiskii's [744] 1970 result (see Theorem 18.1.2) shows that $5 \leq \chi(\mathbb{E}^3)$, which was later improved in 2002 by Nechushtan [682] to $6 \leq \chi(\mathbb{E}^3)$.

In 1989, Székeley and Wormald [850] showed that $\chi(\mathbb{E}^3) \leq 21$. This upper bound of 21 was also given in 1996 by Miklós Bóna and Géza Tóth [123] in a paper that shows that when T is the set of vertices of any right-angled triangle, the Ramsey statement $\mathbb{E}^3 \to (T)_3$ holds (see Theorem 18.2.7 here). The Bóna–Tóth proof uses a simple extension of the Isbell hexagonal packing for $\mathbb{E}^2$ to hexagonal prisms.

In 1997, Coulson [213] showed $\chi(\mathbb{E}^3) \leq 18$ and in 2002 [214] gave the improvement of $\chi(\mathbb{E}^3) \leq 15$. This last result was also obtained independently the following year by Radoičić and Tóth [738]; their proof is based on a tiling of 3-space with truncated octahedra with edge lengths $1/\sqrt{10}$.

In [738], the truncated octahedron is called a "permutahedron", a term described by Ziegler [940, pp. 17–18]. Roughly speaking, a permutahedron is a polytope whose vertices can be represented by permutations (and an edge occurs when two vertices differ only by a transposition). Radoičić and Tóth [738] conjectured that higher dimensional permutahedra can be used to give colourings that improve known upper bounds in d-dimensional space. (For example, using their techniques, they claim a proper 54-colouring of $\mathbb{R}^4$, which is likely not best known.)

In 1970, Raiskii [744] showed that $6 \le \chi(\mathbb{E}^4)$, which was improved by in 2006 by Ivanov [499] to $7 \le \chi(\mathbb{E}^4)$. In 2014, Exoo, Ismailescu, and Lim [327] gave a further improvement of $9 \le \chi(\mathbb{E}^4)$. The best published upper bound is $\chi(\mathbb{E}^4) \le 54$ by Radoičić and Tóth [738]. According to Raigorodskii [743], in 2000, Coulson also announced an upper bound of $\chi(\mathbb{E}^4) \le 49$, but this result does not seem to be published.

The lower bound $16 \le \chi(\mathbb{E}^8)$ was given by Larman and Rogers [591]. (In the same paper, Larman and Rogers proved much more; see Theorem 17.4.14 here.) In 2002, Székely [848] showed $16 \le \chi(\mathbb{E}^9)$ (and hence also $16 \le \chi(\mathbb{E}^{10})$), but it wasn't long before these bounds were surpassed. In 2009, Kupavskii and Raigorodskii [580, 581] showed $21 \le \chi(\mathbb{E}^9)$, $23 \le \chi(\mathbb{E}^{10})$, $23 \le \chi(\mathbb{E}^{11})$, and $25 \le \chi(\mathbb{E}^{12})$. Some lower bounds were proved by Kahle and Taha [516] in 2014, who showed that $19 \le \chi(\mathbb{E}^8)$, $21 \le \chi(\mathbb{E}^9)$ (another proof), $26 \le \chi(\mathbb{E}^{10})$, $32 \le \chi(\mathbb{E}^{11})$, and $32 \le \chi(\mathbb{E}^{12})$.

General bounds are still rather far apart.

Theorem 17.4.14 (Larman–Rogers, 1972 [591])**.** *As $d \to \infty$,*

$$\chi(\mathbb{E}^d) \le (3 + o(1))^d.$$

The chromatic number of the unit-distance graph of $\mathbb{E}^d$ is indeed exponential in d:

Theorem 17.4.15 (Frankl–Wilson, 1981 [347])**.** *As $d \to \infty$,*

$$(1 + o(1))(1.2)^d \le \chi(\mathbb{E}^d).$$

This result was improved upon two decades later:

Theorem 17.4.16 (Raigorodskii, 2001 [741])**.** *As $d \to \infty$,*

$$(1.239 + o(1))^d \le \chi(\mathbb{E}^d).$$

A summary of known bounds for $\chi(\mathbb{E}^d)$ is given in Figure 17.13. For surveys of related results, see [143, Sect. 5.9], [411], [743], and [804].

d	lower bound for $\chi(\mathbb{E}^d)$			upper bound for $\chi(\mathbb{E}^d)$
1		2	2	
2	Nelson (ca. 1950)	4	7	Isbell (ca. 1950)
	Mosers [674] (1961)	4		
	de Grey [242] (2018)	5		
3	Nechushtan [682]	6	18	Coulson [213] (1997)
			15	Coulson [214] (2002)
			15	Radoičić–Tóth [738] (2003)
4	Raiskii [744] (1970)	6	54	Radoičić–Tóth [738] (2003)
	Cantwell [174] (1996)	7	49	Coulson, 2000, unpublished
	Ivanov [499] (2006)	7		
	Cibulka [196] (2008)	7		
	Exoo–Ismailescu–Lim [327] (2014)	9		
5	Cantwell [174] (1996)	9		
6	Cibulka [196] (2008)	11		
7	Raigorodskii [741] (2001)	15		
8	Kahle–Taha [516] (2014)	19		
9	Kupavskii–Raigorodskii [580] (2009)	21		
	Kahle–Taha [516] (2014)	21		
10	Kahle–Taha [516] (2014)	26	63	Kupavskii–Raigorodskii [580] (2009)
11	Kahle–Taha [516] (2014)	32		
12	Exoo–Ismailescu [326] (2014)	36		
13	Exoo–Ismailescu [326] (2014)	36		
14	Exoo–Ismailescu [326] (2014)	36		
15	Székely [848] (2002)	37		
$d \to \infty$	Frankl–Wilson [347] (1981) $(1+o(1))(1.2)^d$			Larman–Rogers [591] (1972) $(3+o(1))^d$
	Raigorodskii [741] (2001) $(1.239+o(1))^d$			

Figure 17.13: Some bounds for $\chi(\mathbb{E}^d)$

17.4.9 Colouring unit distance graphs in rational spaces

If G is the unit-distance graph on $\mathbb{E}^2$, then $\chi(G) \geq 3$ (also written $\chi(E^2) \geq 3$) because the vertices of a unit-distance equilateral triangle cannot be properly 2-coloured.

By Lemma 9.1.1, in the lattice plane $\mathbb{Z}^2$, no equilateral triangles exist as an obstacle to proper 2-colourings. Consequently, the rational plane $\mathbb{Q}^2$ also does not have a unit-distance equilateral triangle.

Since equilateral triangles are obstructions to finding a good 2-colouring of the plane, and in the rational plane equilateral triangles do not exist, can the unit distance graph for the rational plane be properly 2-coloured (no pair of points at distance 1 are coloured the same)? Write $\chi(\mathbb{Q}^d)$ to abbreviate the chromatic number of the unit distance graph in $\mathbb{Q}^d$.

Perhaps surprisingly, $\chi(\mathbb{Q}^2) = 2$; before proving this (Theorem 17.4.18, below), an elementary lemma is presented first.

Lemma 17.4.17. *Let $\frac{a}{b}$ and $\frac{c}{d}$ be rationals in reduced form and suppose that*

$$\left(\frac{a}{b}\right)^2 + \left(\frac{c}{d}\right)^2 = 1. \tag{17.6}$$

Then $b = d$, which is odd, and exactly one of a or c is even.

Proof of Lemma 17.4.17: For integers, the notation $x \mid y$ means that x divides y. Rewriting equation (17.6),

$$a^2d^2 + c^2b^2 = b^2d^2. \tag{17.7}$$

Then $d^2 \mid c^2b^2$ and since d and c are relatively prime, $d^2 \mid b^2$, and so $d \mid b$. Similarly, $b^2 \mid a^2d^2$, and so $b \mid d$. Hence $b = d$. Rewriting equation (17.7) and simplifying gives,

$$a^2 + c^2 = b^2. \tag{17.8}$$

If $b = d$ is even, then both a and c are odd (because of reduced forms), and so counting modulo 4, it follows that equation (17.8) has no solution, so $b = d$ is odd. Also from equation (17.8), since b is odd, it follows that exactly one of a or c is odd. □

The conclusion $b = d$ in Lemma 17.4.17 is not needed below.

Theorem 17.4.18 (Woodall, 1973 [931])**.** *Let G be the unit distance graph on $\mathbb{Q}^2$ (i.e., $V(G) = \mathbb{Q} \times \mathbb{Q}$, and $\{\mathbf{v}, \mathbf{w}\} \in E(G)$ if and only if $d(\mathbf{v}, \mathbf{w}) = 1$). Then $\chi(G) = 2$.*

Proof of Theorem 17.4.18: Define a binary relation on rational points by $(r_1, r_2) \sim (s_1, s_2)$ if and only if both $s_1 - r_1$ and $s_2 - r_2$ have odd denominators in their reduced forms. Since 0 in its reduced form is $\frac{0}{1}$, the relation $\sim$ is reflexive. Also, $\sim$ is trivially symmetric. The sum or difference of two rationals with odd denominators yields another rational with odd denominator (since the sum/difference put over a common denominator has only a product of odd numbers in the denominator even before simplification), and so $\sim$ is also transitive. Therefore, $\sim$ is an equivalence relation and thus $\sim$ partitions the rational plane into equivalence classes. Let C be the equivalence class containing $(0, 0)$; each equivalence class is a translate of C.

By Lemma 17.4.17, two points in different classes are not at distance 1, so to prove the theorem, it suffices to give a good 2-colouring of C. Since $(0, 0) \in C$, any other point in C has coordinates that are rationals with odd denominators in their reduced forms.

Define a colouring $f : C \to \{\text{red, blue}\}$ as follows: $f(\frac{w}{x}, \frac{y}{z}) = \text{red}$ if and only if w and y are either both odd or both even (and x, z are odd), and blue otherwise.

To see that f is indeed a good 2-colouring, some cases need checking. Let o and e denote odd and even. If two points are coloured red, the distance vector is one of the reduced forms:

$$\left(\frac{o}{o}, \frac{o}{o}\right) - \left(\frac{o}{o}, \frac{o}{o}\right) = \left(\frac{e}{o}, \frac{e}{o}\right);$$
$$\left(\frac{o}{o}, \frac{o}{o}\right) - \left(\frac{e}{o}, \frac{e}{o}\right) = \left(\frac{o}{o}, \frac{o}{o}\right);$$
$$\left(\frac{e}{o}, \frac{e}{o}\right) - \left(\frac{e}{o}, \frac{e}{o}\right) = \left(\frac{e}{o}, \frac{e}{o}\right).$$

If two points are coloured blue, there are essentially two cases:

$$\left(\frac{o}{o}, \frac{e}{o}\right) - \left(\frac{o}{o}, \frac{e}{o}\right) = \left(\frac{e}{o}, \frac{e}{o}\right);$$
$$\left(\frac{o}{o}, \frac{e}{o}\right) - \left(\frac{e}{o}, \frac{o}{o}\right) = \left(\frac{o}{o}, \frac{o}{o}\right).$$

In each of the above expressions, the right side $(\frac{a}{b}, \frac{c}{d})$ does not have exactly one of a and c odd, and so by Lemma 17.4.17, identically coloured points are not distance 1 apart. $\square$

Note: In the proof of Theorem 17.4.18, one might ask how many equivalence classes there are. Among only points of the form $(\frac{1}{2^n}, 0)$, no two are in the same class (since the denominator of the difference is a power of 2, not odd). So there are a countably infinite number of classes.

In 1982, Johnson [505] extended Woodall's result to show that $\chi(\mathbb{Q}^3) = 2$ and $\chi(\mathbb{Q}^4) > 2$. In 2000, Benda and Perles [73] showed that $\chi(\mathbb{Q}^4) = 4$. The bounds $\chi(\mathbb{Q}^5) \geq 7$, $\chi(\mathbb{Q}^6) \geq 10$, $\chi(\mathbb{Q}^7) \geq 13$, $\chi(\mathbb{Q}^8) \geq 13$ were given by Mann [632]; Cibulka [196] improved two of these results: $\chi(\mathbb{Q}^5) \geq 8$ and $\chi(\mathbb{Q}^7) \geq 15$.

17.4.10 The Hadwiger-Nelson problem over extensions of $\mathbb{Q}$

The Hadwiger–Nelson problem has also been looked at for fields other than $\mathbb{R}$ or $\mathbb{Q}$. Until 2009, it seems that not much was done; Soifer [803] posed the problem of finding $\chi(\mathbb{Q}(\sqrt{2})^2)$, (where $\mathbb{Q}(\sqrt{2}) \subseteq \mathbb{R}$ is the field extension of $\mathbb{Q}$ formed by adding $\sqrt{2}$).

For example, a recent (2015) paper by David Madore [630] shows that $\chi(\mathbb{Q}(\sqrt{2})^2) = 2$, $\chi(\mathbb{Q}(\sqrt{3})^2) = 3$, $\chi(\mathbb{Q}(\sqrt{7})^2) = 3$, and $4 \leq \chi(\mathbb{Q}(\sqrt{3}, \sqrt{11})^2) \leq 5$. (Thanks to Danylo Radchenko for this reference, who also found that $\chi(\mathbb{Q}(\sqrt{2})^2) = 2$ in 2012, but his results were never published.) Madore gives some references for results in other fields (e.g., p-adic numbers, finite fields), and points out that the only known bounds for the complex numbers are the trivial ones of $4 \leq \chi(\mathbb{C}^2) \leq \infty$. Unfortunately, it seems that Madore was unaware that in 1990, Fischer [334] proved that $\chi(\mathbb{Q}(\sqrt{3})^2) = 3$, Fischer also proved $\chi(\mathbb{Q}(\sqrt{11})^2) \leq 4$.

17.5 Colouring distance graphs on spheres

17.5.1 Notation

For each integer $n \geq 2$, and $\mathbf{x} \in \mathbb{E}^n$, let $S^{n-1}(\mathbf{x}) \subseteq \mathbb{E}^n$ denote the $(n-1)$-dimensional sphere in n-dimensional Euclidean space of radius 1 centered at $\mathbf{x}$. In other words,

$$S^{n-1}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{E}^n : |\mathbf{x} - \mathbf{y}| = 1\}.$$

Extending this definition, for any $r \in \mathbb{R}^+$, let

$$S_r^{n-1}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{E}^n : |\mathbf{x} - \mathbf{y}| = r\},$$

and so $S^{n-1}(\mathbf{x}) = S_1^{n-1}(\mathbf{x})$. Let S^{n-1} (or S_r^{n-1}) denote any sphere isomorphic to some $S^{n-1}(\mathbf{x})$ (or $S_r^{n-1}(\mathbf{x})$).

For $n \geq 2$ and $r \geq \frac{1}{2}$, the unit-distance graph on S_r^{n-1} is the graph $G = (V, E)$ where V is some copy of S^{n-1} and edges are pairs of vertices at distance 1. The distance 1 is not special; for any distance $d > 0$, one can define the "d-distance graph". For work on the chromatic number of the unit-distance graph for spheres, see [742].

17.5.2 Lyusternik–Schnirelman and Borsuk–Ulam theorems

In 1930, Russian mathematicians L. A. Lyusternik and L. G. Shnirelman proved a powerful theorem that relates topology to combinatorics. Three years later, Borsuk [131] proved a similar theorem (now seen to be equivalent) based on a problem of Ulam. It seems as if the Russians' theorem did not receive the most attention in the West, and so the form of the theorem that became popular is often still called the "Borsuk–Ulam" theorem. Different versions of the same theorem(s) are now popular (see [646] for more details).

Note: I found that Borsuk and Ulam also wrote a joint paper [132] in 1933, however I can not determine its content (something about ε-images); this paper is not referenced in [646] or in *MathSciNet*, so perhaps the topic was completely different.

Theorem 17.5.1 (Borsuk, Lyusternik–Schnirelman)**.** *Let S be any sphere in $\mathbb{R}^3$. If S is covered with three closed sets, then there exists a pair of antipodal points in one set.*

The more general result (also proved by, Borsuk [131] et al.) is that if any n closed sets cover a sphere $S^{n-1} \subseteq \mathbb{R}^n$, then one set contains a pair of antipodal points.

See [286] for a proof of the version in $\mathbb{R}^3$ (and many other colouring results), and [646] for a more thorough examination of this theorem, its history, extensions, and applications.

The next theorem extends the Borsuk–Ulam theorem (Theorem 17.5.1).

Theorem 17.5.2 (Hadwiger, 1944 [446])**.** *Let $S^2 \subseteq \mathbb{R}^3$ be any unit sphere (with arbitrary center), and suppose that S^2 is covered by 3 closed sets. Then* *at least one of these closed sets is so that for any distance d, $0 < d \leq \pi$, there is a pair of points at that (spherical) distance.*

When $d = \pi$, the Borsuk–Ulam theorem follows.

17.5.3 Diameters and colouring

For any $V \subseteq \mathbb{E}^n$, define the *diameter* of V to be

$$\mathrm{diam}(V) = \sup\{|\mathbf{x} - \mathbf{y}| : \mathbf{x}, \mathbf{y} \in V\}.$$

The "diameter graph" on V is the graph whose edges are

$$\{\{\mathbf{x}, \mathbf{y}\} : |\mathbf{x} - \mathbf{y}| = \mathrm{diam}(V)\}.$$

So if V is the set of points of a sphere, the edges of the diameter graph consists of pairs of antipodal points. The diameter graph of a sphere is then bipartite (its edges form a perfect matching).

Example 17.5.3 (Rödl, 1983 [757]). *Let $\alpha \leq \sqrt{2}$, and $m \geq 3$. Define $G_{m,\alpha}$ to be the α-distance graph on S^{m-1}. Then the subgraph induced by the $m-1$ vertices*

$$\left(\sqrt{1-\frac{\alpha^2}{2}}, \frac{\alpha}{\sqrt{2}}, 0, \ldots, 0\right), \left(\sqrt{1-\frac{\alpha^2}{2}}, 0, \frac{\alpha}{\sqrt{2}}, 0, \ldots, 0\right), \ldots, \left(\sqrt{1-\frac{\alpha^2}{2}}, 0, \ldots, 0, \frac{\alpha}{\sqrt{2}}\right)$$

is complete and so $\chi(G_{m,\alpha}) \geq m-1$.

Theorem 17.5.4 (Rödl, 1983 [757]). *Let $\alpha \in \mathbb{R}$, $\sqrt{2} < \alpha < 2$, and $n, k \in \mathbb{Z}^+$ satisfy $\alpha \leq 2\sqrt{1-\frac{k}{2(n+k)}}$. If G is the α-distance graph on $S^{2n+k-1} \subset \mathbb{E}^{2n+k}$, then $\chi(G) \geq k+2$.*

Remark 17.5.5. *In Rödl's paper [757], the notation S^m is the unit sphere in m dimensions, which is called S^{m-1} here.*

Theorem 17.5.4 together with Example 17.5.3 gives the following:

Corollary 17.5.6 (Rödl, 1983 [757]). *If $0 < \alpha < 2$, and G_n is the α-distance graph on the unit sphere in $\mathbb{E}^n$, then as $n \to \infty$, so also $\chi(G_n) \to \infty$.*

17.6 Borsuk's problem

In the 1930s, Karol Borsuk solved a problem for 3-dimensional balls, and asked a question about higher dimensional versions.

Theorem 17.6.1 (Borsuk, 1933 [131]). *For $n \geq 3$, an n-dimensional ball in $\mathbb{E}^n$ can be partitioned into $n+1$ (compact) solids, each of which has diameter smaller than the original ball.*

Borsuk also asked if such a phenomenon holds true for shapes other than the ball; due to many attempts to answer this question (with a belief that the answer was yes) Borsuk's problem became known as Borsuk's conjecture.

Problem 17.6.2 (Borsuk, 1933). *Can any bounded subset S of $\mathbb{R}^n$ be partitioned into $n+1$ sets, each of which has a smaller diameter than S?*

Borsuk apparently also showed that n sets are not enough (but I am unaware of his proof).

Borsuk's question was answered in the affirmative for special cases: for $n = 2$, I think that this was done by Borsuk [130] in 1932 (but I have not read the paper); for $n = 3$ by Perkal [718] in 1947 and independently by Eggleston [289] in 1955; for all n for smooth bodies by Hadwiger [448, 449] in 1946; for all n for centrally symmetric bodies by Riesling [751] in 1971; and for all n for bodies of revolution by Dekster [246] in 1995.

However, in general, the answer to Borsuk's problem is "no"'; for large enough dimension n, many more than $n+1$ sets are required.

Theorem 17.6.3 (Kahn–Kalai, 1993 [517]). *For sufficiently large n, if $\alpha(n)$ is the minimum number of sets of smaller diameter required to cover any bounded set, then $\alpha(n) \geq (1.2)^n$.*

In the Kahn–Kalai paper [517], it was mentioned that in 1988, Oded Schramm [780] showed that for any $\epsilon > 0$, for sufficiently large $n = n(\epsilon)$, $\alpha(n) \leq (\sqrt{3/2} + \epsilon)^n$. The method of proof by Kahn–Kalai was probabilistic, using Rödl's "nibble method".

Counterexamples were given in [517] for various values of n, including $1325 \leq n \leq 1560$, but some adjustments needed to be made (see [502] for details).

In 2014, Bondarenko [124] gave a counterexample to Borsuk's conjecture in 65 dimensions giving an example of a 2-distance set (using strongly regular graphs) consisting of 416 points in the unit sphere $S^{64} \subset \mathbb{R}^{65}$ that cannot be partitioned into 83 parts of smaller diameter. (Bondarenko supervised Andriy Prymak, now a professor at University of Manitoba. Thanks to Andriy for providing this reference.) Very soon after Bondarenko's result, Brouwer and Jenrich [154] (also see [501]) gave a minor improvement to 64 dimensions (with a set of 352 points that cannot be divided into fewer than 71 sets of smaller diameter) using strongly regular graphs. The set provided is a subset of the set found by Bondarenko.

For more on Euclidean representations of strongly regular graphs in relation to 2-distance sets, see [153, Ch. 8].

17.7 Voronoi diagrams and Delauney graphs

This section is meant to be only a brief introduction to two major concepts in computational geometry. The reader is invited to create diagrams to go along with the discussions below. For more on Voronoi diagrams and Delauney graphs, see see, e.g., O'Rourke's book [693] on computational geometry. or [248, Ch. 3.2], a treatise on triangulations.

Definition 17.7.1. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbb{R}^d$ be a set of n points. The Voronoi diagram for S is a partition of $\mathbb{R}^d$ into n convex d-polytopes (regions) $R_1, \ldots, R_n$ so that for each $i = 1, \ldots, n$, R_i contains all points of $\mathbb{R}^d$ that are closer to $\mathbf{v}_i$ than any other points of S.

Note that in Definition 17.7.1, the regions do not precisely partition $\mathbb{R}^d$ since points equidistant to two points of S are not included. It seems customary to alter slightly the definition so as to have the polytopes (regions) share common borders (faces).

Definition 17.7.2. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbb{R}^d$ be a set of n points, and let $\mathbb{R}^d$ be the associated Voronoi diagram. The Delauney graph has vertex set S and two vertices $\mathbf{v}_i, \mathbf{v}_j$ are adjacent if and only if R_i and R_j share a common face.

The Delauney graph for S is a topological dual of the Voronoi diagram for S.

Definition 17.7.3. A Delauney graph triangulation of a set $S \subset \mathbb{R}^2$ is a straight-line triangulation on S with all internal faces being triangles so that any three vertices form a face if and only if no other vertex of S appears in the circumscribed circle of these three points.

It is known that if four points occur on a circle, a Delauney triangulation is not unique. It is also known that a Delauney triangulation is a triangulation that maximizes the minimum internal angles.

Chapter 18

Euclidean Ramsey theory

18.1 Introduction and notation

For each positive integer d, the Euclidean space $\mathbb{E}^d$ is $\mathbb{R}^d$ with the usual Euclidean metric defined by

$$d((x_1, \ldots, x_d), (y_1, \ldots, y_d)) = \sqrt{(y_1 - x_1)^2 + \cdots + (y_d - x_d)^2}.$$

In this section, the usual notion of a graph $G = (V, E)$ is strengthened to that of a "geometric graph", that is, G is embedded in some Euclidean space, and distances are prescribed between vertices (also called points). Unless otherwise noted, a geometric graph has all of its edges drawn as straight line segments.

Much of the area called "Euclidean Ramsey theory" is concerned with colouring points of $\mathbb{E}^d$ and looking for a geometric graph whose vertices are coloured the same (monochromatic). Pairs of vertices can also be coloured, in which case one looks for a geometric graph whose edges are monochromatic, but the thrust of the research concentrates on finitely colouring $\mathbb{E}^d$ for $d = 2, 3$. Even then, many problems are still open.

Here are three warm-up exercises.

Exercise 345. *Let the points of $\mathbb{E}^2$ be coloured in two colours. Show that some rectangle has its four points in one colour.*

Exercise 346. *Let the points of the real line be coloured with two colours, say $c : \mathbb{R} \to \{red, blue\}$. Show that there are three points of the same colour so that one is the midpoint of the other two.*

Note that the result of Exercise 346 follows from a case of van der Waerden's theorem (which gives that under any 2-colouring of $\{1, 2, 3, ...9\}$, there exists a monochromatic 3-term arithmetic progression); see, e.g., [410] for details.

Exercise 347. *Is there a 3-colouring of the points in $\mathbb{E}^2$ so that all colours are used and every line receives exactly two colours?*

For $X, Y \subset \mathbb{E}^d$, an *isometry* from X to Y is a distance preserving bijection. The sets X and Y are said to be *congruent* if and only if there is an isometry from X to Y. Let $\text{cong}(X)$ denote the set of all congruent copies of X.

Isometries are also called *rigid motions* . In Euclidean 2-space, it is known that any rigid motion is a composition of translations, reflections, and rotations (in fact, translations are unnecessary since a translation can be obtained by a pair of reflections about parallel lines). Although many of the statements in this section guarantee a congruent copy of some configuration, many of the proofs allow restriction to only translates or to translates with rotation and no reflection.

Definition 18.1.1. For $r \in \mathbb{Z}^+$, say that a set $X \subset \mathbb{E}^d$ is r-Ramsey (in $\mathbb{E}^d$) if for every r-colouring

$$c : \mathbb{E}^d \to [r],$$

there is a monochromatic $Y \in \text{cong}(X)$, that is, a (congruent) copy of X whose vertices are coloured the same. In this case, write the "Ramsey arrow notation":

$$\mathbb{E}^d \to (X)^{\bullet}_r.$$

In Definition 18.1.1, the $^{\bullet}$ indicates vertex-colouring; since most of this section is about vertex colourings, when clear, the simpler notation $\mathbb{E}^d \to (X)_r$ is used.

Exercise 348. *Show that for every 2-colouring of points in the real plane, one of the colour classes contains points at every possible distance d, $0 < d < \infty$*

The result in Exercise 348 remains true even with three colours, in a result that extends Theorem 17.4.10:

Theorem 18.1.2 (Raiskii, 1970 [744])**.** *For $n \geq 2$, if the points of $\mathbb{E}^n$ are coloured with $n + 1$ colours, then there is one colour class so that pairs of points in that colour realize all possible distances.*

In 1973, Woodall [931] gave another proof of Theorem 18.1.2; in particular, Woodall's paper contains a simple proof of the case $n = 2$.

It seems that some sets X need to be considered in a much higher dimensional space before they become r-Ramsey. As a result of such considerations, there is another definition with a strong condition.

Definition 18.1.3. A set $X \subset \mathbb{E}^d$ is *Ramsey* if and only if for every $r \in \mathbb{Z}^+$ there exists a least integer $N_0 = N_0(X, r)$ so that for $N \geq N_0$, $\mathbb{E}^N \rightarrow (X)_r$.

18.2 Basic results

Many of the results given in this section are now folklore, but can be found in a sequence of three surveys written in the 1970s entitled "Euclidean Ramsey Theorems" I, II, III by Erdős, Graham, Montgomery, Rothschild, Spencer, and Straus [308], [309], [310] or the surveys by Graham [404, 405]. For fairly recent (2004, 2011) collections of open problems in Euclidean Ramsey theory, see [406] or [411].

Lemma 18.2.1 ([308])**.** *Let ℓ_2 denote the configuration of two points at distance 1. Then ℓ_2 is Ramsey.*

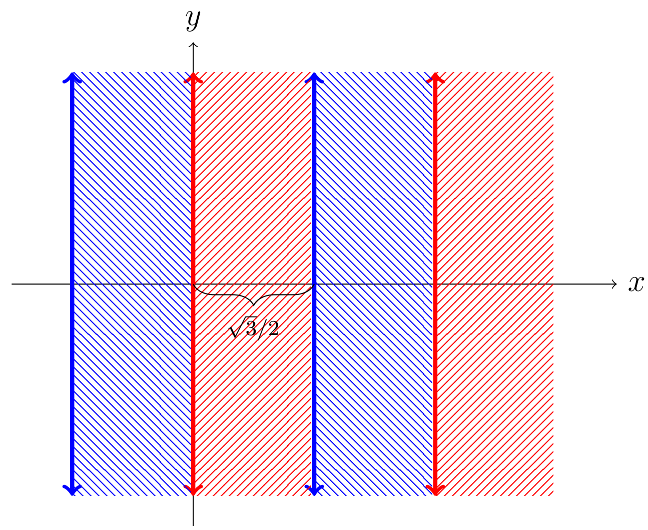
Proof: Let $r \geq 2$ and let S_{r+1} denote a unit equilateral simplex on $r + 1$ vertices in $\mathbb{E}^r$. Any two points in S_{r+1} are at distance 1. If S_{r+1} is partitioned into r parts, then by PHP, two vertices of S_{r+1} are in the same part. Thus $S_{r+1} \longrightarrow (\ell_2)^\bullet_r$, so also $\mathbb{E}^r \longrightarrow (\ell_2)^\bullet_r$. □

Theorem 18.2.2 ([308])**.** *Let $\triangle$ denote the equilateral triangle with side length 1. Then $\mathbb{E}^2 \not\longrightarrow (\triangle)^\bullet_2$,*

Proof: Define the (vertex) colouring $\Delta : \mathbb{E}^2 \rightarrow \{0, 1\}$ by

$$\Delta((x, y)) = \left\lceil \frac{2x}{\sqrt{3}} \right\rceil \pmod 2.$$

If 0 is red and 1 is blue, then the colouring looks like:



The plane is covered with half-open vertical strips, each with width equal to the height of the triangle. Since each $\triangle_1$ straddles two colour classes, the theorem is proved. □

It was also observed that in the half-open strip colouring given in the proof of Theorem 18.2.2, points on the boundaries can be given different colours (e.g., colour each vertical border by alternating half-open unit intervals).

By scaling, Theorem 18.2.2 holds when the side length 1 is replaced by any constant $d > 0$. Theorem 18.2.2 also shows that for any $r \geq 2$,

$$\mathbb{E}^2 \not\longrightarrow (\triangle_1)^{\bullet}_r.$$

However, the strip colouring in the proof of Theorem 18.2.2f only seems to forbid one size of equilateral triangle.

Conjecture 18.2.3 (Erdős et al., 1975, [310]). *If $\mathbb{E}^2$ is 2-coloured so that there is no monochromatic equilateral triangle with side length d, then for any $d' \neq d$, there exists a monochromatic equilateral triangle with side length d'.*

It was conjectured [310, p. 560] that the half-open strip-colouring is essentially (up to colours of points on the boundaries) the only such bad 2-colouring. In 2009, Jelínek, Kynčl, Stolař, and Valla [500] disproved this conjecture by producing another class of colourings that also forbid a unit distance monochromatic equilateral triangle; these colourings are called "zebra colourings", and are formed by wavy stripes with borders that are polygonal. In the same paper [500], the authors showed that the two known "stripe" colourings that forbid a monochromatic equilateral triangle (of a fixed size) are in a class of colourings that guarantee monochromatic triangles of all other types.

Theorem 18.2.4 (Jelínek et al., 2009 [500]). *If the plane is 2-coloured so that the borders of each colour class are made of piece-wise linear "curves", then there are monochromatic copies of all non-equilateral triangles.*

Theorem 18.2.5 (Jelínek et al., 2009 [500]). *Let $\mathbb{E}^2 = C_1 \cup C_2$ be any 2-colouring of the plane so that one C_i is closed (and so the other is open), then there is a monochromatic copy of every triangle.*

See [310] for theorems about what kinds of monochromatic 3 point configurations can be found in any 2-coloured plane; examples are isolated (some of these and more recent results are given below). It is conjectured that the only triangle for which there is no hope is the equilateral one:

Conjecture 18.2.6 (Erdős et al. [308, 310]). *If T is a triangle but not equilateral, then $\mathbb{E}^2 \longrightarrow (T)^{\bullet}_2$.*

Bialostocki and Nielsen [90] showed that for any non-equilateral triangle T, if there is a set $P(T)$ of points so that $P(T) \to (T)_2$, then $|P(T)| \geq 7$ and for the triangle T with internal angles $\frac{\pi}{7}, \frac{2\pi}{7}, \frac{4\pi}{7}$, then 7 points are sufficient.

Shader [788] proved that when T is a right triangle then $\mathbb{E}^2 \to (T)_2$ (this had been proven for certain subcases before, e.g., when the ratio of the legs is rational or the square root of a rational). Shader also gave results for a few other types of triangles.

For right triangles in 3-space and 3-colourings, Bóna [122] showed that if T has angles $\pi/2, \pi/3, \pi/6$, then $\mathbb{E}^3 \to (T)_3$. Bóna and (Géza) Tóth then extended this result to all right-angle triangles:

Theorem 18.2.7 (Bóna–Tóth, 1996 [123]). *For any right-angled triangle T, $\mathbb{E}^3 \to (T)_3$.*

Before answering whether or not $\mathbb{E}^3 \longrightarrow (\triangle^1)^{\bullet}_2$, an easy proof of $\mathbb{E}^4 \longrightarrow (\triangle^1)^{\bullet}_2$ is given.

Theorem 18.2.8 ([308]). *Using notation above, $\mathbb{E}^4 \longrightarrow (\triangle^1)^{\bullet}_2$.*

Proof: Let S be an equilateral simplex in $\mathbb{E}^4$. Since S has five points, under any two colouring of S, by the pigeonhole principle three vertices have the same colour; these three form a triangle in the simplex S. $\square$

Following the proof of Theorem 18.2.8 gives a more general result:

Theorem 18.2.9 ([308])**.** *For each $d \geq 4$, $r \geq 2$, if $d+1 \geq 2r+1$, then $\mathbb{E}^d \longrightarrow (\triangle_1)^{\bullet}_r$.*

Recall that a configuration C is called Ramsey if and only if for every number r of colours, there is a (least) dimension $n = n(r)$ so that $\mathbb{E}^n \to (C)_r$.

Corollary 18.2.10. *An equilateral triangle with unit distance is Ramsey.*

Since the proof above does not rely on distances being 1, an equilateral triangle of any size is also Ramsey.

Theorem 18.2.11 ([308])**.** $\mathbb{E}^3 \longrightarrow (\triangle_1)^{\bullet}_2$.

Proof: Let a red-blue colouring of $\mathbb{E}^3$ be given. There exists two points $\mathbf{v}$ and $\mathbf{w}$ at unit distance coloured the same, say red. If there exists another point at unit distance from both $\mathbf{v}$ and $\mathbf{w}$ that is also coloured red, then the desired monochromatic triangle exists, so suppose that all such points are coloured blue. These points form a blue circle C of radius $\sqrt{3}/2$.

For any two points on C that are also distance 1 apart, if any point with unit distance to both is also coloured blue, again a desired monochromatic circle exists. So assume that for every pair of points with distance 1 on C, all points that have distance one to each such pair are coloured red; such points form a degenerate (no hole) red torus. On this red torus, it is now not difficult to find three points forming an equilateral triangle. (One way to see this is to pick three points on the equator, which has radius $\frac{\sqrt{2}+\sqrt{3}}{2}$ and hence circumference greater than 3, with pairwise distance larger than 1, and then move these points closer to the center.) $\square$

The next result seems quite strong, but has a simple proof (omitted):

Theorem 18.2.12 (Erdős et al., 1975 [310, Thm. 24])**.** *For any 2-colouring of $\mathbb{E}^3$, there is one colour so that an equilateral triangle of every size occurs in that colour.*

By the early 1970s, many types of triangles were shown to be Ramsey. The last stubborn case was for obtuse triangles. This, too, was finally settled:

Theorem 18.2.13 (Frankl–Rödl, 1986 [344]). *All triangles are Ramsey. In other words, for any triangle T and positive integer r there exists an n so that $\mathbb{E}^n \longrightarrow (T)^{\bullet}_r$.*

For the next theorem about unit squares, a lemma from "graph-Ramsey theory" (where no geometry is assumed) is useful:

Lemma 18.2.14 (Chvátal–Harary, 1972 [195]). *Letting C_4 denote the cycle on four vertices, $K_6 \to (C_4)^2_2$; in other words, for any 2-colouring of the edges of K_6, a monochromatic C_4 is guaranteed.*

Proof: Let K_6 have vertices $\{u, v, w, x, y, z\}$, and let a colouring $c : E(K_6) \to \{\text{red}, \text{blue}\}$ be given. By Theorem 7.1.3, $R(3, 3) = 6$, and so there exists a monochromatic triangle T; suppose that T has all red edges and is on vertices u, v, w. If any of x, y, z has two red edges to T, then a red C_4 is formed. So assume that each of x, y, z has at most one red edge to T, that is, each of x, y, z send at least two blue edges to T. If, say, x and y have two blue edges to the same two vertices of T, then a blue C_4 is found; the same is true for the pair x, z and y, z.

So the final case to consider is when all edges of T are red, and each of x, y, z has a pair of blue edges to each of a different pair of vertices in $\{u, v, w\}$ (and a red edge to the remaining one). Without loss of generality, suppose that the edges ux, vy, and wz are red, and uy, uz, vx, vz, wx, wy are blue. If any of xy, yz, xz are red, a red C_4 is formed (e.g., if xy is red, then x, y, v, u is a red C_4). So assume that all edges between x, y, z are blue. In this case, for example, the vertices x, v, z, y is form a blue C_4. □

Theorem 18.2.15 ([308]). *Let S denote the unit square (on four vertices). Then $\mathbb{E}^6 \longrightarrow (S)_2$.*

Proof: Consider the set $X = \{\mathbf{x_1}, \mathbf{x}_2, \ldots, \mathbf{x}_{15}\}$ of points in $\mathbb{E}^6$, each of which have two entries $\frac{1}{\sqrt{2}}$ and the four remaining entries 0. Each of these fifteen points corresponds to an edge of K_6 on $v_1, \ldots, v_6$, where, if $\mathbf{x}_k$ has non-zero edges in positions i and j, then $\mathbf{x}_k$ corresponds to the edge $\{v_i, v_j\}$. Hence, any 2-colouring of $\mathbb{E}^6$ (and hence of X) gives a 2-colouring of the edges of K_6.

By Lemma 18.2.14, any 2-colouring of the edges of K_6 produces a monochromatic cycle on four vertices; these vertices correspond to four points in $\mathbb{E}^6$, all of the same colour, with consecutive points at unit distance. □

Exercise 349. *If S denotes the four points of a unit square, show that $\mathbb{E}^2 \not\to (S)_2$.*

Recall (from Definition 1.9.4) that a cyclic quadrilateral is one whose four vertices lie on a circle. A proof of the following conjecture would have once earned \$100 from Ron Graham (Ron Graham passed away 6 July 2020, and I don't know if someone else has made the same offer).

Conjecture 18.2.16 (see [411, Conj. 4])**.** *Cyclic quadrilaterals are Ramsey.*

Theorem 18.2.17 ([308])**.** *Let T be a triple of points in $\mathbb{E}^2$ in some specified configuration (with fixed distances, including the possibility of three in a row). Then*

$$\mathbb{E}^3 \longrightarrow (T)_2.$$

Proof: Suppose that the distances in T are a, b, c. By Theorem 18.2.11, there exists an equilateral triangle $\triangle ABC$ with side length a, with A, B, C the same colour; suppose that A, B, C are red. As in Figure 18.1, let D, E, and H be points in the plane of $\triangle ABC$ so that $\triangle CBD \cong \triangle ABE \cong \triangle ACH \cong T$, with $|DB| = |BE| = |CH| = b$. Add points F and G as in the diagram. (For the drawing in Figure 18.1, $a = 1$, $b = 1.2$, and the angle in T between sides with lengths a and b is $120°$.)

Suppose that no monochromatic copy of T exists. Then each of D, E, and H are blue. Then $\triangle DEG$ forces G to be red. However, there is no proper way to colour F, since by $\triangle CFG$ it is blue, but by $\triangle EFH$, it is red. So the assumption of no monochromatic copy of T leads to a contradiction. $\square$

So if ℓ_3 denotes three points in a row separated by two unit distances (in fact by any fixed distance), then

$$\mathbb{E}^3 \to (\ell_3)_2. \tag{18.1}$$

Recall that in Exercise 95, any 2-colouring of the points (including all points along the edges) of some equilateral triangle guarantees a right angle triangle with all vertices the same colour; however, in that case, the size or shape of the right angle triangle could not be specified in advance. By Theorem 18.2.17, if all the points in the 3-space are used, one can specify a size of the right triangle desired (but not its orientation). However, a lemma is required.

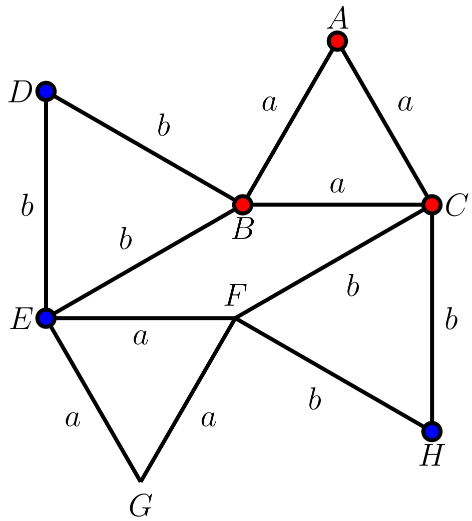


Figure 18.1: A step in proof of Theorem 18.2.17

Lemma 18.2.18 (see [308, Thm. 9])**.** *Let $d \in \mathbb{R}^+$. If T_1, T_2, and T_3 are triangles where T_1 contains a side of length d, T_2 contains a side with length $\sqrt{3}d$, and T_3 contains a side with length $2d$, then for any 2-colouring of $\mathbb{E}^2$, there exists a monochromatic copy of at least one of T_1, T_2, or T_3.*

Proof: Let $\mathbb{E}^2$ be 2-coloured. If one can find a monochromatic equilateral triangle with side length being one of $d, \sqrt{3}d, 2d$, then the proof of Theorem 18.2.17 applies to one of the T_is, yielding a monochromatic T_i.

Suppose that there is no monochromatic copy of any of the three desired equilateral triangles. Look at any equilateral triangle with side length d, and without loss of generality, suppose that the vertices are $O = (0, 0)$, $\mathbf{u} = (d, 0)$ and $\mathbf{v} = (\frac{1}{2}, \frac{3}{2})$. Then two of these vertices are of one colour; say O and $\mathbf{u}$ are red. Then $\mathbf{v}$ is blue, and (the point on the other side of $y = 0$) $\mathbf{u} - \mathbf{v}$ is also blue. (All points used here are on a triangular lattice formed by equilateral triangles of side length d.) Since $\mathbf{v}$ and $\mathbf{u} - \mathbf{v}$ are both blue, and the three points $\mathbf{v}$, $\mathbf{u}-\mathbf{v}$ and $2\mathbf{u}$ form an equilateral triangle with side length $\sqrt{3}d$, then $2\mathbf{u}$ is red. Since the points O, $2\mathbf{u}$ and $2\mathbf{v}$ form an equilateral triangle of side length $2d$, it follows that $2\mathbf{v}$ is blue. Now it is not possible to colour the point $\mathbf{u} + \mathbf{v}$, since $\mathbf{u}, 2\mathbf{u}, \mathbf{u} + \mathbf{v}$ form a triangle (with two red points already) and $\mathbf{v}, 2\mathbf{v}, \mathbf{u}+\mathbf{v}$ form a triangle (with two blue points already). This contradiction implies that the original supposition is false. $\square$

Theorem 18.2.19 (see [308]). *If T is any $30°$–$60°$–$90°$ triangle, then $\mathbb{E}^2 \to (T)_2$.*

Proof: Let T be such a right triangle with legs having length $d, \sqrt{3}d$ (and hypotenuse $2d$). Put $T_1 = T_2 = T_3$ all equal to T, and apply Lemma 18.2.18. □

There are other triangles for which there is a Ramsey theorem in the plane, but in general, the question seems to be open.

Theorem 18.2.20 (Erdős et al., 1973 [308]). *Let L be the configuration of four points in shape of an L, where adjacent points are separated by some fixed distance $c > 0$. (For example, the coordinates of the four points can be taken as $(0,0), (c,0), (2c,0), (2c,c)$.) Then $\mathbb{E}^3 \to (L)_2$.*

Proof: See [308].

Theorem 18.2.21 ([308]). *Let ℓ_k denote the configuration of k points in a line with unit spaces between points. For each $n \geq 1$,*

(i) $\mathbb{E}^n \not\longrightarrow (\ell_3)_4$,

(ii) $\mathbb{E}^n \not\longrightarrow (\ell_4)_3$, *and*

(iii) $\mathbb{E}^n \not\longrightarrow (\ell_6)_2$.

Proof: Fix n; in each of the three cases, a "bad" colouring c of $\mathbb{E}^n$ is given. Only the first case is proved here; see the original paper for the other two cases.

(i) For each $\mathbf{z} \in \mathbb{E}^n$, define $c(\mathbf{z}) = \lfloor \|\mathbf{z}\|^2 \rfloor \pmod 4$. Suppose that some ℓ_3 configuration is monochromatic; that is, for some $\mathbf{x} \in \mathbb{E}^n$, $\mathbf{u} \in \mathbb{E}^n$ with $\|\mathbf{u}\| = 1$, and $r \in \{0,1,2,3\}$, suppose that

$$c(\mathbf{x} - \mathbf{u}) = c(\mathbf{x}) = c(\mathbf{x} + \mathbf{u}) = r.$$

Then there exists $a, b, c \in \mathbb{Z}$ and real numbers $\alpha, \beta, \gamma \in [0,1)$ so that

$$\|\mathbf{x} - \mathbf{u}\|^2 = 4a + r + \alpha, \tag{18.2}$$
$$\|\mathbf{x}\|^2 = 4b + r + \beta, \tag{18.3}$$
$$\|\mathbf{x} + \mathbf{u}\|^2 = 4c + r + \gamma. \tag{18.4}$$

Subtracting (18.2) from (18.3) yields (using $\|\mathbf{w}\|^2 = \mathbf{w} \bullet \mathbf{w}$ and $\|\mathbf{u}\| = 1$)

$$2\mathbf{x} \bullet \mathbf{u} - 1 = 4(b - a) + \beta - \alpha. \tag{18.5}$$

Similarly, equations (18.3) and (18.4) imply

$$2\mathbf{x} \bullet \mathbf{u} + 1 = 4(c - b) + \gamma - \beta. \tag{18.6}$$

Finally, (18.6)-(18.5) gives

$$2 = 4(c - 2b + a) + \gamma - 2\beta + \alpha,$$

which is impossible since $-2 < \gamma - 2\beta + \alpha < 2$, and so the last term is neither 2 nor -2. □

Again, by scaling, Theorem 18.2.21 implies the similar results for lines with points spaced by any fixed distance.

The following extends Theorem 18.2.13. Recall that a trapezoid is convex quadrilateral with two parallel sides; however, in this next theorem, the common definition is used that a trapezoid is a quadrilateral with only one pair of parallel sides.

Theorem 18.2.22 (Kříž, 1992 [577]). *All trapezoids (that are not parallelograms) are Ramsey,*

The next theorem yields many of the configurations that are known to be Ramsey.

Theorem 18.2.23 (Erdős et al., 1973 [308]). *If A and B are Ramsey, then so is $A \times B$.*

A *brick* is a set of vertices of a rectangular parallelepiped. A brick is a product of simplices of the form $\{0, a_i\}$, so applying Theorem 18.2.23 repeatedly gives:

Corollary 18.2.24 (Erdős et al., 1973 [308]). *All bricks are Ramsey.*

Theorem 18.2.25 (Erdős et al., 1973 [308]). *Let $d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34}$ be six distances satisfying $d_{ij}^2 + d_{jk}^2 \geq d_{ik}^2$. Then there is a 6-dimensional brick so that 4 of its points realize these distances.*

Proof idea: First create a 7-dimensional brick satisfying some distances. Then use six equations, one for each tetrahedron determined by four 7-dimensional vertices. Then eliminate one coordinate. □

Exercise 350. *Use Corollary 18.2.24 to show that any acute triangle is Ramsey.*

Exercise 351. *Prove that non-rectangular parallelepipeds are not Ramsey.*

Theorem 18.2.26 ([403]). *Let B be a brick. Then for each r, for sufficiently large dimension n, and sufficiently large radius ρ, for any n-dimensional sphere S with radius ρ, $S \longrightarrow (B)^{\bullet}_{r}$.*

In [308], the authors observe that some simplices can be embedded in a brick, and hence are Ramsey. Implicitly, they ask which simplices are Ramsey. Frankl and Rödl settled this problem.

Theorem 18.2.27 (Frankl–Rödl, 1990 [345]). *For any set A of $d+1$ points in $\mathbb{R}^d$ in general position, there exists $\epsilon = \epsilon(A) > 0$ so that for sufficiently large n, every partition of $\mathbb{R}^n$ into at most $(1+\epsilon)^n$ parts, one of the parts contains a subset congruent to A.*

Since a *simplex* is a set of affinely independent points, Theorem 18.2.27 has the following consequence.

Theorem 18.2.28 ([345]). *All finite simplices are Ramsey.*

Exercise 352. *Show how Theorem 18.2.27 implies Theorem 18.2.28.*

In 2004, Frankl and Rödl [346] gave a strengthening of Theorem 18.2.27: if X is a simplex in $\mathbb{E}^d$ with circumradius ρ, then for every $\delta > 0$, and every $n \geq d$, there exists a positive real σ so that for $r \leq (1+\sigma)^n$, if an n-dimensional sphere of radius $\rho+\delta$ is r-coloured, there exists a monochromatic copy of X.

Say that a set $K = \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ of points in $\mathbb{E}^n$ is *spherical* if and only if K is embeddable in the surface of a sphere in $\mathbb{E}^n$. To show that if a set is not Ramsey then it is not spherical, two lemmas are needed (for proofs, see [308]).

Lemma 18.2.29 (Erdős et al., 1973 [308]). *A set $K = \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ is spherical if and only if there exist scalars $c_1, \ldots, c_k$, not all zero, so that*

$$\sum_{i=1}^{k} c_i(\mathbf{v}_i - \mathbf{v}_0) = \mathbf{0},$$

and

$$\sum_{i=1}^{k} c_i(|\mathbf{v}_i|^2 - |\mathbf{v}_0|^2) \neq 0.$$

The next lemma follows from an extension of Rado's theorem (see [410]) to the reals. The authors provide the proof, which is non-trivial.

Lemma 18.2.30 (Erdős et al., 1973 [308]). *Let $c_1, \ldots, c_k$ be real numbers and $b \neq 0$. Then there exists $r \in \mathbb{Z}^+$ and an r-colouring of the reals so that the equation*

$$\sum_{i=1}^{k} c_i(x_i - x_0) = b$$

has no monochromatic solution $x_0, x_1, \ldots, x_k$.

Theorem 18.2.31 (Erdős et al., 1973 [308]). *If a finite configuration of points K is Ramsey then K is spherical.*

For example, any three collinear points are not spherical, and hence not Ramsey. Graham [404] offered \$1000 for a proof of the following:

Conjecture 18.2.32 (Graham, 1994 [404]). *All spherical sets are Ramsey.*

Observe that since simplices are Ramsey, and Ramsey sets are spherical, it follows that simplices are spherical.

Exercise 353. *Sketch a proof not using Ramsey theory that shows simplices are spherical.*

The notion of a configuration being Ramsey is extended to allow more than one colour. For $\ell \in \mathbb{Z}^+$ say that a set K is ℓ-Ramsey if and only if for any $r \in \mathbb{Z}^+$, there exists n so that if points of $\mathbb{E}^n$ are r-coloured, then there exists a congruent copy of K whose vertices are ℓ-coloured.

Theorem 18.2.33 (Erdős et al., 1973 [308]). *If K cannot be embedded into $\ell - 1$ concentric spheres, then for each $m < \ell$, K is not m-Ramsey.*

Theorem 18.2.34 (Erdős et al., 1973 [308]). *For $i = 1, \ldots, t$, if X_i is a finite set that is r_i-Ramsey, then $X_1 \times X_2 \times \cdots \times X_t$ is $r_1 r_2 \cdots r_t$-Ramsey.*

Igor Kříž gave two powerful results:

Theorem 18.2.35 (Kříž, 1991 [576]). *If $X \subset \mathbb{E}^n$ has a transitive solvable group of isometries, then X is Ramsey.*

As a consequence, the set of vertices of a regular n-gon is Ramsey.

Theorem 18.2.36 (Kříž, 1991 [576]). *If $X \subset \mathbb{E}^n$ has a transitive group of isometries that has a solvable subgroup with at most 2 orbits, then X is Ramsey.*

As a result, the set of vertices of any Platonic solid is Ramsey. Cantwell [175] showed that in fact, all regular polytopes are Ramsey, including the 4-dimensional 120-cell.

Leader, Russell, and Walters [595] proposed that instead of proving that all spherical sets are Ramsey (as in Conjecture 18.2.32) they conjecture that Ramsey sets are precisely the transitive sets (its symmetry group acts transitively).

The result in the next theorem by Spencer might seem natural, but I found it a bit surprising in that it gives a very general "approximate" result. To state this theorem, a definition is useful.

Definition 18.2.37 (Spencer, 1979 [810]). Call a configuration C (in some $\mathbb{E}^m$) *almost Ramsey* if for any $k \in \mathbb{Z}^+$ and any $\epsilon > 0$, there exists n so that for any k-coloring of $\mathbb{E}^n$, there exists a monochromatic configuration C' that can be made into a congruent copy of C by moving each point of C' a distance at most ϵ.

Theorem 18.2.38 (Spencer, 1979 [810]). *Every finite configuration is almost Ramsey.*

In fact, Spencer proved his result by a density argument—any sufficiently large colour class contains the desired "almost-configuration". Spencer's proof uses some technical result about measures of sets on spheres.

18.3 Euclidean Ramsey theorems for cubes

I learned of the following result from Rödl, after he learned of it in a telephone conversation with Hal Kierstead in late 1992 or early 1993. It was so interesting to me that I typed up the proof on 20 January 1993. Kent Cantwell is responsible for this proof; perhaps it appeared in his thesis from Arizona State, but I have not found another source. The following appears as the first non-trivial case mentioned in the paper by Graham and Rothschild where their famous theorem on parameter sets is proved.

Theorem 18.3.1 (Graham–Rothschild, 1971 [409, p. 290])**.** *There exists an n so that if V is the set of 2^n vertices of the n-dimensional hypercube Q_n, and G is the complete graph on V, then for any 2-colouring of the edges of G, there exists a monochromatic complete graph on four co-planar vertices.*

The n from Theorem 18.3.1 is called *Graham's number*, the number that held the record for the largest number ever used in a meaningful proof. In the same paper, it is mentioned that the lower bound for n is 6. (This was later improved by Exoo to 8.) However, after Gowers proved a new bound on the Hales–Jewett function, the upper bound that follows is now much smaller.

The more general theorem of Graham and Rothschild gave a Ramsey result for parameter sets, and was extended (with Leeb) to answering a conjecture by Rota (27 April 1932–18 April 1999). More on this result can be found in [410].

The following idea does not use the complete graph on vertices of Q_n, but only cubes and sub-cubes.

Let Q_q denote the unit hypercube of dimension q, represented by vertices of the form $(0, 1, 1, \ldots, 0)$, where the vector has q positions, or *coordinates*, with entries, or *components* 0's and 1's. Use *q-cube* as shorthand for *q-dimensional cube* (it need not be a unit cube, i.e., have side length 1, nor need it be centered at the origin) and an m-subcube of a cube is called an *m-face*.

The vertices of an m-face of Q_q are given by fixing components in $q-m$ coordinates, and in the remaining coordinates, taking all possibilities of 0's and 1's. Call coordinates *fixed* and *moving* accordingly. In Q_q, for each collection of m moving coordinates, there are 2^{q-m} m-cubes, and so in all, there are $\binom{q}{m}2^{q-m}$ m-faces of Q_q. An m-face is completely determined by which coordinates are moving (or fixed).

The proof of the main theorem begins with a lemma.

Lemma 18.3.2 (Cantwell, 1992)**.** *For all positive integers r, m, and q, there exists s so that for every coloring $\Delta : \binom{\mathbb{R}^s}{Q_m} \longrightarrow [r]$, there is $Q_q^* \in \binom{\mathbb{R}^s}{Q_q}$ so that the colour of any m-face of Q_q^* depends only on the moving coordinates for that face, i.e., for any Q_m^*, $Q_m^{**} \in \binom{Q_q^*}{Q_m}$, $\Delta(Q_m^*) = \Delta(Q_m^{**})$ if the moving coordinates that yield Q_m^*, and Q_m^{**} agree.*

Proof: Rather than working with unit cubes, work with cubes of side length $\sqrt{2}$, and then rescale to obtain result. This is merely a matter of convenience so that pairs of coordinates from the unit cube case are used. For example, in $\mathbb{R}^4$ the points $(1,0,0,0)$ and $(0,1,0,0)$ are $\sqrt{2}$ apart. In $\mathbb{R}^6$, the points $(0,1,0,1,0,0)$, $(0,1,1,0,0,0)$, $(1,0,0,1,0,0)$ and $(1,0,1,0,0,0)$ determine a square, where the first two pairs of coordinates are moving coordinate-pairs, and the last pair is fixed. In $\mathbb{R}^s$ (in fact, in Q^s), look at m-cubes and q-cubes of side length $\sqrt{2}$ determined by such moving coordinate pairs, each pair being either $(0,1)$ or $(1,0)$. For example, an m-cube of interest is determined by $2m$ moving coordinates, $a_1 < b_1 < a_2 < b_2, < \ldots, < a_m < b_m$. All points $\mathbf{p} = (p_1, p_2, \ldots, p_s)$ in such an m-cube satisfy $p_{a_i} + p_{b_i} = 1$ and the remaining $s - 2m$ coordinates are fixed. There are $\binom{s}{2m} 2^{s-2m}$ such m-cubes in Q_s; such an m-cube is denoted by $Q_{(2)m}$.

By Ramsey's Theorem, pick s so large that

$$s \to (2q)^{m+q}_{r^{\binom{m+q}{2m}}}$$

and let $\Delta : \binom{\mathbb{R}^s}{Q_{(2)m}} \longrightarrow [r]$ be given. Letting $S = \{1, 2, \ldots, s\}$, the colouring Δ can be viewed as an r-colouring Δ^* of $[S]^{2m}$ given by $\Delta*(\{i_1, i_2, \ldots, i_{2m}\}) = \Delta(Q^*_{(2)m})$ if the moving coordinates (pairs) of $Q^*_{(2)m}$ are $\{i_1, i_2, \ldots, i_{2m}\}$.

In what follows is a colouring α of each $(m+q)$-subset of S with a function (or pattern of colours), that is,

$$\alpha : [S]^{m+q} \longrightarrow [r^{\binom{m+q}{2m}}].$$

Let $T \in [S]^{m+q}$ and list $[T]^{2m} = \{U_1, U_2, \ldots, U_{\binom{m+q}{2m}}\}$, and for each $i \in \{1, 2, \ldots, \binom{m+q}{2m}\}$, let M_i be the m-cube having m moving coordinate-pairs (occurring in positions $2j-1$ and $2j$) in U_i, and, outside of U_i, the j-th component p_j of points $\mathbf{p} = (p_1, p_2, \ldots, p_s)$ is given by

$$p_j = \begin{cases} 0 & \text{if } j \in S \backslash T, \\ 1 & \text{if } j \in T \backslash U_i. \end{cases}$$

Define $\alpha(T) = (\Delta(M_1), \Delta(M_2), \ldots, \Delta(M_{\binom{m+q}{2m}}))$.

By the choice of s, pick $V \in [S]^{2q}$ monochromatic with respect to α. Dividing elements (coordinates) of V into consecutive pairs gives a q-cube, call it Q^*. Since V is monochromatic with respect to α, it remains to show that this cube is one desired in the statement of the lemma.

For every pair $T_1, T_2 \in [V]^{m+q}$, the $2m$-subsets of T_1 and T_2 have the same colour pattern, that is, if subsets $U_1, U_2 \in [V]^{2m}$ have the same position in both T_1 and T_2, then they are coloured the same by Δ.

Let A and B be two m-cubes determined by the same moving pairs, say on coordinates $U \in [V]^{2m}$, in Q^*, although having different components in the fixed coordinate pairs. It suffices to find $T(A)$, $T(B) \in [V]^{m+q}$, each containing U, so that U is in the same position in both $T(A)$ and $T(B)$. Merely pick the $T(A)$ to be U together with those fixed coordinates having 1's in points from A, and similarly for B. Each coordinate pair has exactly one 1 as a component, so a different fixed component still leaves the moving coordinates in the same relative position. $\square$

Example: $q = 6$, $m = 3$, $|V| = 12$. Let xy represent a moving pair, and let A and B be given by:

A: xy 10 xy xy 01 01

B: xy 01 xy xy 01 10

Then $U = \{1, 2, 5, 6, 7, 8\}$,
$T(A) = \{1, 2,\ 3,\ 5, 6,\ 7, 8,\ 10,\ 12\}$, and
$T(B) = \{1, 2,\ 4,\ 5, 6,\ 7, 8,\ 10,\ 11\}$.

Theorem 18.3.3 (Cantwell, 1992)**.** *For all positive integers r, m, and $n > m$, there exists s so that for every colouring $\Delta : \binom{\mathbb{R}^s}{Q_m} \longrightarrow [r]$, there is a $Q_n^* \in \binom{\mathbb{R}^s}{Q_n}$ so that $\binom{Q_n^*}{Q_m}$ is monochromatic with respect to Δ. If $q \longrightarrow (n)_r^m$, then there are 2^{q-n} such Q_n^*s.*

Proof: Fix r, $m < n$ and by Ramsey's Theorem, pick q so large that $q \longrightarrow (n)_r^m$. By Lemma 18.3.2, fix s so large that for every pair Q_m^*, $Q_m^{**} \in \binom{Q_q^*}{Q_m}$, $\Delta(Q_m^*) = \Delta(Q_m^{**})$ if the moving coordinates which yield Q_m^*, and Q_m^{**} agree. Fix a colouring $\Delta : \binom{\mathbb{R}^s}{Q_m} \longrightarrow [r]$ and fix such a Q_q^*.

Without loss of generality, let Q_q^* be determined by moving coordinates $1, \ldots, q$ (with coordinates $q+1, q+2, \ldots, s$ fixed). Since the colour of each copy of Q_m in Q_q^* depends only on the collection of moving coordinates determining that copy, Δ induces a colouring

$$\Delta^* : [q]^m \longrightarrow [r].$$

By the choice of q, there exists n coordinates $I \in [q]^n$ so that $[I]^m$ is monochromatic with respect to Δ^*. Fixing components in coordinates outside I (which can be done in 2^{q-n} ways), and letting coordinates indexed by I be moving coordinates, obtain a

$$Q_n^* \in \binom{Q_q^*}{Q_n} \subset \binom{\mathbb{R}^s}{Q_n}.$$

Since $[I]^m$ is monochromatic with respect to Δ^*, each $Q_m^* \in \binom{Q_n^*}{Q_m}$ is coloured the same under Δ, for each $Q_m^* \in \binom{Q_n^*}{Q_m}$ is determined by some $J \in [I]^m$, indexing the moving coordinates of Q_m^*, while other coordinates are fixed. $\square$

18.4 Sphere-Ramsey

Let $S_\rho^d(\mathbf{x})$ denote the d-dimensional sphere of radius ρ centered at $\mathbf{x}$ (not including the interior). Say that a set $X \subseteq \mathbb{E}^m$ is *sphere-Ramsey* if and only if for every r there is N and ρ so that for any $S = S_\rho^N(\mathbf{x})$, if S is r-coloured, some colour class contains a congruent copy of X. Note that if a set is sphere-Ramsey, it is also Ramsey, and so also spherical.

Theorem 18.4.1 (Matoušek–Rödl, 1995 [647]). *If $X \subseteq S_1^d(\mathbf{x})$ is a simplex then for all r and all $\epsilon > 0$, $S_{1+\epsilon}^N(\mathbf{x}) \longrightarrow (X)_r^\bullet$. The condition $\epsilon > 0$ is necessary (for almost all configurations).*

For more on sphere-Ramsey problems, see [403].

18.5 Asymmetric Euclidean Ramsey

Just as in the off-diagonal cases of Ramsey numbers, for structures A and B and some dimension d, write $\mathbb{E}^d \longrightarrow (A, B)^\bullet$ to mean that under any red-blue colouring of $\mathbb{E}^d$, either there exists a copy of A all of whose points are red, or there exists a copy of B all of whose points are blue. As usual, if it is clear that points are being coloured, the simpler notation $\mathbb{E}^d \longrightarrow (A, B)$ is used.

Theorem 18.5.1 (Erdős et al., 1975 [309]). $\mathbb{E}^3 \longrightarrow (\ell_2, \ell_4)$.

Proof outline: Let a colouring of $\mathbb{E}^3$ be given. By Theorem 18.2.17, there exists a monochromatic ℓ_3. If these points are red, then a red ℓ_2 is formed, so assume that these three points are blue. Construct a lattice using equilateral unit triangles based on the three blue points of the ℓ_3. Now look at the colours of nearby lattice points. $\square$

With a little more work, Theorem 18.5.1 is improved to the plane.

Theorem 18.5.2 (Erdős et al., 1975 [309]). $\mathbb{E}^2 \longrightarrow (\ell_2, \ell_4)$.

Proof outline: Suppose that $\mathbb{E}^2$ is red-blue coloured and that no red ℓ_2 exists and no blue ℓ_4 exists. Since no blue ℓ_4 exists, there exists a point P that is red.

Consider two circles centered at P, C_1 with radius 1 and C_2 with radius $\sqrt{3}$. Since P is red, all points of C_1 are blue. Inscribe an equilateral triangle $T = \triangle ABC$ on C_2, and let D and E be the intersection points of C_1 and AB. Since both D and E are blue, not both of A and B are blue. Similarly, not both of A and C are blue, and not both B and C are blue. So assume that A and B are red. Let F and G be points on C_2 so that F is unit distance from A in a clockwise direction and G is unit distance from B in a clockwise direction. Then F and G are blue (since they are at unit distance from the red points A and B), and the two points H, I on C_1 intersecting FG are also blue (because they are on C_1). This gives a contradiction since (with a little geometry) FH, HI, and IG are all unit distances. $\square$

Theorem 18.5.3 (Erdős et al., 1975 [309]). *If T denotes vertices of a $(1, 1, \sqrt{2})$ triangle and S denotes the vertices of a unit square, then $\mathbb{E}^3 \longrightarrow (T, S)$.*

Theorem 18.5.4 (Erdős et al., 1975 [309]). *Let T be any triple of points in $\mathbb{E}^2$, and let $d \in \mathbb{R}$. Then for any red-blue colouring of the points of $\mathbb{E}^2$, there exists either two red points at distance 1 or there exists a blue translate of T.*

Proof: Let $T = \{\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3\}$ and let K be the set of seven vertices of a Moser graph (see Figure 17.10), where the bottom left vertex is at $(0, 0)$ and the bottom right is at $(1, 0)$. Let a red-blue colouring of $\mathbb{E}^2$ be given. If there are no pair of red points at unit distance, then each $\mathbf{t}_i + K$ has at most two red points (the top and top right). This means that there is one point of the Moser graph, say $\mathbf{k} \in K$, so that in each of the three $\mathbf{t}_i + K$, the respective points are blue, giving the blue $T + \mathbf{k}$. □

The following two questions were posed as open problems in [309]. For the first, compare Theorem 18.5.3:

Question 18.5.5 ([309])**.** *Let S denote the points of the unit square. Is it true that $\mathbb{E}^2 \longrightarrow (\ell_2, S)$?*

Question 18.5.5 was answered in a strong fashion:

Theorem 18.5.6 (Juhász, 1979 [511])**.** *Let S be any configuration of four points in the plane, and let ℓ_2 denote two points at unit distance. Then $\mathbb{E}^2 \longrightarrow (\ell_2, S)$.*

Juhász also showed that in Theorem 18.5.6 if instead $|S| \geq 12$, then the result fails. On the other hand, the number 12 was brought down to 8 by György Csizmadia and Géza Tóth.

Theorem 18.5.7 (Csizmadia–Tóth, 1994 [232])**.** *Let K denote the configuration of eight vertices formed by a regular heptagon with circumradius 0.9 together with its center. Then $\mathbb{E}^2 \not\longrightarrow (\ell_2, K)$.*

Evidence is given in [232] for the possibility that the 8 can be brought down even further, maybe even down to 5. Recall that Theorem 18.5.2 says $\mathbb{E}^2 \longrightarrow (\ell_2, \ell_4)$.

Question 18.5.8 (Erdős et al., 1975 [309])**.** *Is it true that $\mathbb{E}^2 \longrightarrow (\ell_2, \ell_5)$? How about $\mathbb{E}^3 \longrightarrow (\ell_2, \ell_5)$?*

It is known (see [309, p. 535]) that $\mathbb{E}^4 \longrightarrow (\ell_2, \ell_5)$ by using a 4-dimensional analogue of the Moser graph idea (see Theorem 17.4.12 and Figure 17.10).

According to Juhász, her student answered (in a Master's thesis, in Hungarian, unpublished) the second part of Question 18.5.8 with a stronger result:

Theorem 18.5.9 (Iván, 1979 [498])**.** *For any five point set $S \subseteq \mathbb{E}^3$,*

$$\mathbb{E}^3 \longrightarrow (\ell_2, S).$$

In 2017, an extension of the second part of Question 18.5.8 was also given:

Theorem 18.5.10 (Arman–Tsaturian, 2017 [34]). $\mathbb{E}^3 \longrightarrow (\ell_2, \ell_6)$.

Two months later, the first part of Question 18.5.8 was also answered with a fairly elementary proof:

Theorem 18.5.11 (Tsaturian, 2017 [880]). $\mathbb{E}^2 \longrightarrow (\ell_2, \ell_5)$.

Soon after Theorems 18.5.10 and 18.5.11 were posted, Conlon and Fox proved the following:

Theorem 18.5.12 (Conlon–Fox, 2017 [204]). *For each $k \geq 2$, there exists a maximum number $m(k)$ so that $\mathbb{E}^k \longrightarrow (\ell_2, \ell_{m(k)})$. Furthermore,*

$$(1 + o(1))(1.2)^k < m(k) < 10^{5k}.$$

In 1979, Erdős and Graham [307] reported that, "more or less", $m(2) \leq 10,000,000$.

The lower bound in Theorem 18.5.12 was proved using a construction based on a result of Frankl and Wilson [347] regarding the density of sets containing two points at distance 1.

For $4 \leq k \leq 10$, the lower bound in Theorem 18.5.12 was improved by the following general result:

Theorem 18.5.13 (Arman–Tsaturian, 2017 [35]). *For $k \geq 4$,*

$$\mathbb{E}^k \longrightarrow (\ell_2, \ell_{k+3}).$$

Theorems 18.5.12 and 18.5.13 imply that for each $k \geq 2$, there exists n so that $\mathbb{E}^n \longrightarrow (\ell_2, \ell_k)$. In [308], it is noted that for all n, $\mathbb{E}^n \not\longrightarrow (\ell_6, \ell_6)$. What is the minimal s such that there exists k so that for all n, $\mathbb{E}^n \not\longrightarrow (\ell_s, \ell_k)$? It was conjectured that the minimal such s is 3:

Conjecture 18.5.14 (Arman–Tsaturian [35], Conlon–Fox [204]). *There exists $k \in \mathbb{Z}^+$ so that for every n,*

$$\mathbb{E}^n \not\longrightarrow (\ell_3, \ell_k).$$

Only after Theorem 18.5.13 was proved, a theorem with a simple proof was found that implies some of the above results. (Some of these and related ideas are recorded in [33].)

Theorem 18.5.15 (Szlam, 2001 [854]). *Let $d \geq 2$. If $T \subseteq \mathbb{E}^d$ has $|T| = \chi(\mathbb{E}^d) - 1$ points, then for any red-blue colouring of the points in $\mathbb{E}^d$, there either exists a pair of red points at unit distance or there exists a blue translate of T.*

Proof: Let $d \geq 2$, put $k = \chi(\mathbb{E}^d)$, and let $T = \{\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}\} \subseteq \mathbb{E}^d$. Let the points of $\mathbb{E}^d$ be red-blue coloured.

Suppose that there is no translate of T that is entirely blue; in other words, for each $\mathbf{v} \in \mathbb{E}^{\mathbf{d}}$, the translate $T_{\mathbf{v}} := \mathbf{v}+T = \{\mathbf{v}+\mathbf{x}_i : i = 1, \ldots, k-1\}$ contains at least one red point. Define a $(k-1)$-colouring $\Delta : \mathbb{E}^d \to [k-1]$ as follows: for each $\mathbf{v} \in \mathbb{E}^d$, define $\Delta(\mathbf{v})$ to be the least $i \in [k-1]$ so that translated point $\mathbf{v} + \mathbf{x}_i \in T_{\mathbf{v}}$ is red. Since $k - 1 < \chi(\mathbb{E}^d)$, there exists a $j \in [k-1]$ and two points $\mathbf{v}, \mathbf{w}$ at unit distance with colour j. Then $\mathbf{v} + \mathbf{x}_j$ and $\mathbf{w} + \mathbf{x}_j$, which are also at unit distance, form a red ℓ_2. $\square$

Abusing notation slightly, if T_i denotes an arbitrary configuration of i points in $\mathbb{E}^d$, Theorem 18.5.15 says (in part),

$$\mathbb{E}^d \longrightarrow (\ell_2, T_{\chi(\mathbb{E}^d)-1}).$$

Since $\chi(E^2) \geq 4$ (see Figure 17.13 for the table of bounds on chromatic numbers of $\mathbb{E}^d$), Theorem 18.5.15 says

$$\mathbb{E}^2 \to (\ell_2, \ell_3)$$

and letting Δ^2 denote any 3-point simplex (triangle),

$$\mathbb{E}^2 \to (\ell_2, \Delta^2),$$

both of which are weaker than Juhász's result (Theorem 18.5.6). However, if Theorem 17.4.13 (which says that $\chi(\mathbb{E}^2 \geq 5)$ is confirmed, then Juhász's result can be seen as a corollary of Theorem 18.5.15.

When $d = 3$, since $\chi(\mathbb{E}^3) \geq 6$, Theorem 18.5.15 says that

$$\mathbb{E}^3 \to (\ell_2, \ell_5)$$

and in fact, for any 5 point set $T_5 \subset \mathbb{E}^3$,

$$\mathbb{E}^3 \to (\ell_2, T_5),$$

again, a result weaker than Theorem 18.5.6, but with the monochromatic T_5 being a translate.

With $d = 4$ and the fact that $\chi(\mathbb{E}^4) \geq 9$ (see [327]), Theorem 18.5.13 says that $\mathbb{E}^4 \to (\ell_2, \ell_7)$; however, Theorem 18.5.15 gives a better result (with less work).

Corollary 18.5.16. *If T_8 is any 8 point set in $\mathbb{E}^4$, then for any red-blue colouring of $\mathbb{E}^4$ there is either a red ℓ_2 or a blue translate of T_8. In particular, $\mathbb{E}^4 \to (\ell_2, T_8)$.*

18.6 Some related results

In [309], density type theorems are given that correspond to results in Euclidean Ramsey theory. I give only one, without proof:

Theorem 18.6.1 (Erdős et al., 1975 [309, Thm 4, pp. 536–8])**.** *Let B be a k-dimensional brick with edge lengths $d_1, d_2, \ldots, d_k$. Then for any integer $n > 1$, $m = n^{2^k}$ and $N = n^{2^k-1}$, there exists a set S of N points in $\mathbb{E}^m$ so that every subset of S with $2^k N^{(2^k-2)/(2^k-1)}$ points contains a brick congruent to B.*

Similar theorems were given in [309] for unit squares and $(1, 1, \sqrt{2})$ triangles. In that same paper is a proof for the result in Exercise 230.

Another direction taken in [309] concerns negative results for vector spaces and Hilbert spaces. I only mention one of the many results I found interesting:

Theorem 18.6.2 ([309])**.** *There is a red-blue colouring of $\mathbf{E}^1$ so that no two red points have distance 1 and no set of blue points is congruent to $\mathbb{Q}$.*

18.7 Edge colouring Euclidean Ramsey

For colouring edges and looking for triangles in the plane, see [309, Thm. 24].

Cantwell [173] coined the term "edge-Ramsey" and proved a few such results, some of which came from Cantwell's dissertation [172] in 1992.

Say that a *line colouring* of a geometric graph is a colouring of the edges so that collinear edges have the same colour.

Theorem 18.7.1 (Erdős et al., 1975 [309])**.** *For $t \in \mathbb{Z}$, let E_t be the set of edges of the unit $t \times t$ lattice. If the edges of $\mathbb{E}^3$ are line coloured with two colours, there is a (similar) copy of E_t that has all its edges the same colour.*

Exercise 354 (Károlyi–Pach–Toth, 1997 [527, Thm. 1])**.** *Let K_n be drawn in the plane with points (vertices) in general position and lines (edges) drawn straight. Prove that any red-blue colouring of the edges of this geometric graph gives a monochromatic planar spanning tree. Hint: Try induction.*

Károlyi, Pach, and Toth [527] also prove that there exist $\lfloor (n+1)/3 \rfloor$ pairwise disjoint edges of the same colour, and this bound is tight.

Exercise 355. *Prove that if the edges of a complete geometric graph on $3n-1$ vertices are 2-coloured, there exist n pairwise disjoint edges of the same colour.*

The results in Exercises 354 and 355 were conjectured by Bialostocki and Dierker. Károlyi, Pach, and Toth [527] also improve an earlier result of Larman *et al.* (see [527] for references) that constructs a family of m segments in the plane that has no more than $m \ln(4)/\ln(27)$ members that are either pairwise disjoint or pairwise crossing.

Exercise 356. *Let 6 points be in the plane so that no three are in a line, and all 15 pairwise distances are different. Join all pairs of points with (straight) edges (producing $\binom{6}{3} = 20$ triangles). Show that there is one edge that is both the longest edge of some triangle and the shortest edge of some triangle.*

Chapter 19

Separating sets and shatterring sets

19.1 Separating sets

Recall Helly's theorem (Theorem 5.2.3): For $n \geq 1$, if convex sets $C_1, C_2, ..., C_r$ in $\mathbb{R}^n$ have the property that any $n+1$ of them share a common point, then some point is contained in all of the sets.

By a compactness argument (see, *e.g.* [450, p. 60]) the number of convex sets in Helly's theorem may also be infinite. In a sense, now the opposite situation is considered, when sets do not pairwise intersect.

Definition 19.1.1. Two sets of points in the plane are called *separable* if and only if there is a line that intersects neither, and the two sets are in different half-planes determined by the line.

Exercise 357 (Kirchberger, 1903)**.** *Show that two bounded closed sets in the plane are separable if and only if any two subsets whose union includes at most four points are separable.*

In the next exercise, some special terminology is used (perhaps unnecessarily so, but it helps to make the statement of the exercise and the proof briefer, and it is used in many other related problems).

For $1 \leq q \leq p$, a family of sets is said to have the (p, q) *property* if among every p sets in the family, there are q with non-empty intersection. For the solution of the next exercise, the following observation can be helpful:

Lemma 19.1.2. *Let $2 \leq q \leq p$ and let $\mathcal{F}$ be a family of sets that has the (p, q) property; then for each $r = 1, 2, \ldots, q-1$, $\mathcal{F}$ has the $(p-r, q-r)$ property.*

Proof: Fix r, and suppose that $p-r$ sets are chosen. Extend this family to any p sets by adding r "new sets" (from $\mathcal{F}$ of course). By the (p, q) property, some q of these p sets have non-empty intersection, and so any $q-r$ of these q sets have non-empty intersection; in particular, when omitting the r new sets at least some $q-r$ sets from the original $p-r$ sets have non-empty intersection, proving the lemma. □

The next exercise looks simple; however, even with the help of Lemma 19.1.2 and a hint, solving this problem might be challenging:

Exercise 358. *Let $2 \leq q \leq p$ be integers, and suppose that a collection of (at least p) closed segments of the real line has the (p, q) property. Prove that the segments can be partitioned into $p-q+1$ classes so that sets in each class have non-empty intersection. Hint: Induct on $p-q$.*

19.2 Shattered sets and VC dimension

Let X be a set and let $\mathcal{H} \subseteq \mathcal{P}(X)$ be a family of subsets of X. For any $Y \subset X$, let

$$\mathcal{H}|_Y = \{H \cap Y : H \in \mathcal{H}\},$$

called the *restriction* of $\mathcal{H}$ to Y.

Definition 19.2.1. A set $Y \subset X$ is called *shattered* by $\mathcal{H}$ if and only if $\mathcal{H}|_Y = \mathcal{P}(Y)$, that is, for every subset $W \subset Y$, there exists $H \in \mathcal{H}$ such that $H \cap Y = W$. Define the *VC-dimension* of $\mathcal{H}$ to be

$$\text{VC-dim}(\mathcal{H}) = \sup_{Y \subset X} \{|Y| : Y \text{ is shattered}\}.$$

The notion of shattered sets is often expressed in terms of hypergraphs, and has been used in set theory, combinatorics, combinatorial geometry, and probability. For more on VC-dimension, see, for example, [185], [645], or [699].

Observe that if $\mathcal{G} \subset \mathcal{H}$, any set shattered by $\mathcal{G}$ is certainly shattered by $\mathcal{H}$ and so $\text{VC-dim}(\mathcal{G}) \leq \text{VC-dim}(\mathcal{H})$.

Example 19.2.2 (from [645, p. 238]). *Let $X = \mathbb{R}^2$ and let $\mathcal{H}$ be the set of all closed half-planes. Perhaps surprisingly (by reasoning given below) VC-dim$(\mathcal{H}) = 3$. To see this, any set of three points in general position is shattered, but no four-point set is shattered because*

- *if three of the four points are in a row, the middle point cannot be separated by a half-plane;*
- *if four points are convex, a pair of points at opposite corners can be not be separated; and*
- *if one point lies in the convex hull of the others, then that point can not be separated.*

The result in Example 19.2.2 has a generalization:

Theorem 19.2.3. *Let $d \geq 2$ and let $X = \mathbb{R}^d$. If $\mathcal{H}$ is the family of all closed half-spaces in $\mathbb{R}^d$, then VC-dim$(\mathcal{H}) = d + 1$.*

Exercise 359. *Prove Theorem 19.2.3. Does the same theorem hold for open half-spaces?*

Example 19.2.4. *Let $\mathcal{K}$ be the set of all convex sets in the plane. Then VC-dim$(\mathcal{K}) = \infty$ since any finite set of vertices of a convex polygon (with empty interior) is shattered since for any subset, the convex hull of the subset is a desired convex set with the desired intersection.*

Below, only finite sets X (and hence finite families) are considered. Note that if $\mathcal{H}$ is the empty family, no set is shattered, in which case VC-dim$(\mathcal{H}) = 0$. To eliminate complete trivialities, assume that X is non-empty.

In the early 1970's, the following theorem was proved independently by Sauer [775], Shelah [789], and Vapnik and Chervonenkis [899].

Theorem 19.2.5 (Sauer, Shelah, Vapnik–Chervonenkis). *Let X be a set with $n \geq 1$ elements, and let $\mathcal{H} \subseteq \mathcal{P}(X)$. If VC-dim$(\mathcal{H}) = d$, then*

$$|\mathcal{H}| \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d},$$

and this bound is best possible.

To see that the bound in Theorem 19.2.5 is best possible, let $|X| = n$ and let $\mathcal{H}$ be the family of subsets of X that have at most d elements, that is, let $\mathcal{H} = [X]^{\leq d}$. Then any d-set is shattered, but any $(d+1)$-subset is not, in which case equality in Theorem 19.2.5 holds.

Exercise 360. *Using mathematical induction, prove the inequality in Theorem 19.2.5.*

For more on VC-dimension (and its many relations to computing and discrepancy) see [185].

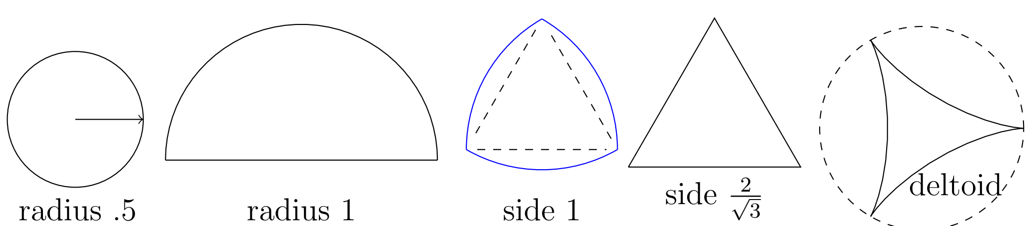
Chapter 20

The Kakeya problem

Problem 20.0.1 (Kakeya 1917 [518]). *What is the smallest area in $\mathbb{R}^2$ that a pin with length one (and thickness 0) can be continuously turned around in?*

For the moment, define a *Kakeya set* to be a connected subset of the plane so that a unit length pin can be continuously turned around while staying inside the set. Note that a Kakeya set contains a unit line segment in every direction, and to show that a set is a Kakeya set, it suffices to show that a unit pin can be continuously turned 180 degrees.

Some basic shapes are Kakeya sets. The circle with radius $\frac{1}{2}$ is a Kakeya set with area $\frac{\pi}{4} \sim .7854$. A semicircle with radius 1 is a Kakeya set with area much larger. An equilateral triangle with side length 1 is not quite large enough to turn the pin in, but a Rouleaux triangle will do, with area $\frac{\sqrt{3}}{4} \sim .7047$. A slightly larger equilateral triangle, one with height 1 is a Kakeya set (such a triangle has side length $\frac{2}{\sqrt{3}}$) with area $\frac{1}{\sqrt{3}} \sim .5774$. In 1921, Pal [710] showed that among convex shapes, this last triangle is optimal. The standard "deltoid" or "3-cusp hypocycloid" solution has area only $\frac{\pi}{8} \sim .39$ (one half of the area of a circle with diameter 1).



Exercise 361. *For reals $0 < r < R$ let $A(r, R)$ denote an annulus centered at the origin with inner radius r and outer radius R (where $R > \frac{1}{2}$); in other words, $A(r, R) = \{\mathbf{x} \in \mathbb{E}^2 : r \leq \|\mathbf{x}\| \leq R\}$. What is the relation between r and R so that a unit-length pin can be turned 180 degrees in $A(r, R)$? (So $R > \frac{1}{2}$.) Does the area of any such $A(r, R)$ have a minimum?*

The deltoid might seem best possible among all shapes. This notion was put forward by, e.g., Osgood and Kubota (see also [79] for other references, including those by Garrett Birkhoff [94, p. 4] and W. B. Ford [339]), but this is far from true!

In a Russian journal in 1920, Besicovitch [78] (unaware of the Kakeya problem) showed that if one deletes "continuously" from the definition of a Kakeya set, a set with a unit segment in every direction exists with Lebesgue measure 0. Pál introduced Besicovitch to the Kakeya problem, who improved one of his previous lemmas and (with the help of Pál), solved the Kakeya problem, the continuous version of his previous result.

In 1926, Besicovitch [79] submitted his paper, which appeared in 1928, answering the original problem. Besicovitch's answer might be seen as somewhat stunning: the area of a Kakeya set can have arbitrarily small area!

In 1928, Perron [719] simplified Besicovitch's construction, using what are now called "Perron triangles". There are many questions regarding the dimension (either Minkowski or Hausdorff), most open for $d \geq 3$. Discussing these problems was not an intended topic for this course, so I skip these issues.

In 1996, Thomas Wolff gave a survey [930] (published in 1999) on work associated with Kakeya's problem. Wolff also proposed a finite field analogue of Kakeya's problem. It is well-known that every finite field has prime power order, and for each such order q, the finite field is unique, denoted $\mathbb{F}_q$. For any positive integer n, define a *Kakeya set* in $\mathbb{F}^n$ to be a subset of $\mathbb{F}^n$ that contains a line in every direction. Recall (see Section 11.5) that the n-dimensional affine geometry over $\mathbb{F}_q$ is denoted by $\mathrm{AG}(n, q)$.

Problem 20.0.2 (Wolff, 1999 [930])**.** *Let q be a power of a prime. Find the size of the smallest Kakeya set in $AG(n, q)$.*

Question 20.0.3 (Wolff, 1999 [930])**.** *For each $n \geq 1$ does there exist a constant $C_n > 0$ (depending only upon n) so that for any prime power q, if K is a Kakeya set in $\mathbb{F}_q^n$, then $|K| > C_n q^n$?*

Wolff gave a lower bound of the form $c_n q^{\frac{n+2}{2}}$. Bounds were found for $n = 3, 4$ (e.g., Tao did $n = 4$ using additive number theory techniques developed by Bourgain; see [283] for references). For nearly a decade, the best general lower bound was of the form $c_n q^{\frac{4}{7}n}$ by Rogers [759].

In 2008, the Finite Field Kakeya (FFK) question was answered in the affirmative with $C_n = \frac{1}{n!}$ by Zeev Dvir [283] using polynomials. So the story goes, he was at the time, a Part 3 student at Cambridge.

Theorem 20.0.4 (Dvir, 2009 [283])**.** *For each $n \geq 1$, for any prime power q, if K is a Kakeya set in $\mathbb{F}_q^n$, then $|K| > \frac{1}{n!}q^n$.*

Soon after Dvir's result was known (in fact, even before Dvir's result was published), improvements on the constant $C_n = \frac{1}{n!}$ were found. For example, Saraf and Sudan [774] improved the constant to $\frac{1}{4^n}$. Dvir then noticed that the 4 could be improved to 2.6. In 2009, Dvir, together with Kopparty, Saraf, and Sudan [285] then improved the constant to $\frac{1}{2^n}$.

Only the case $n = 2$ of the FFK problem has been completely settled. Faber [328] showed that if q is even and $K \subseteq \mathrm{AG}(2, q)$ is Kakeya set, then then $|K| \geq \frac{q(q+1)}{2}$, where equality holds if and only if K is of "hyperoval type". For odd q, Blokhuis and Mazzocca [111] showed that $|K| \geq \frac{q(q+1)}{2} + \frac{q-1}{2}$, where equality holds if and only if E is "of oval type". [I have not seen these proofs, but they may rely on a solid knowledge of arcs and blocking sets in a FPP.]

For more on Kakeya type problems, see Dvir's survey [284].

Chapter 21

Lattices and geometries

This chapter is only an introduction to some of the connections between lattices and finite geometries, including FPPs and those arising from polyhedra or polytopes.

There are many resources for lattices; for example, see any of George Grätzer's (now standard) works on lattice theory (*e.g.*, [413] or [414]), or Davey and Priestly's *Introduction to lattices and order* [238]. Many books on combinatorics also contain the basics (*e.g.*, Aigner's *Combinatorial theory* [12]).

21.1 Posets and lattices

A *partially ordered set* (or simply *poset*) is a pair $(S, <)$ where S is a set and $<$ is a binary relation on S that is irreflexive ($x \not< x$), asymmetric (if $x < y$ then $y \not< x$), and transitive ($x < y$ and $y < z$ imply $x < z$). Often a poset is denoted by $(S, \leq)$, where the relation $\leq$ is reflexive ($x \leq x$), antisymmetric ($x \leq y$ and $y \leq x$ imply $x = y$) and transitive. Some authors call a poset of the form $(S, <)$ a *strict poset*. It is common to refer to a poset $(S, \leq)$ by simply S. (The letter P is commonly used to refer to posets, but since P is almost overused in the context of planes, another letter is often used here.)

For two elements x, y in a poset S, say that y *covers* x if $x < y$ and $x < z \leq y$ implies $z = y$.

An element $x \in S$ is called *minimal* if there is no element $c \in S$ such that $c < x$, or *maximal* if there is no element a such that $x < a$. An element $x \in S$ is called the *greatest element* if for every $u \in S$, $u \leq x$; the analogous definition holds for *least element*. For a subset B of a partially ordered set S, an *upper bound* of B is an element $s \in S$ so that for every $b \in B$, $b \leq s$; an upper bound s for B is called a *least upper bound* (LUB) if for any other upper bound t of B, $s \leq t$ holds. Similar definitions hold for *lower bound* and *greatest lower bound* (GLB). When a poset is infinite, the LUB and GLB are called the *supremum* (or *lim sup*) and *infimum* (or *lim inf*).

If there is an element in a poset which is a (least) upper bound for the entire set, then this upper bound is often denoted by "1", and a lower bound for the poset is denoted by "0". Note that upper or lower bounds need not exist (for example, when the poset is the integers with the natural order, the LUB is ω, which is not an integer). Even if a poset is finite, GLB's or LUB's need not exist (for example, if the poset is on three distinct elements, a, b, c, with only $a \leq b$ and $a \leq c$, then LUB($\{b, c\}$) does not exist). The greatest lower bound of a pair $\{x, y\}$ in a poset is denoted by GLB($\{x, y\}$), or by $x \wedge y$, the *meet* of x and y; the least upper bound LUB($\{x, y\}$) is denoted by $x \vee y$, the *join* of x and y.

A *lattice* is a poset $(L, \leq)$ with the property that for each pair a and b in L, both the meet $a \wedge b$ and the join $a \vee b$ exist in L. A lattice $(L, \leq)$ is often denoted by simply L when no confusion can arise. A lattice L is *bounded* or *complete* if there exists both a greatest element $1 = 1_L$ and a least element $0 = 0_L$ in L (in this text, only finite lattices are considered, in which case these always exist).

Elements that cover 0 in a lattice are called *atoms*. A lattice is called a *point lattice*, or *atomic* if every element is a supremum of atoms; every finite complete lattice is atomic. The *dual* of a lattice is a partial order on the same points however with the relation $\leq$ reversed; the dual of a lattice is again a lattice.

Definition 21.1.1. A lattice L is called *distributive* if for all $a, b, c \in L$,

$$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c),$$

or equivalently,

$$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c).$$

For example, a power set lattice (a lattice consisting of a set and all its subsets with inclusion as the relation) is distributive. In fact, it is known that a lattice is distributive if and only if it is isomorphic to some lattice of sets. For an element x of a bounded lattice (with least element 0 and greatest element 1) define a *complement* of x to be an element x' such that $x \wedge x' = 0$ and $x \vee x' = 1$. In a bounded distributive lattice, it can be shown (straightforward exercise) that if a complement exists, it is unique. A *complemented lattice* is a bounded lattice for which every element has a complement (not necessarily unique).

Two examples of complemented lattices (shown in Figure 21.1) are worthy of mention. Let M_3 be the lattice on the elements $0, a, b, c, 1$ where $0 \leq a \leq 1$, $0 \leq b \leq 1$, and $0 \leq c \leq 1$; this lattice is called the *diamond* (perhaps because it has points in the shape of a baseball diamond with the pitcher's mound in the middle). Let N_5 be the lattice on elements $0, x, y, z, 1$ where $0 \leq x \leq y \leq 1$ and $0 \leq z \leq 1$; this lattice is often called the *pentagon*.

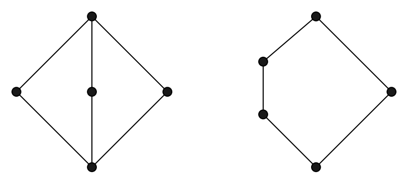


Figure 21.1: Lattices M_3 and N_5

Both M_3 and N_5 are complemented, but complements are not always unique. Furthermore, it can be shown that any lattice is non-distributive if and only if it contains as a sublattice at least one of M_3 or N_5.

21.2 Modular and geometric lattices

Definition 21.2.1. A lattice L is *semimodular* if and only if for every $x, y \in L$ if a covers $a \wedge b$ then $a \vee b$ covers b.

Definition 21.2.2. A lattice L is *modular* if for every $x, y, z \in L$, $x \leq z$ implies

$$z \wedge (x \vee y) = (y \wedge z) \vee x. \tag{21.1}$$

One can check that M_3 is modular, however N_5 is not. Indeed, a lattice is not modular if and only if N_5 appears as a sublattice.

Exercise 362. *Show that a distributive lattice is modular.*

One way of looking at a FPP is as a lattice. Let $\mathcal{P} = (P, \mathcal{L})$ be a FPP and let $L_{\mathcal{P}}$ be the lattice defined on points $1 = \mathcal{P}$, $\mathcal{L}$, P and $0 = \emptyset$, where the relation $<$ is defined by inclusion, that is, if $X \in P$ lies on line ℓ, then in the lattice, $X < \ell$. Call $L_{\mathcal{P}}$ the lattice of the FPP $\mathcal{P}$. For any geometry, one can define the lattice of the geometry in a similar way. For example, a geometry defined by the subspaces of a d-dimensional vector space over a field corresponds to a lattice.

Other simple examples of a lattice come from three dimensional geometries (or polytopes). For example, letting C denote the standard 3-dimensional cube (a platonic solid), the 0-dimensional "subspaces" are the points (vertices), the 1-dimensional subspaces are the edges, the 2-dimensional subspaces are the faces, and the whole cube is the unique 3-dimensional space. Ordering these "subspaces" by inclusion gives a lattice. It might be interesting to observe that since the dual of a cube is the octahedron, the lattices for the cube and for the octahedron are duals (reversals of one another). Similarly, the lattice for a tetrahedron is self-dual, and the lattices for the dodecahedron and for the icosahedron are duals.

Definition 21.2.3. A geometric lattice is a semimodular atomic lattice with no infinite chains.

In the finite case, geometric lattices are the semimodular lattices. (Geometric lattices were sometimes called "matroids"; see [412, p.179] for history of this terminology. Grätzer gives another definition: "A lattice L is called geometric iff L is semimodular, L is algebraic, and the compact elements of L are exactly the finite joins of atoms in L.") Regarding lattices and geometries, a terrific amount is known, and only a very few results are considered here. The interested reader might consult Birkhoff's *Lattice theory* [96, Ch.s VII, VII] or Grätzer's *General lattice theory* [412, pp. 178–224] In the 1930s, Birkhoff [95] and MacLane [629] showed that geometric lattices are complemented.

One connection between FPPs and lattices is the following.

Theorem 21.2.4. *The lattice of a FPP is a complemented modular lattice.*

Proof: Let L be the lattice of a FPP $\mathcal{P} = (P, \mathcal{L})$. Pick a point $A \in P$ and let $\ell \in \mathcal{L}$ be a line not containing A. It is easily seen that A and ℓ are complements in the lattice. Now check that L is modular; certainly this can be done in an easier way, but here it is done from the definition. Fix $x \leq z$ in L. In the case that $x = 0 = 0_L$ equation (21.1) becomes

$$z \wedge (0_L \vee y) = (y \wedge z) \vee 0_L,$$

which is trivially true. Similarly, (21.1) easily holds when $z = \mathcal{P}$. When $x = z$, equation (21.1) becomes

$$z \wedge (z \vee y) = (y \wedge z) \vee z,$$

each side of which is equal to z.

So assume that $\emptyset \neq x < z \neq \mathcal{P}$. Thus x denotes a point X in the FPP and z denotes a line ℓ in the plane containing X. In this case, if $y = \emptyset$, then both sides of equation (21.1) are equal to x; if $y = \mathcal{P}$, then both sides are equal to z. It remains to examine the case when y denotes a point or a line (and $x = X$ is a point and $z = \ell$ is a line containing X).

Suppose that y is the lattice point corresponding to the point Y in the plane. There are two cases depending on whether or not Y lies on ℓ. If $Y \in \ell$, then $x \vee y = z$ and so the left hand side of (21.1) becomes z; since in this case, $y \wedge z = y$, the right hand side of (21.1) becomes $y \vee x$, denoting a line containing both X and Y, namely ℓ (two points uniquely determine a line) so $(y \wedge z) \vee x = z$ as well. On the other hand, if Y does not lie on ℓ, $x \vee y$ is a line containing X, as does z, so their meet is X, that is, the left hand side of (21.1) is x; now since Y does not lie on ℓ, $y \wedge z = \emptyset$ and so the right hand side of (21.1) is $0_L \vee x = x$.

The case where y denotes a line is analogous (and in fact, follows from the fact that the dual lattice of a FPP is a lattice of a FPP). □

Birkhoff [95] also showed that to each projective geometry (e.g., $\mathrm{PG}(n, \mathbb{F})$), there corresponds a complemented modular lattice (the atoms of the lattice are the points of the geometry; elements of height 2 are the lines, etc.). On the other hand,

Theorem 21.2.5 (Frink, 1946 [354])**.** *Every complemented modular lattice is isomorphic to a sublattice of the lattice of all subspaces of a (possibly degenerate) projective space (which may be infinite-dimensional).*

Dilworth [258] showed that a finite dimensional (see Definition 21.3.3 below for "dimension" of a poset) complemented modular lattice is isomorphic to a product of a Boolean algebra and projective geometries. Early work in lattice theory and projective geometries included lattice-theoretic ways to view perspectivities and when Desargues' theorem holds (see, e.g., [412], [510] for references and details).

21.3 Poset dimension and FPPs

Finite geometries can be used as examples for properties of partial orders. In this section, the goal is to give one application of FPPs in dimension theory for posets.

21.3.1 Posets and dimension

Recall that a strict partial order on a set P is a binary relation R that is irreflexive, antisymmetric, and transitive. The set P together with R is called a strict partially ordered set or a strict "poset", often denoted $(P, <)$. Only strict posets are considered in this section. If $(x, y) \in R$, write xRy. If $(P, <)$ is a poset and $x, y \in P$, write $x||y$ if and only if x and y are not comparable.

If R and S are strict partial orders on a set P, say that S *extends* R if and only if $R \subseteq S$, and that (P, S) is a an extension of (P, R).

A *linear extension* of a partial order R is a linear ordering S that extends R.

Theorem 21.3.1 (Szpilrajn, 1930 [857]). *Every strict partial ordering has a linear extension. Also, if $x||y$, there exists a linear extension $<_L$ with $x <_L y$.*

Theorem 21.3.2 (Dushnik–Miller, 1941 [282]). *If (X, R) is a strict poset and $\mathcal{F}$ is the set of linear extensions of (X, R), then $(X, R) = \cap_{L\in\mathcal{F}} L$.*

Definition 21.3.3 (Dushnik–Miller, 1941 [282]). The dimension of a strict poset (X, R) is the minimum size of a family $\mathcal{F}$ of linear extensions of (X, R) so that $\cap_{L\in\mathcal{F}} L = (X, R)$.

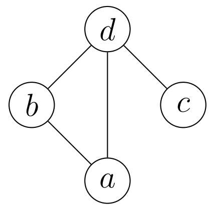
If R is not a linear order, then $\dim(X, R) > 1$. If $x||y$ in R, and $(X, R) = \cap L_i$, then some L_i has $x < y$, and some L_j has $y < x$.

Theorem 21.3.4 (Dilworth, 1950 [257]). *For any strict poset (X, R),*

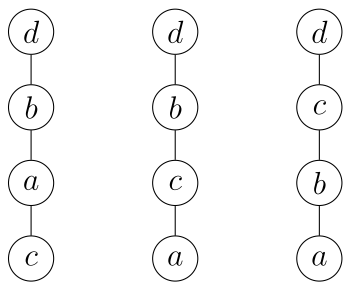
$$\dim(X, R) \leq \textit{width} = \textit{size of largest antichain}.$$

In 1982, it was shown by Yannakakis [934] that testing for $\dim(X, R) \leq t$ is NP-complete.

Example 21.3.5. *Let $P = \{a, b, c, d\}$ and $R = \{(a, b), (b, d), (a, d), (c, d)\}$.*



There are three linear extensions of (P, R). As in the following sketch, let L_1 be given by $c < a < b < d$, L_2 given by $a < c < b < d$ and L_3 given by $a < b < c < d$.



The intersection of any two give (P, R), so $dim(P, R) = 2$.

21.3.2 Two famous examples of posets

One type of poset used in the study of dimensions is called a *crown*. The "4-crown", denoted C_4, is in Figure 21.2.

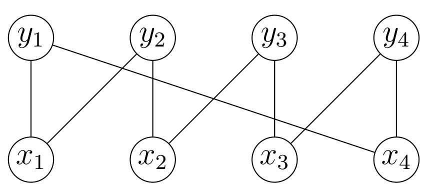


Figure 21.2: The crown poset C_4.

It is not difficult to see that the width (size of a largest antichain) of C_4 is 4.

Exercise 363. *Let C_4 be the crown partial order in Figure 21.2, Show that $dim(C_4) = 3$.*

Another poset is called a "standard example", formed by a complete equibipartite graph and deleting a matching; the standard example S_4 is given in Figure 21.3. The incidence partial order for a MSRS (see Section 14.6) is the complement of some S_d. In general, what is the dimension of S_d?

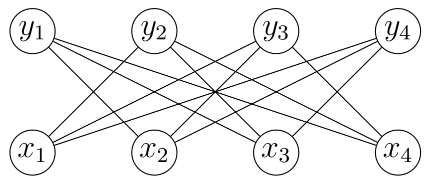


Figure 21.3: The standard example S_4.

Lemma 21.3.6. *If S_4 is the standard example given in Figure 21.3, then $dim(S_4) = 4$.*

Proof: It is not difficult to see that $\dim(S_4) \leq 4$. To show that equality holds, suppose the contrary, that is, there are only 3 linear extensions intersecting to give S_4, say $L_1 \cap L_2 \cap L_3 = S_4$.

For each i, $x_i || y_i$. If $x_1 <_{L_1} y_1$ in L_1, then either $y_1 <_{L_2} x_1$ or $y_1 <_{L_3} x_1$. So, for each i, $x_i < y_i$ in some L_j. But there are four such pairs, and so by PHP, there is L_i, say L_1, and two pairs, say x_1, y_1 and x_2, y_2, so that

$$y_1 <_{L_1} x_1, \quad x_1 <_{L_1} y_2, \quad y_2 <_{L_1} x_2, \quad x_2 <_{L_1} y_1,$$

and so by transitivity, $y_1 <_{L_1} y_1$, a contradiction. $\square$

Theorem 21.3.7 (Hiraguchi, 1955 [476]). *If S is a strict poset with $|S| \leq 2n+1$, then $dim(S) \leq n$.*

Theorem 21.3.8 (Bogart–Trotter, 1973 [113]). *When n is even, equality holds only when $S = S_n$.*

Theorem 21.3.9 (Kimble, 1973 [546]). *When n is odd, equality holds only when $S = S_n$.*

21.3.3 Large dimension does not imply large S_d

Example 21.3.10 (Dushnik–Miller, 1941 [282]). *The inclusion partial order for singletons and pairs from $[n]$ has dimension $\Omega(\ln \ln n)$, but does not even contain S_4.*

(One proof uses the Erdős-Szekeres [320] theorem, which says in any permutation of $1, 2, \ldots, n^2 + 1$, there is a monotonic subsequence with at least $n + 1$ elements.)

It was hoped that "really large" dimension implies "really large" S_d.

Theorem 21.3.11 (Biró–Hamburger–Pór, 2015 [99]). *For every integer $d \geq 2$ and every $\epsilon > 0$, there is n_0 so that for $n > n_0$, if P is a poset with $|P| \leq 2n+1$ and P does not contain S_d, then dim$(P) < \epsilon n$.*

So, if P contains $S_2, \ldots, S_{99}$ but not S_{100}, then $\dim(P)$ is still reasonably small.

Example 21.3.12. *Let Π be a FPP of order q. Define the poset P of height 2 with $|P| = 2(q^2 + q + 1)$, where the minimal points of P are points of Π, and the maximal elements are the lines of Π, and a point x is less than (in P) a line y if and only if x is* not *on y.*

So P in Example 21.3.12 is the "anti-incidence" bipartite (di)graph of Π. Following the proof idea that shows $\dim(S_d) = d$, one can show that $\dim(P) \geq q^2 + q + 1 - q^{3/2}$. Putting $n = q^2 + q + 1$, then $|P| = 2n$ and $\dim(P) = (1 - o(1))n$. How large is any S_d in P?

If the above P contains S_d with $d > 2q^{3/2}$, take the odd indexed minimal points and the even indexed maximal points, and get a P' with more than q^3 relations, contradicting the q^3 theorem. So if S_d is contained in P, then $d \leq 2q^{3/2} = o(n)$. (For another proof that $d \leq q\sqrt{q}+1$, use the upper bound for a MSRS in FPPs (Theorem 14.6.1)—whose partial order is precisely a standard example!)

However, when the dimension is (really) close to n, there exists a (very) large S_d.

Theorem 21.3.13 (Biró–Hamburger–Pór–Trotter, 2016 [100]). *For every integer $c \geq 1$, there is an integer $f(c) = O(c^2)$ so that for large enough n, if P is a poset with $|P| \leq 2n+1$ and dim$(P) \geq n - c$, then for $d \geq n - f(c)$, P contains S_d.*

The FPP example is also used [100] to show that $f(c) = \Omega(c^{4/3})$.

See the Füredi–Kahn [358] paper for a result on dimension of lattice derived from FPP. (Also see [357].)

Chapter 22

Appendix: Number theory

22.1 A miscellany of tools from number theory

In this section, all variables indicate integers.

The Euclidean division algorithm (EDA) is a way to find the greatest common divisor of any two positive integers without knowing the factorizations of either integer. Reversing the steps in EDA (as shown in an example after the lemma) produces a useful result:

Lemma 22.1.1 (Bezout's lemma)**.** *For positive integers a and b, there exist integers k and ℓ so that* $\gcd(a, b) = ka + \ell b$.

For example, to calculate $\gcd(186, 612)$, the EDA is applied to give

$$\begin{aligned} 612 &= 3 \cdot 186 + 54 \\ 186 &= 3 \cdot 54 + 24 \\ 54 &= 2 \cdot 24 + 6, \\ 24 &= 4 \cdot 6 + 0, \end{aligned}$$

from which it is concluded that $\gcd(186, 612) = 6$. To see 6 as a linear combination of 186 and 612, start at the penultimate equation and substitute terms using the prevous equatoins: $6 = 54 - 2 \cdot 24 = 54 - 2(186 - 3 \cdot 54) = 7 \cdot 54 - 2 \cdot 186 = 7(612 - 3 \cdot 186) - 2 \cdot 186 = 7 \cdot 612 - 23 \cdot 186$. $\square$

The next result shows when numbers have inverses modulo n.

Corollary 22.1.2. *If* $\gcd(a, n) = 1$*, then there exists* b *such that* $ab \equiv 1 \pmod{n}$*.*

Proof: By Lemma 22.1.1, there exist k, ℓ so that $\gcd(a, n) = ka + \ell n$, so in this case, use $b = k$. $\square$

Corollary 22.1.3. *Let* p, a, b *be positive integers with* p *prime and* $p \mid ab$*. Then at least one of* $p \mid a$ *or* $p \mid b$ *holds.*

Proof: Suppose that p does not divide a. Since p is prime, $\gcd(a, p) = 1$, and so by Bezout's lemma, let x and y be integers with $1 = xa + yp$. Multiplying this equation by b gives $b = xab + ypb$. Since $p \mid ab$, $p \mid xab$, and trivially, $p \mid ypb$, so $p \mid b$. $\square$

Note that in Corollary 22.1.3, if p is not prime, then the result may fail; for example, 10 divides $100 = 4 \cdot 25$, but 10 does not divide either 4 or 25.

Corollary 22.1.4. *If* $\gcd(a, n) = 1$ *and* $xa \equiv xy \pmod{n}$*, then* $x \equiv y \pmod{n}$*.*

Proof: By Corollary 22.1.2, pick b such that $ab \equiv 1 \pmod{n}$ and multiply each side of $xa \equiv xy \pmod{n}$ by b. $\square$

Theorem 22.1.5 (Fermat's little theorem)**.** *For any prime* p *and positive integer* a *with* $\gcd(a, p) = 1$*,*

$$a^{p-1} \equiv 1 \pmod{p}.$$

Proof: Examine the integers

$$a, 2a, \ldots, (p-1)a.$$

All are relatively prime to p. Also, all are distinct modulo p, which is seen by the following: if for $1 \leq i < j \leq p-1$, $ia \equiv ja \pmod{p}$, then there is k so that $ja = ia + pk$, implying $pk = (j-i)a$, which is impossible since p can neither divide $j - i$ nor a. So all are distinct modulo p, and none are $0 \bmod p$. Hence

$$a \cdot 2a \cdot \cdots \cdot (p-1)a \equiv 1 \cdot 2 \cdot \cdots \cdot (p-1) \pmod{p}.$$

Cancelling $(p-1)!$ from each side (which is possible by Corollary 22.1.4 since $\gcd((p-1)!, p) = 1$) completes the proof. $\square$

Another more combinatorial proof of Fermat's little theorem is outlined as follows: the number of functions $f : [p] \to [n]$ which are not constant is $n^p - n$. To each such function f, there are p distinct shift functions $f = f_0, f_1, \ldots, f_{p-1}$ defined by $f_i(x) = f(x+i)$, where addition is modulo p. So $n^p - n$ is divisible by p. $\square$

For example, with $p = 7$, $2^6 = 64 = 9 \cdot 7 + 1$, $3^6 = 729 = 104 \cdot 7 + 1$, $4^6 = 4096 = 585 \cdot 7 + 1$, and $5^6 = 15625 = 2232 \cdot 7 + 1$, are all congruent to $1 \pmod 7$.

Lemma 22.1.6. *Let p and q be distinct primes. If $x^r \equiv x \pmod p$ and $x^r \equiv x \pmod q$, then $x^r \equiv x \pmod{pq}$.*

Proof: If both p and q divide $x^r - x$, then so does pq. $\square$

Theorem 22.1.7 (Wilson's Theorem)**.** *For a prime number p,*

$$(p-1)! \equiv -1 \pmod p.$$

Proof: For any $x \in \{1, 2, \ldots, p-1\}$, there exists a unique $y = x^{-1} \in \{1, 2, \ldots, p-1\}$ so that $xy \equiv 1 \pmod p$. If $x = x^{-1}$, $x^2 \equiv 1 \pmod p$, implying $x = 1$ or $x \equiv -1 \equiv p-1$. So upon multiplying out $(p-1)!$, all elements except 1 and $p - 1 \equiv -1$ pair up two at a time to produce 1. $\square$

Corollary 22.1.8. *If p is a prime with $p \equiv 1 \pmod 4$, then*

$$\left[\left(\frac{p-1}{2}\right)!\right]^2 + 1 \equiv 0 \pmod p.$$

Proof: Since for each $i = 1, \ldots, p-1$, $-i \equiv p - i \pmod p$,

$$((p-1)/2)!)^2 \equiv (p-1)!(-1)^{(p-1)/2} \pmod p$$

which is congruent to -1 $(\mathrm{mod}\ p)$ by Wilson's theorem. $\square$

For reference, Corollary 22.1.8 is stated in the following form.

Corollary 22.1.9. *If p is a prime with $p \equiv 1 \pmod 4$, then there exists an a so that $a^2 \equiv -1 \pmod p$.*

In other words, if p is a prime with $p \equiv 1 \pmod 4$, then -1 is a" quadratic residue" modulo p (see Section 22.3 for more on quadratic residues).

In fact, the converse of Corollary 22.1.9 holds, that is, if for an odd prime p there is no a such that $a^2 \equiv -1 \pmod p$, then $p \equiv 3 \pmod 4$ so the statement of Corollary 22.1.9 is not completely vacuous. Even though the proof of the converse is easy, it is omitted here. Some proofs (even standard ones) of later claims rely on this converse, yet the sequence of steps given here does not.

Theorem 22.1.10 (Thue)**.** *Fix a prime number p. For any integer a not divisible by p, there exist positive integers x and y less than $\sqrt{p}$ so that*

$$xa \equiv \pm y \pmod p.$$

Proof: Examine the set $A = \{ax - y : 0 \leq x, y < \sqrt{p}\}$. Since the number of pairs (x, y) with $x, y \in [0, \sqrt{p}]$ is greater than p, it follows that $|A| > p$, and so by the pigeonhole principle, there exist two distinct elements of A congruent modulo p, say $ax_1 - y_1$ and $ax_2 - y_2$, with $ax_1 - y_1 \equiv ax_2 - y_2 \pmod p$. Then $a(x_1 - x_2) \equiv y_1 - y_2 \pmod p$. If $x_1 = x_2$ then $y_1 = y_2$, yielding identical members of A. Similarly, if $y_1 = y_2$ get identical elements of A. So $x = |x_1 - x_2|$ and $y = |y_1 - y_2|$ are non-zero and satisfy the theorem. □

Dirichlet's famous theorem yields that there are infinitely many primes of the form $4k+1$. However, what is sometimes needed is that the density of such primes is reasonably large. To state the following result, let $\phi(n)$ denote the Euler totient function applied to n, namely, the number of positive integers smaller than n that are relatively prime to n.

Theorem 22.1.11 (see [645, Theorem 4.2.4])**.** *Let $d, a \in \mathbb{Z}^+$ satisfy $gcd(d, a) = 1$. Then the number of primes of the form $a + kd$ (where $k \in \{0, 1, 2, \ldots\}$) that are at most n is*

$$(1 + o(1))\frac{n}{\phi(n)\ln(n)}.$$

22.2 Finite fields

A field is a structure $(\mathbb{F}, +, \cdot, 0, 1)$ where $\mathbb{F}$ is a set, "+" and "·" are two binary operators on $\mathbb{F}$ called addition and multiplication respectively, and $0, 1 \in \mathbb{F}$, where the following axioms hold for all $a, b, c \in \mathbb{F}$:

- For addition:

  $a + b \in \mathbb{F}$ (closure);

  $a + (b + c) = (a + b) + c$ (associativity);

  $a + b = b + a$ (commutativity);

  $a + 0 = a = 0 + a$ (additive identity exists);

  For each a, there exists $x = -a \in \mathbb{F}$ so that $a + x = 0 = x + a$ (additive inverses exist).

- For multiplication:

  $a \cdot b \in \mathbb{F}$ (closure);

  $a(b \cdot c) = (a \cdot b) \cdot c$ (associativity);

  $a \cdot b = b \cdot a$ (commutativity);

  $a \cdot 1 = a = 1 \cdot a$ (multiplicative identity exists);

  For each $a \neq 0$, there exists $x = a^{-1} \in \mathbb{F}$ so that $a \cdot x = 1 = x \cdot a$ (multiplicative inverses exist).

  if $a \cdot b = 0$ then at least one of $a = 0$ or $b = 0$ (no zero divisors).

- For both:

  $0 \neq 1$;

  $a \cdot (b + c) = a \cdot b + a \cdot c$ (left distributivity);

  $(a + b) \cdot c = a \cdot c + b \cdot c$ (right distributivity).

One often drops the $\cdot$ in $a \cdot b$ and writes ab.

Below is a summary of some of the major or useful theorems about finite fields; see nearly any book on abstract algebra (e.g., [493, V5]) for the details.

Theorem 22.2.1. *Let a $\mathbb{F}$ with $|F| = q < \infty$ elements. Then there exists a prime p and $\alpha \in \mathbb{Z}^+$ so that $q = p^\alpha$.*

One proof of Theorem 22.2.1 begins with the following lemma. The *characteristic* of a field $\mathbb{F}$ is the least positive integer $n = \text{char}(\mathbb{F})$ (if it exists) so that for every $a \in \mathbb{F}$,

$$\underbrace{a + a + \cdots + a}_{n} = 0;$$

if no such n exists, the field has characteristic 0.

Lemma 22.2.2. *If $\mathbb{F}$ is a finite field, then char$(\mathbb{F})$ is a prime number. Also, char$(\mathbb{R})$ = char$(\mathbb{C}) = 0$.*

Theorem 22.2.3. *For each prime power q, there exists a unique field of order q.*

For a prime power q, let $\mathbb{F}_q$ denote the field of order q; another notation is GF(q) (where the GF is short for "Galois field").

If p is a prime, the field of order p is simply $\mathbb{Z}_p = \{0, 1, \ldots, p-1\}$, where all arithmetic is done modulo p. (This field is also denoted $\mathbb{Z}/\mathbb{Z}_p$, whose elements are equivalence classes $[x]$ with arithmetic defined by $[a] + [b] = [a + b]$ and $[a] \cdot [b] = [ab]$.) Perhaps the most used finite field is $\mathbb{Z}_2 = \{0, 1\}$, where $1 + 1 = 0$.

If p is a prime and $k \geq 2$, if $q = p^k$, finding the field $\mathbb{F}_q$ needs some work. Without going into details, one easy construction for $\mathbb{F}_q$ is by using irreducible polynomials, where elements of the field are represented by polynomials (of small degree). See nearly any text on abstract algebra.

Example 22.2.4. *The finite field $\mathbb{F}_4$ is the set $\{0, 1, a, b\}$ with the following addition and multiplication tables:*

$+$	0	1	a	b
0	0	1	a	b
1	1	0	b	a
a	a	b	0	1
b	b	a	1	0

$\cdot$	0	1	a	b
0	0	0	0	0
1	0	1	a	b
a	0	a	b	1
b	0	b	1	a

Lemma 22.2.5. *In a finite field $\mathbb{F}_q$ the multiplicative group $\mathbb{F}_q^* = (\mathbb{F}_q \backslash \{0\}, \cdot)$ is cyclic of order $q - 1$, and so for any $a \in \mathbb{F}_q$, $a^q = a$.*

22.3 Quadratic residues

Definition 22.3.1. Let $\mathbb{F}$ be a field. An element $x \in \mathbb{F}$ is a *quadratic residue* if and only if $x \neq 0$ and there exists $y \in \mathbb{F}$ so that $y^2 = x$. If no such y exists, say that x is a *quadratic non-residue*.

Remark 22.3.2. *Some might argue that a quadratic non-residue might be more appropriately called a "non-quadratic residue", but the tradition is otherwise.*

For example, when $\mathbb{F} = \mathbb{R}$, the quadratic residues are the positive reals. For finite fields, the quadratic residues are harder to find. For example, in $\mathbb{Z}_7$, the squares are $1^2 = 1; 2^2 = 4; 3^2 = 9 = 2; 4^2 = 16 = 2; 5^2 = 25 = 4; 6^2 = 36 = 1$. Hence, the quadratic residues in $\mathbb{Z}_7$ are 1, 2, and 4.

Theorem 22.3.3. *Let p be an odd prime. The number of quadratic residues modulo p (in $\mathbb{Z}_p$) is $(p-1)/2$ (and so also the number of quadratic non-residues is also $(p-1)/2$).*

Theorem 22.3.4. *Let $\alpha \in \mathbb{Z}^+$ and let $q = 2^\alpha$. Then every non-zero element of $\mathbb{F}_q$ is a quadratic residue.*

Proof: If $\alpha = 1$, there is only one non-zero element, namely 1, that is trivially a quadratic residue. So, assume that $\alpha > 1$. If $a \in \mathbb{F}_q \backslash \{0\}$, then by Lemma 22.2.5, $a = a^{2^\alpha} = (a^{2^{\alpha-1}})^2$, and so a is a quadratic residue. □

22.4 Sums of two squares

It is fairly easy to show that if an odd prime p is the sum of two (integer) squares, then $p \equiv 1 \pmod 4$. By Theorem 22.4.3 (below), an odd prime p is a sum of two squares if and only if $p \equiv 1 \pmod 4$. There are many proofs of this fact, some using quadratic residues, characters (probably due to Jacobi), or other techniques. Even Fermat knew this result, so a 'simple' proof is chosen to be presented here (concluding in Theorem 22.4.3).

Apparently, (see [417]) Fermat (1601–1665) knew the proof of the following theorem, however Gauss gives Euler (1707–1783) the credit for the first published version. Various references say different things; some say that Euler discovered this in 1749, as evidenced by a letter to Goldbach.

Theorem 22.4.1 (Fermat and Euler)**.** *A positive integer is a sum of two squares if and only if any prime factor of the form $4k + 3$ occurs to an even power.*

So Theorem 22.4.1 says that an integer is the sum of two squares if and only if its square free part has no prime divisors of the form $4k+3$. Euler used Fermat's method of infinite descent, probably the same proof that Fermat had in mind.

The following proof (roughly following that in [690]) is a sequence of simple steps based on work of the Norwegian mathematician Axel Thue (1863–1922).

Theorem 22.4.2. *Let p be a prime number. If there is an integer a so that*

$$a^2 + 1 \equiv 0 \pmod{p}, \tag{22.1}$$

then p is a sum of two integral squares.

Proof: Let a be as in the statement of the theorem, and by Theorem 22.1.10, fix positive integers x and y less than $\sqrt{p}$ so that $xa \equiv \pm y \pmod{p}$.

Multiply both sides of (22.1) by x^2 to obtain

$$x^2a^2 + x^2 \equiv y^2 + x^2 \equiv 0 \pmod{p},$$

which shows that $x^2 + y^2 = kp$ for some positive integer k. However, $x^2 < p$ and $y^2 < p$, so $kp < 2p$, and hence $k < 2$, that is, $k = 1$. $\square$

Combining Corollary 22.1.9 with Theorem 22.4.2 gives an old result, which I believe was originally due to Fermat (who predates Wilson by a century or so).

Theorem 22.4.3. *An odd prime p is a sum of two (integral) squares if and only if $p \equiv 1 \pmod{4}$.*

Proof: Let p be an odd prime and x, y be integers with $p = x^2 + y^2$. Since p is odd, exactly one of x and y is odd, that is, there are integers a, b, so that $p = (2a+1)^2 + (2b)^2$, which is congruent to 1 modulo 4.

Let $p \equiv 1 \pmod{4}$. By Corollary 22.1.9 and Theorem 22.4.2, p is a sum of two integral squares. $\square$

Theorem 22.4.3 is also proved using the generalized version of Minkowski's theorem (Theorem 9.4.2); see Section 9.4. A proof using Gaussian integers was given by Dedekind. There are more known proofs of Theorem 22.4.3 (see, e.g., [575] for more information).

An integer is said to be *square-free* if every one of its prime factors occurs exactly once, that is, no squares divide the integer.

Theorem 22.4.4. *Let n, z, x, y be integers, n square-free, and*

$$nz^2 = x^2 + y^2.$$

Then all odd prime factors of n are congruent to 1 (mod 4).

Proof: Let p be an odd prime dividing n. Let $d = \gcd(x, y)$ and write $x = dx_1$, $y = dy_1$ where $\gcd(x_1, y_1) = 1$. Since n is square free, and $d^2 \mid nz^2$, then $d^2 \mid z^2$. Dividing each side of $nz^2 = x^2 + y^2$ by d^2 yields

$$n\frac{z^2}{d^2} = x_1^2 + y_1^2,$$

and since p divides n,

$$x_1^2 + y_1^2 \equiv 0 \pmod{p}. \tag{22.2}$$

The numbers x_1 and y_1 are relatively prime, so one, say x_1, is not divisible by p, and thus there is an integer m so that $x_1 m \equiv 1 \pmod{p}$. Multiplying (22.2) by m^2 gives $1 + (y_1 m)^2 \equiv 0 \pmod{p}$. Theorem 22.4.2 now implies that p is a sum of integral squares, and Theorem 22.4.3 says that in this case, $p \equiv 1 \pmod{4}$. □

Theorem 22.4.5. *Let n be square-free. If every prime factor of n is either 2 or of the form $p \equiv 1 \pmod{4}$, then for any integer k, nk^2 is a sum of two integral squares.*

Proof: The product of a sum of two squares with a sum of two squares is again a sum of two squares. A standard identity to show this is

$$(a^2 + b^2)(c^2 + d^2) = (ac + bd)^2 + (ad - bc)^2.$$

Notice that $2 = 1^2 + 1^2$, and $2(a^2+b^2) = (a+b)^2+(a-b)^2$; also, if $n = x^2+y^2$ is a sum of squares, so is $nk^2 = (xk)^2 + (yk)^2$. The result now follows from Theorem 22.4.3. □

Corollary 22.4.6. *An integer n is a sum of two rational squares if and only if n is the sum of two integer squares.*

Proof: If n is a sum of two integral squares, then trivially n is the sum of two rational squares.

So assume that $n = (\frac{a}{b})^2 + (\frac{c}{d})^2$ where a, b, c, d are integers, b and d non-zero. Then

$$n(bd)^2 = (ad)^2 + (cb)^2,$$

and putting $n = mk^2$, m square-free, then

$$m(kbd)^2 = (ad)^2 + (cb)^2.$$

By Theorem 22.4.4 every odd factor of m is congruent to 1 (mod 4). Hence by Theorem 22.4.5, $mk^2 = n$ is a sum of two integral squares. $\square$

Theorem 22.4.7. *For a square-free integer n and any integer m, nm^2 is a sum of two squares if and only if n contains no prime factors congruent to 3 (mod 4).*

Proof: If n contains no factors congruent to 3 (mod 4), Theorem 22.4.5 gives that nm^2 is a sum of two squares.

If nm^2 is a sum of two squares, n square-free, then Theorem 22.4.4 applies to odd factors of n. $\square$

For example, $48 = 3 \cdot 4^2$ is not the sum of two squares, which one can check by checking the possibilities among $1, 4, 9, 16, 25, 36$.

Theorem 22.4.8 (see [645, Lemma 4.2.3])**.** *Let $p_1, \ldots, p_r$ be primes congruent to 1 (mod 4). Then $n = p_1 \cdot p_2 \cdot \cdots \cdot p_r$ is a sum of two squares in 2^r ways.*

Essentially, to prove Theorem 22.4.8, one first shows that each such prime is expressible as a sum of two squares in two ways, and then use the fact that a product of a sum of two squares with another sum of two squares is indeed yet another sum of two squares by either of two well-known identities known by Diophantus,

$$(a^2+b^2)(c^2+d^2) = (ac-bd)^2 + (ad+bc)^2 = (ac+bd)^2 + (ad-bc)^2. \quad (22.3)$$

The following was proved independently by Edmund Landau in 1908 and Srinivasa Ramanujan (see, e.g., [417, p. 22]).

Theorem 22.4.9 (Landau, Ramanujan)**.** *For a positive integer n, let $S(n)$ be the number of integers less than n that are the sum of two squares. There exists a constant $c \sim .76422$ so that*

$$\lim_{n\to} \frac{S(n)\sqrt{\ln n}}{n} = c.$$

22.5 Sums of four squares

The following is often referred to here as simply "Lagrange's theorem", just one of many by Lagrange.

Theorem 22.5.1 (Lagrange, 1770)**.** *Any positive integer is the sum of at most four integral squares.*

A proof of Lagrange's theorem may be found in any number of elementary number theory books (see [273], for example). In Section 9.4 is another proof that uses Minkowski's theorem for general lattices. First, some simple lemmas are given. All variables are integers unless otherwise noted.

The first lemma is a generalization of equation (22.3).

Lemma 22.5.2 (Euler's sum of four squares identity, 1748)**.** *The product of sums of four squares is again a sum of four squares.:*

$$\begin{aligned}(x_1^2+x_2^2+x_3^2+x_4^2)&(y_1^2+y_2^2+y_3^2+y_4^2)\\ &=(x_1y_1+x_2y_2+x_3y_3+x_4y_4)^2\\ &+(x_1y_2-x_2y_1+x_3y_4-x_4y_3)^2\\ &+(x_1y_3-x_2y_1+x_4y_2-x_2y_4)^2\\ &+(x_1y_4-x_4y_1+x_2y_3-x_3y_2)^2.\end{aligned}$$

How was Euler's identity found? One way to find such an identity is to use quaternions..

A *quaternion* is a number of the form $a+bi+cj+dk$, where $a,b,c,d\in\mathbb{R}$, $i^2=j^2=k^2=-1$, $ij=k$, $jk=i$, $ki=j$, $ji=-k$, $kj=-i$, and $ik=-j$. As in the complex numbers, define the conjugate of $z=a+bi+cj+dk$ by $\bar{z}=a-bi-cj-dk$, and observe that $z\bar{z}=a^2+b^2+c^2+d^2$.

Exercise 364. *If both z and w are quaternions, show that $\overline{zw}=\bar{w}\cdot\bar{z}$.*

If n and m are sums of four squares, write $n=z\bar{z}$, $m=w\bar{w}$. Then noting that $z\bar{z}\in\mathbb{R}$, and so commutes with all quaternions,

$$nm=z\bar{z}w\bar{w}=w(z\bar{z})\bar{w}=(wz)(\bar{z}\bar{w})=(wz)(\overline{wz}),$$

again, a sum of four squares.

Lemma 22.5.3. *For any odd prime p, there exist integers m, $0 < m < p$, and x, y, so that*
$$pm = 1 + x^2 + y^2.$$

Proof: Two proofs are given. Let $k = (p-1)/2$. For any $0 \leq i < j \leq k$, then $i^2 \not\equiv j^2 \pmod{p}$, since otherwise it would follow that either $i + j \equiv 0 \pmod{p}$ or $i - j \equiv 0 \pmod{p}$, which are impossible by the choice of k. Similarly, $-1 - i^2 \not\equiv -1 - j^2 \pmod{p}$ for $0 \leq i < j \leq k$. There are $2(k+1)$ integers which are of the form a^2 or $-1 - a^2$, $0 \leq a \leq k$ but $2(k+1) > p$, so there are some $x \leq k$ and $y \leq k$ such that $x^2 \equiv -1 - y^2 \pmod{p}$. That is, there are integers x, y and m, so that $pm = 1 + x^2 + y^2$. Furthermore,

$$pm \leq 1 + 2k^2 \leq 1 + (p-1)^2/2 < 1 + p^2/2 < p^2,$$

so $m < p$, completing the first proof.

Here is a different proof (maybe by Davenport?). Before beginning the proof, some preliminary remarks about sum-sets are given. For any two sets of numbers A and B, define $A + B = \{a + b : a \in A, b \in B\}$. The Cauchy–Davenport theorem [181, 237] says that if $A, B \subseteq \mathbb{Z}_p$, then $|A+B| \geq \min\{p, |A| + |B| - 1\}$. (For more on the Cauchy–Davenport theorem, see [681].)

For $p = 2$, the statement in the theorem is trivially true (pick $a = 1$, $b = 0$). For $p \geq 3$, let A be the set of all perfect squares (quadratic residues) in $\mathbb{Z}_p$. Then $|A| = \frac{p+1}{2}$. By the Cauchy-Davenport theorem, $|A+A| \geq p$, and so $A + A = \mathbb{Z}_p$, and so there are x and y so that $x^2 + y^2 \equiv -1 \pmod{p}$. □

Lemma 22.5.4. *Let $1 < m < p$, where p is an odd prime. If*

$$pm = x^2 + y^2 + z^2 + w^2,$$

then there is a $k < m$ and x_1, y_1, z_1, w_1 such that

$$kp = x_1^2 + y_1^2 + z_1^2 + w_1^2.$$

Proof: If m is even, then x, y, z, w are all even, all odd, or two are even and two are odd, so arrange them so that $x \equiv y \pmod{2}$ and $z \equiv w \pmod{2}$. In this case,

$$\frac{mp}{2} = \left(\frac{x+y}{2}\right)^2 + \left(\frac{x-y}{2}\right)^2 + \left(\frac{z+w}{2}\right)^2 + \left(\frac{z-w}{2}\right)^2,$$

and repeat this process until $m/2^n$ is odd.

So without loss of generality, assume that m is odd. Choose a, b, c, and d, each between $-m/2$ and $m/2$ so that $a \equiv x$, $b \equiv y$, $c \equiv z$, and $d \equiv w \pmod{m}$. Then

$$a^2 + b^2 + c^2 + d^2 \equiv x^2 + y^2 + z^2 + w^2 \pmod{m},$$

that is, there is a k so that $a^2+b^2+c^2+d^2 = km$ (recall that $x^2+y^2+z^2+w^2 = pm$). Since each of a^2, b^2, c^2, and d^2 is less than $m^2/4$, $a^2+b^2+c^2+d^2 < m^2$, and thus $k < m$. Note that $k \neq 0$, since otherwise if $k = 0$, each of a, b, c, d is then zero, implying that each of x, y, z, w is divisible by m, giving that $m^2 | pm$, impossible since $1 < m < p$ and p is prime. So $0 < k < m$.

Then

$$m^2kp = (mp)(km) = (x^2 + y^2 + z^2 + w^2)(a^2 + b^2 + c^2 + d^2),$$

and by Lemma 22.5.2,

$$\begin{aligned} m^2kp &= (xa + yb + zc + wd)^2 + (xb - ya + zd - wc)^2 \\ &\quad + (xc - za + wb - yd)^2 + (xd - wa + yc - zb)^2. \end{aligned}$$

Also,

$$\begin{aligned} xa + yb + zc + wd &\equiv x^2 + y^2 + z^2 + w^2 \equiv 0 \pmod{m}; \\ xb - ya + zd - wc &\equiv xy - yx + zw - wz \equiv 0 \pmod{m}; \\ xc - za + wb - yd &\equiv xz - zx + wy - yw \equiv 0 \pmod{m}; \\ xd - wa + yc - zb &\equiv xw - wx + yz - zy \equiv 0 \pmod{m}. \end{aligned}$$

Putting

$$\begin{aligned} x_1 &= \frac{1}{m}(xa + yb + zc + wd), \\ y_1 &= \frac{1}{m}(xb - ya + zd - wc), \\ z_1 &= \frac{1}{m}(xc - za + wb - yd), \\ w_1 &= \frac{1}{m}(xd - wa + yc - zb), \end{aligned}$$

arrive at $x_1^2 + y_1^2 + z_1^2 + w_1^2 = m^2kp/m^2 = kp$, finishing the proof of the lemma. $\square$

For another proof of Lemma 22.5.4, see [144]. The details are now in place to give a proof of Lagrange's Theorem:

Proof of Lagrange's Theorem: In view of Lemma 22.5.2, it suffices to show that every prime p is the sum of four squares. For $p = 2$, this is clear, so let $p > 2$ be prime. By Lemma 22.5.3, there exists an m, $0 < m < p$ and integers x_0, y_0 so that $pm = 1^2 + x_0^2 + y_0^2 + 0^2$. Repeated application of Lemma 22.5.4 shows that p is a sum of four squares. □

Of related interest is a theorem counting the number of ways a positive integer can be represented as a sum of four squares.

Theorem 22.5.5 (Jacobi)**.** *The number of solutions to $n = a^2 + b^2 + c^2 + d^2$ is 8 times the sum of all divisors of n that are not multiples of 4.*

Thus, for p prime, there are $8(p + 1)$ solutions to $p = a^2 + b^2 + c^2 + d^2$, and it was proved that among them, $p + 1$ of them have a odd and each of b, c, d even.

For more information on representations as sums of squares, see [417].

Chapter 23

Some vector geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$

In a first year course in linear algebra, geometric notions in $\mathbb{R}^2$ or $\mathbb{R}^3$ can be expressed in terms of vectors. In some introductory linear algebra courses, it is convenient to give generalizations to other spaces in terms of general vector spaces. Some of the relevant terminology and basic results for vector spaces are briefly reviewed here. (some of these topics are repeated in other chapters.)

Some elementary linear algebra (e.g., solving systems of equations and multiplying matrices) is assumed. For those with a good command of first year linear algebra and vector geometry, this section can be skipped. For more details on what is outlined below, see nearly any elementary linear algebra text. (Beware: other authors may use slightly different notation than what is used here.)

23.1 Vector space review

There are a vast number of problems in geometry that can be solved by algebra alone, usually in some vector space.

Definition 23.1.1. A *vector space* over a field $\mathbb{F}$ is a non-empty set V together with two operations $+ : V \times V \to V$ (called vector addition) and $\cdot : \mathbb{F} \times V \to V$ (called scalar multiplication) defined so that the following axioms are satisfied:

1. If $\mathbf{u}, \mathbf{v} \in V$, then $\mathbf{u} + \mathbf{v} \in V$.
2. If $\mathbf{u}, \mathbf{v} \in V$, then $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$.
3. If $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, then $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$.
4. There exists a vector in V, denoted $\mathbf{0}$, so that for any vector $\mathbf{v} \in V$, $\mathbf{v} + \mathbf{0} = \mathbf{v}$ holds.
5. For each $\mathbf{v} \in V$, there exists a vector denoted $-\mathbf{v}$ so that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$.
6. If $k \in \mathbb{F}$ and $\mathbf{v} \in V$, then $k \cdot \mathbf{v} \in V$.

For any $k, \ell \in \mathbb{F}$ and $\mathbf{u}, \mathbf{v} \in V$, the following four properties hold:

7. $k(\mathbf{u} + \mathbf{v}) = k\mathbf{u} + k\mathbf{v}$.
8. $(k + \ell)\mathbf{u} = k\mathbf{u} + \ell\mathbf{u}$.
9. $(k\ell)\mathbf{u} = k(\ell\mathbf{v})$.
10. $1_{\mathbb{F}}\mathbf{u} = \mathbf{u}$.

Property (1) is called "closure of addition"; property (2) is called "commutativity of addition"; property (3) is called "associativity of addition". property (6) is called "closure of scalar multiplication". The vector $\mathbf{0} = \mathbf{0}_V$ is called the *zero vector* (see below). The vector $-\mathbf{v}$ is called the *negative* of $\mathbf{v}$. (Instead of writing "$k \cdot \mathbf{v}$", one can simply write "$k\mathbf{v}$".) Note that for any vector $\mathbf{v}$, $0_{\mathbb{F}}\mathbf{v} = \mathbf{0}_V$.

Definition 23.1.2. A *vector* is an element of a vector space.

Note that some authors say that a vector is an object with direction and magnitude. In standard Euclidean geometry (in either 2 or 3 dimensions), if P and Q are two points, the vector from P to Q is denoted by $\overrightarrow{PQ}$ and so magnitude and direction are obvious. However, such objects are only one type of vector; in other situations, some vectors need not have "direction" and magnitude can be defined in various ways. For example, the set of all real-valued 4×4 matrices is a vector space, and it is not clear how to interpret "direction" of a vector in such a space. Therefore, in this text, boldface notation $\mathbf{v}$ is often preferred over arrow notation $\overrightarrow{v}$. [In handwriting a boldface symbol, the notation used by typesetters is convenient—simply put a wavy line under the character.]

A vector space V is a *zero vector space* if and only if V consists of the single vector $\mathbf{0}$. [Note: the zero vector $\mathbf{0}$ is usually identified as being the zero vector of some known vector space, but can be defined as simply anything with the properties that $\mathbf{0}+\mathbf{0}=\mathbf{0}$ and for any real number k, $k\mathbf{0}=\mathbf{0}$. Since no matter what the zero vector really is, it behaves the same, so one often calls a zero vector space "the zero space" when context is clear. Note that there is a significant difference between the number 0 and the vector $\mathbf{0}$.

When $\mathbb{F}=\mathbb{R}$, a vector space over $\mathbb{R}$ is called a *real vector space*. For example, $\mathbb{R}^2$ and $\mathbb{R}^3$ are two standard real vector spaces. The vector space P_n is the set of all polynomial functions with degree at most n and having real coefficients. For positive integers m and n, he vector space $M_{m\times n}(\mathbb{R})$ is the set of all $m\times n$ matrices with real entries, together with matrix addition and scalar multiplication.

A subset W of a vector space V is called a *subspace* of V if and only if W is itself a vector space under the addition and scalar multiplication defined on V.

Definition 23.1.3. Vectors $\mathbf{v}_1,\ldots,\mathbf{v}_k$ in a vector space V over $\mathbb{F}$ are called *linearly independent* if and only if for any scalars $c_1,\ldots,c_k\in\mathbb{F}$, the equation $c_1\mathbf{v}_1+\ldots+c_k\mathbf{v}_k=\mathbf{0}$ is satisfied only when $c_1=\cdots=c_k=0_{\mathbb{F}}$.

Definition 23.1.4. The *span* of a set vectors $S=\mathbf{v}_1,\ldots,\mathbf{v}_k$ in a vector space V over $\mathbb{F}$ is the set of linear combinations

$$\operatorname{span}(S)=\{r_1\mathbf{v}_1+\ldots+r_k\mathbf{v}_k : r_1,\ldots,r_k\in\mathbb{F}\}.$$

If S is an infinite set of vectors in a vector space, $\operatorname{span}(S)$ is the set of all *finite* linear combinations from S.

It is not difficult to show that the span of only finitely many vectors in V is a subspace of V. The span of infinitely many vectors is also a subspace. A *spanning set* for a vector space V is a subset of vectors $S\subseteq V$ for which $\operatorname{span}(S)=V$; in this case, say that S spans V.

A *minimal spanning set* for a vector space V is a subset of vectors $S\subseteq V$ for which $\operatorname{span}(S)=V$ yet for any proper subset $R\subsetneq S$, $\operatorname{span}(R)\neq V$.

Definition 23.1.5. A *basis* of a vector space V is a linearly independent set of vectors in V that span V.

Note that a basis may be infinite. [The term " bases" is the plural form of "basis".] One can show that a basis is a minimal spanning set and any minimal spanning set is a basis.

For a positive integer k, the standard basis for the real vector space $\mathbb{R}^k$ consists of the ordered k-tuples (vectors) $\mathbf{e}_1 = (1, 0, 0, \ldots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)$, $\ldots$, $\mathbf{e}_k = (0, 0, \ldots, 0, 1)$.

A central theorem for vector spaces can be proved using what is called "Zorn's lemma", a topic beyond the scope of these notes:

Theorem 23.1.6. *Every vector space has a basis.*

It is not difficult to prove that any two bases for a vector space have the same cardinality.

Definition 23.1.7. The *dimension* of a vector space V is the number, denoted $\dim(V)$, of elements in a basis for V; if this number is finite or if the space is the zero space, the space is called *finite-dimensional* and otherwise, *infinite-dimensional*.

In finite dimensional vector spaces, it is not difficult to show that a minimal spanning set is a basis.

Definition 23.1.8. Given a vector space V with an ordered basis $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$, for any vector $\mathbf{u} \in \mathbf{V}$, the *coordinate vector of* $\mathbf{u}$ relative to B is $(\mathbf{u})_B = (a_1, \ldots, a_r)$ if and only if $\mathbf{u} = a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r$.

For example, in $\mathbb{R}^2$ with standard basis $B = \{\mathbf{e}_1, \mathbf{e}_2\} = \{(1, 0), (0, 1)\}$ if $\mathbf{v} = 3\mathbf{e}_1 + 4\mathbf{e_2}$, then the coordinate vector of $\mathbf{v}$ relative to B is $\mathbf{v}_B = (3, 4)$, agreeing with standard usage in the cartesian plane. When B is the standard basis for $\mathbb{R}^k$, the subscript B is often omitted.

Definition 23.1.9. An *inner product* on a real vector space V is a function $\langle\cdot,\cdot\rangle$ from pairs of vectors in V to $\mathbb{R}$ that satisfies the following four properties:

1. For all $\mathbf{v} \in V$, $\langle\mathbf{v}, \mathbf{v}\rangle \geq 0$, and, $\langle\mathbf{v}, \mathbf{v}\rangle = 0$ if and only if $\mathbf{v} = \mathbf{0}$.
2. For all $\mathbf{u}, \mathbf{v} \in V$, $\langle\mathbf{u}, \mathbf{v}\rangle = \langle\mathbf{v}, \mathbf{u}\rangle$,
3. For all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, $\langle\mathbf{u}, \mathbf{v} + \mathbf{w}\rangle = \langle\mathbf{u}, \mathbf{v}\rangle + \langle\mathbf{u}, \mathbf{w}\rangle$.
4. For all $\mathbf{u}, \mathbf{v} \in V$ and all $k \in \mathbb{R}$, $\langle k\mathbf{u}, \mathbf{v}\rangle = k\langle\mathbf{u}, \mathbf{v}\rangle$.

An *inner product space* is a vector space with an inner product. A vector space can have more than one inner product, so the term "inner product space" is used with a specific inner product in mind.

For each positive integer k, a standard inner product in a real vector space $\mathbb{R}^k$ is called the *dot product*, defined by

$$(u_1, u_2, \ldots, u_k) \bullet (v_1, v_2, \ldots, v_k) = u_1v_1 + u_2v_2 + \cdots + u_kv_k.$$

This dot product is sometimes called the *standard Euclidean inner product*. In some cases, $\mathbb{R}^k$ together with the dot product is called a Euclidean vector space, denoted by $\mathbb{E}^k$.

Definition 23.1.10. A *norm* on a vector space V is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies the following four properties:

1. For any $\mathbf{v} \in V$, $\|\mathbf{v}\| \geq 0$.
2. $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.
3. For any scalar α and $\mathbf{v} \in V$, $\|\alpha\mathbf{v}\| = |\alpha| \cdot \|\mathbf{v}\|$.
4. For any vectors $\mathbf{u}, \mathbf{v} \in V$,

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|. \tag{23.1}$$

The inequality in (23.1) is called the "triangle inequality". (If vector addition is drawn with arrows tip to tail, the picture corresponds to the fact that the sum of lengths of any two sides of a non-degenerate triangle is greater than the length of the third.)

Definition 23.1.11. Let V be a vector space with an inner product $\langle \cdot, \cdot \rangle$. The *norm with respect to the inner product* of a vector $\mathbf{v} \in V$ is

$$\|\mathbf{v}\| = \langle \mathbf{v}, \mathbf{v} \rangle^{1/2}.$$

If $V = \mathbb{R}^n$ with the standard Euclidean dot product, and $\mathbf{v} = (v_1, v_2, \ldots, v_n) \in V$, then

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}.$$

The norm of a vector is sometimes called its length (or magnitude). The *distance* between two vectors $\mathbf{u}$ and $\mathbf{v}$ in a normed space is $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$. A vector $\mathbf{v}$ in a normed space is called a *unit vector* if and only if $\|\mathbf{v}\| = 1$.

Definition 23.1.12. If V is a vector space with an inner product $\langle \cdot, \cdot \rangle$ two vectors $\mathbf{u}, \mathbf{v} \in V$ are called *orthogonal* if and only if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

Orthogonal vectors are sometimes called "perpendicular". The notation $\mathbf{u} \perp \mathbf{v}$ is sometimes used to denote when $\mathbf{u}$ is orthogonal to $\mathbf{v}$. Two non-zero vectors are said to be *parallel* if one is a scalar multiple of the other.

It is not difficult to see that in any inner product space V, the zero vector is orthogonal to any vector in V.

A basis B for a vector space with an inner product is called *orthogonal* if and only if any two distinct vectors in B are orthogonal. An *orthonormal basis* is an orthogonal basis consisting of unit vectors.

Definition 23.1.13. Given an inner product space V and a subset $W \subseteq V$, the *orthogonal complement of* W *in* V is

$$W^{\perp} = \{\mathbf{v} \in V : \text{ for every } \mathbf{w} \in W, \mathbf{v} \perp \mathbf{w}\}.$$

Theorem 23.1.14 (Cauchy–Schwarz inequality)**.** *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be vectors in an inner product space. Then*

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|. \tag{23.2}$$

Proof: If either of $\mathbf{u}$ or $\mathbf{0}$, both sides of (23.2) are 0, giving equality. So assume that $\mathbf{u}$ and $\mathbf{v}$ are non-zero vectors. The proof given here involves somewhat of a trick: let $t \in \mathbb{R}$ and examine the vector $\mathbf{u} + t\mathbf{v}$. By properties of inner products,

$$\begin{aligned} 0 &\leq \|\mathbf{u} + t\mathbf{v}\|^2 \\ &= \langle \mathbf{u} + t\mathbf{v}, \mathbf{u} + t\mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, t\mathbf{v} \rangle + \langle t\mathbf{v}, \mathbf{u} \rangle + \langle t\mathbf{v}, t\mathbf{v} \rangle \\ &= \|\mathbf{u}\|^2 + t\langle \mathbf{u}, \mathbf{v} \rangle + t\langle \mathbf{v}, \mathbf{u} \rangle + t^2\langle \mathbf{v}, \mathbf{v} \rangle \\ &= \|\mathbf{u}\|^2 + 2t\langle \mathbf{u}, \mathbf{v} \rangle + +t^2\|\mathbf{v}\|^2. \end{aligned}$$

Putting $a = \|\mathbf{v}\|^2$, $b = 2\langle \mathbf{u}, \mathbf{v} \rangle$ and $c = \|\mathbf{u}\|^2$, the above inequality becomes

$$0 \leq at^2 + bt + c,$$

where $a > 0$. By the quadratic formula, if $at^2 + bt + c = 0$, then $t = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$. Since $0 \leq at^2 + bt + c$, the graph of the corresponding parabola opens upward and intersects the x-axis at most once, and so $b^2 - 4ac \leq 0$. Thus $b^2 \leq 4ac$, which says

$$4\langle \mathbf{u}, \mathbf{v}\rangle^2 \leq 4\|\mathbf{v}\|^2 \cdot \|\mathbf{u}\|^2.$$

Cancelling the 4 and taking positive square roots gives (23.2). $\square$

In some texts the Cauchy–Schwarz inequality is given as $\langle \mathbf{u}, \mathbf{v}\rangle \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|$, which follows from (23.2) since $\langle \mathbf{u}, \mathbf{v}\rangle \leq |\langle \mathbf{u}, \mathbf{v}\rangle|$.

In $\mathbb{R}^k$ (with the standard dot product) if two vectors are $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_k)$, then the Cauchy–Schwarz inequality becomes (after squaring each side)

$$\left(\sum_{i=1}^{k} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{k} x_i^2\right)\left(\sum_{i=1}^{k} y_i^2\right).$$

In $\mathbb{R}^2$ or $\mathbb{R}^3$, the notion of an angle is familiar, however, angles in arbitrary inner product spaces are also defined. Note that by the Cauchy-Schwarz inequality,

$$-1 \leq \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|} \leq 1.$$

Definition 23.1.15. For non-zero vectors $\mathbf{u}$ and $\mathbf{v}$ in some inner product space, define the *angle between* $\mathbf{u}$ *and* $\mathbf{v}$ to be the unique $\theta \in [0, \pi]$ satisfying

$$\cos(\theta) = \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|}.$$

Observe that $\theta \in [0, \pi/2]$ if and only if $\langle \mathbf{u}, \mathbf{v}\rangle \geq 0$.

23.2 Some vector geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$

For this section, $\mathbb{R}^2$ and $\mathbb{R}^3$ are considered as the Euclidean vector spaces $\mathbb{E}^2$ and $\mathbb{E}^3$. In either, the origin is denoted by either a capital O or $\mathbf{0}$. A point P has the same coordinates as the vector $\overrightarrow{OP}$, and so points can be considered as vectors. (Sometimes to emphasize the role as a vector, a boldface $\mathbf{P}$ might be used.)

If $X = (x_1, x_2)$ and $Y = (y_1, y_2)$ are points (or vectors) in $\mathbb{R}^2$, either XY or $\overrightarrow{XY}$ is used to denote the vector from X to Y and is calculated coordinate-wise: $XY = Y - X = (y_1, y_2) - (x_1, y_2) = (y_1 - x_1, y_2 - x_2)$. One must take care to clarify notation, because, for example, the notation XY might indicate a vector, whereas in a geometry without mention of vectors, it could denote the line segment $\overline{XY}$, or it could denote the product of numbers or matrices. Some authors do not use boldface letters for vectors, which can make interpreting some expressions rather challenging. If $\mathbf{u}$ and $\mathbf{v}$ are interpreted as vectors, the vector from $\mathbf{u}$ to $\mathbf{v}$ is denoted by either $\mathbf{uv}$ or $\mathbf{v} - \mathbf{u}$. (There are two types of products for vectors and neither of these use the notation $\mathbf{uv}$, so the meaning of $\mathbf{uv}$ can only be the vector from $\mathbf{u}$ to $\mathbf{v}$.)

23.2.1 Lines in $\mathbb{R}^2$ and $\mathbb{R}^3$

A line in $\mathbb{R}^2$ or $\mathbb{R}^3$ can be represented in a number of ways.

First, lines in $\mathbb{R}^2$ are considered. Recall that if (x_1, y_1) and (x_2, y_2) are points on a line ℓ in $\mathbb{R}^2$, the slope of ℓ is

$$m = \frac{\text{rise}}{\text{run}} = \frac{y_2 - y_1}{x_2 - x_1}$$

when the line is not vertical, and the slope is said to be ∞ when the line is vertical (in which case $x_1 = x_2$). A line in $\mathbb{R}^2$ with slope m containing a point (x_0, y_0) can be given by

$$\left\{ (x, y) : \frac{y - y_0}{x - x_0} = m \right\},$$

called the *point-slope form*.

For example, the line with slope 2 passing through the point $(6, 7)$ has point-slope equation

$$\frac{y - 7}{x - 6} = 2. \tag{23.3}$$

By solving the point-slope form for y, any non-vertical line in $\mathbb{R}^2$ can be written in *slope-intercept* form $y = mx + b$, where m is the slope and b is the y-intercept. For example, the slope-intercept form (23.3) yields $y - 7 = 2(x - 6)$ and so $y = 2x - 5$, where the y-intercept -5 can also be found by setting $x = 0$.

If a line is vertical, it can be represented by $x = c$, where c is a constant.

Let $x_1 \neq x_2$ and suppose that (x_1, y_1) and (x_2, y_2) are two points on a line ℓ. Then another way to write the point-slope form of the equation for ℓ is is $\frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1}$.

Another way to represent a line in $\mathbb{R}^2$ is called the "point-vector" equation. One advantage of this representation is that the slope is not explicitly used, and so is a strong candidate to be generalized to three (or more) dimensions. For any line ℓ in $\mathbb{R}^2$, a "direction vector" is any vector parallel to ℓ. For example, the line ℓ given by $y = 2x + 5$ contains points $P = (0, 5)$ and $Q = (1, 7)$; the vector from P to Q is $\mathbf{w} = \overrightarrow{PQ} = Q - P = (1, 2)$ is a direction vector for ℓ. Any point on ℓ can be found by starting at P, say, and adding an appropriate multiple of $\mathbf{w}$. Inn other words, line ℓ is the set of points (vectors)

$$\{P + t\mathbf{w} : t \in \mathbb{R}\}.$$

To each point on ℓ there is a unique $t \in \mathbb{R}$.

So for any line containing $\mathbf{u}$ and having direction vector $\mathbf{w}$, points on the line are of the form $\mathbf{v} = \mathbf{u} + t\mathbf{w}$ for some $t \in \mathbb{R}$. This *point-vector* notation is used here for 2 and 3 dimensions, but in fact, can describe a line in any $\mathbb{R}^k$.

Example 23.2.1. *(Intersecting lines in $\mathbb{R}^3$) If ℓ_1 is a line in $\mathbb{R}^3$ given by*

$$(x, y, z) = (1, 2, 0) + s(-1, 4, 6)$$

and if ℓ_2 is a line given by

$$(x, y, z) = (0, 0, 2) + t(-1, 7, 8).$$

Do ℓ_1 and ℓ_2 intersect?

If so, there is (x, y, z) that satisfies both equations simultaneously. (Note: different parameters s and t are used since they may vary independently.) Examining the three coordinates, one needs s and t so that

$$\begin{aligned} 1 - s &= 0 - t; \\ 2 + 4s &= 0 + 7t; \\ 0 + 6s &= 2 + 8t. \end{aligned}$$

There are different ways to solve this system. One can move all variables to one side and solve by Guass–Jordan elimination, or one can just solve a system of two of these equations and see if the result satisfies the third. It turns out that there is the unique solution $s = 3$ and $t = 2$, both giving the point $(-2, 14, 18)$, so the lines indeed intersect. $\square$

Two lines in $\mathbb{R}^3$ that do not intersect and are not parallel are called *skew lines*.

In $\mathbb{R}^3$, a line ℓ has point-vector equation

$$(x, y, z) = (x_0, y_0, z_0) + t(a, b, c),$$

where (x_0, y_0, z_0) is a point on ℓ and (a, b, c) is a vector parallel to ℓ (where x_0, y_0, z_0, a, b, c are constants and t is a parameter, allowed to vary over all real numbers). Examining the three coordinates of the point-vector equation gives what are called *parametric equations*:

$$\begin{aligned} x &= x_0 + at, \\ y &= y_0 + bt, \\ z &= z_0 + ct. \end{aligned}$$

Solving each of the above three equations for t gives what are called *symmetric equations* (when defined):

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}.$$

Recall from high school that if a line ℓ in $\mathbb{R}^2$ has slope m, then any line perpendicular to ℓ has slope $\frac{-1}{m}$. This fact can also be achieved by vectors and dot products as follows. A line with slope m has direction vector $(1, m)$. Any vector (r, s) that is orthogonal to $(1, n)$ satisfies $(1, m) \bullet (r, s) = 0$, which says $r + ms = 0$ and so $s = -r/m$. Then the vector $(r, s) = (r, -r/m)$ has slope $-1/m$. $\square$

23.2.2 Projection in $\mathbb{R}^2$ or $\mathbb{R}^3$

Projections are based on the following fact: if $\mathbf{u}$ and $\mathbf{v}$ are vectors in an inner product space, then $\mathbf{u}$ can be expressed as a sum $\mathbf{u} = \mathbf{v}_0 + \mathbf{w}$, where $\mathbf{v}_0$ is parallel to $\mathbf{v}$ and $\mathbf{w}$ is orthogonal to $\mathbf{v}$ (see Figure 23.2.2.) The vector $\mathbf{v}_0$ is called the projection of $\mathbf{u}$ onto $\mathbf{v}$, denoted $\mathbf{v}_0 = \mathrm{proj}_{\mathbf{v}}\mathbf{u}$.

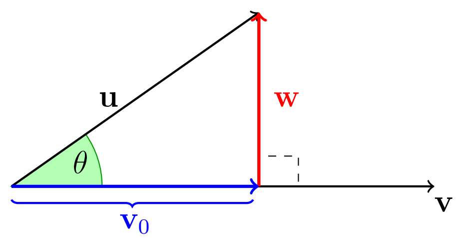


Figure 23.1: Projecting $\mathbf{u}$ onto $\mathbf{v}$

Theorem 23.2.2. *For non-zero vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in an inner product space, the projection of* $\mathbf{u}$ *onto* $\mathbf{v}$ *is*

$$\mathit{proj}_{\mathbf{v}}\mathbf{u} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2}\mathbf{v}. \tag{23.4}$$

Proof: Let θ be the angle between $\mathbf{u}$ and $\mathbf{v}$. If $\theta \in [0, \pi/2]$,

$$\|\mathrm{proj}_{\mathbf{v}}\mathbf{u}\| = \cos(\theta)\|\mathbf{u}\| = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|}\|\mathbf{u}\| = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|}.$$

Since $\frac{1}{\|\mathbf{v}\|}\mathbf{v}$ is a unit vector in the direction of $\mathbf{v}$ and so of $\mathrm{proj}_{\mathbf{v}}\mathbf{u}$, it follows that

$$\mathrm{proj}_{\mathbf{v}}\mathbf{u} = \|\mathrm{proj}_{\mathbf{v}}\mathbf{u}\| \left(\frac{1}{\|\mathbf{v}\|}\mathbf{v}\right) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|} \left(\frac{1}{\|\mathbf{v}\|}\mathbf{v}\right) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2}\mathbf{v}.$$

Similarly, when $\theta \in (\pi/2, \pi]$, $\cos(\theta) < 0$ and so $\|\mathrm{proj}_{\mathbf{v}}\mathbf{u}\| = \frac{-\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|}$ and $\frac{-1}{\|\mathbf{v}\|}\mathbf{v}$ is a unit vector in the direction of $\mathrm{proj}_{\mathbf{v}}\mathbf{u}$. Thus,

$$\mathrm{proj}_{\mathbf{v}}\mathbf{u} = \|\mathrm{proj}_{\mathbf{v}}\mathbf{u}\| \left(\frac{1}{\|\mathbf{v}\|}\mathbf{v}\right) = -\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|} \left(\frac{-1}{\|\mathbf{v}\|}\mathbf{v}\right) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2}\mathbf{v}.$$

□

23.2.3 The cross product

Definition 23.2.3. Let $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$ be vectors in $\mathbb{R}^3$. The *cross product* of $\mathbf{u}$ with $\mathbf{v}$ is the vector

$$\mathbf{u} \times \mathbf{v} = (u_2v_3 - u_3v_2, u_3v_1 - u_1v_3, u_1v_2 - u_2v_1). \tag{23.5}$$

One way to remember (23.5) is by using determinants:

$$\mathbf{u} \times \mathbf{v} = \left(\begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix}, - \begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix}, \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \right).$$

Using standard unit basis vectors $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$, and $\mathbf{e}_3 = (0, 0, 1)$, some authors abuse notation and use

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix},$$

to remember the formula for the cross product (where the "determinant" is expanded along the first row).

It is left to the reader to verify the following properties of the cross product:

- $\mathbf{u} \times \mathbf{v}$ is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$;
- $\mathbf{v} \times \mathbf{u} = -\mathbf{u} \times \mathbf{v}$;
- If $\mathbf{u}$ and $\mathbf{v}$ are parallel, then $\mathbf{u} \times \mathbf{v} = (0, 0, 0)$;
- For any scalar k, $(k\mathbf{u}) \times \mathbf{v} = k(\mathbf{u} \times \mathbf{v})$; For any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$,
$$\mathbf{u} \times (\mathbf{v} + \mathbf{w}) = (\mathbf{u} \times \mathbf{v}) + (\mathbf{u} \times \mathbf{w}).$$

The following is a useful fact:

Theorem 23.2.4. *If $\theta \in [0, \pi]$ is the angle between $\mathbf{u}$ and $\mathbf{v}$, then $\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\| \cdot \|\mathbf{v}\| \sin(\theta)$*

Proof outline: One can verify (with a tedious calculation) that

$$\|\mathbf{u} \times \mathbf{v}\|^2 = \|\mathbf{u}|^2 \|\mathbf{v}\|^2 - (\mathbf{u} \bullet \mathbf{v})^2.$$

It follows that

$$\|\mathbf{u} \times \mathbf{v}\|^2 = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 (1 - \cos^2(\theta)) = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \sin^2(\theta),$$

and taking positive square roots yields the desired result. $\square$

So Theorem 23.2.4 says that the magnitude of the cross product is the area of the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$. Thus, the area of the triangle determined by $\mathbf{u}$ and $\mathbf{v}$ is $\frac{1}{2}\|\mathbf{u} \times \mathbf{v}\|$.

Note that the cross product is only defined in $\mathbb{R}^3$. However, the cross product can be adapted to apply in $\mathbb{R}^2$ for various computations by adding 0 as a third coordinate. For example, to find the area A of the parallelogram determined by the two vectors $\mathbf{u} = (2,3)$ and $\mathbf{v} = (-1,5)$. Then

$$A = \|(2,3,0) \times (-1,5,0)\| = \|(0,0,13)\| = 13.$$

To double check this answer, let $\theta \in [0,\pi]$ be the angle between $\mathbf{u}$ and $\mathbf{v}$. Then $\cos(\theta) = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\cdot\|\mathbf{v}\|} = \frac{13}{\sqrt{13}\sqrt{26}} = \frac{1}{\sqrt{2}}$. Thus $\sin(\theta) = \frac{1}{\sqrt{2}}$ and so

$$\begin{aligned} A &= \sin(\theta)\|\mathbf{u}\| \cdot \|\mathbf{v}\| \\ &= \frac{1}{\sqrt{2}}\sqrt{13}\sqrt{26} \\ &= 13. \end{aligned}$$

There are modified "cross products" in other dimensions, but such products are not discussed here.

23.2.4 Scalar triple products

Definition 23.2.5. For vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ in $\mathbb{R}^2$, define the real number

$$(\mathbf{u} \times \mathbf{v}) \bullet \mathbf{w}$$

to be the *scalar triple product* of $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$.

Observe that the scalar triple product changes sign if $\mathbf{u}$ and $\mathbf{v}$ are reversed; however, the dot product with $\mathbf{w}$ can be written on either side.

Lemma 23.2.6. *Let a parallelepiped P be determined by three linearly independent vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^3$ Then*

$$vol(P) = |(\mathbf{u} \times \mathbf{v}) \bullet \mathbf{w}|.$$

Proof: Without loss of generality, let the scalar triple product be positive (if not, reverse the order of $\mathbf{u}$ and $\mathbf{v}$). let $\mathbf{u}$ and $\mathbf{v}$. Furthermore, put $\mathbf{u} \times \mathbf{v} = \mathbf{n}$.

assume that $\mathbf{u}$ and $\mathbf{v}$ determine the base of the parallelepiped, in which case, the area of the base is $\|\mathbf{n}\| = |\mathbf{u} \times \mathbf{v}|$.

The volume of P is the area of its base times the height measured along $\mathbf{n} = \mathbf{u} \times \mathbf{v}$ (which is orthogonal to the base). The height is

$$h = \|\text{proj}_{\mathbf{n}} \mathbf{v}\| = \frac{|\mathbf{n} \bullet \mathbf{v}|}{\|\mathbf{n}\|}.$$

Hence the volume of P is $|\mathbf{n} \bullet \mathbf{v}| = (\mathbf{u} \times \mathbf{v}) \bullet \mathbf{v}$ as desired. $\square$

Finding the scalar triple product can be done by computing the determinant of matrix. If $\mathbf{u} = (u_1, u_2, u_3)$, $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$, the reader can verify that

$$(\mathbf{u} \times \mathbf{v}) \bullet \mathbf{w} = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}.$$

The above representation agrees with the interpretation of the determinant (or its absolute value) of a 3×3 matrix as a volume in 3 dimensions. (Some authors define the determinant as "volume" in other dimensions.) Note that if the three vectors are not linearly independent, the scalar triple product is zero. From the matrix interpretation, one sees that the scalar triple products of three vectors in any order differ only by the sign.

23.2.5 Planes in $\mathbb{R}^3$

From high school, a plane in $\mathbb{R}^3$ can be represented by an equation $ax + by + cz + d = 0$ (or $ax + by + cz = d'$, where $d' = -d$).

Example 23.2.7. *Consider the equation $3x + 2y - 4y = 1$. To find a point (x_0, y_0, z_0) on this plane, one can arbitrarily choose x and y and solve for z. For example, with $x = 0$ and $y = 0$, find that $z = -1/4$, showing that $(x_0, y_0, z_0) = (0, 0, -1/4)$ is a point on the plane. Similarly, if $x = 2$ and $z = 3$, then $6 + 2y - 12 = 1$ gives $y = \frac{7}{2}$, and so $(2, \frac{7}{2}, 3)$ is a point on the plane.*

Another way to describe a plane is by way of a point and a "normal" vector. Let $\mathbf{n} = (a, b, c)$ be a non-zero vector and let $P = (x_0, y_0, z_0)$ be a fixed point in $\mathbb{R}^3$. Then the set of points $Q = (x, y, z)$ so that $\overrightarrow{PQ}$ is orthogonal to $\mathbf{n}$ forms a plane. In this case, $\mathbf{n}$ is called a normal to the plane containing P. Writing $\overrightarrow{PQ} \bullet \mathbf{n} = 0$ gives $(x - x_0, y - y_0, z - z_0) \bullet (a, b, c) = 0$, and so

$$a(x - x_0) + b(y - y_0) + c(y - y_0) = 0, \tag{23.6}$$

which is called the *point-normal* form for the equation of a plane.

From Example 23.2.7, using $\mathbf{n} = (a, b, c) = (3, 2, -4)$ and $P = (2, \frac{7}{2}, 3)$, the point-normal form of the equation $3x + 2y - 4y = 1$ is

$$3(x - 2) + 2\left(y - \frac{7}{2}\right) - 4(z - 3) = 0.$$

If $ax + by + cz = d$ is a plane and R is some point not on the plane, how does one find the distance from R to the plane? This distance is measured along a vector orthogonal to the plane, so pick any point Q on the plane and project $\overrightarrow{RQ}$ onto the normal vector $\mathbf{n} = (a, b, c)$ and take the magnitude of this projection.

Example 23.2.8. *Let $2x + y + 5z = -7$ be the equation of a plane and let $R = (4, -1, 3)$ be a point not on the plane. Picking $Q = (1, 1, -2)$ on the plane,*

$$proj_{\mathbf{n}} \overrightarrow{RQ} = \frac{(-3, 2, -5) \bullet (2, 1, 5)}{\|(2, 1, 5)\|^2}(2, 1, 5) = \frac{-29}{30}(2, 1, 5).$$

The distance from R to the plane is the magnitude of this projection, namely $\frac{29}{\sqrt{30}}$. Furthermore, the point S on the plane closest to R can be computed by adding the projected vector to $R = \overrightarrow{OR}$:

$$S = (4, -1, 3) + \frac{-29}{30}(2, 1, 5) = \left(\frac{62}{30}, \frac{-59}{30}, \frac{-55}{30}\right).$$

As a check to see if S lies on the plane,

$$2 \cdot \frac{62}{30} + 1 \cdot \frac{-59}{30} + 5 \cdot \frac{-55}{30} = \frac{124 - 59 - 55}{30} = -7,$$

as desired.

Consider two planes $a_1x + b_1y + c_1z = d_1$ and $a_2x + b_2y + c_2z = d_2$. If $\mathbf{n}_1 = (a_1, b_1, c_1)$ is parallel to $\mathbf{n}_2 = (a_2, b_2, c_2)$, then either the planes are the same or they are parallel. If these two normals are not multiples of one another, then the planes intersect in a line (with direction vector $\mathbf{n}_1 \times \mathbf{n}_2$). When two different planes intersect, two "dihedral" angles are formed between them. (Such dihedral angles are formed in a plane orthogonal to both planes.) How does one compute these dihedral angles?

If α is the angle between the normals, and β is a dihedral angle, then a quadrilateral is formed with angles α, $\pi/2$, β, and $\pi/2$. Since the sum of the angles in a convex quadrilateral is 2π, it follows that $\alpha + \beta = \pi$. In this case, α is also a dihedral angle (the supplementary angle of β). If $\alpha > \pi/2$, then β is the smaller of the two dihedral angles; if $\alpha < \pi/2$ then α is the smaller of the two dihedral angles. So it suffices to calculate

$$\cos(\alpha) = \frac{\mathbf{n}_1 \bullet \mathbf{n}_2}{\|\mathbf{n}_1\| \cdot \|\mathbf{n}_2\|}.$$

23.2.6 Some review exercises

This section contains only a small sample of geometrical problems in $\mathbb{R}^2$ or $\mathbb{R}^3$ that have solutions using standard vector space techniques.

Exercise 365. *Describe geometrically the intersection of the three planes* $2x - 2y + z = 0$, $3x - y + 2z = 0$, $x + y + z = 0$.

Exercise 366. *Find the point two-thirds of the way from the point* $(3, 4)$ *to the point* $(6, -5)$.

Exercise 367. *Compute the distance between the point* $(2, 1, 3)$ *and the line* $(x, y, z) = (1, 1, 1) + t(2, -2, 1)$ *showing all work and reasoning.*

Exercise 368. *Find the line of intersection of the planes* $x - 4y + 2z = 2$ *and* $2x - 7y + 3z = 5$ *by solving the system, and verify that the direction vector is correct by using a cross product.*

Exercise 369. *Find the line orthogonal to the plane* $x + 3y - 7z = 10$ *that passes through the point* $(0, 6, 6)$.

Exercise 370. *Let* ℓ *be the line given by* $(x, y, z) = (1, 4, 2) + s(0, 3, 2)$ *and let* m *be the line given by* $(x, y, z) = (1, 0, -2) + t(-1, 0, 1)$. *Do these lines intersect? If not find the minimum distance between the lines. Added problem: if the lines do not intersect, find a point on* ℓ *and a point on* m *that are at minimum distance from each other. (This minimum distance is called the* distance between lines.*)*

Exercise 371. *Find the (dihedral) angle between the planes* $-y + z = 3$ *and* $2z = 7 + 2x$.

Bibliography

[*] **Note:** Numbers in the box following each entry are page numbers in this book where the entry is cited.

[1] T. G. Abbott, Z. Abel, D. Charlton, E. D. Demaine, M. L. Demaine, and S. D. Kominers, Hinged dissections exist, *Discrete Comput. Geom.* **47** (2012), 150–186. 243

[2] H. Abbott and A. Liu, Property $B(s)$ and projective planes, *Ars Combin.* **20** (1985), 217–220. 359

[3] B. M. Ábrego, O. Aichholzer, S. Fernández-Merchant, P. Ramos, and G. Salazar, Shellable drawings and the cylindrical crossing number of K_n, *Discrete Comput. Geom.* **52** (2014), 743–753. 435, 436

[4] B. M. Ábrego and S. Fernández-Merchant, A lower bound for the rectilinear crossing number, *Graphs and Combinatorics* **21** (2005), 293–300. 438, 439

[5] B. M. Ábrego, S. Fernández-Merchant, J. Leaños, Jesús; and G. Salazar, The maximum number of halving lines and the rectilinear crossing number of K_n for $n \leq 27$, The IV Latin-American Algorithms, Graphs, and Optimization Symposium, *Electron. Notes Discrete Math.* **30** (2008), 261–266. 438, 439

[6] E. Ackerman, On topological graphs with at most four crossings per edge, manuscript, 29 October 2013, 35 pages. 442

[7] E. Ackerman and R. Pinchasi, On the degenerate crossing number, *Discrete and Computational Geometry* **49** (2013), 695–702. 442

[8] P. Agarwal and B. Aronov, Counting facets and incidences, *Discrete and Computational Geometry* **7** (1992), 359–369. 400

[9] R. Aharoni, personal communication, Emory University, 27 March 1994. 360, 361

[10] O. Aichholzer, *On the rectilinear crossing number*, webpage, `http://www.ist.tugraz.at/staff/aichholzer/crossings.html`, last updated 26 May 2015, (accessed 29 Dec 2017). 439

[11] M. Aigner, Turán's graph theorem, *Amer. Math. Monthly* **102** (1995), 808–816. 424

[12] M. Aigner, *Combinatorial theory,* Classics in Mathematics, Springer, Berlin, 1997. (Reprint of the 1979 edition, originally published by Springer as Vol. 234 of Grundlehren der mathematischen Wissenschaften). 503

[13] M. Aigner and G. Ziegler, *Proofs from the book*, 3rd ed., Springer, 2004. [As of 2018, this highly praised book is now in its sixth edition!] 247, 249, 251, 253, 422, 423, 424, 441

[14] M. Ajtai, V. Chvátal, M. Newborn, and E. Szemerédi, Crossing-free subgraphs, in *Theory and practice of combinatorics: a collection of articles honoring Anton Kotzig on the occasion of his sixtieth birthday* (eds. P. L. Hammer, A. Rosa, G. Sabidussi, J. Turgeon), *Ann. Discrete Math.* **12** (1982), 9–12. 441, 442

[15] V. A. Aksenov, On continuation of 3-colouring of planar graphs (in Russian), *Diskret. Anal. Novosibirsk* **26** (1974) 3–19. 403

[16] R. M. Alain, *A course in p-adic analysis*, Graduate Texts in Mathematics **198**, Springer, 2000. 253

[17] A. A. Albert and R. Sandler, *An introduction to finite projective planes*, Dover 2015. [Republication of 1968 edition originally published by Holt, Rinehart and Winston, Inc.] 315, 316

[18] J. M. Aldous and R. J. Wilson, *Graphs and applications: an introductory approach*, Springer, 2000. 147

[19] G. L. Alexanderson, L. F. Klosinski and L. C. Larson, eds., *The William Lowell Putnam mathematical competition. Problems and solutions: 1965-1984*, Mathematical Association of America, Washington, D.C., 1986. 14

[20] R. B. J. T. Allenby and A. Slomson, *How to count: An Introduction to combinatorics*, 2nd ed., CRC Press, 2011. 760

[21] N. Alon, H. Last, R. Pinchasi, and M. Sharir, On the complexity of arrangements of circles in the plane, *Discrete Comput. Geom.* **26** (2001), 465–492. 385

[22] N. Alon and Z. Füredi, Legitimate colorings of projective planes, *Graphs and combinatorics* **5** (1989), 95–106. 374

[23] N. Alon, L. Rónyai, and T. Szabó, Norm-graphs: variations and applications, *J. Combin. Theory Ser. B* 76 (1999), 280–290. 321

[24] N. Alon and J. Spencer, *The probabilistic method* (2nd ed.), Wily-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc., New York, 2000. 220, 221, 364, 366, 370, 373

[25] C. Alsina and R. B. Nelsen, A visual proof of the Erdős-Mordell inequality, *Forum Geometricorum* **7** (2007), 99–102. 124

[26] S. Anderson, *Squaring.net*, webpage, 2014, `http://www.squaring.net/index.html`. 290, 299, 300, 624

[27] T. Andreescu and R. Gelca, *Putnam and beyond*, Springer, 2007. 14, 685, 696, 711

[28] T. Andreescu, C. Mortici, and M. Tetiva, *Mathematical bridges*, Birkhäuser, 2017. 691

[29] T. Andreescu and S. Savchev, *Mathematical miniatures*, MAA, 2003. 761

[30] G. Antonick, *Numberplay: Bugs on a square*, New York Times blog post, https://wordplay.blogs.nytimes.com/2014/09/08/bugs/, 8 September 2014. (Accessed 11 Oct. 2017.) 189

[31] K. Appel and W. Haken, Every planar map is four colorable, Part I: Discharging, *Illinois Journal of Mathematics* **21** (1977), 429–490. 146

[32] K. Appel, W. Haken, and J. Koch, Every planar map is four colorable, Part II: Reducibility, *Illinois Journal of Mathematics* **21** (1977), 491–567. 146

[33] A. Arman, D. S. Gunderson, and S. Tsaturian, *Equally spaced points and simplices in Euclidean Ramsey theory*, in preparation, 11 May 2017. 491

[34] A. Arman and S. Tsaturian, *A result in asymmetric Euclidean Ramsey theory*, *Discrete Math.* **341** (2018), 1502–1508. (Preprint also available at arXiv:1702.04799v1 [math.CO] 15 Feb 2017.) 491

[35] A. Arman and S. Tsaturian, Two-colourings of Euclidean space, preprint, 19 May 2017. 491

[36] M. Artebani and I. Dolgachev, *The Hesse pencil of plane cubic curves*. Available at www.math.lsa.umich.edu/~idolga/hesse09.pdf, accessed 26 July 2017. 393, 394

[37] E. Artin, *Galois theory*, 2nd ed., Notre Dame Mathematical Lectures, number 2, Universtiy Notre Dame, 1944. [Printed by Edwards Brothers, Inc., Ann Arbor, Michigan.] 104

[38] L. Auslander, An account of the theory of crystallographic groups, *Proc. Amer. Math. Soc.* **16** (1965), 1230–1236. 80

[39] L. Babai and P. Frankl, *Linear algebra methods in combinatorics with applications to geometry and computer science*, Preliminary version 2, September 1992, University of Chicago. (Available at www.matthewkahle.org/download/file/fid/499.) 360

[40] R. Baer, Polarities in finite projective planes, *Bulletin of the American Math. Soc.* **52** (1946), 77–93. 344, 345, 346, 347

[41] J. Baez, *The Fano plane*, website, `http://math.ucr.edu/home/baez/octonions/node4.html`, 2001. (Last accessed December 2017). 315

[42] D. Bailey and J. Borwein, *Mathematics by experiment: Plausible reasoning in the 21st century*, 2nd ed., A. K. Peters, Ltd., Wellesley, Massachusetts. 69

[43] A. Bálintová and V. Bálint, On the number of circles determined by n points in the Euclidean plane, *Acta Math. Hungar.* **63** (1994), 283–289. 385

[44] A. Baker, Squaring the circle; transcendence, logarithmic forms and Diophantine analysis, *The Mathematical Gazette* **80**, no. 487, (March 1996), 71–77. 104

[45] K. Ball, *An elementary introduction to modern convex geometry*, MSRI Publications **31**, 1997, 58 pages. Available at `http://library.msri.org/books/Book31/files/ball.pdf`; also reprinted as the first chapter in [611]. 167, 193

[46] S. Ball, A. Blokhuis, and F. Mazzocca, Maximal arcs in Desarguesian planes of odd order do not exist, *Combinatorica* **17** (1997), 31–41. 343

[47] J. Balogh, B. Lidický, and G. Salazar, Closing in on Hill's conjecture, preprint available at `arXiv:1711.089558v1` [math.Co], 27 November, 2017 (accessed December 2017). 437

[48] S. Banach and A. Tarski, Sur la decomposition des ensembles de points en parties respectivement congruents, *Fund. Math.* **6** (1924), 244–277. 261, 262

[49] L. Bankoff, An elementary proof of the Erdős–Mordell theorem, *Amer. Math. Monthly* **65**, 521. 124

[50] R. B. Bapat, A constructive proof of a permutation-based generalization of Sperner's lemma, *Mathematical Programming* **44** 1989), 113–120. 253

[51] A. Baragar, *A survey of classical and modern geometries: with computer activities*, Prentice Hall, 2001. 370 pages. 305, 315, 316

[52] I. Bárány, A short proof of Kneser's conjecture, *J. Combin. Theory Ser. A* **25** (1978), 325–326. 213

[53] I. Bárány, A generalization of Carathéodory's theorem, *Discrete Math.* **40** (1982), 141–152. 202

[54] I. Bárány, A note on Sylvester's four point problem, *Studia Scientiarum Mathematicarum Hungarica* **38** (2001), 73–77. (Available at `www.renyi.hu/~barany/cikkek/83.pdf`.) 115, 183

[55] I. Bárány and P. Frankl, How (not) to cut your cheese, *Amer. Math. Monthly* **128** (May 2021), 543–552. 401

[56] I. Bárány and P. Valtr, A positive fraction Erdős–Szekeres theorem, *Discrete Comput. Geom.* **19** (1998), 335–342. 228

[57] J.-É Barbier, Note sur le probléme de l'aiguille et le jeu du joint couvert, *Journal des Mathématiques Pures et Appliquées, 2nd ser.* **5** (1860), 273–286. 182

[58] G. Barequet, A lower bound for Heilbronn's triangle problem in d dimensions, *SIAM J. Discrete Math.* **14** (2001), 230–236. 259

[59] G. Barequet, The on-line Heilbronn's problem, *Discrete Math.* **283** (2004), 7–14. 259

[60] D. Barnette, Trees in polyhedral graphs, *Canad. J. Math.* **18** (1966), 731–736. 153

[61] D. F. Barrow, Solution to Problem 3740, *Amer. Math. Monthly* **44** (1937), 252–254. 124

[62] A. Barvinok, The beauty and mystery of the symmetric moment curve, pp. 19–22 in *Discrete geometry and convexity in honour of Imre Bárány*, G. Ambrus, K. Böröcky, and Z. Füredi,

eds., Proceedings of conference "Discrete geometry and convexity Bárány 70", 19–23 June 2017, Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest, published by Typotex Electronical Publishing Ltd., Hungary, 2017. 205, 206

[63] H. Bateman, *Differential equations*, Longmans, Green and Co., London, 1918. 188

[64] L. M. Batten, A dual approach to embedding the complement of two lines in a finite projective plane, *J. Austral. Math. Soc. Ser. A* **51** (1991), 426–435. 355

[65] L. M. Batten, *Combinatorics of finite geometries*, 2nd ed., Cambridge University Press, Cambridge, 1997. 353

[66] A. Beck and M. N. Bleicher, Packing convex sets into a similar set, *Acta Mathematica Academiae Scientiarum Hungaricae* **22** (1971), 283–303. 128, 708

[67] A. Beck, M. N. Bleicher, and D. W. Crowe, *Excursions into mathematics*, A. K. Peters, Natick. Mass., 2000. 177

[68] J. Beck, On the lattice property of the plane and some problems of Dirac, Motzkin and Erdős in combinatorial geoemetry *Combinatorica* **3** (1983), 281–297. 399

[69] J. Beck and T. Fiala, "Integer making" theorems, *Discrete Applied Math.* **3** (1981), 1–8. 366

[70] J. Beck and V. Sós, Discrepancy theory, in *Handbook of Combinatorics*, (Vol. II, 1405–1446), R. L. Graham, M. Grötschel, L. Lovász, eds, Elsevier Science B. V., and MIT Press, 1995 364, 370, 373

[71] M Beermann and M. Reitzner, Monotonicity of functionals of random polytopes, pp. 23–28 in *Discrete geometry and convexity in honour of Imre Bárány*, G. Ambrus, K. Böröcky, and Z. Füredi, eds. (Proc. of conference "Discrete geometry and convexity Bárány 70", 19–23 June 2017, Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest),

published by Typotex Electronical Publishing Ltd., Hungary, 2017. 184

[72] L. Beineke and R. Wilson, The early history of the brick factory problem, *The Mathematical Intelligencer* **32** (June 2010), 41–48. 433, 435, 437, 756

[73] M. Benda and M. Perles, Coloring of metric spaces, *Geombinatorics* **9** (2000), 113–126. 464

[74] L. Berardi and F. Eugeni, Blocking sets in the projective plane of order four, *Annals of Discrete Mathematics* **37** (1988), 43–50. 352

[75] L. Berardi and S. Innamorati, Irreducible blocking sets in the projective plane of order five, *Atti Sem. Mat. Fis. Univ. Modena* **38** (1990), no. 2, 293–311. 355

[76] E. Berlekamp and J. P. Buhler, Puzzles column, *Emissary*, MSRI, Spring 2014, page 15. 703

[77] C. Bertram-Kretzberg, T. Hofmeister, and H. Lefmann, An algorithm for Heilbronn's problem, *SIAM J. Comput.* **30** (2000), 383–390. 257, 259

[78] A. Besicovitch, Sur deux questions d'integrabilite des fonctions, *J. Soc. Phys. Math.* **2** (1920), 105–123. 500

[79] A. S. Besicovitch, On Kakeya's problem and a similar one, *Mathematische Zeitschrift* **27** (1928), 312–320. 500

[80] A. Beutelspacher and U. Rosenbaum, *Projective geometry: from foundations to applications*, Cambridge University Press, 1998. 303, 304

[81] W. A. Beyer and A. Zardecki, (2004), The early history of the ham sandwich theorem, *American Mathematical Monthly* **111** (2004), 58–61, 413

[82] A. Bezdek, On the density of dual circle coverings, *Geometriae Dedicata* **33** (1990), 227–238. 384, 747

[83] A. Bezdek, On the intersection points of unit circles, *Amer. Math. Monthly* **99** (1992), 779–780. 385

[84] A. Bezdek, Incidence problems for points and unit circles, in *Paul Erdős and His Mathematics* (A. Sali, M. Simonovits and V. T. Sós, Eds.), J. Bolyai Math. Soc., Budapest, 1999, pp. 33–36. 384, 385

[85] A. Bezdek, On unit circles which avoid all but two points of a given point-set, *European J. Combin.* **23** (2002) 11–13. 385

[86] A. Bezdek, F. Fodor, and I. Talata, Sylvester-type theorems for unit circles, *Discrete Math.* **241** (2001) 97–101. 385

[87] K. Bezdek and R. Connelly, Intersection Points, *Annales Univ. Sci. Budapest, Sect. Math.* **31** (1988), 115–127. 384

[88] M. Bhargava, K. Kedlaya, and L. Ng, , and *Solutions to the 82nd William Lowell Putnam Mathematical Competition, Saturday, December 4, 2021*, `https://kskedlaya.org/putnam-archive/`, accessed 13 Dec. 2021. 720

[89] A. Bialostocki, P. Dierker, and B. Voxman, Some notes on the Erdős-Szekeres theorem, *Discrete Math.* **91** (1991), 231–238. 228

[90] A. Bialostocki and M. Nielsen, Minimum sets forcing monochromatic triangles, *Ars. Combin.* **81** (2006), 297–303. 475

[91] D. Bienstock and N. Dean, Bounds for rectilinear crossing numbers, *J. Graph Theory* **17**17 (1993), 333–348. 439

[92] J. Bierbrauer, Blocking sets of 16 points in projective planes of order 10, II, *Quart. J. Math. Oxford Ser. (2)* **36** (1985), 383–391. 355

[93] J. Bierbrauer, Blocking sets of 16 points in projective planes of order 10, III, *Rend. Sem. Math. Univ. Padova* **74** (1985), 163–174. 355

[94] G. D. Birkhoff, *The origin, nature and influence of relativity*, 1925. 500

[95] G. Birkhoff, Combinatorial relations in projective geometries *Annals of Mathematics Second series* **36** (July 1935), 743–748. 506, 507

[96] G. Birkhoff, *Lattice theory*, revised edition, American Mathematical Society Colloquium Publications, **XXV**, 1961. 506

[97] G. Birkhoff, Tres observaciones sobre el algebra lineal, *Univ. Nac. Tucumán Rev, Ser. A* **5** (1946), 147–151. 214

[98] C. Biró, *personal communication*, 5 October 2016. 356

[99] C. Biró, P. Hamburger, and A. Pór, Standard examples as subposets of posets, *Order* **32** (2015), 293–299. 511

[100] C. Biró, P. Hamburger, A. Pór, and W. T. Trotter, Forcing posets with large dimension to contain large standard examples, *Graphs and Combinatorics* **32** (2016), 861–880. Preprint available at arXiv:1402.5113v3 [math.CO] 24 Aug 2015. (Accessed 29 September 2016.) 356, 359, 511, 512

[101] T. Bisztriczky, A construction for periodically-cyclic Gale $2m$-polytopes, *Beiträge zur Algebra und Geometrie (Contributions to Algebra and Geometry)* **42** (2001), 89–101. 210

[102] T. Bisztriczky, Characterizations of cyclic polytopes, *J. Geometry*, **84** (2005), 30–36. 209, 210

[103] T. Bisztriczky, K. Böröczky and D. S. Gunderson, Cyclic polytopes, hyperplanes, and Gray codes, *J. Geom.* **78** (2003), 25–49. 211

[104] T. Bisztriczky and G. Károlyi, Subpolytopes of cyclic polytopes, *Europ. J. Combinatorics* **21** (2000), 13–17. 210

[105] T. Bisztriczky and G. Fejes Tóth, A generalization of the Erdős–Szekeres convex n-gon theorem, *J. Reine Angew. Math.* **395** (1989), 167–170. 228

[106] M. Black, *Critical thinking*, 1946. [The 2nd revised edition, 459 pages, was published by Prentice-Hall, 1980.] 284

[107] J. Blažek and M. Koman, A minimal problem concerning complete plane graphs, pp. 113–117 in *Theory of Graphs and its Applications* (Proc. Sympos. Smolenice, 1963), Publ. House Czechoslovak Acad. Sci., Prague, 1964. 436

[108] H. F. Blichfeldt, A new principle in the geometry of numbers, with some applications, *Trans. Amer. Math. Soc.* **15** (1914), 227–235. 273, 274, 604

[109] A. Blokhuis, Extremal problems in finite geometries, in *Extremal problems for finite sets (Visegrád, 1991)*, 111–135, *Bolyai Soc. Math. Stud.* **3**, János Bolyai Math. Soc., Budapest, 1994. 301, 353, 356

[110] A. Blokhuis and A. E. Brouwer, Blocking sets in Desarguesian projective planes, *Bull. London Math. Soc.* 18 (1986), 132–134. 355

[111] A. Blokhuis and F. Mazzocca, The finite field Kakeya problem, *Bridges Between Mathematics and Computer Science*, Bolyai Society Mathematical Studies **19**, M. Grötschel, G. Katona (Eds.), Springer, 2008, pp. 205–218. (See also `arXiv:0911.4370[math.CO]`; accessed 20 June 2016.) 501

[112] A. Blokhuis and K. Metsch, Large minimal blocking sets, strong representative systems, and partial unitals, in *Finite Geometry and Combinatorics* (A. Beutelspacher *et al.* eds.), proceedings of The Second International Conference at Deinze, London Mathematical Society Lecture Note Series **191**, Cambridge University Press, 1993, 37–52. 355

[113] K. P. Bogart and W. T. Trotter, Maximal dimensional partially ordered sets II. Characterization of $2n$-element posets with dimension n, *Discrete Math.* **5** (1973), 33–43. 510

[114] J. Bokowski and A. Eggert, Toutes les réalisations du tore de Möbius avec sept sommets, *Topologie Struct.* **17** (1991), 59–78. 153

[115] B. Bollobás, On generalized graphs, *Acta Math. Acad. Sci. Hungar.* **16** (1965), 447–452. 360

[116] B. Bollobás, *Combinatorics: set systems, hypergraphs, families of vectors and combinatorial probability*, Cambridge University Press, 1986. 200, 360, 361, 730

[117] B. Bollobás, *Modern graph theory*, Graduate texts in mathematics **184**, Springer, 1998. 135, 140, 220, 288, 292, 429

[118] B. Bollobás, *The art of mathematics: Coffee time in Memphis*, Cambridge University Press, 2006. 55, 190, 199, 377, 709, 715, 730

[119] V. G. Boltianskii, *Hilbert's third problem* (transl. by R. Silverman), Winston, 1978. 260

[120] V. Boltyanski, H. Martini, and P. S. Soltan, *Excursions into combinatorial geometry*, Springer, 1997. 203, 453

[121] F. Bolyai, Tentamen iuventutem studiosam in elementa matheseos purae dementaris ac sublimioris methodo intuitiva evidentiaque huic propria introducendi, cum appendice triplici, (in Latin) *Acad. Scient. Hungar.*, Budapest, 2nd edition, 1832. 232

[122] M. Bóna, A Euclidean Ramsey theorem, *Discrete Math.* **122** (1993), 349–352. 475

[123] M. Bóna and G. Tóth, A Ramsey-type problem on right-angled triangles in space, *Discrete Math.* **150** (1996), 61–67. 459, 475

[124] A. V. Bondarenko, On Borsuk's conjecture for 2-distance sets, *Disc. Comput. Geom.* **51** (2014), 509–515. 467

[125] J. A. Bondy and U. S. R. Murty, *Graph theory with applications*, The Macmillan Press Ltd., and American Elsevier Publishing Co. Inc., New York, 1976. 135, 402, 447, 757, 758

[126] W. E. Bonnice, On convex polygons determined by a finite planar set, *Amer. Math. Monthly* **87** (7) (1974), 749–752. 224, 732

[127] O. V. Borodin, A new proof of Grünbaum's 3 color theorem, *Discrete Mathematics* **169** (1997), 177–183. 403

[128] O. V. Borodin, Z. Dvorak, A. Kostochka, B. Lidický, and M. Yancey, *On a conjecture by Gallai and a question by Erdős*, slides for a talk at *Erdős Centennial*, Budapest, 2013. `www.renyi.hu/conferences/erdos100/slides/kostochka.pdf` (Accessed 16 June 2016) Results in these slides now seem to be published by the same set of authors in "Planar 4-critical graphs with four triangles", *Eur. J. Comb.* **41** (2014), 138–151. (Also available on Kostochka's website). 403

[129] E. Boros, $\mathrm{PG}(2, p^s)$, $p > 2$ has property $\mathrm{B}(p+2)$, *Ars Combinatoria* **25** (1988), 111–113. 360

[130] K. Borsuk, Über Schnitte der n-dimensionalen Euklidischen Räume. (German) *Math. Ann.* **106** (1932), no. 1, 239–248. 467

[131] K. Borsuk, Drei Sätze über die n-dimensionale euklidische Sphäre, *Fund. Math.* **20** (1933), 177-190. (In German) 465, 467

[132] K. Borsuk and S. Ulam, Über gewisse Invarianten der ε-Abbildungen, *Math. Ann.* **108** (1933), no. 1, 311–318. (In German) 465

[133] P. Borwein, On Sylvester's problem and Haar spaces, *Pacific J. Math.* **109** (1983), 275–278. 383

[134] P. Borwein, Sylvester's problem and Motzkin's theorem for countable and compact sets, *Proc. of the Amer. Math. Soc.* **90** (1984), 580–584. 383

[135] P. Borwein and M. Edelstein, A conjecture related to Sylvester's problems, *Amer. Math. Monthly* **90** (1983), 389–390. 383

[136] P. Borwein and W. Moser, A survey of Sylvester's problem and its generalizations, *Aequationes Mathematicae* **40** (1990), 111–135. 376

[137] R. C. Bose, Mathematical theory of the symmetrical factorial design, *Sankhya* **8** (1947), 107–166. 342

[138] R. C. Bose, A note on Fisher's inequality for balanced incomplete block designs, *Annals of Mathematical Statistics* **40** (1949), 619–620. 330

[139] J. Botermans and P. van Delft, *Creative puzzles of the world*, ADM International bv, Amsterdam, 1978. 239

[140] O. Bottema and R. Erne, Triangles with two equal angle bisectors, pp. 1–3 in *Topics in Elementary Geometry*, Springer, New York, NY, 2008 702

[141] C.J. Bowkamp, Packing a rectangular box with the twelve solid pentominoes, *J. Combin. Theory* **7** (1969), 278–280. 287

[142] P. Brass, Erdős distance problems in normed spaces, *Combinatorial Geometry* **6** (1996), 195–214. 453

[143] P. Brass, W. Moser, and J. Pach, *Research problems in discrete geometry*, Springer, 2005. 375, 460

[144] A. Brauer and T. L. Reynolds, On a theorem of Aubry-Thue, *Canad. J. Math.* **3** (1951), 367–374. 526

[145] C. A. Bretschneider, Untersuchung der trigonometrischen Relationen des geradlinigen Viereckes, article XXIII in *Archiv der Mathematik und Physik* **2** (1842), 225—261. 62

[146] H. Brocard, Problem 250, *Nouvelle Correspondance Mathématique* **3** (1877), edited by E. Catalan, Brussels, p. 175. 188

[147] A. Brodsky, S. Durocher, and E. Gethner, The rectilinear crossing number of K_{10} is 62, *Electronic J. Combin.* **8** (2001), #R23, 30 pages. 438

[148] A. Brodsky, S. Durocher, and E. Gethner, Toward the rectilinear crossing number of K_n: new drawings, upper bounds, and asymptotics, *Discrete Math.* **262** (2003), 59–77. 438

[149] U. Bronfenbrenner, (1944). The graphic presentation of sociometric data, *Sociometry* **7** (1944), 283–289. 433

[150] M. Brooke, Fibonacci numbers: their history through 1900, *Fibonacci Quarterly* **2** (April 1964), 149–153. 94

[151] R. L. Brooks, C. A. B. Smith, A. H. Stone, and W. T. Tutte, The dissection of rectangles into squares, *Duke Math. J.* **7** (1940), 312–340. 292

[152] A. E. Brouwer, Block designs, in *Handbook of combinatorics,* R. Graham, M. Grötschel, and L. Lovász, eds., Elsevier Science B. V. (1995), 693–745. 325

[153] A. E. Brouwer and W. H. Haemers, *Spectra of graphs*, Universitext, Springer, 2012. 468

[154] A. E. Brouwer and T. Jenrich, A 64-dimensional counterexample to Borsuk's conjecture, *Electron. J. Combin.* **21** (2014), Paper 4.29, 3 pages. 468

[155] L. Brouwer, Über Abbildungen von Mannigfaltigkeiten, *Math. Annalen* **71** (1912), 97–115. 250

[156] W. G. Brown, On graphs that do not contain a Thomsen graph, *Canad. Math. Bull.* **9** (1966), 281–285. 321

[157] R. H. Bruck, Difference sets in a finite group, *Trans. Amer. Math. Soc.* **78** (1955), 464–481. 316

[158] R. H. Bruck and H. J. Ryser, The non-existence of certain finite projective planes, *Canad. J. Math.* **1** (1949), 88–93. 309, 331

[159] A. A. Bruen, Baer subplanes and blocking sets, *Bull. Amer. Math. Soc.* **76** (1970), 342–344. 353

[160] A. A. Bruen, Blocking sets in finite projective planes, *SIAM J. Appl. Math.* **21** (1971), 380–392. 352, 353, 355, 747

[161] A. Bruen and J. C. Fisher, Blocking sets and complete arcs, *Pacific J. Math.* **53** (1974), 73–84. 359

[162] A. A. Bruen and R. Silverman, Arcs and blocking sets II, *Eur. J. Comb.* **8** (1987), 351–356. 353

[163] A. A. Bruen and J. A. Thas, Blocking sets, *Geom. Dedicata* **6** (1977), 193–203. 352, 355

[164] N. G. de Bruijn, Problem 109, *Matematikai Lapok* **12** (1959), 110–112, 113. 288

[165] N. G. de Bruijn, Problem 119, *Matematikai Lapok* **12** (1961), 314–317. 288

[166] G. Buffon, Editor's note concerning a lecture given 1733 by Mr. Le Clerc de Buffon to the Royal Academy of Sciences in Paris, *Histoire de lÁcad. Roy. des Sci.*,1733, pp. 43–45. 181

[167] G. Buffon, Essai d'arithmétique morale, *Histoire naturelle, générale er particuliére, Supplément* **4** (1777), 100–104. 181, 182

[168] S. A. Burr, B. Grünbaum, and N. J. A. Sloane, The orchard problem, *Geometriae Dedicata* **2** (1974), 497–424. 386

[169] S. Cabello and B. Mohar, Adding one edge to planar graphs makes crossing number and 1-planarity hard, *SIAM J. Comput.* **42**(5) (2013), 1803–1829. 430

[170] P. J. Cameron, Finite geometries, in *Handbook of combinatorics*, R. Graham, M. Grötschel, and L. Lovász, eds., Elsevier Science B. V. (1995), 647–691. 301

[171] P. J. Cameron and J. H. van Lint, *Designs, graphs, codes and their links*, London Mathematical Society Student Texts **22**, Cambridge University Press, Cambridge, 1991. 325

[172] K. Cantwell, *Euclidean Ramsey theory*, dissertation, Arizona State, ProQuest Dissertations Publishing, 1992. 493

[173] K. Cantwell, Edge-Ramsey theory, *Discrete Comput. Geom.* **15** (1996), 341–352. 493

[174] K. Cantwell, Finite Euclidean Ramsey theory, *J. Combin. Theory Ser. A* **73** (1996), 273–285. 461

[175] K. Cantwell, All polytopes are Ramsey, *J. Comb. Theory Ser. A* **114** (2007), 555–562. 484

[176] C. Carathéodory, Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werts nicht annahmen, *Math. Ann.* **64** (1907), 95–115. 198

[177] C. Carathéodory, Über den Variabilitätsbereich der Fourierschen Koeffizienten von Positiven harmonischen Funktionenen, *Rendiconti del Circolo Matimatico di Palermo* **32** (1911), 193–217. 205

[178] B. Casselman, *On the dissecting table*, Henry Perigal 1801–1898, `https://plus.maths.org/content/dissecting-table`, 2000 (accessed 13 April 2016). 236

[179] J.-D. Cassini, Une nouvelle progression de nombres, *Histoire de l'Académie Royale des Sciences*, Paris, **1** (1733), 201. [Proceedings from 1680] 266

[180] A. L. Cauchy, Sur les polygones et les polyédres, seconde memoire, *J. École Polytech .* **9** (1813), 87–98. 422

[181] A. L. Cauchy, Recherches sur les nombres, *J. École Polytech.* **9** (1813), 99–116. 524

[182] A. Cayley, On a tactical theorem relating to the triads of fifteen things, *Collected mathematical papers*, vol. 5 [No. 323]. 389

[183] A. Cayley, On the notion and boundaries of algebra, *Quarterly Journal of Mathematics* **6** (1864), 382–384. See also *Collected mathematical papers*, vol. 5 [No. 347]. 391

[184] G. Chartrand, L. Lesniak, and P. Zhang, *Graphs and digraphs*, 5th ed., CRC Press, Boca Raton, 2011. 135, 145, 438

[185] B. Chazelle, *The discrepancy method, randomness and complexity*, Cambridge University Press, 2000 (paperback in 2001). 364, 366, 496, 498

[186] L. Chen and Z. Zeng, On the Heilbronn optimal configuration of seven points in the square, in *Automated deduction in geometry*, 196–224, Lecture Notes in Comput. Sci. **6301**, Lecture Notes in Artificial Intelligence, Springer, Heidelberg, 2011 255

[187] S. Chowla, and H. J. Ryser, Combinatorial problems, *Canadian J. Math.* **2** (1950), 93–99, 331

[188] I.-P. Chu and R. Johnsonbaugh, Tiling deficient boards with trominoes, *Math. Mag.* **59** (1986), 34–40. 744, 745

[189] F. Chung, The number of different distances determined by n points in the plane, *J. Combin. Theory, Ser. A* **36** (1984), 342–354. 450

[190] F. R. K. Chung and R. L. Graham, Forced convex n-gons in the plane, *Discrete and Computational Geometry* **19** (1998), 367–371. 226

[191] F. Chung and S. Sternberg, Mathematics and the buckyball, *American Scientist* **81** (1993), 56–71. (Published by Sigma Xi.) 177

[192] F. Chung, E. Szemerédi, and W. T. Trotter, The number of different distances determined by a set of points in the Euclidean plane, *Discrete and Computational Geometry* **7** (1992), 342–354, 450

[193] V. Chvátal, The minimality of the Mycielski graph, *Graphs and combinatorics* (Proc. Capital Conf., George Washington Univ., Washington, D.C., 1973), Lecture Notes in Mathematics **406**, Springer-Verlag, Berlin, 1974, pp. 243–246. 402

[194] V. Chvátal, A combinatorial theorem in plane geometry, *J. Combin. Theory Ser. A* **18** (1975), 39–41. 179

[195] V. Chvátal and F. Harary, Generalized Ramsey theory for graphs. II. Small diagonal numbers, *Proc. Amer. Math. Soc.* **32** (1972), 389–394. 477

[196] J. Cibulka, On the chromatic number of real and rational spaces, *Geombinatorics* **18** (2008), 53–66. 461, 464

[197] K. Clarkson, H. Edelsbrunner, L. Guibas, M. Sharir, and E. Welzl, Combinatorial complexity bounds for arrangements of curves and spheres, *Discrete Comp. Geom.* **5** (1990), 99–160. 450

[198] S. T. Coffin, *The puzzling world of polyhedral dissections*, Oxford University Press, 1990. 260

[199] D. I. A. Cohen, On the Sperner lemma, *J. Combin. Theory* **2** (1967), 585–587. 253

[200] H. Cohn, A conceptual breakthrough in sphere packing, *Notices of the A.M.S.* **64** (February 2017), 102–115. 17

[201] H. Cohn, A. Kumar, S. D. Miller, D. Radchenko, and M. Viazovska, *The sphere packing problem in dimension 24*, preprint, version 2 (11 January 2017) available at `arXiv:1603.06518v2`. (Last accessed 8 Oct. 2017.) 17

[202] F. Comellas and J. L. A. Yebra, New lower bounds for Heilbronn numbers, *The Electronic Journal of Combinatorics* **9** (2002), #R6 (10 pages). 255, 257

[203] D. Conlon, A new upper bound for diagonal Ramsey numbers, *Annals of Mathematics* **170** (2) (2009), 941–960. 220

[204] D. Conlon and J. Fox, Lines in Euclidean Ramsey theory, preprint, `https://arxiv.org/abs/1705.02166`, posted 5 May 2017. 491

[205] R. Connelly, Conjectures and open problems in rigidity, in *Proceedings of the International Congress of Mathematicians*, Helsinki, pp. 407–414. 422

[206] R. Connelly, A counter-example to the rigidity conjecture for polyhedra, *Publications Math de l'Institut des Hautes Études Scientifiques* **47** (1978), 333–338. 422

[207] R. Connelly, The rigidity of polyhedral surfaces, *Mathematics Magazine* **52** (1979), 275–283. 422

[208] J. H. Conway, Mrs. Perkins's quilt, *Proc. Cambridge Phil. Soc.* **60** (1964), 363–368. 290

[209] J. H. Conway and P. Doyle, email conversation from 15 Dec 1997 on proofs of Heron's theorem, `https://math.dartmouth.edu/~doyle/docs/heron/heron.txt`, accessed 16 January 2017. 46

[210] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*, 3rd ed., Springer-Verlag, New York, 1999. 80

[211] R. J. Cook and G.V. Wood, Feynman's Triangle, *Mathematical Gazette* **88** (2004), 299–302. 126

[212] J. L. Coolidge, A historically interesting formula for the area of a quadrilateral, *Amer. Math. Monthly* **67** (1939), 345–347. 62

[213] D. A. Coulson, An 18-colouring of 3-space omitting distance one, *Discrete Math.* **170** (1997), 241–247. 459, 461

[214] D. A. Coulson, A 15-colouring of 3-space omitting distance one, *Discrete Math.* **256** (2002), 83–90. 459, 461

[215] R. Courant and H. Robbins, *What is mathematics? An elementary approach to ideas and methods*, Oxford University Press, Oxford, England, 1941. Second, third, and fourth editions with corrections in 1943, 1945, 1947. [My copies are the 9th printing of hardback in 1958 and 4th printing of paperback in 1981, which seem to be equivalent.] 93, 102, 114, 115, 116, 117, 118, 119, 411

[216] D. A. Cox, *Galois theory*, Wiley, 2004. 104

[217] H. S. M. Coxeter, A problem of collinear points, *Amer. Math. Monthly* **55** (1948), 26–28. 376

[218] H. S. M. Coxeter, *Regular polytopes, 3rd ed.*, Dover Publications, 1948.

[219] H. S. M. Coxeter, Self-dual configurations and regular graphs, *Bulletin of the American Mathematical Society* **56** (1950), 413–455. 215

[220] H. S. M. Coxeter, *Introduction to geometry* (2nd ed.), Wiley, New York, 1969. 395

[221] H. S. M. Coxeter, *Regular polytopes*, 3rd ed., Dover Publications, Inc., New York, 1973 (a reprint of the 1963 original). 69, 102, 127, 266

[222] H. S. M. Coxeter, *Regular complex polytopes*, 2nd ed., Cambridge University Press, 1991. 203

[223] H. S. M. Coxeter, *The real projective plane*, 3rd ed., Springer–Verlag, 1992. 203

[224] H. S. M. Coxeter and S. L. Greitzer, *Geometry revisited*, Mathematical Association of America, Washington, D. C., 1967. 376, 380

[225] H. T. Croft, 9-point and 7-point configurations in 3-space, *Proc. London Math. Soc.* (3) **12** (1962), 400–424. 266

[226] H. T. Croft, K. J. Falconer, and R. K. Guy, *Unsolved problems in geometry*, Springer, 1991. 454

[227] P. Cromwell, *Polyhedra*, Cambridge University Press, 1997. 453

[228] D. Crowe and T. McKee, Sylvester's problem on collinear points, *Math. Mag.* **41** (1968), 30–34. 422

[229] A. Császár, A polyhedron without diagonals, *Acta Sci. Math. Szeged* **13** (1949), 140–142. 382, 747

[230] J. Csima and E. Sawyer, There exist $6n/13$ ordinary points, *Discrete and Computational Geometry* **9** (1993), 187–202. 153

[231] Prof. Csirmaz, Problem sheet from Karen Gunderson's notes for course "Proof and Conjecture", given in Budapest, summer 2009. 382

[232] G. Csizmadia and G. Tóth, Note on a Ramsey-type problem in geometry, *J. Combin. Theory Ser. A* **65** (1994), 302–306. 191, 692, 720, 728, 742, 743, 745

[233] L. Curchin and R. Herz-Fischler, De quand date le premier rapprochement entre la suite de Fibonacci et la division en extréme et moyenne raison? *Centaurus* **28** (1985), 129–138. 490

[234] Cut the Knot, *Golden ratio in Geometry*, `www.cut-the-knot.org/do_you_know/GoldenRatio.shtml`, 18 pages. (Accessed 19 May 2015). 94

[235] L. Danzer, B. Grünbaum, and V. Klee, Helly's theorem and its relatives, pp. 101–180 in *Convexity*, Proc. Symposia in Pure Maths, vol. VII, Amer. Math. Soc., 1963. 94

[236] T. O. Dao, T. D. Nguyen, and N. M Pham, A strengthened version of the Erdős–Mordell inequality, *Forum Geom.* **16** (2016), 317–321. 200

[237] H. Davenport, On the addition of residue classes, *J. London Math. Soc.* **S1-10** (1935), 30–32. 124

[238] B. A. Davey and H. A. Priestly, *Introduction to lattices and order*, Cambridge University Press, 1990. 524

[239] N. G. de Bruijn, Filling boxes with bricks, *Amer. Math. Monthly* **76** (1969), 37–40. 503

[240] N. G. de Bruijn and P. Erdős, On a combinatiorial [sic] problem, *Indag. Math.* **10** (1948), 421–423. 288

376, 378, 380, 381

[241] N. G. de Bruijn and P. Erdős, A colour problem for infinite graphs and a problem in the theory of relations, *Nederl. Akad. Wetensch. Proc. Ser. A.* **54**, No. 5, and *Indag. Math.* **13** (1951), 369–373. 457

[242] A. D. N. J. de Grey, *The chromatic number of the plane is at least 5*, preprint, `arXiv:1804.02385`, submitted 8 April 2018; v3 submitted 30 May 2018. 458, 461

[243] M. Dehn, Über raumgleiche Polyeder, *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse* (1900), 345–354. 260

[244] M. Dehn, Über den Rauminhalt, *Math. Ann.* **55** (1902), 465–478. 260

[245] M. Dehn, Über die Zerlegung von Rechtecken in Rechtecke, *Math. Ann.* **57** (1903), 314–322. 289

[246] B. Dekster, The Borsuk conjecture holds for bodies of revolution, *Journal of Geometry* **52** (1995), 64–73. 467

[247] J. A. de Loera, E. Petersen, and F. E. Su, A polytopal generalization of Sperner's lemma, *J. Combin. Theory, Ser. A* **100** (2002), 1–26. 253

[248] J. A. De Loera, J. Rambau, and F. Santos, *Triangulations*, Springer, 2010. 253, 468

[249] P. Delsarte, J. M. Goethals, and J. J. Seidel, Spherical codes and designs, *Geometrica Dedicata* **6** (1977), 363–388. 455

[250] P. Dembowski, *Finite Geometries*, Springer-Verlag, New York, 1968. 301, 342, 343

[251] R. H. F. Denniston, Some maximal arcs in finite projective planes, *J. Combin. Theory* **6** (1969), 317–319, 343

[252] M. J. de Resmini and G. Migliori, A 78-set of type (2,6) in PG(2,16), *Ars Combinatoria* **22** (1986), 73–75. 359

[253] R. Descartes, Progymnasmata de solidorum elementis, in *Oeuvres de Descartes*, vol. X, 1630, pp. 265–276. 164

[254] S. L. Devadoss and J. O'Rourke, *Discrete and computational geometry*, Princeton University Press, Princeton NJ and Oxford, 2011. 180

[255] M. de Villiers, Feedback: Feynman's triangle, *Mathematical Gazette* **89** (2005) (March), 107. 127

[256] M. de Villiers, *Feynman's triangle*, website, `http://dynamicmathematicslearning.com/feynman.html`, 2009. (Accessed 9 October 2017.) 127

[257] R. P. Dilworth, A decomposition theorem for partially ordered sets, *Ann. of Math.* (2) **51** (1950), 161–166. 508

[258] R. P. Dilworth, Proof of a conjecture on finite modular lattices, *Ann. of Math* (2), **60** (Sept. 1954), 359–364. 507

[259] J. H. Dinitz and D. R. Stinson, eds., *Contemporary design theory, a collection of surveys*, John Wiley & Sons, Inc., New York, 1992. 325

[260] G. A. Dirac, Collinearity properties of sets of points, *Quart. J. Math.* **2** (1951), 221–227. 381

[261] J. P. G. L. Dirichlet, Notes prepared by R. Dedekind, 1879. See *Werke* **1**, Springer, 1889. 275

[262] D. Djukić, V. Janković, I. Matić, and N. Petrović, *The IMO compendium: A collection of problems suggested for the International Mathematical Olympiads: 1959–2009*, second edition, Springer, 2011. 14, 701, 741, 759

[263] M. Dobbins, A. Holmsen, and A. Hubard, The Erdős–Szekeres problem for non-crossing convex sets, *Mathematika* **60** (2014), 463–484. 228

[264] A. A. Dobrynin and L. S. Mel'nikov, 4-chromatic Koester graphs, *Discussiones Mathematicae Graph Theory* **32** (2012), 617–627. 404

[265] H. Dörrie, (transl. by D. Antin) *100 great problems of elementary mathematics: their history and solution*, Dover Publications, Inc., New York, 1965. [This edition is a translation by D. Antin of the original *Triumph der Mathematik*, Physica-Verlag, Würzburg, 1958.] 102

[266] D. Dowling and R. Dowling, The lasting legacy of Ludoph Lehmus, *Manitoba Math Links* Volume II, Issue 3, Spring 2002. 702

[267] J. A. Dossey, A. D. Otto, L. E. Spence, and C. Vanden Eynden, *Discrete Mathematics*, 2nd ed., Harper Collins College Publishers, New York, 1993. 744

[268] A. W. M. Dress, L. Yang, and Z. B. Zeng, Heilbronn problem for six points in a planar convex body, in *Combinatorics and graph theory '95, Vol. 1* (Hefei), 97–118, World Sci. Publ., River Edge, NJ, 1995. 254, 258

[269] A. W. M. Dress, L. Yang, and Z. B. Zeng, Heilbronn problem for six points in a planar convex body, *Minimax and applications*, 173–190, *Nonconvex Optim. Appl.* **4**, Kluwer acad. Publ., Dordrecht, 1995. 258

[270] H. E. Dudeney, *The Canterbury puzzles and other curious problems*, Thomas Nelson and Sons, London, 1907. [Revised edition by Dover Publications, 1958.] 242, 289, 386, 387, 388

[271] H. E. Dudeney, *Amusements in mathematics*, Thomas Nelson & Sons, Limited, 1917. [Recently available is a 1970 reprint from Dover Publications, Inc., New York.] 126, 237, 242, 290

[272] H. E. Dudeney, *536 Puzzles & curious problems*, edited by Martin Gardner, Charles Scribner's Sons, New York, 1967. 122, 239, 242, 290

[273] U. Dudley, *Elementary number theory*, W. H. Freeman, 1978. 523

[274] J. Dugundji, *Topology*, Allyn & Bacon, Boston and London, 1966. [This has been reprinted by Prentice Hall.] 78

[275] A. J. W. Duijvestijn, Simple perfect square of lowest order, *J. Combin. Theory Ser. B* **25** (1978), 240–243. 293

[276] A. J. W. Duijvestijn, P. J. Federico, and P. Leeuw, Compound perfect squares, *Amer. Math. Monthly* **89** (1982), 15–32. 293

[277] W. Dunham, *Journey through genius*, Penguin Books, 1991. 46, 689

[278] A. F. Dunn, ed., *Mathematical bafflers*, McGraw-Hill, 1964. 695

[279] A. F. Dunn, ed., *Second book of mathematical bafflers*, Dover Publications, 1983. A collection of puzzles from *Aviation week* and *Electronic news*, 1959–1971; many puzzles also appeared in pamphlets *Problematic recreations* published by Litton Industries, 1965–1970. (or in Canada, published by General Publishing Company, Ltd., Toronto). 704, 742

[280] P. L. Duren, *Theory of H^p spaces*, Dover Publications, 2000. (Originally published by Academic Press, San Diego CA, 1970.) 33

[281] S. Durocher, *personal communication*, 11 Jan. 2018. 430

[282] B. Dushnik and E. Miller, Partially ordered sets, *American Journal of Mathematics* **63** (1941), 63, 600–610. 508, 511

[283] Z. Dvir, On the size of Kakeya sets in finite fields, *J. Amer. Math. Soc.* **22** (2009), 1093–1097. 501

[284] Z. Dvir, *Incidence theorems and their applications*, preprint, arXiv:1208.5073v2[math.CO], 27 Aug 2013, 105 pages, (accessed 26 May 2016). 383, 400, 501

[285] Z. Dvir, S. Kopparty, S. Saraf, and M. Sudan, Extensions to the method of multiplicities, with applications to Kakeya sets and mergers, *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 181–190, IEEE Computer Society, Washington, DC, USA, 2009. 501

[286] E. B. Dynkin and V. A. Uspenskii, *Mathematical conversations: muliticolor problems in the theory of numbers and random walks*, Dover Publications, Mineola N.Y., 2006. 465

[287] H. Edelsbrunner, Lower bounds for many cells, section 6.5 in *Algorithms in combinatorial geometry*, Springer Verlag, 1987. 400

[288] H. Edelsbrunner and R. Waupotitsch, Computing a ham sandwich cut in two dimensions, *Journal of Symbolic Computation* **2** (1986), 171–178. 413

[289] H. G. Eggleston, Covering a three-dimensional set with sets of smaller diameter, *Journal of the London Mathematical Society* **30** (1955), 11–24. 467

[290] R. B. Eggleton and R. K. Guy, The crossing number of the n-cube, *Amer. Math. Soc. Notices* **17** (1970), 757. 433, 437

[291] G. Elekes, n points in the plane can determine $n^{3/2}$ circles, *Combinatorica* **4** (1984), 131. 401

[292] Gy. Elekes, Sums versus products in number theory, algebra and Erdős geometry, pp. 241–290 in *Paul Erdős and his mathematics, II* (Budapest 1999), *Bolyai Soc. Math. Stud.* **11**, János Bolyai Math. Soc., Budapest, 2002. 400

[293] J. Ellenberg, *How not to be wrong: the power of mathematical thinking*, Penguin books, 2014. 181

[294] A. Engel, *Problem-solving strategies*, Problem books in mathematics, Springer, 1997. 14, 380, 408, 688, 703, 750, 751, 752

[295] P. Erdős, Problem 3740, *Amer. Math. Monthly* **42** (1935), 396. 123

[296] P. Erdős, Problems for solution, number 4065, *Amer. Math. Monthly* **50** (1943), 65. 375, 574

[297] P. Erdős, Integral distances, *Bull. Amer. Math. Soc.* **51** (1945), 996. 747

[298] P. Erdős, On sets of distances of n points, *Amer. Math. Monthly* **53** (1946), 248–250. 448, 449, 450, 451, 758

[299] P. Erdős, Some remarks on the theory of graphs, *Bull. Amer. Math. Soc.* **53** (1947), 292–294. 221

[300] P. Erdős, Remarks on a theorem by Ramsay [sic], *Bulletin of the Research Council of Israel* **7F** (1957), 21–24. 220

[301] P. Erdős, On sets of distances of n points in Euclidean space, *Publ. Math. Inst. Hungar. Acad. Sci.* **5** (1960), 165–169. 453

[302] P. Erdős, Some unsolved problems, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961), 221–254. 451, 455

[303] P. Erdős, On some problems of elementary and combinatorial geometry, *Annali di Matematica pura ed applicata* (IV), **103** (1975), 99–108. 375, 383

[304] P. Erdős, Personal reminiscences and remarks on the mathematical work of Tibor Gallai, *Combinatorica* **2**(3)(1982), 207–212. 375, 376

[305] P. Erdős, Problems and results in combinatorial geometry, discrete geometry and convexity (New York, 1982), *Annals of New York Acad. Sci.* **440** (1985) , 1–11. 399, 453

[306] P. Erdős and R. L. Graham, On packing squares with equal squares, J. Combin. Theory Ser A. **19** (1975), 119–123. 128, 129

[307] P. Erdős and R. L. Graham, Old and new problems and results in combinatorial number theory: van der Waerden's theorem and related topics, *Enseign. Math.* **25** (1979), 325–344. 491

[308] P. Erdős, R. L. Graham, P. Montgomery, B. L. Rothschild, J. Spencer, and E. G. Straus, Euclidean Ramsey theory I, *J. Combin. Theory Ser. A* **14** (1973), 341–363. 473, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 491, 761

[309] P. Erdős, R. L. Graham, P. Montgomery, B. L. Rothschild, J. Spencer, and E. G. Straus, Euclidean Ramsey theory II, pp. 529–557 in *Infinite and finite sets*, (Proc. of Colloq. Keszthely, 1973) Vol. I, published by *Colloq. Math. Soc. Janos Bolyai*, Vol. **10**, North–Holland, Amsterdam, 1975. 473, 489, 490, 493, 738

[310] P. Erdős, R. L. Graham, P. Montgomery, B. L. Rothschild, J. Spencer, and E. G. Straus, Euclidean Ramsey theory III, pp. 559–583 in *Infinite and finite sets*, (Proc. of Colloq. Keszthely, 1973) Vol. I, published by *Colloq. Math. Soc. Janos Bolyai*, Vol. **10**, North–Holland, Amsterdam, 1975. 473, 474, 475, 476

[311] P. Erdős and R. K. Guy, Crossing number problems, *Amer. Math. Monthly* **80** (1973), 52–58. 429, 430, 433, 434, 435, 437, 441

[312] P. Erdős and G. Purdy, Some extremal problems in geometry, *J. Combin. Theory* **10** (1971), 246–252. 749

[313] P. Erdős and G. Purdy, Extremal problems in combinatorial geometry, *Handbook of Combinatorics* (R. Graham, M. Grötschel, and L. Lovász, eds.), 809–874, Elsevier, 1995. 383, 386

[314] P. Erdős and R. Rado, A combinatorial theorem, *J. London Math. Soc.* **25** (1950), 249–255. 217

[315] P. Erdős, A. Rényi, and V. T. Sós, On a problem of graph theory, *Studia Sci. Math. Hungar.* **1** (1966), 215–235. 321, 348, 349

[316] P. Erdős, R. Silverman and A. Stein, Intersection properties of families containing sets of nearly the same size, *Ars Combinatoria* **15** (1983), 247–259. 359

[317] P. Erdős and J. Spencer, *Probabilistic methods in combinatorics*, Academic Press New York, 1974. 220

[318] P. Erdős and R. Steinberg, Three point collinearity, solution to problem 4065, [includes editor's note containing Gallai's solution as submitted by Erdős in original problem in [296]], *Amer. Math. Monthly* **51** (3) (1944), 169–171. 375

[319] P. Erdős and J. Surányai, *Topics in the theory of numbers*, 2nd ed., (translated by Barry Guiduli), Springer, 2003. [Available at `en.bookfi.org/book/1217821`, downloaded 18 Jan 2014.] 279

[320] P. Erdős and G. Szekeres, A combinatorial problem in geometry, *Compositio Math.* **2** (1935), 463–470. 217, 219, 222, 224, 226, 227, 511, 732

[321] P. Erdős and G. Szekeres, On some extremum problems in elementary geometry, *Ann. Univ. Sci. Budapest* **3** (1960), 53–62. 224, 226

[322] F. Eugeni and S. Innamorati, On sets of fixed parity in Steiner systems, *Annals of Discrete Mathematics* **37** (1988), 157–167. 355

[323] L. Euler, Enumratio modorum, quibus figurae planae rectilinae per diagonales diuiduntur in triangula, auctore I. A. de Segner, Pag. 203, *Novi Comm. Acad. Sci. Imp. Pet.* **7** (1761), 13–15. 64, 65, 67

[324] L. Euler, *Recherches sur une nouvelle espece de quarres magiques*, 1782. 306

[325] H. W. Eves, *College Geometry*, Jones & Bartlett Learning, 1995. 116

[326] G. Exoo and D. Ismailescu, *On the chromatic number of* $\mathbb{R}^n$ *for small values of* n, preprint, `arXiv:1408.2002v1` [math.CO] 9 Aug 2014. 461

[327] G. Exoo, D. Ismailescu,, and M. Lim, On the chromatic number of R^4, *Discrete Comput. Geom.* **52** (2014), 416–423. 460, 461, 493

[328] X. W. C. Faber, On the finite field Kakeya problem in two dimensions, *J. Number Theory* **117** (2006), 471–481. 501

[329] K. J. Falconer, The realization of distances in measurable subsets covering $\mathbf{R}^n$, *J. Combin. Theory Ser. A* **31** (1981), 184–189. 459

[330] I. Fáry, On straight line representations of planar graphs, *Acta Sci. Math. (Szeged)* **11** (1948), 229–233. 438

[331] E. Fedorov (1891) Simmetrija na ploskosti (Symmetry in the plane), *Proceedings of the Imperial St. Petersburg Mineralogical Society, series 2* **28** (1891), 345–390 (in Russian). 185

[332] S. Felsner, F. Hurtado, M. Noy, and I. Streinu, Hamiltonicity and colorings of arrangement graphs. Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 2000), 155–164, ACM, New York, 2000. 404

[333] K. Feuerbach, *Eigenschaften einiger merkwürdigen Punkte des geradlinigen Dreiecks und mehrerer durch sie bestimmten Linien und Figuren. Eine analytisch-trigonometrische Abhandlung* (Properties of some special points in the plane of a triangle, and various lines and figures determined by these points: an analytic-trigonometric treatment), Wiessner, Nürnberg, 1822. [Parts were edited and translated by R. A. Johnson as “Feuerbach on the theorem which bears his name”, in D. E. Smith, ed., *A Source Book in Mathematics*, New York-London, 1929; paperback reprint, New York, 1959.] 75, 76

[334] K. Fischer, Additive k-colorable extensions of the rational plane, *Discrete Math.* **82** (1990), 181–195. 464

[335] R. A. Fisher, An examination of the different possible solutions of a problem in incomplete blocks, *Annals of Eugenics* **10** (1940), 52–75. 330

[336] S. Fisk, A short proof of Chvátal's watchman theorem, *J. Combin. Theory Ser. B* **24** (1978), 374. 180

[337] H. Flaschka, *Some geometry in high-dimensional spaces*, online notes for MATH 527A, circuit.ucsd.edu/~massimo/ECE287C/Handouts_files/Geometry.pdf, 2010. (Accessed August 2017.) 175

[338] A. Flores, *Dudeney's triangle into square*, http://www.public.asu.edu/~aaafp/hinged/Dudeneytrianglesquare.html. (Accessed 30 June 2016.) 243

[339] W. B. Ford, On Kakeya's minimum area problem, *Bull. Amer. Math. Soc.* **28** (1922), 45–53. 500

[340] L. R. Ford, *Differential equations*, 2nd ed., McGraw-Hill Book Company, Inc., New York, 1955. (First edition published in 1933.) 188

[341] L. R. Ford, Jr., D. R. Fulkerson, and D. C. B. Marsh, Solution to problem E1279: cutting an $n \times n \times n$ cube into n^3 unit cubes, Elementary Problems and Solutions, *Amer. Math. Monthly* **65** (no. 3, March, 1958), p. 210. 416, 417

[342] J. Fox, J. Pach, B. Sudakov, and A. Suk, Erdős–Szekeres-type theorems for monotone paths and convex bodies, *Proc. Lond. Math. Soc.* **105** (2012), 953–982. 228

[343] O. Frank, F. Harary, and M. Plantholt, The line-distinguishing chromatic number of a graph, *Ars Combin.* **14** (1982), 241–252. 374

[344] P. Frankl and V. Rödl, All triangles are Ramsey, *Trans. Amer. Math. Soc.* **297** (1986), 777–779. 477

[345] P. Frankl and V. Rödl, A partition property of simplices in Euclidean space, *J. Amer. Math. Soc.* **3** (1990), 1–7. 482

[346] P. Frankl and V. Rödl, Strong Ramsey properties of simplices, *Israel J. Math.* **139** (2004), 215–236. 482

[347] P. Frankl and R. M. Wilson, Intersection theorems with geometric consequences, *Combinatorica* **1** (1981), 357–368. 460, 461, 491

[348] G. N. Frederickson, *Dissections: Plane & fancy*, Cambridge University Press, 1997. [Paperback edition 2002.] xii+310 pp. 232, 239, 263

[349] G. N. Frederickson, *Hinged dissections: Swinging & Twisting*, Cambridge University Press, 2002. 238, 240, 241, 242, 577

[350] G. N. Frederickson, Website for book "Hinged dissections" [349], `https://www.cs.purdue.edu/homes/gnf/book2/Booknews2/gundm.html`, last updated 24 May 2004, last accessed 27 June 2016. 242, 243

[351] G. N. Frederickson, *Piano-hinged dissections: Time to fold!*, A. K. Peters, 2006. 240, 243

[352] H. Freudenthal and W. Hurewicz, Dehnungen, Verkürzungen, Isometrien, *Fund. Math.* **26** (1936), 120–122. 78

[353] E. Friedman, *The Heilbronn problem*, `http://www2.stetson.edu/~efriedma/heilbronn/` (accessed August 2017). 255

[354] O. Frink, Complemented modular lattices and projective spaces of infinite dimension, *Trans. Amer. Math. Soc.* **60** (1946), 452–467. 507

[355] Z. Füredi, An upper bound on Zarankiewicz' problem, *Combin. Probab. Comput.* **5** (1966), 29–33. 426

[356] Z. Füredi, Maximal τ-critical linear hypergraphs, *Graphs and Combinatorics* **17** (2001), 73–78. 362

[357] Z. Füredi and J. Kahn, On the dimensions of ordered sets of bounded degree, *Order* **3** (1986), 15–20. 512

[358] Z. Füredi and J. Kahn, Dimension versus size, *Order* **5** (1988), 17–20. 512

[359] Z. Füredi and I. Palásti, Arrangements of lines with a large number of triangles, *Proc. of the A.M.S.* **92** (1984), 561–566. 386

[360] A. Gagarin, W. Kocay, and D. Neilson, Embeddings of small graphs on the torus, *CUBO* **5** (2003), 351–371. 445

[361] D. Gale, Neighboring vertices on a convex polyhedron, pp. 255–263 in *Linear Inequalities and Related Systems* (H. W. Khun and A. W. Tucker, Eds.), Annals of Mathematical Studies **38**, Princeton University Press, Princeton, N. J., 1956. 213

[362] D. Gale, Neighborly and cyclic polytopes, in *Convexity* (Seattle 1961), V. Klee ed., *Proc. Symp. Pure Math.* **7** (1963), 225–232. (Published by Amer. Math. Soc.) 213

[363] T. Gallai, Critical graphs, *Theory of Graphs and its Applications* (Proc. Symp. Smolenice 1963), Publ. House Czechoslovak Acad. Sci, Prague 1964, 43–45. 403

[364] M. Gardner, *Mathematics, magic and mystery*, Dover, 1956. 263, 265

[365] M. Gardner, Mathematical Games, Nine titillating puzzles, the answers to which will be given next month, *Scientific American* **197** (November 1957), 140–146. 416

[366] M. Gardner, *The Scientific American book of puzzles & games*, Simon and Schuster, New York, 1959. (A reprint, with a new afterward, is published by University of Chicago Press, 1988 under the title *Hexaflexagons and other mathematical diversions: the first Scientific American book of mathematical puzzles and games*.) 416, 753

[367] M. Gardner, Mathematical games, *Scientific American* **203/4** (1960), 180. 456

[368] M. Gardner, *The Second Scientific American book of mathematical puzzles & diversions: A new selection*, Simon and Schuster, New York, 1961. 242, 299, 412, 413, 746

[369] M. Gardner, Mathematical games, *Scientific American*, September 1966. 290

[370] M. Gardner, Mathematical games, *Scientific American* **225** (1971), No. 5 (November), 114–121. 133, 235

[371] M. Gardner, Mathematical games, *Scientific American* **225** (1971), No. 6 (December), 96–99. 235

[372] M. Gardner, *Aha! Insight!* Scientific American, W. H. Freeman, 1978. 416, 753

[373] M. Gardner, *The sixth book of mathematical games from Scientific American*, University of Chicago Press, Chicago, Il., 1984. 69

[374] M. Gardner, *Mathematical circus*, MAA, 1992. 38

[375] M. Gardner, Mr. Apollinax comes to New York, Chapter 11 of *New mathematical diversions*, revised edition, The Mathematical Association of America, Washington, DC, 1995. 267

[376] M. Gardner, Gardner, Martin (1997), *Penrose Tiles to Trapdoor Ciphers*, Cambridge University Press, 1997. (First published by W. H. Freeman, New York, 1989.) Chapter 1 (pp. 1–18) is a reprint of Gardner's article "Extraordinary nonperiodic tiling that enriches the theory of tiles", *Scientific American* **236** (January 1977),110–121. 98

[377] M. Gardner, Curves of Constant Width, in Chapter 18 of *Knots and Borromean Rings, Rep-Tiles, and Eight Queens* The New Martin Gardner Mathematical Library **4**, Cambridge University Press, 2014, pp. 223–245 92, 93

[378] R. J. Gardner and S. Wagon, At long last, the circle has been squared, *Amer. Math. Soc. Notices*, **36** (1989), 1338–1343. 262

[379] M. R. Garey and D. S. Johnson, Crossing number is NP-complete, *SIAM J. Alg. Discrete Methods* **4** (1983), 312–316. 430

[380] B. Gartner and G. Ziegler, Randomized simplex algorithms on Klee-Minty cubes, *Proceedings 35th Annual Symposium on Foundations of Computer Science(FOCS), Santa Fe, NM, USA, 1994*, 502–510. Published by IEEE, Computer Society Digital Library, FOCS, 1994. (https://www.computer.org/csdi/proceedings/focs/1994). 208

[381] C. F. Gauss, *Disquisitiones Arithmeticae*, Leipzig, Germany, 1801. Translated by A. A. Clarke. New Haven, CT: Yale University Press, 1965. 103

[382] J. P. Georges, Non-hamiltonian bicubic graphs, *J. Combin. Theory B* **46** (1989), 121–124. 398

[383] E. H.-A. Gerbracht, *Eleven unit distance embeddings of the Heawood graph*, preprint available at arXiv:0912.5395 2009. 456

[384] J. Geelen, A. Guo, and D. Mckinnon, Straight line embeddings of cubic planar graphs with integer edge lengths, *J. Graph Theory* **58** (2008), 270–274. 438

[385] T. Gerken, Empty convex hexagons in planar point sets, *Discrete and Computational Geometry* **39** (2008), 239–272. 228

[386] P. Gerwien, Zerschneidung jeder beliebigen Anzahl von gleichen geradlinigen Figuren in dieselben Stücke. Zerschneidung jeder beliebigen Menge verschieden gestalteter Figuren von gleichem Inhalt auf der Kugelfläche in dieselben Stücke, *Crelle's Journal* (*Journal für die reine u. angewandte Mathematik*) **10** (1833), 228–240. 232

[387] A. Ghouila-Houri, Caractérisation des matrices totalement unimodulaires, *C. R. Acad. Sci. Paris* **254** (1962), 1192–1194. 365

[388] P. Glaister, Golden earrings, *Mathematical Gazette* **80** (1996), 224–225. 693

[389] A. M. Gleason, Finite Fano planes, *American Journal of Mathematics* **78** (1956), 797–807. 390

[390] A. M. Gleason and G. E. Greenwood, Combinatorial relations and chromatic graphs, *Canad. J. Math.* **7** (1955), 1–17. [Note: The authors are sometimes given in opposite order.] 219

[391] A. M. Gleason, R. E. Greenwood, and L. M. Kelly, *The William Lowell Putnam Mathematical Competition, Problems and Solutions, 1938-1964*, Mathematical Association of America, Washington, D.C., 1980. 14

[392] D. G. Glynn, On the anti-Pappian 10_3 and its construction, *Geometriae Dedicata* **77** (1999), 71–75. 397

[393] M. Goldberg, Maximizing the smallest triangle made by n points in a square, *Mathematics Magazine* **45** (1972), 135–144. 254, 255, 257, 258

[394] M. Goldberg, Unstable polyhedral structures, *Mathematics Magazine* **51** (May 1978), 165–170. 422

[395] S. W. Golomb, Advanced problems and solutions: Isosceles n points, *Amer. Math. Monthly* **55** (1948), 513–514. 455, 760

[396] S. W. Golomb, Checker boards and polyominoes, *Amer. Math. Monthly* **61** (no. 10, Dec.) (1954), 675–682. 743, 744

[397] S. W. Golomb, *Polyominoes*, Scribner's, New York, 1965. 287, 743

[398] L. I. Golovina and I. M. Yaglom, *Induction in geometry*, Little mathematics library, MIR Publishers, Moscow, 1979. 417, 753

[399] J. E. Goodman, Pseudoline arrangements, Chapter 5 (pp. 83–109) of *Handbook of discrete and computational geometry*, J. E. Goodman and J. O'Rourke, eds., CRC Press, Boca Raton, 1997. 382

[400] V. D. Goppa, *Geometry and Codes,* MIA Soviet series **24**, Kluwer Academic Publishers, Dordrecht, 1988. 342

[401] P. Gorroochurn, *Classic problems in probability*, Wiley, 2012. 180

[402] L. A. Graham, *Ingenious problems and methods*, Dover, New York, 1959. 38, 122, 126, 714

[403] R. L. Graham, Euclidean Ramsey theorems on the n-sphere, *J. Graph Theory* **7** (1983), 105–114. 482, 488

[404] R. L. Graham, Recent trends in Euclidean Ramsey theory, *Discrete Math.* **136** (1994), 119–127. 473, 483

[405] R. L. Graham, Euclidean Ramsey theory, pp. 239–254 in *Handbook of discrete and computational geometry* (J. E. Goodman and J. O'Rourke, Eds.), 2nd ed., Chapman & Hall/CRC, Boca Raton, FL, 2004. 473

[406] R. L. Graham, Open problems in Euclidean Ramsey theory, *Geombinatorics* **13** (2004), 165–177. 473

[407] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics: a foundation for computer science* (2nd ed.), Addison-Wesley, Reading, MA, 1994. 266, 415, 753

[408] R. L. Graham and B. L. Rothschild, Rota's geometric analog to Ramsey's theorem, *Proc. Sympos. Pure Math.* **19** (1971), 101–104. 228

[409] R. L. Graham and B. L. Rothschild, Ramsey's theorem for n-parameter sets, *Trans. Amer. Math. Soc.* **159** (1971), 257–292. 485

[410] R. L. Graham, B. L. Rothschild and J. H. Spencer, *Ramsey theory*, Wiley-Interscience Ser. in Discrete Math., New York, 1980. 217, 226, 472, 483, 485

[411] R. Graham and E. Tressler, Open problems in Euclidean Ramsey theory, *Ramsey theory* (A. Soifer, ed.), 115–120, Progr. Math. **285**, Birkhäuser/Springer, New York, 2011. 460, 473, 478

[412] G. Grätzer, *General lattice theory*, 2nd ed., Birkhäuser Verlag, 1978. 506, 507

[413] G. Grätzer, *General lattice theory*, 2nd ed., Birkhäuser Verlag, 2003. 503

[414] G. Grätzer, *Lattice theory: first concepts and distributive lattices*, Dover Publications, 2009. [First published in 1971 by W. H. Freeman and Company, San Francisco.] 503

[415] B. Green and T. Tao, On sets defining few ordinary lines, *Discrete and Computational Geometry* **50** (2013), 409–468. 383, 386

[416] J. Gross and T. Tucker, *Topological Graph Theory*, John Wiley and Sons, New York, 1987. 445

[417] E. Grosswald, Representations of integers as sums of squares, Springer-Verlag, New York, 1985. 519, 522, 526

[418] M. Grötschel, R. Graham, and L. Lovász, eds., *Handbook of combinatorics*, Elsevier, 1994. 453

[419] H. Grötzsch, Ein Dreifarbensatz für dreikreisfreie Netze auf der Kugel, *Wiss. Z. Martin-Luther-U., Halle-Wittenburg, Math.-Nat. Reihe* **8** (1959), 109–120. [Some references have 1958.] 402

[420] P. M. Gruber and C. G. Lekkerkerker, *Geometry of numbers*, North-Holland, Amsterdam, 1987. 273

[421] P. M. Gruber and J. M. Wills (eds.), *Handbook of convex geometry* (2 vols), North-Holland, 1993. 153

[422] P. M. Gruber, *Convex and Discrete Geometry*, Springer-Verlag, Berlin, 2007. 215

[423] P. M. Gruber, *Convex and discrete geometry: ideas, problems and results*, preprint, available at `www.dmg.tuwien.ac.at/gruber/gruber_arbeiten/convex_and_discrete_geometry.pdf`, (accessed 30 Sep. 2017). 215

[424] B. Grünbaum, A proof of Vázsonyi's conjecture, *Bull. Res. Council Israel Sect. A* **6** (1956), 77–78. 448

[425] B. Grünbaum, Grötzsch's theorem on 3-colorings, *Michigan Math. J.* **10** (1963), 303–310. 403

[426] B. Grünbaum, *Convex polytopes*, Pure and Applied Mathematics **16**, Interscience publishers, John Wiley & Sons, New York, 1967. 204, 211

[427] B. Grünbaum, *Convex polytopes*, 2nd ed., Springer, New York, 2003. 197, 203, 227

[428] B. Grünbaum, A 3-connected configuration (n_3) with no Hamiltonian circuit, *Bulletin of the ICA* **46** (2006), 15–26. 398

[429] B. Grünbaum, *Configurations of points and lines*, Graduate Studies in Mathematics **103**, American Mathematical Society, 2009. 388, 394, 397

[430] B. Grünbaum and G. C. Shephard, *Tilings and Patterns*, W. H. Freeman, New York, 1987. 98

[431] B. Grünbaum and G. C. Shephard, Pick's theorem, *Amer. Math. Monthly* **100** (1993), 150–161. 69, 80

[432] P. Guilianas, Orthologic triangles and Miquel's theorem, *The Mathematical Gazette* **92** (2008), 125–128. 73

[433] D. S. Gunderson, *Handbook of mathematical induction: theory and applications*, CRC Press, 2010. 30, 94, 707

[434] D. S. Gunderson, Finite projective planes and applications in combinatorics, working draft, 2016, 207 pages. 325

[435] D. S. Gunderson, *Notes on advanced graph theory*, in progress, 405 pages, 24 July 2017. 135

[436] L. Guth and N. H. Katz, On the Erdős distinct distances problem in the plane, *Annals of Mathematics* **181** (2015), 155–190, 450

[437] R. K. Guy, A combinatorial problem, *Nabla* (*Bulletin of the Malayan Mathematical Society*) **7** (1960), 68–72. 435, 436

[438] R. K. Guy, The decline and fall of Zarankiewicz's theorem, pp. 63–69 in *Proof Techniques in Graph Theory* (Proc. Second Ann Arbor Graph Theory Conf., Ann Arbor, Mich., 1968), Academic Press, New York, 1969. 435, 437

[439] R. K. Guy, Latest results on crossing numbers, in *Recent trends in graph theory*, Springer, N.Y., 1971. 436

[440] R. K. Guy, Crossing numbers of graphs, pp. 111–124 in *Graph theory and applications* (Proc. Conf., Western Michigan Univ., Kalamazoo, Mich., 1972; dedicated to the memory of J. W. T. Youngs), Lecture Notes in Math., Vol. 303, Springer, Berlin, 1972. 436

[441] R. K. Guy, The strong law of small numbers, *Amer. Math. Monthly* **95** (October 1988), 697–712. 414

[442] R. K. Guy, The second strong law of small numbers, *Math. Mag.* **63** (1990), 3–20. 414

[443] R. K. Guy and T. A. Jenkyns, The toroidal crossing number of $K_{m,n}$, *J. Combin. Theory* **6** (1969), 235–250. 436, 445

[444] R. K. Guy, T. A. Jenkyns, and J. Schaer, The toroidal crossing number of the complete graph, *J. Combin. Theory* **4** (1968), 376–390. 445

[445] H. Hadwiger, Ein Überdeckungssätze für den Euklidischen Raum, *Portugalae Math.* **4** (1944), 140–144. 455

[446] H. Hadwiger, Eine Bemerkung zum Borsukschen Antipodensatz (in German), *Vierteljschr. Naturforsch. Ges. Zürich* **89** (1944), 211–214. 465

[447] H. Hadwiger, Überdeckung des Euklidischen Raumes durch kongruente Mengen, *Portugalae Math.* **4** (1945), 238–242. 455

[448] H. Hadwiger, Überdeckung einer Menge durch Mengen kleineren Durchmessers, *Commentarii Mathematici Helvetici* **18** (1945), 73–75. 467

[449] H. Hadwiger, Mitteilung betreffend meine Note: Überdeckung einer Menge durch Mengen kleineren Durchmessers, *Commentarii Mathematici Helvetici* **19** (1946), 72–73. 467

[450] H. Hadwiger, H. Debrunner, and V. Klee, *Combinatorial geometry in the plane*, Holt, Rinehart and Winston, New York, 1964. 200, 455, 495, 730, 731, 747, 758, 761

[451] M. Hall, Jr., Projective planes, *Transactions of the American Mathematical Society* **54** (1943), 229–277, 306

[452] M. Hall, Jr., The uniqueness of the projective plane with 57 points, *Proc. Amer. Math. Soc.* **4** (1953), 912–916. 310

[453] M. Hall, Jr., Correction to "The uniqueness of the projective plane with 57 points", *Proc. Amer. Math. Soc.* **5** (1954), 994–997. 310

[454] M. Hall, Jr., *Combinatorial Theory*, 2nd ed., John Wiley and Sons, New York, 1986. 325, 338

[455] F. Harary and A. Hill, On the number of crossings in a complete graph, *Proceedings of the Edinburgh Mathematical Society* **13** (Dec. 1963), 333–338. 434, 435, 436, 437, 438, 439, 440

[456] H. Harborth, Konvexe Fünfecke in ebenen Punktmengen, *Elemente der Math.* **33** (1978), 116–118. 228

[457] H. Harborth, A. Kemnitz, M. Möller, and A. Sussenbach, Ganzzahlige planare Darstellungen der platonischen Korper, *Elem. Math.* **42** (1987), 118–122. 438

[458] H. Harborth and M. Möller, The Esther-Klein-Problem in the projective plane, *J. Combin. Math. Combin. Comput.* **15** (1994), 171–180. 227

[459] G. Hart, Virtual polyhedra; The intersection of the five cubes, `http://www.georgehart.com/virtual-polyhedra/five-cube-intersection.html`, copyright 1996. (Accessed 21 June 2016.) 151

[460] C. B. Haselgrove and J. Haselgrove, A computer crogram for pentominoes, *Eureka* **23** (October 1960), 16–18. 287

[461] F. Hawthorne, Problem Q-12, *Mathematics Magazine* **24** (1) (Sept.–Oct. 1950), p. 53. 416

[462] T. L. Heath, *The thirteen books of Euclid's elements*, Cambridge University Press, 1926. 42

[463] E. Helly, Über Mengen konvexer Körper mit gemeinschaftlichen Punkten, *Jahrb. Deut. Math. Verein* **32** (1923), 175–176. 200

[464] A. Heppes, Beweis einer Vermutung von A. Vázsonyi, *Acta Math. Acad. Sci. Hungar* **7** (1957), 463–466. 448

[465] J. Herman, R. Kučera, and J. Šimša, translated by Karl Dilcher, *Counting and configurations; Problems in combinatorics, arithmetic, and geometry*, CMS Books in Mathematics, Springer, 2003. 692, 700, 701, 709, 711, 727, 738, 742

[466] D. Hess, *The population explosion and other mathematical puzzles*, World Scientific, Singapore, 2016. 300, 412, 721

[467] O. Hesse, Über die Elimination der Variabeln aus drei algebraischen Gleichungen vom zweiten Grade mit zwei Variabeln, (in German), *Journal für die Reine und Angewandte Mathematik* **28** (1844), 68–96. 394

[468] D. HiIlbert, *Grundlagen der Geometrie* (in German), Leipzig, 1899. [See also a 1902 translation by Townsend [471] and a 1971 translation by Unger [473].] 231

[469] D. Hilbert, *Göttinger Nachrichten*, 1900, 253–297 (in German). 259

[470] D. Hilbert, *Archiv der Mathematik und Physik, 3d ser.* **1** (1901), 44–63, 213–237 (in German) 259

[471] D. Hilbert, *Foundations of geometry*, translation by E. J. Townsend, The Open Court Publishing Co., Chicago, 1902. 232, 587

[472] D. Hilbert, Mathematical problems, *Bull. Amer. Math. Society*, **8**, (1902), 437–479. (English translation of the original 23 problems, translated by Mary Frances Winston Newson.) 259, 260

[473] D. Hilbert, *Foundations of geometry*, translated from German by L. Unger, and revised and enlarged by P. Bernays, Open Court, Chicago, 1971. 232, 587

[474] D. Hilbert and S. Cohn-Vossen, *Geometry and the imagination* (1952 translation by Paul F. Nemenyi of *Anschauliche Geometrie*, 1932), American Mathematical Society, 1999. 375

[475] P. Hliněný, Crossing number is hard for cubic graphs, *J. Combin. Theory Ser. B.* **96** (2006), 455–471. 433

[476] T. Hiraguchi, On the dimension of orders, *Sci. Rep. Kanazawa Univ.* **4** (1955), 1–20. 510

[477] J. W. P. Hirschfeld, *Projective geometries over finite fields*, Clarendon Press, Oxford, 1979. 301

[478] J. W. P. Hirschfeld and J. A. Thas, *General Galois geometries*, Clarendon Press, Oxford; Oxford University Press, New York, 1991. 301

[479] . E. W. Hobson, On geometrical constructions by means of the compass, *Mathematical Gazette* **7** (1913), 49–54. Also reprinted in Chapter 2.4 of [734]. 102

[480] K. Hoffman and R. Kunze, *Linear algebra*, 2nd ed., Prentice-Hall, Inc., Englewood Cliffs, N.J., 1971. 169

[481] V. E. Hoggatt, Jr., *Fibonacci and Lucas numbers*, Houghton Mifflin, 1969. 38

[482] M. H. Holt, Mystery puzzler and phi, *Fibonacci Quarterly* **3** (April 1965), 135–138. 38

[483] R. Honsberger, *Mathematical Gems*, Mathematical Association of America, Washington, D. C., 1973. 284

[484] R. Honsberger, *Mathematical Gems III*, Mathematical Association of America, Washington, D. C., 1985. 123, 710, 744

[485] R. Honsberger, *From Erdős To Kiev*, MAA, 1996. 760

[486] J. Hopcroft and R. E. Tarjan, Efficient planarity testing, *Journal of the Association for Computing Machinery* **21** (1974), 549–568. 430

[487] H. Hopf and E. Pannwitz, Problem 167, *Jahresber.Deuthsch. Math.-Verein* **43** (1934), 2 Abt., 114. 758

[488] J. D. Horton, Sets with no empty convex 7-gons, *Canadian Math. Bull.* **26** (1983), 482–484. 228

[489] D. Hughes, A class of non-desarguesian projective planes, *Canad. J. Math.* **9** (1957), 378–388. 310

[490] D. R. Hughes and F. C. Piper, *Projective Planes*, Graduate Texts in Mathematics **6**, Springer-Verlag, 1973. 317, 338, 342, 343, 344, 345, 349, 350

[491] D. R. Hughes, and F. C. Piper, *Design Theory*, Cambridge University Press, Cambridge, 1985. 325

[492] N. Hungerbühler and M. Nüsken, Delian metamorphoses, *Elem. Math.* **61** (2006), 1–19. [*Elemente der Mathematik* is published by the Swiss Mathematical Society.] 232, 260

[493] T. W. Hungerford, *Algebra*, Springer-Verlag, 1974. 517

[494] T. Illés, T. Szőnyi, and F. Wettl, Blocking sets and maximal strong representative systems in finite projective planes, Proceedings of the First International Conference on Blocking Sets (Giessen, 1989), *Mitt. Math. Sem. Giessen* **201** (1991), 97–107. 363

[495] E. L. Ince, *Ordinary differential equations*, Dover, New York, 1944. 188

[496] E. Ionascu, *Problem 4195* Solutions, *Crux Mathematicorum* **43** (2017), p. 453. 722

[497] I. M. Isaacs, *Geometry for college students*, Brooks/Cole, 2001. 59, 129, 709

[498] L. Iván, *Monochromatic point sets in the plane and in space*, Master's thesis, University of Szeged, 1979. 490

[499] L. L. Ivanov, An estimate for the chromatic number of the space $\mathbb{R}^4$, *Russian Math. Surveys* **61** (2006), 984–986. 460, 461

[500] V. Jelínek, J. Kynčl, R. Stolař, and T. Valla, Monochromatic triangles in two-colored plane, *Combinatorica* **29** (2009), 699–718. 474, 475

[501] T. Jenrich, A 64-dimensional two-distance counterexample to Borsuk's conjecture, preprint `arXiv:1308.0206`, 2014. 468

[502] T. Jenrich, On the counterexamples to Borsuk's conjecture by Kahn and Kalai, `arXiv:1809.09612v4`, 2018. 467

[503] T. R. Jensen and B. Toft, *Graph coloring problems*, Wiley-Interscience, New York, 1995. 374

[504] T. Jiang, M. Li, and P. Vitányi, The average-case area of Heilbronn-type triangles, *Random Structures and Algorithms* **20** (2002), 206–219. 257

[505] P. D. Johnson, Coloring abelian groups, *Discrete Math.* **40** (1982), 219–223. 464

[506] R. Johnson, A circle theorem, *Amer. Math. Monthly* **23** (1916), 161–162. 749

[507] R. Johnson, *Modern geometry—an elementary treatise on the geometry of the triangle and the circle*, Houghton Mifflin, 1929. 749

[508] S. Johnson, A new proof of the Erdős Szekeres convex k-gon result, *J. Combin. Theory Ser. A* **42** (1986), 318–319. 226

[509] R. T. Jones and B. B. Peterson, *Mathematics Magazine* **47** (Sept. 1974), 180–189. 38

[510] B. Jónsson, Modular lattices and Desargues' theorem, *Math. Scand.* **2** (1954), 295–314.

[511] R. Juhász, Ramsey type theorems in the plane, *J. Combin. Theory Ser. A* **27** (1979), 152–160. 507

[512] S. Jukna, *Extremal combinatorics with applications in computer science*, 2nd ed., Springer, 2011. 490

[513] H. W. E. Jung, Über die kleinste Kugel, die eine räumliche Figur einschliesst, *J. Reine Angew. Math.* **123** (1901), 241–257. 759

[514] H. W. E. Jung, Über den kleinsten Kreis, der eine ebene Figur einschliesst, *J. Reine Angew. Math.* **137** (1910), 310–313. 91

[515] L. Kadison and M. T. Kromann, *Projective geometry and modern algebra*, Birkhäuser, 1996 (208 pages). 91

[516] M. Kahle and B. Taha, New lower bounds on $\chi(\mathbb{R}^d)$ for $d = 8, \ldots, 12$, preprint, `arXiv:1409.1278v1` [math.CO], 3 Sep 2014. 315

[517] J. Kahn and G. Kalai, A counterexample to Borsuk's conjecture, *Bull. Amer. Math. Soc. (N.S.)* **29** (1993), 60–62. (A preprint is also available at `arXiv:math/9307229`.) 460, 461

[518] S. Kakeya, Some problems on maximum and minimum regarding ovals, *Tóhoku Science Reports* **6** (1917), 71–78. 467

[519] G. Kalai, Seven problems around Tverberg's theorem, in *Combinatorics and more*, `https://gilkalai.wordpress.com/2008/12/23/seven-problems-around-tverbergs-theorem/` (posted 23 December 2008, accessed 19 January 2016). 499

[520] G. Kalai, *Aubrey de Grey: The chromatic number of the plane is at least 5*, blog, `https://gilkalai.wordpress.com/2018/04/10/aubrey-de-grey-the-chromatic-number-of-the-plane-is-at-least-5/`, 2018. 202

[521] S Kalajdzievski, in collaboration with R. Padmanabhan, *Math and art: an introduction to visual mathematics*, CRC Press, 2008. 458

[522] J. D. Kalbfleisch, J. G. Kalbfleisch, and R. G. Stanton, A combinatorial problem on convex n-gons, *Proc. Louisana Conf. on Combinatorics, Graph Theory, and Computing* (1970), 180–188. 102

[523] S. Kantor, Die Konfigurationen $(3,3)_{10}$, *Sitz.-Ber. Math.-Nat. Cl. Akad. Wiss. Wien* 84 (1881), 1291–1314. 224

[524] S. Kantor, (1882), Über die Configurationen (3, 3) mit den Indices 8, 9 und ihren Zusammenhang mit den Curven dritter Ordnung, *Sitzungsberichte der Mathematisch-Naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften, Wien* **84** (1882), 915–932. 396

[525] J. Kappraff, *Connections: the geometric bridge between art and science*, McGraw-Hill, Inc., 1991. 395

[526] G. Károlyi, Ramsey-remainder for convex sets and the Erdős–Szekeres theorem, *Discrete Appl. Math.* **109** (2001), 163–175. 153, 160

[527] G. Károlyi, J. Pach, and G. Tóth, Ramsey-type results for geometric graphs. I., in ACM Symposium on Computational Geometry (Philadelphia, PA, 1996); *Discrete Comput. Geom.* **18** (1997), 247–255. 228

[528] G. Károlyi and P. Valtr, Point configurations in d-space without large subsets in convex position, *Discrete Comput. Geom.* **30** (2003) 277–286. 494

[529] M. Katchalski and A. Liu, A problem of geometry in R^n, *Proc. Amer. Math. Soc.* **75** (1979), 284–288. 228

[530] G. O. H. Katona, Solution of a Problem of A. Ehrenfeucht and J. Mycielski, *J. Combin. Theory, Ser. A.* **17** (1974), 265–266. 201

[531] N. H. Katz and G. Tardos, (2004), A new entropy inequality for the Erdős distance problem, pp. 119–126 in J. Pach, *Towards a theory of geometric graphs*, Contemporary Mathematics 342, American Mathematical Society, Providence, RI. 360

[532] D. A. Kay, *College geometry*, Holt, Rinehart, and Winston, New York, 1969. 450

[533] D. K. Kazarinoff, A simple proof of the Erdős–Mordell inequality for triangles, *Michigan Mathematical Journall* **4** (1957), 97–98, 75, 127

[534] N. D. Kazarinoff and R. Weitzenkamp, Squaring rectangles and squares, *Amer. Math. Monthly* **80** (1973), 877–888. 124

[535] K. S. Kedlaya, *The Putnam Archive*, website, `https://kskedlaya.org/putnam-archive/`, last accessed 9 December 2019. 299

[536] K. S. Kedlaya, B. Poonen, and R. Vakil, *The William Lowell Putnam Mathematical Competition, 1985–2000: Problems, Solutions, and Commentary*, Mathematical Association of America, Washington, D.C., 2002. 722

[537] L. M. Kelly, Elementary problems and solutions, Isosceles n-points. *Amer. Math. Monthly* **54** (1947), 227–229. 14

[538] L. M. Kelly, A resolution of the Sylvester–Gallai problem of J.-P. Serre, *Discrete Comput. Geom.* **1** (1986), 101–104. 453 376

[539] L. M. Kelly and W. O. J. Moser, On the number of ordinary lines determined by n points, *Canad. J. Math.* **1** (1958), 210–219. 381, 382

[540] L. M. Kelly and R. Rottenberg, Simple points in pseudoline arrangements, *Pacific J. Math.* **40** (1972), 617–622. 382

[541] A. B. Kempe, *How to draw a straight line*, Macmillan and Co., London, 1877. [A small book] 116

[542] P. M. H. Kendall and G. M. Thomas, *Mathematical puzzles for the connoisseur*, Charles Griffin and Company Limited, London, 1962. 188, 416, 714

[543] J. Kepler, Letter to Joachim Tancke (12 May 1608), in his *Gesamelte Werke*, volume 16, 154–165. 266

[544] R. B Killgrove, The k-configuration, preliminary report, abstracts for The January Meeting in San Francisco January 23-26, 1968, abstract number 653-28, in *A. M. S. Notices* **15** (Jan. 1968), p. 85. 390

[545] J. H. Kim, and V. H. Vu, Small complete arcs in projective planes, *Combinatorica* **23** (2) (2003), 311–363. 342

[546] R. J. Kimble, *Extremal problems in dimension theory for partially ordered sets*, Ph.D. Thesis, Massachusetts Institute of Technology, 1973. 510

[547] T. P. Kirkman, On a problem in combinations, *Cambridge and Dublin Math.* **2** (1847), 191–204. 325, 389

[548] T. P. Kirkman, Query VI, *Lady's and Gentleman's Diary*, 1950, p. 48. 389

[549] T. P. Kirkman, Note on an unanswered prize question, *Cambridge Dublin Math. J.* **5** (1850), 255–262. 389

[550] M. S. Klamkin and A. Liu, Three more proofs of Routh's theorem, *Crux Mathematicorum* **7** (7) (Aug–Sep, 1981), 199–203. 128

[551] M. S. Klamkin and D. J. Newman, Cyclic pursuit or "The three bugs problem", *Amer. Math. Monthly* **78** (1971), 631–639. 188, 189

[552] M. Kleber, R. Vakil, and S. Stein, Cutting a polygon into triangles of equal areas, *The Mathematical Intelligencer* **26** (2004), 17. 253

[553] V. Klee and S. Wagon, *Old and new unsolved problems in plane geometry and number theory*, Dolciani mathematical expositions **11**, Cambridge University Press, 1991, p. 21. 92

[554] D. J. Kleitman, The crossing number of $K_{5,n}$, *J. Combin. Th.* **9** (1970), 315–323. 435

[555] D. J. Kleitman, A Note on the parity of the numbers of crossings of a graph, *J. Combin. Th., Ser. B* **21** (1976), 88–89. 435

[556] D. J. Kleitman and L. Pachter, Finding convex sets among points on the plane, *Discrete and Computational Geometry* **19** (1998), 405–410. 226

[557] W. L. Kocay, A note on the Georges configuration and a paper of Grunbaum, *Bulletin of the I.C.A* **58** (2010), 103–111. 398

[558] W. L. Kocay, Non-Fano quads in finite projective planes, *Ars Combinatoria* **83** (2007), 65–92.310

[559] W. L. Kocay, One-point extensions in n_3 configurations, *Ars Math. Contemp.* **10** (2016), 291–322, `https://amc-journal.eu/index.php/amc/article/view/758`. 398

[560] W. L. Kocay, Coordinatizing n_3 configurations, *Ars Mathematica Contemporanea* **15** (2018), 127–145. 388

[561] G. Koester, Bemurkung zu einem Problem von H. Grötzsch, *Wiss. Z. Univ. Halle* **33** (1984), M.H.S.5.S.129. 403

[562] G. Koester, Coloring problems on a class of 4-regular planar graphs, *Graphs, hypergraphs and applications* (Eyba, 1984), Teubner-Texte Math. **73**, Teubner, Leipzig, 1985, pp.102–105. 403

[563] G. Koester, Note to a problem of T. Gallai and G. A. Dirac, *Combinatorica* **5** (1985), 227–228. 403

[564] J. Kollár, L. Rónyai, and T. Szabó, Norm-graphs and bipartite Turán numbers, *Combinatorica* **16** (1996), 399–406. 321

[565] M. Koman, On the crossing numbers of graphs, *Acta Univ. Carolinae Math. Phys.* **10** (1969), 9–46. 445

[566] P. Komjath, Ramsey-type results for metric spaces, *J. Combin. Theory Ser. A* **45** (1987), 323–324. 228

[567] J. Komlós, J. Pintz, and E. Szemerédi, On Heilbronn's triangle problem, *J. London Math. Soc.* **24** (1981), 385–396. 257

[568] J. Komlós, J. Pintz, and E. Szemerédi, A lower bound for Heilbronn's problem, *J. London Math. Soc.* **25** (1982), 13–24. 257

[569] D. König, *Theorie der endlichen und unendlichen Graphen*, Chelsea Publishing Company, New York, 1935. 214

[570] B. A. Kordemsky, *The Moscow puzzles; 359 mathematical recreations*, edited and with an introduction by Martin Gardner, Charles Scribner's Sons, New York, 1972. 239

[571] A. N. Kostovskii, *Geometrical constructions using compasses only*, Translated from the Russian by Halina Moss, Blaisdell Pub. Co., New York and London, 1961. 102

[572] T. Kővári, V. T. Sós, and P. Turán, On a problem of K. Zarankiewicz, *Coll. Math.* **3** (1954), 50–57. 427

[573] M. Kraitchik, *La Mathématique des Jeux*, Brussels, 1930. 290

[574] M. Kraitchik, *Mathematical recreations*, W. W. Norton Company, Inc., New York, 1942. 231, 292

[575] T. Kreid, On representations of natural numbers as sums of 2 squares, *Demonstratio Math.* **9** (1976), 367–378. 520

[576] I. Kříž, Permutation groups in Euclidean Ramsey theory, *Proc. Amer. Math. Soc.* **112** (1991), 899–907. 484

[577] I. Kříž, All trapezoids are Ramsey, *Discrete Math.* **108** (1992), 59–62. 481

[578] H. W. Kuhn, Some Combinatorial lemmas in topology, *IBM Journal of Research and Development* **4** (1980) (5), 518–524. 253

[579] A. Kupavskii, J. Pach, and G. Tardos, *Triangles: Erdős, Tutte, and butterflies* talk given at the 2nd Russian-Hungarian combinatorial workshop, June 27–29, 2018, Budapest. (See `https://www.renyi.hu//conferences/hunrus18/participants.html` for abstract.) 300

[580] A. B. Kupavskii and A. M. Raigorodskii, On the chromatic number of $\mathbb{R}^9$, *Journal of Mathematical Sciences* **163** (2009), 720–731. 460, 461

[581] A. B. Kupavskii and A. M. Raigorodskii, On the chromatic numbers of small-dimensional Euclidean spaces, in "European Conference on Combinatorics, Graph Theory and Applications (Euro- Comb 2009)", *Electron. Notes Discrete Math* **34** (2009), 435–439. 460

[582] Y. S. Kupitz, H. Martini, and M. A. Perles, Ball polytopes and the Vazsonyi problem, `arXiv:0905.1528`, 2009 (80 pages). 448

[583] K. Kuratowski, Sur le problème des courbes gauches en topologie, *Fund. Math.* **15** (1930), 271–283. 145

[584] M. Laczkovich, Equidecomposability and discrepancy; a solution of Tarski's circle-squaring problem, *J. Reine Angew. Math. (Crelle's Journal)* **404** (1990), 77–117. 262

[585] C. W. Lam, The search for a finite projective plane of order 10, *Amer. Math. Monthly* **98** (April 1991) 305–318. 310

[586] C. W. Lam, J. McKay, S. Swiercz, and L. Thiel, The nonexistence of ovals in a projective plane of order 10, *Discrete Math.* **45** (1983), 319–321. 342

[587] C. W. H. Lam, G. Kolesova, and L. Thiel, A computer search for finite projective planes of order 9, *Discrete Math.* **92** (1991), 187–195. 310

[588] P. S. de Laplace, *Théorie analytique des probabilités*, Paris, V. Courcier, 1812. 183

[589] C. W. Lam, S. Swiercz, and L. Thiel, The non-existence of finite projective planes of order 10, *Can. J. Math.* **XLI** (1989), 1117–1123. 310

[590] D, G. Larman, On the realization of distances within coverings of an n-sphere, *Mathematika* **14** (1967), 203–206. 456

[591] D. G. Larman and C. A. Rogers, The realization of distances within sets in Euclidean space, *Mathematika* **19** (1972), 1–24. 460, 461

[592] D. G. Larman, C. A. Rogers, and J. J. Seidel, Ont two-distance sets in Euclidean space, *Bull. London Math. Soc.* **9** (1977), 261–267. 454

[593] M. E. Larsen, The eternal triangle—a history of a counting problem, *The College Math. J.* **20** (1989), 370–384. 698, 700

[594] S. R. Lay, *Convex sets and their applications*, Dover Publications, Mineola, New York, 1982. 93, 193, 194, 730

[595] I. Leader, P. A. Russell, and M. Walters, Transitive sets in Euclidean Ramsey theory, `arXiv:1012.1350v1` [Math.CO] 6 Dec. 2010. 484

[596] H. Lee, Another proof of the Erdős–Mordell theorem, *Forum Geom.* **1** (2001), 7–8. 124

[597] H. Lefmann and N. Schmitt, A deterministic algorithm for Heilbronn's problem in dimension three (extended abstract), *Proceedings of the 5th Latin American Theoretical Informatics*, LATIN'02, Springer, LNCS 2286, ed. S. Rajsbaum, 2002, 165–180. 259

[598] H. Lefmann, On Heilbronn's problem in higher dimensions, *Combinatorica* **23** (2003), 669–680. 259

[599] H. Lefmann, Distributions of points in d dimensions and large k-point simplices, *Discrete and Computational Geometry* **40** (2008), 401–413. 259

[600] H. Lefmann, Point sets in the unit square and large areas of convex hulls of subsets of points, *J. Comb. Optim* **16** (2008), 182–195. 258

[601] H. Lefmann, *personal communication*, 9 Sept. 2017. 258

[602] J. Lehel, Course notes for *Combinatorial Geometry*, University of Memphis, TN, 2009. 752

[603] I. Lehmann and A. S. Posamentier, *The fabulous Fibonacci numbers*, Prometheus Books, New York, 2007. 94, 707

[604] F. T. Leighton, *Complexity Issues in VLSI*, Foundations of Computing Series, MIT Press, Cambridge MA, 1983. 441

[605] F. T. Leighton, New lower bound techniques for VLSI, *Math. Systems Theory* **17** (1984), 47–70. 441

[606] H.-C. Lenhard, Verallgemeinerung und Verschärfung der Erdős–Mordellschen Ungleichung für Polygone, *Archiv für Mathematische Logik und Grundlagenforschung* **12** (1961), 311–314, 124

[607] F. Levi, Die Teilung der projektiven Ebene durch Gerade oder Pseudogerade, *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Math-Physik Classe* **78** (1926), 256–267. 382

[608] F. W. Levi, *Finite geometrical systems*, University of Calcutta, Calcutta, 1942, 51 pages. [These notes are based on a series of six lectures given in Calcutta in 1940.] 380, 382

[609] H. Levy, *Projective and related geometries*, The Macmillan Co., New York; Collier-Macmillan, Ltd., London, 1964. 382

[610] L. S. Levy, *Geometry: Modern mathematics via the Euclidean plane*, Prindle, Weber & Schmidt, Inc., Boston, MA, 1970. 89, 382

[611] S. Levy, *Flavors of geometry*, MSRI Berkeley series, Cambridge University Press, 1997. 382, 549

[612] X. B. Lin, Another brief proof of the Sylvester Theorem, *Amer. Math. Monthly* **95** (Dec. 1988), 932–933. 376

[613] H. Lindgren, *Recreational problems in geometric dissections & how to solve them*, Dover, 1972. 239

[614] J. E. Littlewood, *Littlewood's Miscellany*, B. Bollobás ed., Cambridge University Press, 1986, vi+200 pp. 190

[615] M. Livio, *The golden ratio: the story of PHI, the world's most astonishing number*, Broadway Books, New York, 2002. 94

[616] C.-Y. Lo, J. Matousek, and W. L. Steiger, Algorithms for ham-sandwich cuts, *Discrete and Computational Geometry* **11** (1994), 433–452. 413

[617] P-S Loh, 55th International Mathematical Olympiad, *Mathematics Magazine* **87** (2015), 310–317. 704, 751

[618] J. Q. Longyear, Large tactical configurations, *Discrete Math.* **4** (1973) 379–382. 391

[619] J. Q. Longyear and T. D. Parsons, The friendship theorem, *Indagationes Mathematicae (Proceedings)* **75** (1972), 257–262. 322

[620] E. S. Loomis, *The Pythagorean proposition*, National Council of Teachers of Mathematics, Reston, VA, 1968. 41

[621] N. Lord, Balancing and golden rectangles, *The Mathematical Gazette* **79** (1995), 573–574. 101

[622] L. Lovász, *Combinatorial problems and exercises*, Akadémiai Kiadó, Budapest, 1993. 224, 718

[623] Lowry, Solution to question 269, [proposed] by Mr. W. Wallace. In *New series of the Mathematical Repository* (T. Leybourn, ed.) **3** (1814) part I, 44–46. Published by W. Glendinning, London. 232

[624] D. Lubell, A short proof of Sperner's lemma, *J. Combin. Theory* **7** (1966), 299. (Also reprinted in: *Classic Papers in Combinatorics*, I. Gessel, G-C. Rota, eds., Birkhauser, Boston, 1987.) 361

[625] E. Lucas, Problem 251, *Nouvelle Correspondance Mathématique* **3** (1877), edited by E. Catalan, Brussels, pp. 175–176. 189

[626] F. H. Lutz, Császár's torus, *Elect. Geom. Models*, 2001.02.069, 2001. 153

[627] J. Lützen, Why was Wantzel overlooked for a century? The changing importance of an impossibility result, *Historia Mathematica* **36** (Nov. 2009), 374–394. 104

[628] D. Mackenzie, A tisket, a tasket, an Apollonium gasket, pp. 13–27 in *The best writing on mathematics, 2011*, edited by M. Pitici, Princeton University Press, Princeton and Oxford, 2012. [This article first appeared in *American Scientist*, but I do not have the reference.] 75

[629] S. MacLane, A lattice formulation for transcendence degrees and p-bases, *Duke Math. J.* **4** (1935), 455–468. 506

[630] D. A. Madore, *The Hadwiger–Nelson problem over certain fields*, preprint, available at `https://arxiv.org/abs/1509.07023v1`. (Accessed 8 Oct. 2017). 464

[631] J. Malkevitch, *Linkages: From Fingers to Robot Arms (I) Introduction*, Feature column from the AMS, September 2002, `http://www.ams.org/publicoutreach/feature-column/fcarc-linkages1`, last accessed 13 June 2021. 116

[632] M. Mann, A new bound for the chromatic number of the rational five-space, *Geombinatorics* **11** (2001), 49–53. 464

[633] W. Mantel, Solution to Problem 28, by H. Gouwentak, W. Mantel, J. Teixeira de Mattes, F. Schuh, and W. A. Wythoff, *Wiskundige Opgaven* **10** (1907), 60–61. 423

[634] D. Manušič and T. Pisanski, ; The remarkable generalized Petersen graph G(8,3), *Mathematica Slovaca* **50** (2000), 117–121. 395

[635] E. Maor, *e: The story of a number*, Princeton University Press, Princeton, New Jersey, 1994. 189

[636] E. A. Marchisotto, The theorem of Pappus: A bridge between algebra and geometry, *Amer. Math. Monthly* **109** (May 2002), 497–516. 56

[637] D. S. Marinescu and M. Monea, About a strengthened version of the Erdős–Mordell inequality, *form Geom.* **17** (2017), 197–202. 124

[638] F. Marko and S. Litvinov, Routh's theorem for tetrahedra, *Geom. Dedicata* **174** (2015), 155–167. 128

[639] F. Marko and S. Litvinov, On the Steiner-Routh theorem for simplices, *Amer. Math. Monthly* **124** (2017), 422–435. 127, 128

[640] G. Markowsky, Misconceptions about the golden ratio, *The College Mathematics Journal* **23** (1992), 2–19. 94

[641] G. E. Martin, *Transformation geometry; an introduction to symmetry*, Springer, 1982. 80, 185, 624

[642] G. E. Martin, *Polyominoes: Puzzles and problems in tiling*, Mathematical Association of America, Washington, D. C., 1991. 287

[643] L. Mascheroni, *Geometria del Compasso*, 1797. 102

[644] C. Maslanka, Chris Maslanka's puzzles, in *The Guardian*, London, 24 June 2017, p. 55. 124

[645] J. Matoušek, *Lectures on discrete geometry*, Graduate Texts in Mathematics **212**, Springer, 2002. 15, 201, 202, 203, 228, 273, 274, 279, 418, 420, 441, 442, 444, 451, 496, 497, 516, 522, 604

[646] J. Matoušek, *Using the Borsuk-Ulam Theorem: Lectures on topological methods in combinatorics and geometry*, Springer, 2003. 465

[647] J. Matoušek and V. Rödl, On Ramsey sets in spheres. *J. Combin. Theory Ser. A* **70** (1995), 30–44. 488

[648] R. D. Mauldin, ed., *The Scottish book*, Birkäuser, Boston, 1981. 413

[649] V. R. Maurer, G. L. Musser, and L. E. Trimpe, *College Geometry: a problem solving approach with applications*, 2nd ed., Pearson Prentice Hall, Upper Saddle River, New Jersey 07458. [Note: The order of authors given in this text is Musser, Trimpe, Maurer.] 102

[650] S. Mazur and S. M. Ulam, Sur les transformations isométriques des espaces vectoriels normés, *C. R. Acad. Sci. Paris* **194** (1932), 946–948. 90

[651] P. McMullen, The maximum number of faces of a convex polytope, *Mathematika* **17** (1970), 179–184. 210

[652] E. Melchior, Über Vielseite der projektiven Ebene, *Deutsche Math.* **5** (1941), 461–474. [Some sources say 1940.] 381

[653] K. Mellinger, R. Vaughn, and O. Vega, Graphs embedded into finite projective planes, *Contributions to Discrete Mathematics*, **10** (2015), 113–125. 323, 399, 747

[654] Z. A. Melzak, *Companion to concrete mathematics*, John Wiley & Sons, New York, 1973. 127

[655] N. S. Mendelsohn, Tiling with dominoes, *The College Mathematics Journal*, **35** (2004), 115–120. (Published by the Mathematical Association of America.) 284

[656] H. Meschkowski, *Noneuclidean geometry*, Academic Press, 1964. [Translated from the original 1961 version by A. Shenitzer.] 114

[657] J. G. Mikusiński, Sur quelques propriétés du triangle, (in French) *Ann. Univ. M. Curie-Skłodowska* **1** (1946), 45–50. 126

[658] H. Minkowski, *Geometrie der Zahlen*, G. Teubner, Leipzig and Berlin, 1910. [Matousek [645] gives the publication date of 1896, but Blichfeldt [108] gives 1910.] 272, 273, 274, 277

[659] D. G. Mixon, *Polymath16, first thread: Simplifying de Grey's graph*, https://dustingmixon.wordpress.com/2018/04/14/polymath16-first-thread-simplifying-de-greys-graph/, 14 April 2018. 459

[660] A. F. Möbius, Kann von zwei dreiseitigen Pyramiden eine jede in Bezug auf die andere um—und eingeschrieben zugleich heissen?, *Journal für die reine und angewandte Mathematik* **3** (1828), 273–278. (Also in *Gesammelte Werke* (1886), vol. 1, pp. 439–446.) 395

[661] B. Mohar and C. Thomassen, *Graphs on Surfaces*, John Hopkins Studies in the Mathematical Sciences, The John Hopkins University Press, 2001. 446

[662] G. Mohr, *Euclides Danicus*, 1672. 102

[663] P. Monsky, On dividing a square into triangles, *Amer. Math. Monthly* **77** (1970), 161–164. 253

[664] H. L. Montgomery, I. Niven, and H. S. Zuckerman, *An introduction to the theory of numbers*, 5th ed., John Wiley & Sons, Inc., 1991. 279

[665] E. H. Moore, Tactical memoranda I–III, *Am. J. Math.* **18** (1896), 264–333. 391

[666] T. L. Moore, Using Euler's formula to solve plane separation problems, *The College Mathematics Journal* **22** (1991), 125–130. 408

[667] L. J. Mordell Solution to 3740, *Amer. Math. Monthly* **44** (1937), 252–254. 123

[668] F. V. Morley, Linkages, *The Scientific Monthly* **9** (1919 Oct.), 366–378. 116, 119

[669] Z. Moroń, O rozkładach prostokatów na kwadraty, *Przeglad Mat. Fiz.* **3** (1935), 151–152. 290

[670] W. Morris and V. Soltan, The Erdős-Szekeres problem on points in convex position—a survey, *Bull. Amer. Math. Soc.* **37** (2000), 437–458. 228

[671] I. Moscovich, *The puzzle universe: A history of mathematics in 315 puzzles*, Firefly books, Richmond Hill, Ontario, 2015. 396 pp. 239

[672] L. Moser, On the different differences determined by n points, *Amer. Math. Monthly* **59** (no. 2) (Feb. 1952), 85–91. 450, 451

[673] L. Moser, Problems and questions, Problem 102, *Math. Magazine* **25** (no. 4) (March-April 1952), 219. 416

[674] L. Moser and W. Moser, Solution to problem 10, *Canad. Math. Bull.* **4** (1961), 187–189. 457, 461

[675] T. Motzkin, The lines and planes connecting the points of a finite set, *Trans. Amer. Math. Soc.* **70** (1951), 451–464. 376, 381

[676] T. S Motzkin, Nonmixed connecting lines, *Notices of the American Mathematical Society* **14** (1967), 837. 383

[677] T. Motzkin, Sets for which no point lies on many connecting lines, *J. Combin. Theory A* **18** (1975), 345–348. 383

[678] T. S. Motzkin and E. G. Straus, Maxima for graphs and a new proof of a theorem of Turán, *Canad. J. Math.* **17** (1965), 533–540. 424

[679] F. R. Moulton, A simple non-Desarguesian geometry, *Trans. Amer. Math. Soc.* **3** (1902), 192–195. 390

[680] D. Mumford, C. Series, and D. Wright, *Indra's pearls: the vision of Felix Klein*, Cambridge University Press, 2002. 116

[681] M. Nathanson, *Additive number theory; inverse problems and the geometry of sumsets*, Graduate Texts in Mathematics **165**, Springer, 1996. 524

[682] O. Nechushtan, On the space chromatic number, *Discrete Math.* **256** (2002), 499–507. 459, 461

[683] C. M. Nicolás, The empty hexagon theorem, *Discrete and Computational Geometry* **39** (2008), 389–397. 228

[684] D. Niederman, *The puzzler's dilemma*, Perigee, New York, 2012. 284

[685] N. Nilakantan, Extremal problems related to the Sylvester–Gallai Theorem, pp. 479–494 in *Combinatorial and computational geometry*, (Jacob E. Goodman, Janos Pach, Emo Welzl, eds) MSRI publications **52**, 2005 (Published by Cambridge University Press). Also available online at `library.msri.org/books/Book52/files/25nila.pdf`, (accessed 22 July 2016). 377, 381, 383

[686] I. Niven, A new proof of Routh's theorem, *Math. Mag.* **49** (1976), 25–27. 127

[687] I. Niven and H. S. Zuckerman, Lattice points and polygonal area, *Amer. Math. Monthly* **74** (1967), 1195. 69

[688] T. H. O'Beirne, *Puzzles and paradoxes: fascinating excursions in recreational mathematics*, New York, Oxford University press,1965. (Recently reprinted by Dover.) 268

[689] E. Olson and J. Spencer, Balancing families of sets, *J. Combin. Theory Ser. A* **25** (1978), 29–37. 366

[690] O. Ore, *Number theory and its history*, McGraw-Hill Book Company, New York, 1948. (Reprinted by Dover Publications, Mineola, N. Y., 1988.) 519

[691] O. Ore, *The four-color problem*, Pure and Applied Mathematics **27**, Academic Press, New York and London, 1967. 146

[692] J. O'Rourke, *Art gallery theorems and algorithms*, Oxford University Press, 1987. Out of print, but available at `http://cs.smith.edu/~orourke//books/ArtGalleryTheorems/art.html` 180

[693] J. O'Rourke, *Computational geometry in C*, 2nd ed., Cambridge University Press, 1998. 180, 468

[694] A. Ostermann and G. Wanner, *Geometry by its history*, Springer, 2012. 73

[695] T. G. Ostrom, Double transitivity in finite projective planes, *Canad. J. Math* **8** (1956), 563–567. [See correction for case $n = 9$ in *Canad. J. Math.* **10** (1958), 507–512.] 316

[696] T. G. Ostrom, Dual transitivity in finite projective planes, *Proc. Amer. Math. Soc.* **9** (1958), 55–56. 315

[697] T. G. Ostrom, Vector spaces and construction of finite projective planes, *Arch. Math.* **19** (1968), 1–25. 315

[698] T. G. Ostrom and A. Wagner, On projective and affine planes with transitive collineation groups, *Math. Zeitschr.* **71** (1959), 186–189. 315

[699] J. Pach and P. K. Agarwal, *Combinatorial geometry*, Wiley, New York, 1995. [This text is based on notes taken by Agarwal from lectures by Pach.] 496

[700] J. Pach, R. Radoičić, G. Tardos, and G. Tóth, Improving the crossing lemma by finding more crossings in sparse graphs, *Disc. Comput. Geometry*, **36** (2006), 527–552. 442

[701] J. Pach and J. Solymosi, Canonical theorems for convex sets, *Discrete Comput. Geom.* **19** (1998), 427–435. 228

[702] J. Pach, J. Spencer, and G. Tóth, New bounds on crossing numbers, *Discrete Comput. Geom.* **24** (2000), 623–644. 445

[703] J. Pach and G. Tóth, Graphs drawn with few crossings per edge, *Combinatorica* **17** (1997), 427–439. 442

[704] J. Pach and G. Tóth, Which crossing number is it anyway? *J. Combin. Theory Ser. B* **80** (2000), 225–246. 429

[705] J. Pach and G. Tóth, Degenerate crossing numbers, *Discrete Comput. Geom.* **41** (2009), 376–384. 440, 442

[706] R. Padmanabhan, *personal communication*, 11 July 2016. 318

[707] D. Pagni, Regular pentagon construction, *Mathematics in School* **29**(1) (2000), 14–15. Reprinted in [734], pp. 296–298. 112

[708] I. Pak, History of Catalan numbers, Appendix B, pages 177–189 in [819], 2015. 63

[709] I. Pak, *Catalan numbers page*, website, `https://www.math.ucla.edu/~pak/lectures/Cat/pakcat.htm`, last accessed 29 December 2019. 63

[710] J. Pál, Ein Minimumproblem für Ovale, *Math. Annalen* **83** (1921), 311–319. 499

[711] S. Pan and R. B. Richter, The crossing number of K_{11} is 100, *J. Graph Theory* **56** (2007), 128–134. 437

[712] E. T. Parker and R. B. Killgrove, A note on projective planes of order 9, *Mathematics of Computation* **18** (1964), 506–508. (Published by the A.M.S.) 310

[713] E. Pascal, *Repertorium der Hoheren Mathematik, Bd.2: Geometrie*, Teubniger, Leipzig, 1910. 394

[714] R. G. Pawley, 5-con triangles, *Mathematics Teacher* **60** (May 1967), 438–443. 38

[715] D. Pedoe, *Geometry, a comprehensive course*, Dover Publications, Inc., Mineola, N. Y., 1988. (This is a corrected republication of the work first published by Cambridge University Press, 1970.) 73, 127

[716] D. Pedoe, *Circles: a mathematical view*, Mathematical Association of America, 1995. 76

[717] H. Perigal, On geometric dissections and transformations, *Messenger of Mathematics* (second series) **2** (1872), 103–106. [The article was published in November 1872, however is in a volume with 1873 on the front. Perigal's article ends partway down p. 105, where 1872 indicates the end, but some comments were added by the editor J. W. L. Glaisher, who noted that Perigal's proof does slightly more than just the Pythagorean theorem; JWLG's comments conclude partway down p. 106.] 236

[718] J. Perkal, Sur la subdivision des ensembles en parties de diamètre inférieur, *Colloquium Mathematicum* **2** (1947), 45. 467

[719] O. Perron, Über eine Satz von Besicovitch, *Mathematische Zeitschrift* **28** (1928), 383–386. 500

[720] C. Peters, An interesting problem to "bug" your students with, *Alabama Journal of Mathematics* **3** (2), 1979. Available at `https://faculty.missouri.edu/~casazzap/pdf/teach/bug.pdf`. (Accessed 9 October, 2017.) 189

[721] R. E. Pfiefer, The historical development of J. J. Sylvester's four point problem, *Math. Mag.* **62** (1989), 309–317. 183

[722] π *in the sky*, December 2000, The Pacific Institute for the Mathematical Sciences (PIMS). 673

[723] 2017 Math Quickies, Problem 8, π *in the sky* **21** (Spring 2019), p. 26, The Pacific Institute for the Mathematical Sciences (PIMS). 740

[724] G. Pick, Geometrisches zur Zahlentheorie, *Lotos (Prague)* **19** (1899), 311–319. 68, 69

[725] R. Pinchasi, Gallai–Sylvester theorem for pairwise intersecting unit circles, *Discrete Comput. Geom.* **28** (2002), 607–624. 384, 385

[726] T. Pisanski and B. Servatius, *Configurations from a graphical viewpoint*, Birkhäuser Advanced Texts, Birkkäuser, New York, 2013. 388

[727] B. Polster, Viviani à la Kawasaki: Take two, *Mathematics Magazine* **87** (2014), 280–283. 122

[728] P. Polito and O. Polverino, On linear blocking sets in PG$(2, q^4)$, Combinatorics '98 (Palermo), *Discrete Math.* **255** (2002) nos.1–3, 343–348. 355

[729] O. Polverino and L. Storme, Small minimal blocking sets in PG$(2, q^3)$, *European J. Combin.* **23** (2002), 83–92. 355

[730] G. Pólya, Über die Analogie der Kristallsymmetrie in der Ebene, *Zeitschrift für Kristallographie* **60** (1924), 278–282. 185

[731] G. Pólya, *Induction and analogy in mathematics*, (Volume I: mathematics and plausible reasoning; Volume II: patterns of plausible inference), Princeton University Press, Princeton, New Jersey, 1954. 692, 754

[732] A. Pór and P. Valtr, The partitioned version of the Erdős-Szekeres theorem, *Discrete Comput. Geom.* **28** (2002), 625–637. 228

[733] A. S. Posamentier and C. T. Salkind, *Challenging problems in algebra*, Dover Publications, Inc., New York, 1996. 751

[734] C. Pritchard, ed., *The changing shape of geometry: celebrating a century of geometry and geometry teaching*, Cambridge University Press, 2003. 588, 608

[735] D. Pritikin, All unit-distance graphs of order 6197 are 6-colorable, *J. Combin. Theory, Ser. B* **73** (1998), 159–163. 458

[736] A. Prymak, *personal communication*, 15 May 2018. 123

[737] E. J. Putzer and R. W. Loewen, *On the optimum method of cutting a rectangular box into unit cubes*, Research Memorandum, Convair Scientific Research Laboratory, San Diego, 1958. 417

[738] R. Radoičić and G. Tóth, Note on the chromatic number of the space, in "Discrete and Computational Geometry: The

Goodman–Pollack Festschrift”, *Algorithms and Combinatorics* **25** (2003), 695–698. Published by Springer, Berlin. 459, 460, 461

[739] J. Radon, Mengen konvexer Körper, die einen gemeinsamen Punkt enthalten, *Math. Ann.* **83** (1921), 113–115. 199

[740] S. P. Radziszowski, Small Ramsey numbers, *Electronic J. Combin.* **1** Dynamic survey DS1, `http://www.combinatorics.org/Surveys/` (last updated March 2017). 220

[741] A. M. Raigorodskii, The Borsuk problems and chromatic number of some metric spaces (in Russian), *Uspekhi Mat. Nauk* **56** (2001), 107–146; translation in *Russian Math. Surveys* **56** (2001), 103–139. 460, 461

[742] A. M. Raigorodskii, On the chromatic numbers of spheres in $\mathbb{R}^n$, *Combinatorica* **32** (2012), 111–123. Preprint available at `arXiv:1010.0384v1` [math.CO] 3 Oct. 2010. 465

[743] A. M. Raigorodskii, Coloring distance graphs and graph diameters, pp. 429–460 in J. Pach (ed.), *Thirty Essays on Geometric Graph Theory*, Springer, 2013. 460

[744] D. E. Raiskii, Realization of all distances in a decomposition of the space R^n into $(n+1)$ parts, *Mat. Zametki* **7** (1970), 319–323 (Russian) and *Math. Notes* **7** (1970), 194–196 (English translation). 459, 460, 461, 472

[745] F. P. Ramsey, On a problem of formal logic, *Proc. London Math. Soc.*, **30** (1930), 264–286. 217

[746] D. K. Ray-Chaudhuri and R. M. Wilson, Solution of Kirkman’s schoolgirl problem, *Amer. Math. Soc. Symp. Pure Math.* **19** (1971), 187–204. 389

[747] H. Reichard and H. Toepken, Aufgaben und Lösungen, *Jahresbericht der Deutschen Math. Verinigung* **50** (1940), 13–14. 292

[748] I. Reiman, Uber ein Problem von K. Zarankiewicz, *Acta Math. Acad. Sci. Hung.* **9** (1959), 269–279. 321

[749] F. Richman and J. Thomas, Problem 5479, *Amer. Math. Monthly* **74** (1967), 329. 253

[750] R. B. Richter and G. Salazar, Crossing numbers, pp. 133–150 in *Topics in topological graph theory* (L. W. Beineke and R. J. Wilson, eds.), Encyclopedia Math. Appl. **128**, Cambridge University Press, 2009. 429

[751] A. S. Riesling, (in Russian, title translated) Borsuk's problem in three-dimensional spaces of constant curvature, *Ukr. Geom. Sbornik* **11** (1971), 78–83. 467

[752] J. Riordan, *Combinatorial identities*, John Wiley, New York, 1968. 63

[753] F. S. Roberts, *Applied Combinatorics*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984. 64, 330

[754] F. S. Roberts and B. Tesman, *Applied combinatorics*, 2nd ed., CRC Press, 2009. 64, 65

[755] J. Robertson and W. Webb, *Cake cutting algorithms*, A. K. Peters, Natick, MA, 1998. 412

[756] N. Robertson, D. Sanders, P. Seymour, and R. Thomas, The four-colour theorem, *J. Combin. Th. Ser. B* **70** (1997), 2–44. 146

[757] V. Rödl, On a problem in combinatorial geometry, *Discrete Math.* **45** (1983), 129–131. 466

[758] V. Rödl, A note on the discrepancy of a finite projective plane, manuscript, 1994. 368

[759] K. M. Rogers, The finite field Kakeya problem, *Amer. Math. Monthly* **108** (8) (2001), 756–759. 501

[760] L. J. Rogers, Biography of Henry Perigal, in appendix of "On certain Regular Polygons in Modular Network", *Proceedings London Math. Soc.* **s1-29** (1897), 732–735. http://plms.oxfordjournals.org/content/s1-29/1/706.full.pdf 236

[761] S. Roman, *Field theory*, Springer-Verlag, 1995. 104

[762] K. H. Rosen, *Discrete mathematics and its applications*, 6th ed., McGraw-Hill, New York, 2007. 69

[763] K. F. Roth, On a problem of Heilbronn, *J. London Math. Soc.* **26** (1951), 198–204. 254, 255, 257

[764] K. F. Roth, Remark concerning integer sequences, *Acta Arith.* **9** (1964), 257–260. 365

[765] K. F. Roth, On a problem of Heilbronn II, *Proc. London Math Society* (3) **25** (1972), 193–212. 254

[766] K. F. Roth, On a problem of Heilbronn III, *Proc. London Math Society* (3) **25** (1972), 543–549. 254

[767] K. F. Roth, Estimation of the area of the smallest triangle obtained by selecting three out of n points in a disc of unit area, pp. 251–262 in *Proc. Symp. Pure Mathematics* **24**, AMS, Providence, 1973. 254

[768] K. F. Roth, Developments in Heilbronn's triangle problem, *Advances in Math.* **22** (1976), 364–385. 254

[769] J. Rotman, *Galois theory*, 2nd ed., Springer, 1998. 104

[770] W. W. Rouse Ball, revised by H. S. M. Coxeter, *Mathematical recreations & essays*, 11th edition, MacMillan and Co., Limited, London, 1947. [Coxeter revised the 10th edition.] 232, 235, 238, 239, 242, 290, 325

[771] E. J. Routh, *A treatise on analytical statics, with numerous examples*, 2nd ed., Vol. 1, Cambridge University Press, London, 1896. 127

[772] D. Rudenko, *Arithmetic of 3-valent graphs and equidissections of flat surfaces*, preprint, arXiv:1411.0285v2 20 October 2015. (Accessed 12 March 2017.) 253

[773] C. Salkind, More-than-similar triangles, *Mathematics Teacher* **47** (1954), 561–562. 38

[774] S. Saraf and M. Sudan, Improved lower bound on the size of Kakeya sets over finite fields, *Analysis and PDE*, **1** (3) (2008), 375–379. 501

[775] N. Sauer, On the density of families of sets, *J. Combin. Theory Ser. A* **13** (1972), 145–147. 497

[776] M. Schaefer, The graph crossing number and its variants: a survey, *Electronic J. Combin.* 2014, #DS21, 100 pages. 429, 430, 436

[777] E. R. Scheinerman and H. S. Wilf, The rectilinear crossing number of a complete graph and Sylvester's "four point problem" of geometric probability, *Amer. Math. Monthly* **101** (1994), 939–943. 184, 444, 445

[778] W. M. Schmidt, On a problem of Heilbronn, *J. Lond. Math. Soc.* (2) **4** (1972), 545–550. 254, 257

[779] S. Schonefeld, *Amorous bugs and pursuit problems*, online article, rfdz.ph-noe.ac.at/fileadmin/Mathematik_Uploads/ACDCA/.../DES.../Schonefeld.pdf, (accessed 11 October 2017). 189

[780] O. Schramm, Illuminating sets of constant width, *Mathematika* **35** (1988), 180–189. 467

[781] H. Schroeter, Ueber lineare Konstructionen zur Herstellung der Konfigurationen n_3, *Nachr. Gesellschaft der Wissenschaften Göttingen* **1888** (1888), 237–253. Available at https://eudml.org/doc/180195. 397

[782] H. Schroeter, Ueber die Bildungsweise und geometrische Konstruction der Konfigurationen 10_3, *Nachr. Gesellschaft der Wissenschaften Göttingen* **1889** (1889), 193–236. 397

[783] J. A. Segner, Enumeratio modorum quibus figurae planae rectilinae per doagonales dividuntur in triangula, *Novi Comment. Acad. Sci. Imp. Petropol.* **7** (1758/59, published in 1761), 203–210. Available at Pak's website `http://www.math.ucla.edu/~pak/lectures/Cat/pakcat.htm` 64

[784] B. Segre, Ovals in a finite projective plane, *Canad. J. Math.* **7** (1955), 414–416. 350

[785] B. Segre, Le geometrie di Galoi, *Ann. Mat. Pura Appl.* **48** (1959), 1–97. 341

[786] B. Segre, Introduction to Galois geometry, ed. J. W. P. Hirschfeld, *Mem. Accad. Naz* **8** (1967), 133-263. 341

[787] E. Seneta, K. H. Parshall, and F. Jongmans, Nineteenth-Century Developments in Geometric Probability: J. J. Sylvester, M. W. Crofton, J.-É. Barbier, and J. Bertrand, *Archive for History of Exact Sciences* **55** (6) (June 2001), 501–524 180, 183

[788] L. E. Shader, All right triangles are Ramsey in $\mathbb{E}^2$! *J. Combin. Theory Ser. A* **20** (1976), 385–389. 475

[789] S. Shelah, A combinatorial problem, stability and order for models and theories in infinitary languages, *Pacific J. Mathematics* **41** (1972), 247–261. 497

[790] I. Shemer, Neighborly Polytopes, *Israel J. Math.* **43** (1982), 291–314. 210, 213

[791] G. C. Shephard, A theorem on cyclic polytopes, *Israel J. Math.* **6** (1968), 368–372. 211, 213

[792] I. D. Shkredov, On some problems of Euclidean Ramsey theory, preprint, `arXiv:1507.02727v2` [Math.CO] 22 July 2015 (accessed 27 May 2016). 459

[793] A. V Shubnikov and V. A. Koptsik, *Symmetry in science and art*, translated by David Harker, Plenum Press, New York and London, 1974. 80

[794] W. Sierpinski, *Pythagorean triples*, translated by A. Shamar, first published by Graduate school of science, Yeshiva University, New York, in 1962 (number 9 in its Scripta Mathematica studies), Dover Publications, Inc., 2003. 43

[795] M. Simonovits, Extremal graph theory, in *Selected Topics in Graph Theory 2*, (eds. L. W. Beineke, R. J. Wilson), Academic Press, New York, 1983, 161–200. 424

[796] R. Simson, An explication of an obscure passage in Albert Girard's commentary upon Simon Stevin's works, *Phil. Trans. Royal Society of London* **48** (1753), 368–376. 266

[797] Singapore Kangaroo Contest, Student (JC1, JC2, Poly) 2015, available at `http://mathkangaroo.sg/samples/` (Accessed 12 March 2017.) 710

[798] D. A. Singer, *The rectilinear crossing number of certain graphs*, manuscript, 1971. Available at `http://www.cwru.edu/artsci/math/singer/publish/Rectilinear_crossings.pdf` (accessed 18 Dec. 2017). 438, 445

[799] J. D. Skinner II, *Squared squares—who is who, and what is what*, privately published, Lincoln, Nebraska, 1993, 167 pages. 299

[800] N. J. A. Sloane, *A handbook of integer sequences*, Academic Press, 1973. 414

[801] J. R. Smart, *Modern geometries*, 5th ed., Brooks/Cole Publishing Company, 1998. 114

[802] A. Soifer, *How does one cut a triangle?*, Center for Excellence in Mathematical Education, Colorado Springs, 1990. 726

[803] A. Soifer, *The mathematical coloring book; mathematics of coloring and the colorful life of its creators*, Springer, 2009. 456, 457, 459, 464

[804] A. Soifer, Chromatic number of the plane & its relatives, history, problems and results: an essay in 11 parts, *Ramsey theory* (A. Soifer, ed.), 115–120, Progr. Math. **285**, Birkhäuser/Springer, New York, 2011. 460

[805] R. M. Solovay, A model of set-theory in which every set of reals is Lebesgue measurable, *Annals of Mathematics. Second Series* **92** (1970), 1–56. 459

[806] J. Solymosi and T. Tao, An incidence theorem in higher dimensions, *Discrete and Computational Geometry* **48** (2012), 255–280. 400

[807] J. Solymosi and Cs. Tóth, Distinct distances in the plane, *Discrete Comput. Geom.* **25** (2001), 629–634. 450

[808] J. Solymosi and V. Vu, Near optimal bounds for the Erdős distinct distances problem in high dimensions, *Combinatorica* **28** (2008), 113–125. 450

[809] I. S. Sominskiĭ, *The method of mathematical induction*, (Translated from the Russian by Halina Moss; translation editor Ian N. Sneddon), Pergamon Press, Oxford, 1961. (Also published by Blaisdell Publishing Co., a division of Random House, New York, London, 1961. vii+57 pp.) 664, 672, 673, 674, 675, 678, 680, 681, 683, 754

[810] J. H. Spencer, All finite configurations are almost Ramsey, *J. Combin. Theory, Ser. A* **27** (1979), 401–403. 484

[811] J. Spencer, Six standard deviations suffice, *Trans. Amer. Math. Soc.* **289** (1985), 679–706. 367, 369

[812] J. Spencer, Coloring the projective plane, *Discrete Math.* **73** (1988/89), 213–220. 367, 373

[813] J. Spencer, E. Szemerédi, and W. T. Trotter, Unit distances in the Euclidean plane, pp. 293–308 in *Graph theory and Combinatorics* (B. Bollobás, ed.), Academic Press, 1984. 451, 452

[814] E. Sperner, Ein Satz über Untermengen einer endlichen Menge, *Mathematische Zeitschrift* **27** (1928), 544–548 (in German). 246

[815] E. Sperner, Neuer Beweis für die Invarianz der Dimensionszahl und des Gebietes, *Abh. Math. Sem. Hamburg* **6** (1928), 265–272. 248

[816] E. Sperner, Fifty years of further development of a combinatorial lemma, Numerical solution of highly nonlinear problems, pp. 183–197, 199–217, in *Sympos. Fixed Point Algorithms and Complementarity Problems*, (Univ. Southampton, Southampton, 1979), published by North-Holland, Amsterdam-New York, 1980. 253

[817] R. Sprague, *Mathematische Zeitschrift* **45** (1939), 607–608. 291

[818] R. Sprague, Über die Zerlegung von Rechtecken in lauter verschiedene Quadrate, *J. für die Reine und Angewandte Mathematick* **182** (1940), 60–64. 292

[819] R. P. Stanley, *Catalan numbers*, Cambridge University Press, New York, 2015. 608

[820] S. Stein, A generalized conjecture about cutting a polygon into triangles of equal areas, *Discrete Comput. Geom.* **24** (2000), 141–145. 253

[821] S. Stein, Cutting a polygon into triangles of equal areas, *Math. Intelligencer* **26** (2004), 17–21. 253

[822] S. K. Stein, Convex maps, *Proc. Amer. Math. Soc.* **2** (1951), 464–466. 438

[823] R. Steinberg, The state of the three colour problem, in "Quo Vadis, Graph Theory?", J. Gimbel, J. W. Kennedy, and L. V. Quintas (eds.), *Ann. Discrete Math.* **55** (1993), 211–248. 403

[824] J. Steiner, Einige Gesetze uber die Theilung der Ebene und Raumes, *J. Reine Angew. Math.* (also called "Crelle's journal") **1** (1826), 349–364. [Also reprinted in his *Gesammelte Werke*, Volume 1, 77–94.] 408

[825] H. Steinhaus, A note on the ham sandwich theorem, *Mathesis Polska* **9** (1938), 26–28. 413

[826] H. Steinhaus, *Mathematical snapshots*, 3rd ed., Dover, New York, 1999. [Paperback, with preface by Morris Kline, first published by Oxford University Press in 1983.] 69, 93, 126, 188, 239, 242, 290

[827] E. Steinitz, Polyeder und Raumeinteilungen, *Encyclop"adie der mathematischen Wissenschaften*, Band 3 (Geometries) (IIIAB12), 1922, pp. 1–139, "Abgeschlossen am 31 August 1916". 153

[828] F. W. Stevenson, *Projective planes*, Freeman, New York, 1972. 310

[829] I. Stewart, *The problems of mathematics*, 2nd ed., Oxford University Press, 1987. 242

[830] I. Stewart, *Galois theory*, 3rd ed., Chapman & Hall/CRC, 2004. 104

[831] J. Stillwell, *Numbers and geometry*, Springer, 1998. 260

[832] D. R. Stinson, A survey of Kirkman triple systems and related designs, *Discrete Math.* **92** (1991), 371–393. 389

[833] A. H. Stone and J. W. Tukey, Generalized "sandwich" theorems, *Duke Mathematical Journal* **9** (1942), 356–359. 413

[834] S. Straszewicz, Sur un problème géométrique de P. Erdős, *Bull. Acad. Polon. Sci. Cl.* **III 5** (1957), 39–40. 448

[835] B. Sturmfels, Cyclic polytopes and d-order curves, *Geometriae Dedicata* **24** (1987), 103–107. 210

[836] B. Sturmfels and N. White, Rational realizations of 11_3- and 12_3-configurations, pp. 92ff (article VII) in *Symbolic Computations in Geometry*, IMA preprint series **389**, 1988. [This volume contains 7 papers by Crapo, Havel, Whitely, and White,

given at a workshop *Symbolic computations in projective geometry*, IMA, 12–16 October 1987.] Available from the University of Minnesota Digital Conservancy, `http://hdl.handle.net/11299/4762`. (In the same volume, pp. 27–28 of article II by Sturmfels is also relevant, studying the relations between complex, real, and rational coordinatizations.) 397

[837] B. Sturmfels and N. White, All 11_3 and 12_3-configurations are rational, *Aequationes Math.* **39** (1990), 254–260. 397

[838] F. E. Su, Borsuk-Ulam implies Brouwer: a direct construction, *Amer. Math. Monthly* **104** (Nov. 1997), 855–859. 252

[839] A. Suk, A note on order-type homogeneous point sets, *Mathematika* **60** (2014), 37–42. 228

[840] A. Suk, On the Erdős–Szekeres convex polygon problem, `arXiv:1604.08657v2` [math.CO], 27 August 2016. 226, 227

[841] J. W. Sutherland, Lösing der Aufgabe 167, *Jahresber.Deuthsch. Math.-Verein* **45** (1935), 2 Abt., 33. 758

[842] J. Suzuki, A brief history of impossibility, *Math. Magazine* **81** (2008) , 27–38. 104

[843] K. J. Swanepoel, A new proof of Vázsonyi's conjecture, *J. Combin. Theory Ser. A* **115** (2008), 888–892. 448

[844] K. J. Swanepoel, Unit distances and diameters in Euclidean spaces, *Discrete Comput. Geom.* **41** (2009), 1–27. 448

[845] J. J. Sylvester, Mathematical question 1491, *Educational Times* (April 1864), London. 183

[846] J. J. Sylvester, Mathematical question 11851, *Educational Times* **59** (1893), 98. 375

[847] L. A. Székely, Crossing numbers and hard Erdős problems in discrete geometry, *Combin. Prob. Comput.* **6** (1997), 353–358. 399, 442, 443, 444, 450, 452

[848] L. A. Székely, Erdős on unit distances and Szemerédi–Trotter theorems, in G. Halász, ed., *Paul Erdős and his mathematics II* (based on conference, Budapest, Hungary, July 4–11, 1999), *Bolyai Soc. Math. Stud.* **11**, Springer, Berlin, 2002, pp. 649–666. 444, 460, 461

[849] L. A. Székely, *Ramsey theory*, preprint, 13 pages, 24 May 2015, available at `www.math.sc.edu/~laszlo/Ramsey.pdf`. 761

[850] L. A. Székely and N. C. Wormald, Bounds on the measurable chromatic number of R^n, *Discrete Math.* **75** (1989), 343–372. 459

[851] E. Szemerédi and W. T. Trotter, Extremal problems in discrete geometry, *Combinatorica* **3** (1983), 381–392. 399

[852] L. Szilassi, Extremblis tulajdonsigfi toroidok, *Acta Acad. Paed. Szeged*, 1983, pp. 43–56 (in Hungarian). 153

[853] L. Szilassi, Regular toroids, *Structural topology* **13** (1986), 69–80. 153

[854] A. D. Szlam, Monochromatic translates of configurations in the plane, *J. Combin. Theory Ser. A* **93** (2001), 173–276. 492

[855] T. Szőnyi, Note on the existence of large minimal blocking sets in Galois planes, *Combinatorica* **12** (1992), 227–235. 353, 355

[856] T. Szőnyi, Blocking sets in Desarguesian affine and projective planes, *Finite Fields Appl.* **3** (1997), 187–202. 355

[857] E. Szpilrajn, Sur l'extension de l'ordre partial, *Fundamenta Mathematicae* **1** (1930), 386–389. 508

[858] G. Tallini, On blocking sets in finite projective and affine spaces, *Annals of Discrete Mathematics* **37** (1988), 433–350. 352

[859] T. Tantau, *TikZ & PFG, Manual for versin 2.10*, available at `http://sourceforge.net/projects/pgf`, 2010. 35

[860] T. Tao, *The crossing number inequality*, blog posting, `http://terrytao.wordpress.com/2007/09/18/the-crossing-number-inequality/`, 2007 (accessed Nov. 2017). 429

[861] G. Tardos, On distinct sums and distinct distances, *Advances in Mathematics* **180** (2003), 275–289, 450

[862] G. Tarry, Le Probléme de 36 Officiers. *Compte Rendu de l'Association Française pour l'Avancement de Science Naturel. Secrétariat de l'Association* **1** (1900), 122–123. 306

[863] G. Tarry, Le Probléme de 36 Officiers, *Compte Rendu de l'Association Française pour l'Avancement de Science Naturel. Secrétariat de l'Association* **2** (1901), 170–203. 306, 310

[864] A. Tarski, Proléme 38, *Fund. Math.* **7** (1925), 381. 262

[865] V. Thebault, Problem E 1162, *American Mathematical Monthly* **62** (1955), pages 255 and 729–730. 38

[866] G. Theobald and E. W. Weisstein, *Dissection*, from Mathworld—a Wolfram Web Resource, `http://mathworld.wolfram.com/Dissection.html`. (Accessed 29 June 2016.) 239

[867] J. Thomas, A dissection problem, *Math. Mag.* **41** (1968), 187–190. 253

[868] R. R. Thomas, *Lectures in geometric combinatorics*, Student Mathematical Library **33**, American Mathematical Society, Institute for Advanced Study, 2006. 203

[869] M. J. Todd, *The Computation of Fixed Points and Applications*, Springer, Berlin, 1976. 253

[870] G. Tomkowicz and S. Wagon, *The Banach–Tarski paradox*, 2nd ed., Encyclopedia of Mathematics and its Applications **163**, Cambridge University Press, 2016. 261, 262, 263

[871] Cs. Tóth, The Szemerédi–Trotter theorem in the complex plane, *Combinatorica* **35** (2015), 95–126. 400

[872] G. Tóth, Note on geometric graphs, *J. Combin. Theory Ser. A* **89** (2000), 126–132. 428

[873] G. Tóth and P. Valtr, Note on the Erdős-Szekeres theorem, *Discrete Comput. Geom.* **19** (3) (1998, special issue), 457–459. 226

[874] G. Tóth and P. Valtr, Note on the Erdős-Szekeres theorem, *Discrete Comput. Geom.* **22** (1999), 633–642. 226, 428

[875] G. Tóth and P. Valtr, The Erdős-Szekeres theorem; upper bounds and related results, in *Combinatorial and Computational Geometry* (J. E. Goodman, J. Pach, and E. Welzl, eds.), MSRI Publications, Vol. 52, Cambridge University Press, Cambridge, 2005, pp. 557–568. 226

[876] J. Travers, *Education Outlook*, 1933, p. 145. 239

[877] D. Trim, *Calculus for engineers*, Prentice-Hall Canada, Inc., 1998. [In 2008, the 4th edition was published.] 686

[878] Trinity Mathematical Society, *The squared square*, `https://www.srcf.ucam.org/tms/about-the-tms/the-squared-square/` website accessed 17 August 2014. 293

[879] G. B. Trustum, Mrs. Perkins's quilt, *Proc. Cambridge Phil. Soc.* **61** (1965), 7–11. 290

[880] S. Tsaturian, *A Euclidean Ramsey result in the plane*, preprint, `arXiv:1703.10723v2`, 4 April 2017. 491

[881] P. Turán, Eine Extremalaufgave aus der Graphentheorie (in Hungarian) [Egygráfemèleti szélsöértekfeladatrol], *Math. Fiz Lapook* **48** (1941), 436–453. 424

[882] P. Turán, On the theory of graphs, *Colloquium Math.* **3** (1954), 19–30. 424

[883] P. Turán, A note of welcome, *J. Graph Theory* **1** (1977), 7–9. 424, 433

[884] W. T. Tutte, The dissection of equilateral triangles into equilateral triangles, *Proc. Camb. Phil. Soc.* **44** (1948), 464–482. 300

[885] W. T. Tutte, Squaring the square, *Canadian J. Math.* **2** (1950), 197–209. 292, 299

[886] W. T. Tutte, "Squaring the square", in Martin Gardner's column "Mathematical games", *Scientific American*, Nov. 1958. Reprinted in Gardner's *The 2nd Scientific American book of Mathematical Puzzles and Diversions*, Chapter 17, Simon and Schuster, New York, 1961. Also available in [26]. 299, 746

[887] W. T. Tutte, The quest of the perfect square, *Amer. Math. Monthly* **27** (1965), 29–35. 299

[888] H. Tverberg, A generalization of Radon's theorem, *J. London Math. Soc.* **41** (1966), 123–128. 202

[889] H. Tverberg, A generalization of Radon's theorem. II, *Bull. Austral. Math. Soc.* **24** (1981), 321–325 202

[890] R. Umble, *Transformational geometry*, Chapter 1, Isometries, 38 pages, available at `sites.millersville.edu/rumble/Math.355/Book/Chapter%201.pdf` (Accessed 25 May 2018.) [These notes share many examples and exercises found in [641].] 78

[891] University of Dallas at Texas Math Camp, see *The Dallas Morning News*, 21 July 2010. `http://www.dallasnews.com/news/education/headlines/20100721-utd-program-challenges-the-best-and-brightest-young-mathletes.ece` (Accessed 3 June 2016.) 708, 711, 714

[892] J. Urrutia, Art gallery and illumination problems, Ch. 22, pp. 973–1027 in *Handbook of computational geometry*, (eds, Jörg-Rüdiger Sack, Jorge Urratio) Elsevier, 2000. 180

[893] P. Valtr, Convex independent sets and 7-holes in restricted planar point sets, *Discrete and Computational Geometry* **7** (1992), 135–152. 228

[894] P. Valtr, Sets in R^d with no large empty convex subsets, *Discrete Math.* **108** (1992), 115–124. 228

[895] P. Valtr, Probability that n points are in convex position, *Discrete Comput. Geom.* **13** (1995), 637–643. 184

[896] P. Valtr, Holes in planar point sets, pp. 111–113 in *Discrete geometry and convexity in honour of Imre Bárány*, G. Ambrus, K. Böröcky, and Z. Füredi, eds., Proceedings of conference "Discrete geometry and convexity Bárány 70", 19–23 June 2017, Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest, published by Typotex Electronical Publishing Ltd, Hungary, 2017. 229

[897] E. P. Vance, *An introduction to modern mathematics*, (Chapter 11, Mathematical Induction), Addison-Wesley Publishing, Inc., 1968. [Also published separately in "Mathematical induction and conic sections", chapters 11 and 12 of the text, Addison-Wesley, 1971.] 664, 666, 668, 669, 670

[898] J. H. van Lint and R. M. Wilson, *A course in combinatorics*, 2nd ed., Cambridge University Press, 2001. 306

[899] V. N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and Its Applications* **16** (1971), 264–280. 497

[900] O. Veblen and J. H. Maclagan-Wedderburn, Non-desarguesian and non-pascalian geometries, *Trans. Amer. Math. Soc.* **8** (1907), 379–388. 310

[901] O. Veblen and J. W. Young, *Projective geometry*, 2 vols, Ginn & Co., Boston, 1916. 304

[902] V. G. Vizing, On an estimate of the chromatic class of a p-graph, *Diskret. Analiz* **3** (1964), 25–30. (In Russian) 140

[903] J. von Neumann, A certain zero-sum two-person equivalent to the optimal assignment problem, *Contributions to the theory of*

games II (H.W. Kuhn, ed.), *Ann. Math. Studies* **28**, Princeton University Press, 1953, pp. 5–12. 214

[904] A. N. Voropaev, Counting k-gons in finite projective planes, *Siberian Electronic Mathematical Reports* **10** (2013), 241–270. (In Russian.) 323

[905] A. Wagner, On collineation groups of finite projective spaces, *Math. Z.* **76** (1961), 421–426. 315

[906] A. Wagner, On finite affine line transitive planes, *Math. Z.* **87** (1965), 1–11. 315

[907] K. Wagner, Bemerkungen zum Vierfarbenproblem, *Jahresbericht der Deutschen Mathematiker-Vereinigung* **46** (1936), 26–32. 438

[908] K. Wagner, Über eine Eigenschaft der ebenen Komplexe, *Math. Ann.* **114** (1937), 570–590. 146

[909] S. Wagon, Fourteen proofs of a result about tiling a rectangle, *Amer. Math. Monthly* **94** (1987), 601–617. 289

[910] W. Wallace, ed., *Elements of geometry*, 8th ed., Bell & Bradfute, Edinburgh, 1831. (First six books of Euclid, with a supplement by John Playfair.) 232

[911] E. C. Wallace and S. F. West, *Roads to geometry*, Prentice Hall, 1992. 102

[912] P. L. Wantzel, Recherches sur les moyens de reconnaître si un Probleme de Géométrie peut se résoudre avec la règle et le compas, *Journal de Mathématiques Pures et Appliquées* **2** (1837), 366–372. 103, 104

[913] K. Watson, Sylvester's problem for spreads of curves, *Canad. J. Math.* **32** (1981), 219–239. 385

[914] E. W. Weisstein, *Bretschneider's Formula*, Wolfram Mathworld, `http://mathworld.wolfram.com/BretschneidersFormula.html` (accessed 17 January 2017). 62

[915] E. W. Weisstein, *Heilbronn triangle problem*, from *Mathworld–a* Wolfram Web Resourse, http://mathworld.wolfram.com/HeilbronnTriangleProblem.html (accessed 9 September 2017). 255, 258

[916] D. G. Wells, *The Penguin dictionary of curious and interesting numbers*, Penguin Books, 1986. 707

[917] D. G. Wells, *Hidden connections, double meanings*, Cambridge University Press, 1988. 414

[918] D. G. Wells, *The Penguin dictionary of curious and interesting geometry*, Penguin, London, 1991. 69, 188, 189, 242, 714

[919] D. B. West, *Introduction to graph theory*, 2nd ed., Prentice Hall, Inc., 2001. 135, 758

[920] F. Wettl, Nuclei in finite non-Desarguesian projective planes, in *Finite Geometry and Combinatorics* (A.Beutelspacher *et al.* eds.), proceedings of The Second International Conference at Deinze, London Mathematical Society Lecture Note Series **191**, Cambridge University Press, 1993, 405–412. 352

[921] J. E. Wetzel, On the division of the plane by lines, *Amer. Math. Monthly* **85** (Oct. 1978), 647–656. 409

[922] H. Weyl, *Symmetry*, Princeton University Press, 1982. 80

[923] A. T. White, *Graphs, groups and surfaces*, Completely revised and enlarged edition, North-Holland Mathematical Studies **8**, North-Holland, 1984. 206, 432, 446, 754

[924] H. S. Wilf, The friendship theorem, pp. 307–309 in *Combinatorial mathematics and its applications*, (D. J. A. Welsh, ed.) Academic Press, New York–London, 1971. 322, 349

[925] H. S. Wilf, On crossing numbers, and some unsolved problems, pp. 557–562 in *Combinatorics, Probability, Cambridge 1993* (B. Bollobás, A. G. Thomason, eds.), Cambridge University Press, Cambridge, 1997. 444

[926] T. H. Willcocks, Problem 7795 and solution, *Fairy Chess Rev.* **7** (1948), 106. 293

[927] T. H. Willcocks, A note on some perfect squared squares, *Canadian J. Math.* **2** (1950), 197–209. 293

[928] P. Wilson and J. Wiseman, A Sylvester theorem for conic sections, *Discrete Comp. Geom.* **3** (1988), 295–305. 385

[929] R. Wilson, *Four colours suffice*, Penguin Books, 2002. 146

[930] T. Wolff, Recent work connected with the Kakeya problem, *Prospects in mathematics*, (Princeton, N. J., 1996), pp. 129–162, 1999. 500

[931] D. R. Woodall, Distances realized by sets covering the plane, *J. Combin. Theory Ser. A* **14** (1973), 187–200. 462, 473

[932] D. R. Woodall, Cyclic-order graphs and Zarankiewicz's crossing-number conjecture, *J. Graph Th.* **17** (1993), 657–671. 435

[933] W. S. B. Woolhouse, Some additional observations on the four-point problem, *Mathematical questions and their solutions, from the Educational Times*, Vol. 7, p. 81. Published by F. Hodgson and Son, London, 1867. 183

[934] M. Yannakakis, The complexity of the partial order dimension problem, *SIAM Journal on Algebraic and Discrete Methods* **3**(3) (1982), 351–358. 508

[935] I. Yashchenko, *Invitation to a mathematical festival*, MSRI/AMS, 2013. 760

[936] B. K Youse, *Mathematical Induction*, Prentice Hall, 1964. [Note: there is no period after "K" in Bevan's name. Harry S Truman is another such example.] 661

[937] K. Zarankiewicz, Problem P101, *Colloq. Math* **2** (1951), 116–131. 427

[938] K. Zarankiewicz, The solution of a certain problem on graphs of P. Turan, *Bull. Acad. Polon. Sci. Cl. III* **1** (1953), 167–168. [This paper announced his results for crossing numbers, with proofs given in [939] a year or so later.] 434

[939] K. Zarankiewicz, On a problem of P. Turán concerning graphs, *Fundamenta Mathematicae* **41** (1954), 137–145. [In many sources, the year 1954 is given for this article, but 1955 is given on the Polish Institute website.] 434, 629

[940] G. M. Ziegler, *Lectures on polytopes*, Springer, 1995. 153, 203, 205, 214, 460

[941] C. Zong, Geometry of numbers in Vienna, *The Mathematical Intelligencer* **31** (3) (2009), 25–31. 272

[942] P. Zvengrowski, *PMAT 320 notes*, University of Calgary, 1977. 47

[943] A. A. Zykov, On some properties of linear complexes, *Mat. Sbornik N. S.* **24** (66) (1949), 163–188. 424

Index