Topics in Undergraduate Mathematics

Christoph Börgers
Department of Mathematics
Tufts University

Contents

Preface

These are notes on miscellaneous topics in undergraduate mathematics that I find interesting and worth knowing about, but that aren't always covered in the undergraduate curriculum. My notes are intended for a class alternating between lectures and discussions. Each chapter is intended for a 75-minute lecture. The questions in the last section of each chapter are suggestions for a 75-minute session of small group discussions following the 75-minute lecture.

Christoph Börgers
Tufts University
Spring 2025

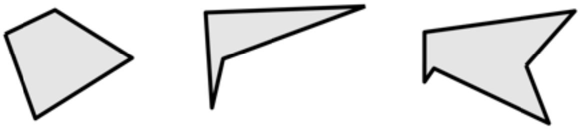
Chapter 1

The Platonic solids

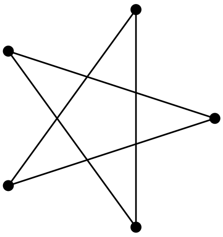
Everybody knows the regular polygons in the plane: the equilateral triangle, the square, the regular pentagon, the regular hexagon, and so on. But how does this generalize to three dimensions? This question was asked and answered by the ancient Greek mathematicians. The three-dimensional analogue of the regular polygons are the *Platonic solids*. Amazingly, there are only five of them.

1.1 Polygons

A *polyg*

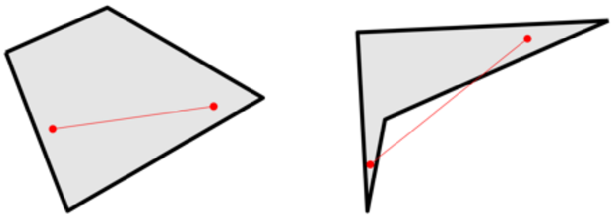


We assume that two sides can only meet at a point that is an end point of both of them. For ir ed the *pentagram*.



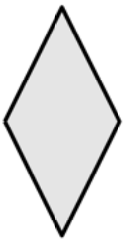
(Only the black dots counts as vertices of the pentagram.)

We count both the boundary and the interior as part of the polygon, and we assume that the boundary is topologically equivalent to a circle; this means it could be bent into a circle without tearing it (and without leaving the plane). The polygon is *convex* if for any two points in it, the straight line connecting them does not leave the polygon. Below you see a convex and a non-convex polygon.



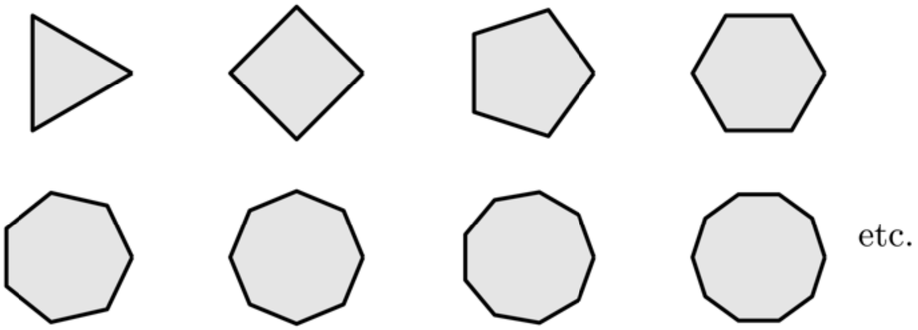
Definition 1.1. *A convex polygon is called* regular *if (1) all sides have the same length, and (2) all interior angles are the same.*

Note that condition (1) does not imply condition (2). Here is a convex quadrilateral in which all sides are of equal length, but not all interior angles are the same.

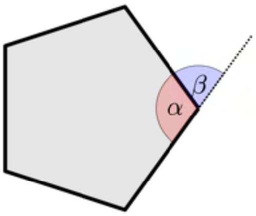


Condition (2) does not imply condition (1) either: In a rectangle that is not a square, all interior angles are the same, but not all sides have the same length.

For each $n \geq 3$, there is exactly one regular polygon with n vertices, called a *regular n-gon*, (A 3-gon is a triangle, a 4-gon is a quadrilateral, a 5-gon is a pentagon, and a 6-gon is a hexagon.) When we say "there is exactly one", we always mean "exactly one up to scaling and rigid body motion". Throughout this chapter, two shapes are considered the same if one can obtained from the other by



Each vertex of a regular n-gon is associated with an *interior angle* α and an *exterior angle* β, with $\alpha + \beta = \pi$.

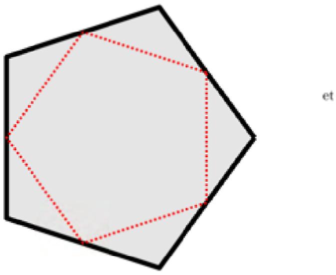


We have $n\beta = 2\pi$, and therefore

$$\alpha = \pi - \beta = \pi - \frac{2\pi}{n} = \frac{n-2}{n}\,\pi. \tag{1.1}$$

You probably learned this formula in school, but perhaps you had forgotten how it comes about. For example, the interior angles in a pentagon ($n = 5$) are of size $\frac{3}{5}\pi$, which means 108^o.

The midpoints of a regular polygon form another regular polygon. It is called the *dual* polygon. This is not very interesting because the dual of a regular n-gon is a regular n-gon. However, it has more interesting generalizations to higher dimensions, as you will see.



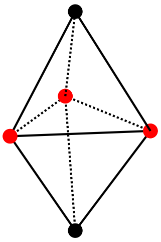
1.2 Platonic and Archimedean solids

Polyhedra are the three-dimensional analogues of polygons. We assume that the surface is made up of finitely many flat pieces, the *faces*, and that it is topologically equivalent to (can without tearing be bent into the shape of) a sphere. We assume that the intersection of any two faces is either an edge of both of them, or a vertex of both of them, or empty. The surface and the interior are both considered part of the polyhedron. We will only consider *convex* polyhedra, that is, ones in which for any pair of points, the straight line connecting them does not leave the polyhedron.

Definition 1.2. *A convex polyhedron is called* regular *or a* Platonic solid *if (1) all faces are regular polygons that are congruent to each other, and (2) in all vertices the same number of faces come together at the same angles; in other words, all vertices are congruent (look alike).*

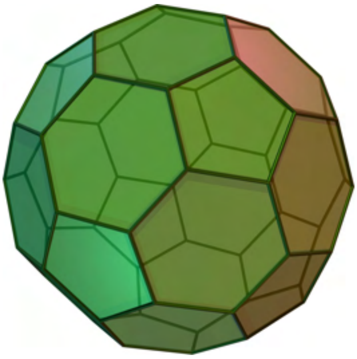
There are exactly five Platonic solids. This was known to the Greek mathematicians more than 2000 years ago. You will see two proofs in this chapter.

Conditio). In fact, here is a polyhedron s



This polyhedron has six faces, all of them equilateral triangles of the same size. However, it has two different kinds of vertices, indicated with two different colors in the figure. Three faces join together in one kind of vertices (black), four in the other (red). So this polyhedron satisfies condition (1) but not condition (2).

Likewise, condition (2) does not imply condition (1). Here is an example of a polyhedron that satisfies (2) but not (1):



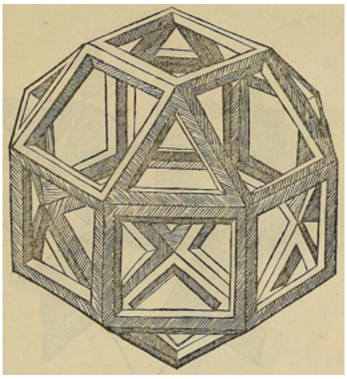
Created by POV-Ray,
Creative Commons license BY-SA 3.0,
via Wikimedia Commons.

It's called a *truncated icosahedron.* (The name will be explained in Section 1.6.2.) This polyhedron has hexagonal and pentagonal faces. If you bend the surface into a sphere and color the pentagons black, the hexagons white, it looks like this:

It's a soccer ball! In each vertex, two hexagons and one pentagon join together. Condition (2) is satisfied, but condition (1) isn't. Here is a toy that belongs to my neighbors' dog:

It's a truncated icosahedron.

Convex polyhedra in which the vertices are all congruent to each other, and the faces are regular polygons, but not all faces have the same number of vertices, are called *Archimedean solids.* There are exactly 13 of them. Archimedes knew them more than 2000 years ago. We won't have time to study them in this chapter, but here is another one of the 13:

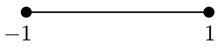


Leonardo da Vinci, via Wikimedia Commons,
public domain

1.3 The five Platonic solids

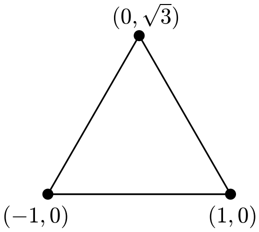
1.3.1 The tetrahedron

The regular tetrahedr construct it, first we
construct the coordin with the line segment
in the x-axis between



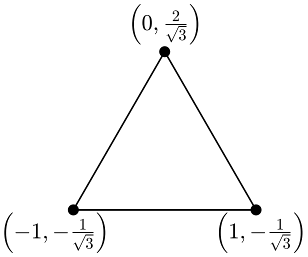
We add a second axis, the y-axis, so we now identify -1 and 1 with $(-1,0)$ and $(1,0)$, and add a third vertex at $(0,h)$, with $h>0$ chosen so that the distance between $(1,0)$ and $(0,h)$, which is clearly equal to the distance between $(-1,0)$ and $(0,h)$, is also equal to the distance between $(-1,0)$ and $(1,0)$:

$$\sqrt{1^2+h^2}=2 \quad \Leftrightarrow \quad h=\sqrt{3}.$$



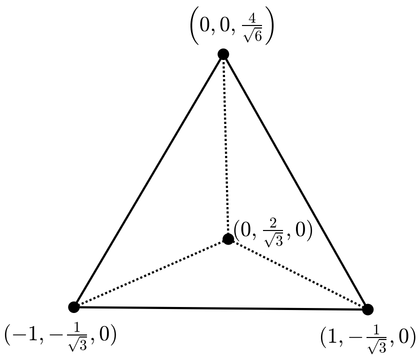
I led you through a laborious construction of an equilateral triangle for a reason. The construction that produces an equilateral triangle starting from a line segment also produces a regular tetrahedron from an equilateral triangle. The same construction can produce the regular tetrahedra in dimensions greater than three.

We shift our equilateral triangle downwards so that its centroid (the average of its three vertices) becomes $(0,0)$:

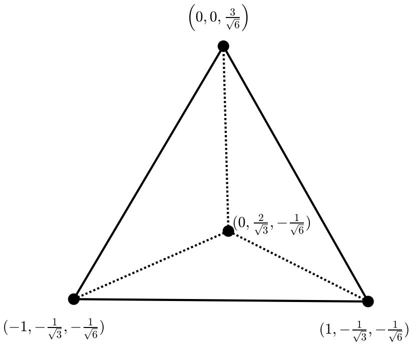

$\left(0,\frac{2}{\sqrt{3}}\right)$
$\left(-1,-\frac{1}{\sqrt{3}}\right)$
$\left(1,-\frac{1}{\sqrt{3}}\right)$

We now add a third coordinate axis, the z-axis, append zeros to the coordinate pairs of the three points that we have already constructed, and add a fourth point at $(0,0,k)$ with $k>0$. Regardless of how we choose k, the new point $(0,0,k)$ is equally far from the three points already constructed. We choose k so that the distance between $\left(0,\frac{2}{\sqrt{3}},0\right)$ and $(0,0,k)$ equals the side length of the equilateral triangle, which is 2:

$$\sqrt{\left(\frac{2}{\sqrt{3}}\right)^2+k^2}=2 \quad \Leftrightarrow \quad k=\sqrt{\frac{8}{3}}=\frac{4}{\sqrt{6}}.$$


$\left(0,0,\frac{4}{\sqrt{6}}\right)$
$(0,\frac{2}{\sqrt{3}},0)$
$(-1,-\frac{1}{\sqrt{3}},0)$
$(1,-\frac{1}{\sqrt{3}},0)$

We can shift thi

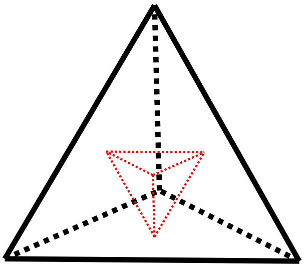

$\left(0,0,\frac{3}{\sqrt{6}}\right)$
$(0,\frac{2}{\sqrt{3}},-\frac{1}{\sqrt{6}})$
$(-1,-\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{6}})$
$(1,-\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{6}})$

In each vertex of a tetrahedron, three equilateral triangles come together.

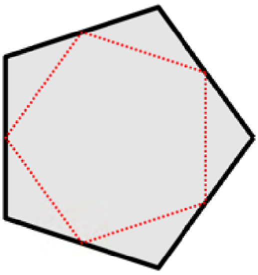
Definition 1.3. *A Platonic solid with faces that are regular n-gons is said to have the* Schläfli symbol (n,q) *if in each vertex, q faces come together.*

So a tetrahedron has Schläfli symbol (3,3). The Schläfli symbol is named after the Swiss mathematician Ludwig Schläfli (1814–1895), who investigated higher-dimensional an

The centr ahedron.

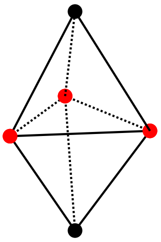


This is called the *dual* of the tetrahedron, analogous to the dual of a polygon:

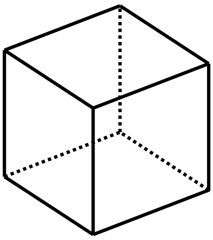


It’s still not very interesting — the dual of a regular tetrahedron is a regular tetrahedron again. One says that the regular tetrahedron is *self-dual*

If you gl polyhedron that we discussed e

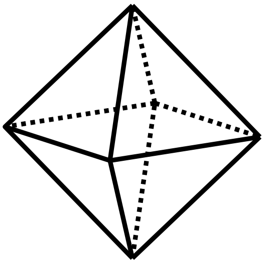


1.3.2 The cube



The cube needs no introduction. It is evidently a Platonic solid. Its Schläfli symbol is $(4,3)$.

1.3.3 The oc

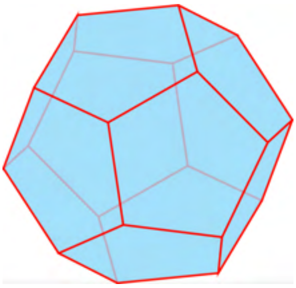


The regular octahedron has eight faces (hence *octa*hedron), each of them an equilateral triangle. At each vertex, four faces come together. So the Schläfli symbol is $(3, 4)$. It is the transpose of the Schäfli symbol of the cube. The octahedron and the cube are *duals* of each other: The centroids of the faces of a cube form an octahedron, and also the centroids of the faces of an octahedron form a cube. Note that this is the first example of an interesting duality. Taking the dual of the cube, you do not get a cube, you get a new Platonic solid, namely the octahedron.

For each face of the cube, the octahedron has a vertex, and for each vertex of the cube, the octahedron has a face.

1.3.4 The dodecahedron

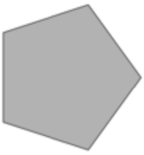
The next Platonic solid, the dodecahedron, is my favorite.



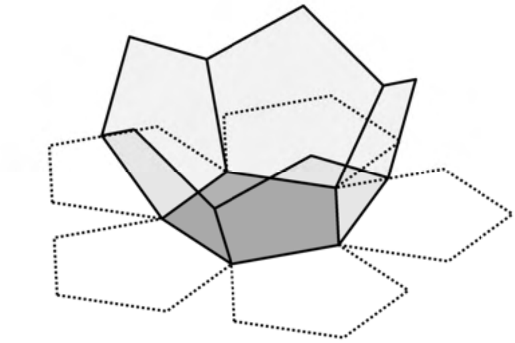
by Efraín Soto Apolinar

It has twelve pentagonal faces. In each vertex, three of them come together. There-
fore the

Y hether it is
an opti her in this
way? I tart with a
regular



Attach to each of its sides a copy of itself:



It is clear that this can be done. The angle by which you have to rotate the flaps up so that they touch each other turns out to be

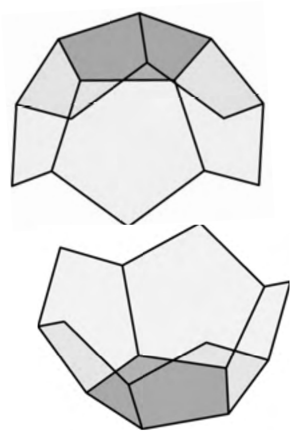
$$\pi - 2\arctan\frac{1+\sqrt{5}}{2} \approx 63^{\text{o}}.$$

I mention this only because it is surprising — notice that $\frac{1+\sqrt{5}}{2}$ is the golden ratio — but I won't prove it here. The *interior* angle along an edge of the dodecahedron is

$$2\arctan\frac{1+\sqrt{5}}{2} \approx 117^{\text{o}}.$$

This is also called the *dihedral angle* of the dodecahedron.

You can make two bowls of this kind, turn one upside down, and use it as a lid for the other.

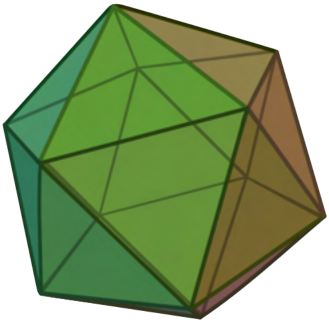


You thereby obtain a polyhedron with 12 pentagonal faces. With very slight hesitation, I say that “it is clear that” all vertices will look identical. An unquestionably possible but unappealing explicit computation would prove that they do.

So indeed the dodecahedron exists. Here is a ball that I found in my house, I think it used to belong to my son:

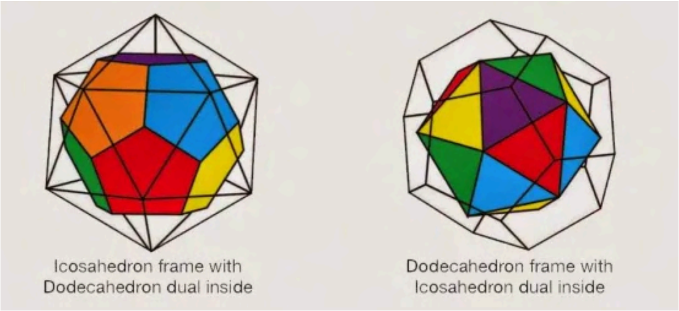
1.3.5 The icosahedron

Here is yet another platonic solid, the icosahedron:



By DTR, Creative Commons License BY-SA 3.0,
via Wikimedia Commons.

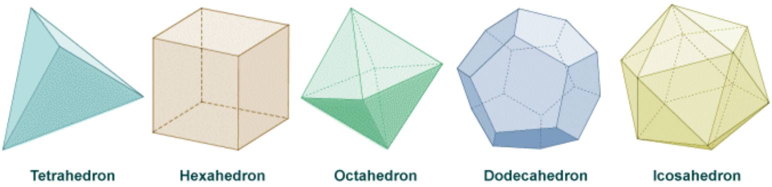
It has twenty triangular faces. In each vertex, five of them come together. Therefore the Schläfli symbol is $(3,5)$. The icosahedron is the dual of the dodecahedron, and vice versa:



from cosmic-core.org

This is why the Schläfli symbols of the dodecahedron and icosahedron are transposes of each other: $(5,3)$ and $(3,5)$. For each face of the dodecahedron, the icosahedron has a vertex, and for each vertex of the dodechaedron, the icosahedron has a face.

1.3.6 The complete lineup

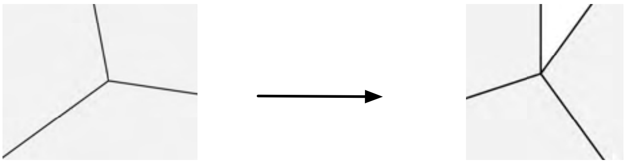


1.4 First proof that there are only five Platonic solids

Theorem 1.4. *Any Platonic solid must have one of the following five Schläfli symbols:*

$$(3,3), \quad (4,3), \quad (3,4), \quad (5,3), \quad (3,5). \tag{1.2}$$

Proof. Suppose that (n,q) is the Schläfli symbol of a Platonic solid. So q regular n-gons come together in each vertex. The interior angles of a regular n-gons are $\frac{n-2}{n}\pi$; see (1.1). The sum of the interior angles of the regular n-gons coming together in a vertex must be less than 2π. You can see this by imagining the vertex flattened out:



So we must have

$$q\frac{n-2}{n} < 2.$$

It is easy to verify that the five pairs listed in (1.2) are the only pairs of integers $n \geq 3$ and $q \geq 3$ that satisfy this inequality. □

Two Platonic solids with the same Schläfli symbol must be the same up to scaling and rigid body motion. (I assert with slight trepidation that that's obvious.) Therefore we have:

Corollary 1.5. *There are exactly five Platonic solids: Tetrahedron, cube, octahedron, dodecahedron, and icoashedron.*

1.5 Second proof that there are only five Platonic solids

1.5.1 Euler's polyhedron formula

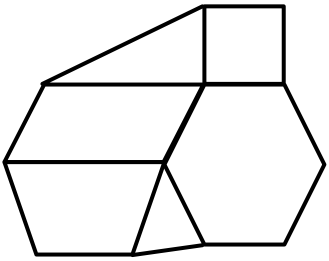
Recall that we defined a *polyhedron* to be a three-dimensional object with a surface made up of finitely many flat pieces, so that the surface is topologically equivalent to (can without tearing be bent into the shape of) a sphere.

Theorem 1.6 (Euler's polyhedron formula). *For any polyhedron, the numbers $v, f,$ and e of vertices, faces, and edges satisfy*

$$v - e + f = 2. \tag{1.3}$$

The polyhedron need not be Platonic, and it needn't be convex. All that matters is that the surface is topologically equivalent to a sphere. The number 2 on the right-hand side of (1.3) can therefore be viewed as a property of the sphere. It is called the *Euler characteristic* of the sphere.

Proof. Remove one face from the polyhedron. Then stretch and flatten the rest into a polygonal net in the plane:

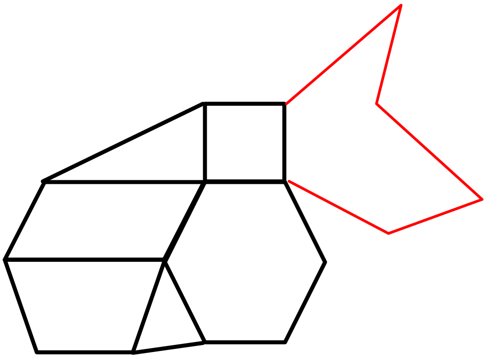


We must prove that for any such net, consisting of finitely many polygons covering a finite area in the plane, the numbers v, e, and f of vertices, edges, and faces satisfies

$$v - e + f = 1. \tag{1.4}$$

(Not 2 but 1, since we removed one face from the original polyhedron.)

When $f = 1$, (1.4) is clear: A polygon has as many vertices as edges. When we attach to a polygonal net one extra polygon, we raise raise e by some number x, and raise v by $x - 1$ and f by 1, so that $v + f - e$ does not change at all. Here is an example with $x = 5$:



The red polygon raises v by 4, e by 5, and f by 1. By induction on f, we conclude (1.4) for any polygonal net, and therefore Euler's polyhedron formula follows. □

For the five Platonic solids, (v, e, f) are as follows:

tetrahedron:	$(4, 6, 4)$
cube:	$(8, 12, 6)$
octahedron:	$(6, 12, 8)$
dodecahedron:	$(20, 30, 12)$
icosahedron:	$(12, 30, 20)$

You can verify $v - e + f = 2$ in each case.

1.5.2 The possible Schläfli symbols of Platonic solids

Consider a Platonic solid with Schläfli symbol (n, q), and with v vertices, e edges, and f faces. Every face has n edges. So you might think that $e = nf$, but that would double-count the edges, since in each edge, two faces come together:

$$e = \frac{nf}{2}.$$

Each face has n vertices, so you might think that $v = nf$, but that would count each vertex q times — since it is shared by q faces. So

$$v = \frac{nf}{q}.$$

Euler's polyhedron formula now becomes

$$\frac{nf}{q} - \frac{nf}{2} + f = 2,$$

or equivalently,

$$n\left(\frac{1}{q} - \frac{1}{2}\right) f + f = 2. \tag{1.5}$$

This equation turns out to constrain the Schläfli symbol (n, q) severely.

Since $q \geq 3$, we have

$$n\left(\frac{1}{q} - \frac{1}{2}\right) f + f \leq \left(1 - \frac{n}{6}\right) f$$

Therefore (1.5) implies $n < 6$. So n can only be 3, 4, or 5. When $n = 3$, (1.5) becomes

$$\frac{3}{q}\left(1 - \frac{q}{6}\right) f = 2,$$

so $q < 6$, so $q = 3, 4$, or 5. When $n = 4$, (1.5) becomes

$$\frac{4}{q}\left(1 - \frac{q}{4}\right) f = 2,$$

so $q < 4$ and therefore $q = 3$. When $n = 5$, (1.5) becomes

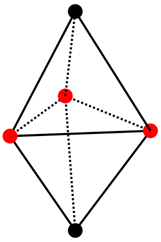
$$\frac{5}{q}\left(1 - \frac{3q}{10}\right) f = 2,$$

so $q < 10/3$ and therefore $q = 3$ again. This completes the second proof that the only possible Schläfli symbols of a Platonic solid are $(3, 3)$, $(3, 4)$, $(3, 5)$, $(4, 3)$, and $(5, 3)$.

1.6 Questions and extensions

1.6.1 Mor

Remember the edra together:



Is this an Archimedean solid?

1.6.2 More about the soccer ball

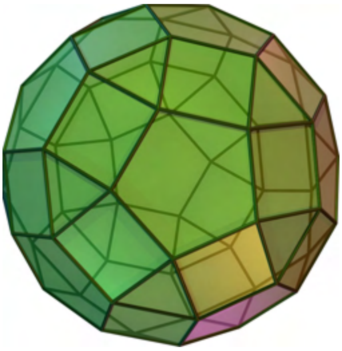
Convince yourself that the truncated icosahedron (the soccer ball) is obtained by truncating the vertices of the icosahedron. How many faces, edges, and vertices does the truncated icosahedron have?

1.6.3 Other truncated Platonic solids

Each of the Platonic solids can be truncated, and the result is an Archimedean solid in each case. If you start out with the icosahedron, you get the truncated icosahedron or soccer ball. Each of the truncated Platonic solid has two kinds of faces. Which kinds of faces, and how many of each kind? For each of them, compute how many faces, edges, and vertices they have.

1.6.4 An Archimedean solid with three different kinds of faces

The following Archimedean solid is called the *rhombicosidodecahedron* (no joke).



CC BY-SA 3.0, via Wikimedia Commons

How many faces of which kind does it have? How many edges and vertices?

1.6.5 Are there Archimedean solids with more than three different kinds of faces?

The rhombicosidodecahedron has three different kinds of faces — triangles, squares, and pentagons. Prove that there are no Archimedean solids with more than three different kinds of faces.

1.6.6 The four-dimensional hypercube

The *four-dimensional hypercube* is the set

$$Q_4 = [-1,1]^4 = \big\{(x,y,z,u) \in \mathbb{R}^4 \ : \ -1 \le x,y,z,u \le 1\big\}\,.$$

It is a *four-dimensional Platonic solid*: It is a polytope (the four- or higher-dimensional generalization of a polyhedron) that “looks alike from all sides”. We won’t make that rigorous here, but it is probably clear to you intuitively in which sense that’s true.

Just as three-dimensional polyhedra have vertices, edges, and faces, the four-dimensional hypercube has vertices, edges, faces, and “cells”. The boundary of Q_4 is the union of its cells.

(a) The cells of the four-dimensional hypercube are three-dimensional cubes. We denote their number by c. What is c?

(b) The boundaries of the cells are the faces. They are two-dimensional squares. We denote their number by f. What is f?

(c) The boundaries of the faces are the edges. We call their number e. What is e?

(d) The boundaries of the edges are the vertices. We call their number v. What is v?

(e) What is $v - e + f - c$? That’s the Euler characteristic of the three-sphere

$$S^3 = \big\{(x,y,z,u) \in \mathbb{R}^4 \ : \ x^2+y^2+z^2+u^2 = 1\big\}\,.$$

(The 3-sphere is embedded in $\mathbb{R}^4$ here. The superscript reflects the dimension of the sphere, not the dimension of the space that it is embedded in.)

1.6.7 Visualizing the four-dimensional hypercube

This section is for you only if you like coding.

To visualize the four-dimensional hypercube, we must project into three dimensions. We can do this by subjecting the hypercube to a rotation, then chopping off the fourth coordinate, and plotting the resulting three-dimensional object.

Write a program that generates a random 4 × 4 rotation matrix. Here is a simple way of doing this in Matlab:

```
A=randn(4,4); [Q,R]=qr(A);
if det(Q)<0, q=Q(:,1); Q(:,1)=Q(:,2); Q(:,2)=q; end
```

First, we define a random 4×4 matrix A. Then we subject it to a QR-decomposition. This means Q is orthogonal, and R is upper triangular, so that $A = QR$. We throw away R; what we are after is just the orthogonal matrix Q. Its determinant is $+1$ or -1. Rotation matrices are the orthogonal matrices with determinant equal to $+1$. If the computed Q has determinant -1, we swap the first two columns; that changes the sign of the determinant.

Using this program, plot some nice examples of three-dimensional projections of $[-1,1]^4$. It may be easier to understand the plots if you color the images of the cell

$$\{(x,y,z,-1) \ : \ -1 \le x,y,z \le 1\}$$

in red, and the image of the cell

$$\{(x,y,z,1) \ : \ 1 \le x,y,z \le 1\}$$

in blue.

1.6.8 The d-dimensional hypercube

For any $d \ge 5$, the *d-dimensional hypercube* is the set

$$Q_d = [-1,1]^d = \left\{(x_1,x_2,\ldots,x_d) \in \mathbb{R}^d \ : \ -1 \le x_1,x_2,\ldots,x_d \le 1\right\}.$$

It is a *d-dimensional Platonic solid.* Its boundary is composed of pieces of the form

$$\{-1,x_2,x_3,\ldots,x_d\}, \quad \{1,x_2,x_3,\ldots,x_d\},$$

$$\{x_1,-1,x_3,\ldots,x_d\}, \quad \{x_1,1,x_3,\ldots,x_d\},$$

and so on. All of these are $(d-1)$-dimensional hypercubes. They are called the *$(d-1)$-faces* of Q_d. The boundary of the $(d-1)$-faces is composed of $(d-2)$-faces, and so on. The 3-faces, 2-faces, 1-faces, and 0-faces are what we have called cells, faces, edges, and vertices until now. The number of j-faces is denoted by f_j, $0 \le j \le d-1$. The number

$$f_0 - f_1 + f_2 - \ldots + (-1)^{d-1} f_{d-1}$$

is the Euler characteristic of the $(d-1)$-sphere

$$S^{d-1} = \left\{(x_1,x_2,\ldots,x_d) \ : \ x_1^2 + x_2^2 + \ldots + x_d^2 = 1\right\}.$$

(a) Show that the Euler characteristic of a sphere is 2 when the dimension of the sphere is even, and 0 when the dimension of the sphere is odd.

(b) Verify directly that S^1 has Euler characteristic 0.

1.6.9 The d-dimensional hypertetrahedron

Just as we constructed the tetrahedron starting with the equilateral triangle with centroid at $(0,0)$, we can construct a four-dimensional hypertetrahedron starting with the three-dimensional tetrahedron with centroid at $(0,0,0)$. What are its vertices? How many 3-faces does it have? What kind of Platonic solids are the 3-faces?

Similarly, we can construct a five-dimensional hypertetrahedron starting with the four-dimensional hpertetrahedron with centroid at $(0,0,0,0)$, and so on.

1.6.10 The d-dimensional hyperoctahedron

The $(d-1)$-faces of the hypercube Q_d have the centroids

$$(0,\ldots,0,\pm1,0,\ldots,0),$$

where the ±1 can appear in any one of the d possible positions. These are the vertices of a d-dimensional Platonic solid, the dual of the hypercube. It is called the *hyperoctahedron*. How many 3-faces does the four-dimensional hyperoctahedron have? What kind of Platonic solids are they?

1.6.11 How many higher-dimensional Platonic solids are there?

dimension d	2	3	4	5	6	7	$\cdots$
number of Platonic solids	∞	5	6	3	3	3	$\cdots$

In dimensions $d > 4$, there are only three Platonic solids: The hypercube, the hypertetrahedron, and the hyperoctahedron. We know that there are five in three dimensions. It turns out that there are six in four dimensions — three additional ones that we have not discussed.

You could, if you wanted to, look up the proof that these are the numbers of Platonic solids in higher dimensions, but I haven't done that myself.

Chapter 2

Conditional probability

Conditional probability is about how the occurrence of one event affects the likelihood that another event will occur. This is so fundamentally important that everybody should learn about it in high school. (It is certainly far more important to most people than solving quadratic equations.)

2.1 Random events and their probabilities, for the mathematically inclined

2.1.1 Basic definitions

A random event is, intuitively, what you think it should be. For instance, if I roll a die, "I get a number greater than 3" is an event. It seems reasonable to assign a probability of 1/2 to this event, assuming the die is not loaded. "It will rain tomorrow" is also a random event, but it is less clear how to assign a meaningful probability to it. Nonetheless my smart speaker thinks it knows, and makes astonishingly specific claims such as "There is only a 14% chance of rain tomorrow."

It is important to understand that the probabilist appears on the scene only after the assignment of probabilities to events has been made. It is not the probabilist's business to make the assignments of probabilities to events, but to work out consequences of those assignments.

Here is how you turn the above discussion into precise mathematics. Let Ω be a non-empty set. Think of it as "the set of all possible outcomes of a random experiment." For instance, if your experiment were to throw a dart at a circular target, then Ω might be the set of all points on the target where the dart might land. This is just the intuition. Mathematically, Ω is a non-empty set, nothing more and nothing less. It is called the *sample space* in this context, although it might better be called the *outcome set.*

Events are subsets $E \subseteq \Omega$. Think of the event that the outcome of our random experiment lies in E, for instance that the dart lands in E. However, instead of saying, awkwardly, "the event that the outcome of our random experiment lies in

E", we usually say more briefly "the event E".

For technical reasons that cannot and need not be discussed here, we do not necessarily consider *all* subsets of Ω. Some subsets $E \subseteq \Omega$ may be so complicated or so awkward that we decline to talk about them. The collection of all events we consider may therefore not be the power set of Ω, but some subset Σ of the power set. So Σ is a collection of subsets of Ω, and we call it the collection of events. We assume:

1. $\Omega \in \Sigma$,
2. $E \in \Sigma \Rightarrow E^c \in \Sigma$,
3. $E_1, E_2, \ldots \in \Sigma \Rightarrow E_1 \cup E_2 \cup \ldots \in \Sigma$.

(In condition 3, "$E_1, E_2, \ldots$" may stand for a finite or an infinite sequence of events.)

Definition 2.1. *A collection Σ of subsets of Ω that satisfies conditions 1–3 is called a* σ-algebra of subsets of Ω.

Condition 1 is very natural: We should be allowed to talk about the event "The outcome lies in Ω." In fact, it always does — so that's an event that's guaranteed to occur. Condition 2 is natural as well: If we are allowed to talk about the event that the outcome of the random experiment lies in E, then we should also be allowed to talk about the event that the outcome of the experiment *does not* lie in E! Notice that conditions 1 and 2 together imply $\emptyset \in \Sigma$. Of course, $\emptyset$ is an event that's guaranteed not to occur.

Condition 3 is the most important one. The union of a sequence of events is again an event. We say that Σ *is closed under countable unions.* Unions are reminiscent of sums, the word sum starts with an s, and the Greek letter σ is reminiscent of s. This tortured reasoning explains the phrase *σ-algebra*, and the letter Σ. Notice that from conditions 2 and 3, we can also deduce:

$$\begin{array}{ll} & E_1, E_2, \ldots \in \Sigma \\ \Rightarrow & E_1^c, E_2^c, \ldots \in \Sigma \\ \Rightarrow & E_1^c \cup E_2^c \cup \ldots \in \Sigma \\ \Leftrightarrow & (E_1^c \cup E_2^c \cup \ldots)^c \in \Sigma \\ \Leftrightarrow & E_1 \cap E_2 \cap \ldots \in \Sigma. \end{array}$$

So Σ is also closed under countable intersections.

Given Ω and Σ, we assume that to each $E \in \Sigma$, a "probability" $P(E) \in [0,1]$ has been assigned. As I wrote before, how the function

$$P: \ \Sigma \to [0.1]$$

is defined is not the probabilist's business. The probabilist appears on the scene after this has been done. The assignment of probabilities can be viewed as *mathematical modeling*, but it isn't mathematics. We assume, however, that it is done so that the following conditions are satisfied:

(a) $P(\Omega) = 1$,

(b) $0 \leq P(E) \leq 1$ for all E,

(c) if $E_1, E_2, \ldots \in \Sigma$ are disjoint, then

$$P(E_1 \cup E_2 \cup \ldots) = \sum_{k=1}^{\infty} P(E_k).$$

Condition (a) is quite obvious: Since Ω is the set of all possible outcomes of the random experiment, the probability that the outcome lies in Ω is 1, or 100%. Probability (b) is equally obvious: Probabilities should lie between 0 and 1. A probability $p \in [0, 1]$ is a probability of $p \cdot 100\%$. Condition (c) is called *countable additivity.*

Definition 2.2. *If (a)–(c) are satisfied, then the function $P : \Sigma \to [0, 1]$ is called a* probability measure on Σ.

If $E \in \Sigma$, then $E^c \in \Sigma$, and if P is a probability measure on Σ, then

$$1 = P(\Omega) = P(E \cup E^c) = P(E) + P(E^c),$$

so

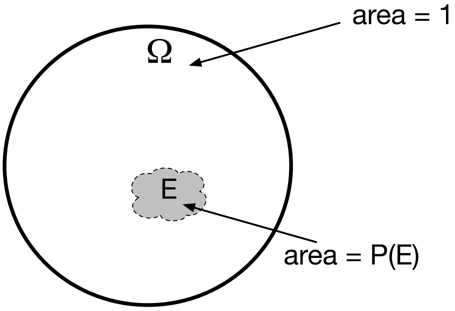
$$P(E^c) = 1 - P(E).$$

This is how it should be: If event E has the probability 0.2 for instance (20% probability), then event E^c, the event that E does not occur, has probability 1-0.2 = 0.8 (80% probability).

Definition 2.3. *If Ω, Σ, and P satisfy all the conditions spelled out above, we say that the triple (Ω, Σ, P) is a* probability space.

All discussions in rigorous probability theory assume an underlying probability space, sometimes referred to as "*the* underlying probability space" without being specified explicitly. We will from now on assume that, and won't write "Let (Ω, Σ, P) be a probability space" over and over again. It turns out that (Ω, Σ, P) fades into the background, and most statements in probability do not depend on exactly how it is defined, as long as it is a probability space of course.

Although our definitions are abstract, it is often useful for intuition to think of Ω as being a disk with area 1, say, and of $P(E)$ as being the area of $E \subseteq \Omega$.



2.1.2 Conditional probability of events

Definition 2.4. *Let A and E be events (that is, members of Σ) with $P(E) > 0$. Then the* probability of A, given E, *in symbols $P(A|E)$, is defined by*

$$P(A|E) = \frac{P(A \cap E)}{P(E)}. \tag{2.1}$$

Notice that the condition $P(E) > 0$ is important: We divide by $P(E)$ in (2.1). So $P(A|E)$ is the fraction of E also occupied by A, where the sizes of sets are measured using the probability measure P.

2.1.3 Bayes' formula

Theorem 2.5 (Bayes' formula). *Let A and E be events (that is, members of Σ) with $P(A) > 0$ and $P(E) > 0$. Then*

$$P(A|E) = P(E|A)\,\frac{P(A)}{P(E)}. \tag{2.2}$$

Proof. The right-hand side is

$$\frac{P(A \cap E)}{P(E)},$$

and the left-hand side is

$$\frac{P(E \cap A)}{P(A)}\,\frac{P(A)}{P(E)}.$$

These are evidently the same. □

This rather straightforward-looking fact turns out to be one of the most important in all of probability. I will devote a separate chapter to it. Here I will restrict myself to a few sentences. Think of $P(A)$ as "what you thought the probability of A was, until you realized the evidence E". (This is why E is called E.) Bayes' formula tells you how you should *update* your beliefs about A in light of the evidence. To highlight this, write (2.2) like this:

$$P(A|E) = \frac{P(E|A)}{P(E)}\,P(A).$$

2.1.4 Independence of events

Intuitively, two events E and F are called independent if knowledge that one occurs does not help you guess whether the other occurs:

$$P(F|E) = P(F) \quad \text{and} \quad P(E|F) = P(E).$$

As written, these conditions make sense only if $P(E) > 0$ and $P(F) > 0$. But we can re-write the conditions as

$$\frac{P(F \cap E)}{P(E)} = P(F) \quad \text{and} \quad \frac{P(E \cap F)}{P(F)} = P(E).$$

Both of these are equivalent to

$$P(E \cap F) = P(E)P(F).$$

This condition makes sense even if $P(E) = 0$ or $P(F) = 0$ (or both).

Definition 2.6. *Events E and F are called* independent *if*

$$P(E \cap F) = P(E)P(F).$$

We are done with the abstract formalism for now, and will now turn to examples for which an intuitive understanding of events, probabilities, and conditional probabilities will be good enough.

2.2 The horrific story of Sally Clark

2.2.1 What happened

Sally Clark was an English solicitor. (That's one kind of legal practitioner in the English system.) In 1996, her infant son died within weeks of his birth. In 1998, another infant son died similarly. Clark was arrested and accused of murder. The defense argued that both instances had been cases of Sudden Infant Death Syndrome (SIDS). However, at the trial, the distinguished pediatrician Sir Roy Meadow testified for the prosecution. He explained that the probability of a child suffering SIDS in an affluent non-smoking household was 1 in 8,543. Therefore, he continued, the probability of two children suffering SIDS was $1/8543^2$, about 1 in 73 million.[1]

This argument went uncontested. Clark was found guilty of murder, and sentenced to life imprisonment.[2] She appealed, but the appeal was rejected, and she went to jail. About a year later, the Royal Statistical Society issued a statement that there was "no statistical basis" for the 1 in 73 million figure. At the same time, it became known that evidence that the second child had died of a bacterial infection had been withheld at the original trial. Clark appealed again, and after more than three years in jail, her conviction was overturned, and she was released. She died a few years later, at age 42, of alcohol poisoning.

Roy Meadow's license to practice medicine was revoked by the General Medical Council. However, he appealed this ruling successfully. A few years later, he voluntarily relinquished his license. He is currently 91 years old.

[1] In fairness, he provided pages from a book that made this argument. It wasn't his own argument.

[2] It appears that the conviction was not *only* based on statistics, but, perhaps largely, on medical evidence.

Sally Clark, *Guardian*, November 2007

Sir Roy Meadow,
from the journal *Clinical Negligence, Law, and Ethics,* November 2020

2.2.2 Was anything wrong with the statistics?

There's the obvious issue of multiplying the probabilities. That's valid if and only if the two deaths are independent events. They aren't — SIDS runs in families.

But there is a more important question. Roy Meadow may have suggested at trial that the relevant probability here is

$$p = P\,(\text{two SIDS deaths})\,. \tag{2.3}$$

That's extremely small, even if not as small as 1 in 73 million. However, the probability of Sally Clark's innocence would, in the absence of further evidence[3], be

$$q = P\,(\text{two SIDS death} \mid \text{two infant deaths in the family})\,. \tag{2.4}$$

Notice that

$$q = \frac{p}{P\,(\text{two infant deaths in the family})}\,.$$

The probability of two infant deaths in the family is, thankfully, very small. Therefore q is very much larger than p.

2.3 O. J. Simpson's murder trial

2.3.1 The story

In 1995, the great football player O. J. Simpson was on trial for the murders of his ex-wife Nicole Brown Simpson and her friend Ron Goldman. The evidence against Simpson seemed overwhelming. However, he hired a high-profile defense team, and was found not guilty. In 1997, Simpson was found responsible for both deaths in a civil lawsuit. (The standard of evidence is lower in civil lawsuits.)

[3] There was a lot of further evidence, and I have not studied the facts carefully enough to have a firm opinion about what to believe.

2.3.2 An argument by Alan Dershowitz

During the trial, evidence was presented that O. J. Simpson had abused Nicole Brown Simpson during their marriage. Alan Dershowitz, one of the star lawyers on Simpson's defense team, argued that this should not be considered evidence in the trial, since only a very small fraction of wife batterers murder their wives. There appears to be some confusion about exactly what that percentage is, but a figure that seems consistent with Dershowitz's assertions is that a wife batterer has a probability of 1/2000 of murdering his wife in a given year.

2.3.3 I. J. Good's response to Alan Dershowitz

While the trial was still ongoing, in June of 1995, *Nature* printed an article by the distinguished statistician I. J. Good.[4] In this article, Good argued that Alan Dershowitz's observation was grossly misleading. He added, perhaps to add insult to injury, that conditional probability should be taught in high school because of its importance in law, medicine, and science. Good refined his argument in a second *Nature* publication in 1996. To explain his reasoning, wite

$$\begin{aligned} B &= \text{event that a wife is battered by her husband in 1994,} \\ M &= \text{event that she is murdered in 1994,} \\ G &= \text{event that she is murdered by her husband in 1994.} \end{aligned}$$

("G" as in "guilty".) Notice that $G \subseteq M$. Dershowitz called attention to the fact that

$$P(G|B) = 1/2000.$$

However, Good pointed out that the relevant probability is clearly not that, but

$$P(G|B \cap M),$$

since Nicole Brown Simpson is known to have been battered by her husband, and to have been murdered in 1994. These two conditional probabilities are vastly different:

Lemma 2.7. *If B, M, and G are events with $G \subseteq M$, then*

$$P(G|B \cap M) = \frac{P(G|B)}{P(M|B)}. \tag{2.5}$$

Proof.

$$P(G|B \cap M) = \frac{P(G \cap B \cap M)}{P(B \cap M)} = \frac{P(G \cap B)}{P(B \cap M)} =$$

$$\frac{P(G \cap B)}{P(B)} \cdot \frac{P(B)}{P(B \cap M)} = P(G|B) \cdot \frac{P(B)}{P(B \cap M)} = \frac{P(G|B)}{P(M|B)}.$$

[4] I. J. stood for Isadore Jacob at the time of his birth. He later changed that to Irving John, but signed his publications as I. J. Good only.

□

Notice that $P(M|B)$ is likely very small — only a small fraction of battered wives are murdered (by anybody). So the correction matters very much. Good proposed an upper bound on $P(M|B)$. First we note that

$$P(M|B) = \frac{P(B \cap M)}{P(B)} = \frac{P(B \cap M \cap G) + P(B \cap M \cap G^c)}{P(B)} =$$

$$\frac{P(B \cap G) + P(B \cap M \cap G^c)}{P(B)}.$$

In the last step, we used $G \subseteq M$. Now

$$\frac{P(B \cap G) + P(B \cap M \cap G^c)}{P(B)} = P(G|B) + P(M \cap G^c|B). \tag{2.6}$$

The probability $P(G|B)$ equals 1/2000 according to Dershowitz. The tricky term is $P(M \cap G^c|B)$. Here Good assumes that being battered by her husband does not affect a wife's probability of being murdered by somebody who isn't her husband, so

$$P(M \cap G^c|B) = P(M \cap G^c).$$

I doubt that, and unfortunately I think the inequality should be "$>$", not the desired "$<$". But following Good, we'll ignore that, and hope that it isn't too large an effect. Now we must estimate $P(M \cap G^c)$, the probability that a wife was murdered by somebody other than her husband in 1994. Based on easily available statistics, Good estimated the probability that a wife was murdered by *somebody* — husband or not — to be $1/20,000$. It appears that he then assumed that only a negligible fraction of murdered wives are murdered by their husbands. That isn't correct. In fact, one-third of murdered women are murdered by an intimate partner. Therefore I am going to take the probability that a wife was murdered by somebody other than her husband in 1994 to be 1/30,000, not 1/20,000. The expression in (2.6) now becomes

$$1/2000 + 1/30,000 = 1/1875.$$

That's our estimate of $P(M|B)$. So the probability cited by Dershowitz was 1875 times smaller than the truly relevant probability. The truly relevant probability is

$$\frac{1}{2000} \cdot 1875 = 0.9375.$$

As pointed out above, the factor 1875 may be too large, because it may be true that

$$P(M \cap G^c|B) > P(M \cap G^c).$$

However, Good's principal point remains valid: Dershowitz reported $P(G|B)$, but the relevant conditional probability should be $P(G|B \cap M)$, which is surely much greater.

2.4 Does the Covid vaccine send you to the ICU?

2.4.1 The anti-vaxxer's worst fears

The following numbers are from the New South Wales Respiratory Surveillance Report of the last week of 2022:

# doses of Covid vaccine	4	3	2	1	0
# ICU admissions with Covid	58	29	17	1	0

The vaccine skeptic's worst fears are confirmed. If you have been vaccinated many times, then you wind up in the ICU more often!

2.4.2 The sensible rebuttal

The table in fact tells you:

$$P\left(\text{4 Covid shots} \mid \text{ICU admission}\right) = \frac{58}{58+29+17+1} \approx 0.55.$$

But you don't want to know that probability. What you want to know is

$$P\left(\text{ICU admission} \mid \text{4 Covid shots}\right).$$

That's totally different! In fact imagine that *all* citizens of New South Wales had had 4 Covid shots. Then the table would have looked like this:

# doses of Covid vaccine	4	3	2	1	0
# ICU admissions with Covid	105	0	0	0	0

And yet it wouldn't have meant that Covid shots send you to the ICU.

2.4.3 The sensible rebuttal holds no water

Define the events

$$I: \text{ patient got admitted to the ICU, and}$$

$$F: \text{ patient got four Covid shots,}$$

$$N: \text{ patient got no Covid shots.}$$

To support our faith in vaccines, we would like to confirm that

$$P(I|F) \ll P(I|N).$$

That is, by Bayes' formula,

$$P(F|I)\frac{P(I)}{P(F)} \ll P(N|I)\frac{P(I)}{P(N)},$$

or

$$\frac{P(F|I)}{P(F)} \ll \frac{P(N|I)}{P(N)}.$$

Unfortunately, that's not true based on the data. In fact, based on the data, $P(N|I)$ appears to be zero.

2.5 The two-child puzzle

2.5.1 Martin Gardner's puzzle

The following puzzle appeared in Martin Gardner's column in the *Scientific American* in 1959, and has since appeared in most elementary books on probability.

> Mr. Smith has two children. He has a boy (at least one). How probable is it that he has two boys?

I should first point out that this puzzle is based on two falsehoods. We assume that each child can be identified unambiguously as "boy" or "girl", and that each of these two possibilities is equally likely. I have tried to find a similarly memorable way of phrasing the puzzle that avoids these falsehoods, but I couldn't.

You are now supposed to answer "1/2". Then the person who asked you triumphantly says "Oh no, the answer is 1/3, as anybody with half a brain can see." (The truth is, they didn't see it either. Somebody explained the two-child puzzle to them.)

Why is the answer 1/3? *A priori* there were four possible birth orders, all equally likely:

(boy,boy) (boy, girl) (girl, boy) (girl, girl)

When I told you that Mr. Smith has a boy, I ruled out (girl, girl). The other three options remained. These three are still equally likely, and in two of those three equally likely cases, Mr. Smith has only one boy.

You may have to puzzle over this for a while, but there's no trick; it is correct. You can convince yourself by computer simulation for instance. But before you go humiliate somebody else by making them answer 1/2, you should read on.

2.5.2 It gets stranger

Suppose that you knew that Mr. Smith had two children. Now you meet him in the grocery store. He is in company of one of his two children, whom he introduces as "my son". How probable is it that Mr. Smith has two sons? You are now asking "How probable is it that the child at home, the one not standing right here, is a boy?" The answer must clearly be 1/2.

So the answer to Martin Gardner's question depends on how you come to know that Mr. Smith has a boy. If Mr. Gardner tells you that Mr. Smith has a boy, the answer is 1/3, as long as you don't ask any questions. If you ask "Mr. Gardner, how do you know that he has a boy?", and Mr. Gardner answers "He was ahead of me in line at the pharmacy, and I overheard him say his son had a fever", then the answer becomes 1/2.

This illustrates that the connection between real life and probability is difficult. We must think very carefully about exactly which random experiment we are talking about, and how probabilities are to be assigned to events to match reality. (Formally speaking, we must think very carefully about what (Ω, Σ, P) should be.)

Take the example of the encounter in the grocery store. Let's assume that Mr. Smith, when he left the house to go to the grocery store, selected one of his two children at random to accompany him, with each child equally likely. There are three things that get chosen at random here. First, Fate determines the sex of the first child at random. Then, Fate determines the sex of the second child at random. Finally, Mr. Smith determines at random whether to take child 1 or child 2 on the excursion to the grocery store. We therefore have eight possible scenarios, all equally likely:

(boy, boy, 1) (boy, boy, 2) (boy, girl, 1) (boy, girl, 2)

(girl, boy, 1) (girl, boy, 2) (girl, girl, 1) (girl, girl, 2)

After your encounter in the grocery store, you know that in fact, one of the black scenarios happened. In half of those scenarios, Mr. Smith has two boys.

The example in the pharmacy is mathematically the same. Here Mr. Smith does not select a random child for a trip to the grocery store, but Fate selects a random child to be stricken by a fever.

2.5.3 The Adam puzzle

> Mr. Smith has two children. He has a boy named Adam. How probable is it that he has two boys?

As in Martin Gardner's puzzle, you know that he has a boy named Adam because your math teacher told you so. So how in the world could it matter whether the boy's name is Adam, Eric, or Sam? But it does.

To understand, assume that a family will call a boy Adam with a small probability p, unless they already have an Adam. The probability that a two-child family has first an Adam, then another boy is

$$\frac{1}{2}p \cdot \frac{1}{2}.$$

The probability that they have first an Adam, then a girl, is the same, also

$$\frac{1}{2}p \cdot \frac{1}{2}.$$

The probability that they have first a boy who isn't called Adam, then an Adam, is

$$\frac{1}{2}(1-p) \cdot \frac{1}{2}p.$$

The probability that they have first a girl, then an Adam is

$$\frac{1}{2} \cdot \frac{1}{2}p.$$

Therefore the probability that a two-child family has an Adam is

$$\frac{1}{2}p \cdot \frac{1}{2} + \frac{1}{2}p \cdot \frac{1}{2}. + \frac{1}{2}(1-p) \cdot \frac{1}{2}p + \frac{1}{2} \cdot \frac{1}{2}p = p - \frac{p^2}{4}.$$

The probability that a two-child family has an Adam and another boy is

$$\frac{1}{2}p \cdot \frac{1}{2} + \frac{1}{2}(1-p) \cdot \frac{1}{2}p = \frac{p}{2} - \frac{p^2}{4}.$$

The fraction of two-child families with an Adam who have another boy is therefore

$$\frac{\frac{p}{2} - \frac{p^2}{4}}{p - \frac{p^2}{4}} = \frac{2-p}{4-p} = 1 - \frac{2}{4-p}.$$

For any $p > 0$, this probability is always less than $1/2$, but for p near 0 it is almost $1/2$. Of course, if $p = 1$, that is, if a family will call a boy Adam unless they already have an Adam, the Adam puzzle becomes Martin Gardner's original puzzle, and the answer is $1/3$.

2.5.4 Why is the Adam puzzle so confusing?

In a 2024 paper in the *Mathematical Intelligencer*, Samer Nour-Eddine and I discussed the Adam puzzle and variations. There we argued that the Adam puzzle is confusing because we are inclined to misunderstand the statement

$$P(F|E) = r \tag{2.7}$$

as meaning

$$E \quad \Rightarrow \quad P(F) = r. \tag{2.8}$$

But (2.8) doesn't mean anything very clear, and certainly isn't the same as (2.7). If F is the event "Mr. Smith has two boys", and E_k is the event "Mr. Smith has a boy whose name is the k-th possible name for a boy", then (2.8) would lead us to believe that

$$\forall k \quad P(F|E_k) = r \qquad \Rightarrow \qquad P(F) = r.$$

This would be true if the E_k were disjoint. (Not independent, but disjoint.) However, the events "Mr. Smith has a boy named Adam" and "Mr. Smith has a boy named Sam" are not disjoint. The same family can have both an Adam and a Sam.

2.6 Questions and extensions

2.6.1 Roy Meadow's response

In an article in the *British Medical Journal*, volume 324, pages 41–43 (2002), Roy Meadow responded to the allegations that he gave misleading testimony in the Clark case. He stated that there was much *medical* evidence that indicated murder, not SIDS, in the Clark case. According to Meadow, it was not a matter of statistics.

If we wanted to know whether this is *actually* an example of a wrongful conviction based on a lack of understanding of conditional probabilities, we would need to study the evidence more carefully.

2.6.2 Cancer screening

Gerd Gigerenzer, director of the Harding Center for Risk Literacy in Berlin, gave a series of statistics workshops to more than 1,000 practising doctors in 2006 and 2007. He started every session with the same question:

> *A 50-year-old woman without symptoms participates in routine mammography screening. She tests positive, is alarmed, and wants to know from you whether she has breast cancer for certain or what the chances are. Apart from the screening results, you know nothing else about this woman. How many women who test positive actually have breast cancer? Which is the best answer?*
>
> - *nine in ten*
> - *eight in ten*
> - *one in ten*
> - *one in a hundred*
>
> *To help you:*
>
> - *1% of 50-year-old women have breast cancer.*
> - *90% of women with breast cancer test positive in mammography screening.*
> - *9% of women without breast cancer test positive in mammography screening.*

What do you say? Which is the best answer? You could start talking about $P(A|E)$ and $P(E|A)$, but Gigerenzer recommends (and I agree) that it is much clearer to simply think about 1000 women and ask how many do and don't have cancer, and how many do and don't test positive. In other words, Gigerenzer recommends that you think in terms of the *frequentist* interpretation of probability — a 10% probability means that it happens 10 times in 100 cases, typically.

The answers of the medical doctors were worse than they would likely have been based on random guessing.

2.6.3 Conditional probabilities vs. implications

We mentioned that

$$\forall k \quad P(F|E_k) = r \qquad \Rightarrow \qquad P(F) = r$$

would be true if the E_k were disjoint and covered all possibilities, in the sense that their union is Ω. Why is that true?

2.6.4 Mr. Smith at the information session for boys' sports

Mr. Smith has two school-age children. You are at the parent meeting at school. The principal says "Would those families who have a boy please go to room 333 to hear about the athletics programs for boys?" You see Mr. Smith get up and head towards room 333. He has a boy! What is the probability that Mr. Smith's other child is also a boy?

2.6.5 Mr. Smith at the military draft office

Mr. Smith has two children. One day the government announces that all families who have at least one boy must send a boy, of their choice if they have more than one, to register for the draft. You see Mr. Smith arrive at the registration office in company of a boy, who introduces himself as "Adam Smith".

Let's assume that the families with two boys choose the boy who gets sent in for draft registration by flipping a fair coin, paying no attention to their names. What is the probability that Mr. Smith has two boys?

2.6.6 Mr. Smith as a loyal follower of Great Leader Adam

We continue with the previous story, so we are still at the draft office, where Mr. Smith has arrived in company of his son Adam. This time, however, the country has an authoritarian leader, lovingly referred to as Great Leader Adam. All families name their first boy Adam, to prove their loyalty. (It is a matter of life or death to prove one's loyalty to the Great Leader.) In other words, $p = 1$ in our previous notation. What is the probability that Mr. Smith has two boys?

2.6.7 Updating probabilities based on two pieces of evidence

Suppose A and E are events. As we pointed out, Bayes' formula can be written like this:

$$P(A|E) = \frac{P(E|A)}{P(E)}\, P(A). \tag{2.9}$$

The "Bayesian" point of view is that *all* probabilities are conditional — conditioned on what we know. When we learn something new, we update our probabilities. To emphasize the viewpoint that all probabilities are conditional, we could write (2.9) as follows:

$$P(A|E) = \frac{P(E|A)}{P(E|\Omega)}\, P(A|\Omega). \tag{2.10}$$

(Note that $P(A|\Omega) = P(A)$.) Now suppose that E_1 and E_2 are events. With $E = E_1 \cap E_2$, (2.10) becomes

$$P(A|E_1 \cap E_2) = \frac{P(E_1 \cap E_2|A)}{P(E_1 \cap E_2|\Omega)} P(A|\Omega). \tag{2.11}$$

But what if news about E_1 having occurred reaches us first? We update from $P(A|\Omega)$ to

$$P(A|E_1) = \frac{P(E_1|A)}{P(E_1|\Omega)} P(A|\Omega). \tag{2.12}$$

Then news about E_2 having occured reaches us. We should now update from $P(A|E_1)$ to

$$P(A|E_1 \cap E_2) = \frac{P(E_2|A \cap E_1)}{P(E_2|E_1)} P(A|E_1),$$

With (2.12), this becomes

$$P(A|E_1 \cap E_2) = \frac{P(E_2|A \cap E_1)}{P(E_2|E_1)} \frac{P(E_1|A)}{P(E_1|\Omega)} P(A|\Omega). \tag{2.13}$$

Are (2.11) and (2.13) the same?

2.6.8 Updating twice based on the same piece of evidence

Again, we think of Bayes' formula,

$$P(A|E) = \frac{P(E|A)}{P(E)} P(A),$$

as a way of going from $P(A)$, our belief about the probability of A before we know anything, to $P(A|E)$, our belief about the probability of A after seeing the evidence E.

Suppose we now updated based on E again. Are we allowed to do that? Now we go from $P(A|E)$ to

$$\frac{P(E|A \cap E)}{P(E|E)} P(A|E).$$

Is that the same as $P(A|E)$?

Chapter 3

How confident should we be that the sun will rise tomorrow?

3.1 Why Laplace found this silly question interesting

Pierre Simon Laplace (1749–1827) was one of the great mathematicians of his time. In his 1814 book “A Philosophical Essay on Probabilities”, he discussed the question how confident we can be that the sun will rise tomorrow, given that — as he wrote — it has risen each morning in the past 5000 years. Obviously Laplace realized that the question, taken literally, is silly. His discussion was, in reality, about the following question.

How should our observations affect our expectations about the future?

That question isn't silly at all. It is fundamental to science.

It is also interesting to consider the generalization to a hypothetical situation where you see a few exceptional days on which the sun does *not* rise. For instance, if you see a vacuum cleaner on Amazon that has 145 good reviews (sunrises) and 15 bad ones (sunrise failures), how confident should you be about that particular vacuum cleaner, compared with one that has 7 good reviews and 0 bad ones for instance?

3.2 Coin tossing

3.2.1 What does coin tossing have to do with the sunrise?

Imagine Fate tossing a coin each morning to decide whether or not the sun should rise. The sun rises if the coin toss yields heads. Fate always uses the same coin, but the probability of heads is not necessarily 1/2. It may be a biased coin — the probability of heads is some number $\beta \in (0, 1)$. Fate does not choose a new β every morning; β has been chosen once and for all.

3.2.2 A biased coin with a known probability of heads

Proposition 3.1. *Suppose we repeatedly toss a biased coin which has a probability $\beta \in (0,1)$ of heads. If we toss the coin n times, the probability of getting heads exactly s times, where s is an integer in $\{0, 1, \ldots, n\}$, equals*

$$\left(\begin{array}{c} n \\ s \end{array} \right) \beta^s (1-\beta)^{n-s}. \tag{3.1}$$

Proof. If I told you a specific set of s tosses on which I want to get heads, and I wanted to get tails on the remaining $n - s$ tosses, then I would have to get heads on those s tosses — that has probability β^s — and tails on the remaining $n - s$ tosses — that has probability $(1 - \beta)^{n-s}$. The probabilities are multiplied because the coin tosses are independent of each other. There are, however, many ways of choosing s tosses out of n, namely $\left(\begin{array}{c} n \\ s \end{array} \right)$ ways. That's why there is a factor of $\left(\begin{array}{c} n \\ s \end{array} \right)$. □

3.2.3 When the probability of heads is itself random

Now suppose that the probability of heads is not known. We think of it as random, but it's not chosen at random before each toss, but rather just once and for all before we start tossing. We use capital letters for random quantities, so now the probability of getting heads is not β any more, but B.

Since we are not acquainted with Fate's way of thinking, we know nothing about B, and it seems reasonable therefore to assume, at least before we gather more information about sunrises, that B is uniformly distributed. This means that for any numbers c, d with $0 < c < d < 1$, the probability that $B \in [c, d]$ equals $d - c$. The number B is as likely to be anywhere in $(0, 1)$ as anywhere else.

Reading on will be more fun if you first try to guess what will happen. If you choose the probability of heads at random in $(0, 1)$ with uniform distribution, then toss the coin n times, which is more likely: that you get $n/2$ heads (assume n is even), or that you get no heads at all?

Proposition 3.2. *Suppose we repeatedly toss a biased coin which has a probability B of heads, where B is uniformly distributed in $(0,1)$, chosen once and for all before the first toss. If we toss the coin n times, the probability of getting heads exactly s times, where s is an integer in $\{0,1,\ldots,n\}$, equals*

$$\frac{1}{n+1}.$$

That is, the number of heads takes on each of its $n+1$ possible values $0,1,2,\ldots$, n with equal probability.

Proof. To find the probability of getting s heads, we simply average the probability of getting s heads if $B=\beta$, over all possible values of β. So we compute

$$\int_0^1 \binom{n}{s} \beta^s (1-\beta)^{n-s}\, d\beta.$$

We call this quantity p_s. For $1 \le s \le n$, we have

$$p_s = \int_0^1 \binom{n}{s} \beta^s (1-\beta)^{n-s}\, d\beta = \quad \text{(integration by parts)}$$

$$\binom{n}{s}\left[-\beta^s \frac{(1-\beta)^{n-s+1}}{n-s+1}\right]_{\beta=0}^{\beta=1} + \binom{n}{s}\int_0^1 s\beta^{s-1}\frac{(1-\beta)^{n-s+1}}{n-s+1}\, ds =$$

$$\binom{n}{s}\int_0^1 s\beta^{s-1}\frac{(1-\beta)^{n-s+1}}{n-s+1}\, ds =$$

$$\frac{n!}{s!(n-s)!}\;\frac{s}{n-s+1}\;\int_0^1 \beta^{s-1}(1-\beta)^{n-s+1}\, ds =$$

$$\frac{n!}{(s-1)!(n-s+1)!}\;\int_0^1 \beta^{s-1}(1-\beta)^{n-s+1}\, ds = p_{s-1}.$$

So all p_s, $0 \le s \le n$, are the same. Their sum is the probability that the number of heads is *some* number in $\{0,1,\ldots,n\}$, and that is certain. Therefore their sum is 1, so their common value is $\frac{1}{n+1}$. □

3.3 Statistics of B conditioned on s heads in n tosses

3.3.1 Conditional distribution function

Suppose that $x \in (0,1)$ is a fixed number, and think about the event

$$A: \quad B \le x.$$

This is an event that may or may not occur at the start, when B is chosen at random. Using our assumption that B is uniformly distributed, the probability of A equals x. But now let's analyze the *conditional* probability of A, given

$$E: \quad \text{On } n \text{ tosses, } s \text{ heads were obtained.}$$

Here Bayes' formula comes in:

$$P(A \,|\, E) = P(E \,|\, A) \frac{P(A)}{P(E)}.$$

We evaluate the right-hand side, using our assumption that $B \in (0,1)$ is uniformly distributed. We have $P(A) = x$, as mentioned already, and $P(E) = \frac{1}{n+1}$ by Proposition 3.2.

The somewhat trickier part is to calculate $P(E \,|\, A)$. So we assume now that $B \leq x$, and ask how likely it is that we get s heads in n tosses, given that. To find the answer, we must average

$$\begin{pmatrix} n \\ s \end{pmatrix} \beta^s (1-\beta)^{n-s}$$

over the interval from 0 to x. This yields

$$\frac{1}{x} \int_0^x \begin{pmatrix} n \\ s \end{pmatrix} \beta^s (1-\beta)^{n-s} \, d\beta.$$

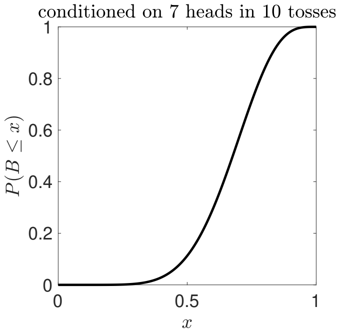
Putting it all together, we conclude:

Lemma 3.3. *Given the notation and assumptions above,*

$$P(B \leq x \mid \text{ s heads on n tosses}) = \int_0^x (n+1) \begin{pmatrix} n \\ s \end{pmatrix} \beta^s (1-\beta)^{n-s} \, d\beta. \quad (3.2)$$

Definition 3.4. *We call* (3.2)*, seen as a function of* $x \in (0,1)$*, the* distribution function of B, given s heads on n tosses.

Here is an example:



Without the condition that on n tosses, s heads were observed, the distribution function of B would simply be $P(B \le x) = x$ for $x \in [0,1]$, since B is assumed to be uniformly distributed.

3.3.2 Conditional probability density

Let $x \in (0,1)$ and $I = (0,x]$. Then (3.2) can also be written like this:

$$P(B \in I \mid \ s \text{ heads on } n \text{ tosses}) = \int_I (n+1) \begin{pmatrix} n \\ s \end{pmatrix} \beta^s (1-\beta)^{n-s} \, d\beta.$$

It is easy to convince yourself that this must be true for *all* intervals $I \subseteq (0,1)$ if it is true for intervals of the form $I = (0,x]$. For instance,

$$\begin{aligned} & P\left(B \in (0.5, 0.75] \mid s \text{ heads on } n \text{ tosses}\right) \\ = \ & P\left(B \in (0, 0.75] \mid \ s \text{ heads on } n \text{ tosses}\right) - P\left(B \in (0, 0.5] \mid \ s \text{ heads on } n \text{ tosses}\right) \\ = \ & \int_{(0,0.75]} (n+1) \begin{pmatrix} n \\ s \end{pmatrix} \beta^s (1-\beta)^{n-s} \, d\beta - \int_{(0,0.5]} (n+1) \begin{pmatrix} n \\ s \end{pmatrix} \beta^s (1-\beta)^{n-s} \, d\beta \\ = \ & \int_{(0.5,0.75]} (n+1) \begin{pmatrix} n \\ s \end{pmatrix} \beta^s (1-\beta)^{n-s} \, d\beta. \end{aligned}$$

In summary, to obtain the probability of $B \in I$, one just needs to integrate the function

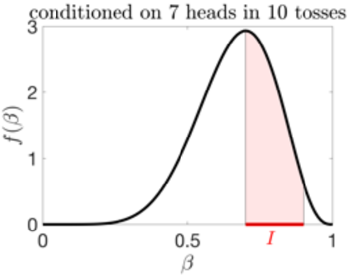
$$f(\beta) = (n+1) \begin{pmatrix} n \\ s \end{pmatrix} \beta^s (1-\beta)^{n-s} \tag{3.3}$$

over I.

Definition 3.5. *We call the function f defined in* (3.3) *the* probability density function of B, given s heads on n tosses.

Given s heads on n tosses, B is more likely to be where f is larger than where it is smaller. It is not hard to verify that f takes on its maximum at $\beta = s/n$. We say therefore that the *most likely value* of B is s/n. This has to be taken with a grain of salt: For any one particular β, the probability that $B = \beta$ is zero. But B is more likely to be *near* s/n than near any other value.

Here is an example:



The probability that B will like in I is the pink area.

3.3.3 Expected value of B given s heads on n tosses

Read this brief section only if you know some probability theory, and therefore understand what expected values are, and how you compute them from density functions.

Theorem 3.6. *The expected value of B, conditioned on s heads in n tosses, equals $\frac{s+1}{n+2}$ — precisely the probability that the $(n+1)$-st toss will yield heads.*

3.4 Laplace's law of succession

We now assume that Fate has already tossed the coin n times, and got s heads. It is about to toss again, for the $(n+1)$-st time. If $B = \beta$, the probability that Fate will get heads on the $(n+1)$-st toss equals β. Therefore, given that Fate got s heads in n tosses, the probability that Fate will get heads on the $(n+1)$-st trial equals

$$\int_0^1 \beta \cdot (n+1) \left(\begin{array}{c} n \\ s \end{array} \right) \beta^s (1-\beta)^{n-s} \, dx. \tag{3.4}$$

We average β over the interval $(0,1)$, but we weigh each value by the value of the conditional probability density at β.

The expression in (3.4) can be written as

$$(n+1) \left(\begin{array}{c} n \\ s \end{array} \right) \int_0^1 \beta^{s+1} (1-\beta)^{(n+1)-(s+1)} \, d\beta =$$

$$(n+1) \frac{\left(\begin{array}{c} n \\ s \end{array} \right)}{\left(\begin{array}{c} n+1 \\ s+1 \end{array} \right)} \int_0^1 \left(\begin{array}{c} n+1 \\ s+1 \end{array} \right) \beta^{s+1} (1-\beta)^{(n+1)-(s+1)} \, d\beta. \tag{3.5}$$

By Proposition 3.2, the integral in (3.5) is $\frac{1}{n+2}$. Therefore (3.5) equals

$$\frac{n+1}{n+2} \frac{\frac{n!}{s!(n-s)!}}{\frac{(n+1)!}{(s+1)!(n-s)!}} = \frac{s+1}{n+2}.$$

So here is the theorem that we have derived.

Theorem 3.7 (Laplace's Rule of Succession). *Suppose that $B \in (0,1)$ is random with uniform distribution, and a coin with probability of heads equal to B is tossed n times. Given that heads comes up s times, the probability that heads will come up on the $(n+1)$-st toss equals*

$$\frac{s+1}{n+2}.$$

Laplace estimated that the sun has risen every day in the past 5000 years, or 1,826,213 days. (This is the number of days in 5000 years, taking into account that a year is a leap year if it is divisible by 4, but not by 100 unless it is also divisible by 400.) He concluded that the likelihood of a sunrise tomorrow morning equals

$$\frac{1,826,214}{1,826,215} \approx 0.99999945.$$

After the first morning on Earth, having observed one night and one sunrise ($n = s = 1$), Adam and Eve should have concluded that the *most likely* value of B was $s/n = 1$, but the probability of seeing another sunrise on the next day was only $\frac{s+1}{n+2} = \frac{2}{3}$.

Suppose you see a vacuum cleaner on Amazon that has 145 good reviews (sunrises) and 15 bad ones (sunrise failures). Here $s = 145$ (that's the number of good reviews, or heads, or sunrises), and $n = 160$ (that's the total number of reviews, or tosses, or days). Your chance of having a good experience is

$$\frac{s+1}{n+2} = \frac{146}{162} \approx 0.90.$$

If you see another vacuum cleaner with 7 good reviews and 0 bad ones, your chance of having a good experience with that one is

$$\frac{8}{9} \approx 0.89.$$

By this analysis, you should go with the one that has 145 good reviews and 15 bad ones. (But it's a close call.)

My colleague Loring Tu points out that we shouldn't trust 7 good reviews and 0 bad reviews anyway, since the 7 good reviews might be fraudulent, placed by the sellers themselves. Laplace didn't know about that possibility.

Proof. The conditional expected value is

$$\int_0^1 \beta f(\beta)\, d\beta = \int_0^1 \beta(n+1) \left(\begin{array}{c} n \\ s \end{array} \right) \beta^s (1-\beta)^{n-s}\, d\beta.$$

This is precisely the integral we evaluated in Section 3.4 to obtain $\frac{s+1}{n+2}$. □

3.5 History

3.5.1 Bayes, 1761 or earlier

Among the most significant papers in the history of science is

> Thomas Bayes, An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London*, 1763.

The paper was submitted by Bayes' friend Richard Price[5]. Bayes had died in 1761. Price had found Bayes' notes, and Price wrote an introduction and an extensive concluding section himself. In his notes, Bayes discusses the problem that we have discussed here, and derived the conditional density of B. Here is a quote from Price's introduction:

> *Every judicious person will be sensible that the problem now mentioned is by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasoning concerning past facts, and what is likely to be hereafter.*

In this paper, Bayes used what is now called Bayes' formula, and that's why the formula bears his name. Richard Price (not Bayes) used the sunrise as an example in his concluding section.

3.5.2 Laplace, 1774

> Pierre Simon Laplace, Mémoire sur la probabilité des causes par les évènements[6], *Mémoires de l'Académie Royale des Sciences de Paris*, 1774.

In this paper, Laplace formulated his rule of succession. (It's the solution to "Problem 1" in the paper.)

3.5.3 Laplace, 1814

Laplace's whimsical discussion of the application of the law of succession to the sunrise appears in

> Pierre Simon Laplace, Essai philosophique sur les probabilités[7], Paris, 1814.

[5] Price was a fascinating character. Look him up on Wikipedia.

[6] Memoir on the probability of the causes of events.

[7] A philosopical essay on probabilities.

3.6 Questions and extensions

3.6.1 The most likely value of B

Convince yourself that $\frac{s}{n}$ is the value at which the conditional probability density f takes on its largest value.

3.6.2 One coin toss

Suppose you start out, as we did throughout this chapter, assuming that $B \in (0, 1)$ is uniformly distributed. Fate flips the coin once, and gets heads. If that's all you see, what is the conditional density of B?

3.6.3 Two coin tosses

Again we start out assuming that $B \in (0, 1)$ is uniformly distributed. Fate tosses the coin once, and gets heads, then another time, and gets tails. If that's all you see, what is the conditional density of B?

3.6.4 Updating after each toss

Suppose you start out using the *conditional* distribution of Section 3.6.2 for B. Fate tosses the coin once and gets tails. If that's all you see, what is the conditional density of B now? Is it the same as the one you got in Section 3.6.3?

3.6.5 Starting out certain that the coin is biased towards tails

Suppose you start out assuming that $B \in (0, 1/2)$ is uniformly distributed. So you believe that $B < 1/2$ with probability 1. Fate tosses the coin n times and gets heads every single time. (This might shake your belief that $B < 1/2$, especially when n is large.) What is the conditional density of B?

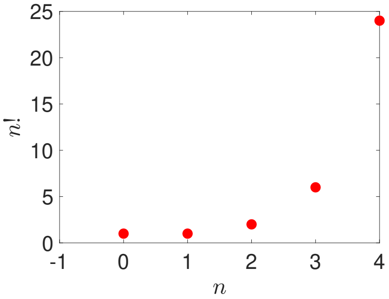
Chapter 4

Factorials of non-integers

4.1 Interpolants of the factorials

4.1.1 Naive interpolants

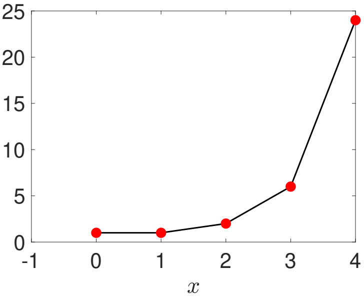
Everybody knows the factorials of integers: $0! = 1$, $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$.



Our question is this:

Is there a natural, appealing, and useful way of defining $x!$ for non-integer x?

We'll go for "natural and appealing" first, setting aside "useful" for now. We start with the simplest idea, piecewise linear interpolation:



We call this an *interpolant* of the factorials: A function $F = F(x)$ defined for all $x \geq 0$ that is equal to the factorials when x is an integer.

A central fact about factorials of integers is the recursion formula:

$$n! = n(n-1)! \quad \text{for } n \geq 1.$$

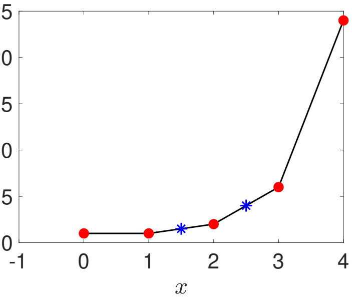
It would be pleasing (well, to me anyway) to preserve that formula:

We'll try to define an interpolant $F(x)$ of the factorials such that

$$F(x) = xF(x-1) \quad \textit{for all real numbers } x \geq 1. \tag{4.1}$$

The piecewise linear interpolant violates (4.1). In fact, if F denotes the piecewise linear interpolant, then for instance

$$F(2.5) = \frac{2! + 3!}{2} = \mathbf{4} \neq 2.5 \cdot F(1.5) = 2.5 \cdot \frac{1! + 2!}{2} = 2.5 \cdot 1.5 = \mathbf{3.75}.$$



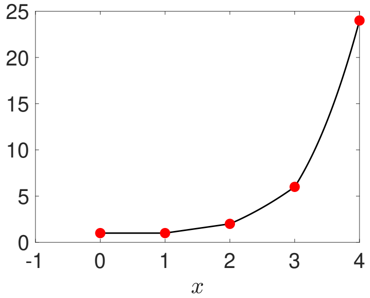
If we want (4.1) to hold, piecewise linear interpolation is not an option. But if preserving the recursion formula were really our only concern, we would still have infinitely many options. We would just need to settle on any function $f = f(x)$, $0 \leq x \leq 1$, with $f(0) = f(1) = 1$, and define $F(x) = f(x)$ for $0 \leq x \leq 1$, then use the recursion formula to define $x!$ for $x > 1$. For example

$$F(4.5) = 4.5F(3.5) = 4.5 \cdot 3.5F(2.5) = 4.5 \cdot 3.5 \cdot 2.5F(1.5) = 4.5 \cdot 3.5 \cdot 2.5 \cdot 1.5F(0.5).$$

Once we know what we want $F(0.5)$ to be, we also know what $F(4.5)$ should be, by the recursion formula. The recursion formula, together with $F(0) = F(1) = 1$, implies that $F(n) = n!$ for all integers $n \geq 0$. Of course, we will want f to be a continuous function, so that F becomes continuous.

Each continuous function $f = f(x)$ defined for $0 \leq x \leq 1$ with $f(0) = f(1) = 1$ gives rise to a continuous interpolant of the factorials that preserves the recursion formula.

Here is what we get with $f(x) = 1$:



This function isn't differentiable at $x = 1$, though: It is 1 for $0 < x < 1$ and x for $1 < x < 2$, so the derivative jumps from 0 to 1 at $x = 1$. This sharp corner translates into sharp corners at 2, 3, 4, etc. by the recursion formula.

Can we choose a differentiable $f(x)$, $0 \le x \le 1$, so that the resulting interpolant $F(x)$ of the factorials has no sharp corner at $x = 1$? Since $F(x) = xf(x-1)$ for $1 \le x \le 2$, we would need

$$f'(x) = \frac{d}{dx}\left(xf(x-1)\right) \quad \text{at } x = 1.$$

Using the product rule,

$$f'(1) = f(0) + f'(0).$$

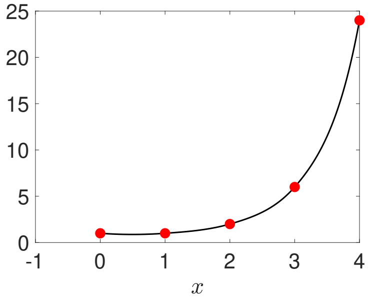
Using $f(0) = 0$, this becomes

$$f'(1) = 1 + f'(0). \tag{4.2}$$

For instance,

$$f(x) = 1 - \frac{x(1-x)}{2}$$

would work, since here $f'(0) = -1/2$ and $f'(1) = 1/2$. That gives this interpolant:

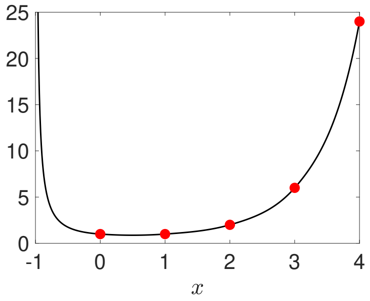


This function is differentiable at $x = 1$, and because of the recursion formula, differentiable for all $x \ge 0$.

Notice that we can also define $F(x)$ for $-1 < x < 0$ using the recursion formula:

$$F(x) = \frac{F(x+1)}{x+1} \quad \text{for } -1 < x < 0.$$

Then the interpolant of $x!$ looks like this.



This seems to be a pretty natural and appealing interpolant of the factorials. I will give it a name.

Notation. *Throughout this chapter, I will denote the particular interpolant of the factorials that we just defined by $F_0 = F_0(x)$. So*

$$F_0(x) = f(x) = 1 - \frac{x(1-x)}{2} \quad \textit{for } x \in [0,1],$$

and $F_0(x)$ is extended to all of $(-1,\infty)$ using the recursion formula.

By design, F_0 is once differentiable everywhere. However, I will note that F_0 is not twice differentiable at the integers n, $n \geq 0$.

Proposition 4.1. *The interpolant F_0 is once but not twice differentiable at the integers $n \geq 0$.*

Proof. We will show that the left-sided second derivative at $x = 1$ does not mach the right-sided one. For $0 \leq x \leq 1$, we have

$$F_0(x) = 1 - \frac{x(1-x)}{2}.$$

This function has the second derivative 1 everywhere. For $1 \leq x \leq 2$, we have

$$F_0(x) = xF_0(x-1) = x\left(1 - \frac{(x-1)(2-x)}{2}\right) = \frac{x^3 - 3x^2 + 4x}{2}.$$

The second derivative of F_0 for $1 \leq x \leq 2$ is therefore

$$\frac{d^2}{dx^2}\frac{x^3 - 3x^2 + 4x}{2} = 3x - 3,$$

and at $x = 1$ that's 0. So the second derivative of F_0 jumps at $x = 1$, from 1 to 0. From this and the recursion formula, we can now conclude that the second derivative of F_0 jumps in all integers $n \geq 0$. □

4.1.2 Euler's interpolant

Leonhard Euler (1707–1783) was one of the greatest figures in the history of mathematics.

By Jakob Emanuel Handmann. Derived from Leonhard Euler.jpg. Edited by BammeskOriginal. Source: Kunstmuseum Basel, Public Domain, via Wikimedia.

Euler observed that

$$\int_0^\infty \frac{t^n}{n!} e^{-t}\, dt$$

is the same for all n. You can see this by integrating by parts:

$$\int_0^\infty \frac{t^{n+1}}{(n+1)!} e^{-t}\, dt = \left[-\frac{t^{n+1}}{(n+1)!} e^{-t}\right]_{t=0}^\infty + \int_0^\infty \frac{t^n}{n!} e^{-t}\, dt = \int_0^\infty \frac{t^n}{n!} e^{-t}\, dt.$$

Since for $n = 0$, we get $\int_0^\infty e^{-t}\, dt = 1$, the common value must be 1. Thereby we have proved:

Observation. *For all integers $n \geq 0$,*

$$n! = \int_0^\infty t^n\, e^{-t}\, dt.$$

The integral $\int_0^\infty t^n e^{-t} dt$ is in fact well-defined for all *real* numbers $n > -1$. Why is $n = -1$ problematic? The integral $\int_0^\infty t^{-1} e^{-t}\, dt = \infty$ diverges because of the behavior near $t = 0$. In fact, $\int_0^1 t^{-1} e^{-t}\, dt$ is already infinite:

$$\int_0^1 t^{-1} e^{-t}\, dt \geq \int_0^1 t^{-1} e^{-1}\, dt = e^{-1} \left[\ln t\right]_0^1 = \infty.$$

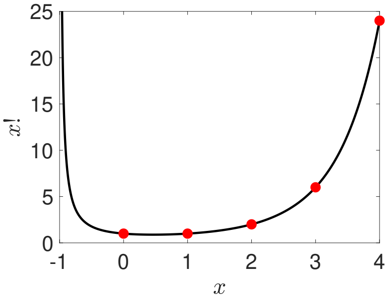
Definition 4.2 (Euler, 1738). *For any real number $x > -1$,*

$$x! = \int_0^\infty t^x e^{-t}\,dt. \tag{4.3}$$

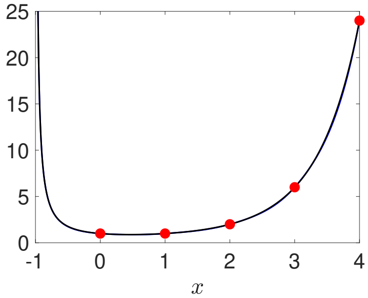
This is a fact if x is an integer, and a definition if it isn't.

We will use this definition from here on, so "$x!$" always means (4.3). Euler gave this definition in a 1738 paper titled, in English translation, "On transcendental progressions, that is, those whose general terms cannot be given algebraically". (He wrote in Latin, as was the custom in those days.) The paper is easy to find online, both in the original Latin and in English translation.

The graph looks like this:



(This is computed numerically. For most non-integer x, it is not possible to compute $x!$ explicitly.) The graph looks awfully familiar! In fact, the interpolant F_0 that we arrived at in Section 4.1.1 looks identical! To emphasize how similar they are, I am going to plot them both — F_0 in blue, and Euler's in black on top of it.



Do you see the two curves, one black and one blue? No? That's my point. They are strikingly similar. Perhaps, if you strain your eyes, you can see a slight deviation between $x = 3$ and $x = 4$. We will now show, however, that Euler's interpolant has advantages, both aesthetically and practically.

First, we will verify that Euler's definition satisfies the recursion formula:

Lemma 4.3. *For all $x > -1$,*

$$(x+1)! = (x+1)x!. \tag{4.4}$$

Proof. By integration by parts:

$$(x+1)! = \int_0^\infty t^{x+1}e^{-t}dt = \left[-t^{x+1}e^{-t}\right]_{t=0}^{\infty} + \int_0^\infty (x+1)t^x e^{-t}\,dt = (x+1)x!\ .$$

□

We also note that Euler's function $x!$ is arbitrarily often differentiable:

Lemma 4.4. *The function $x!$, $x > -1$, is arbitrarily often differentiable. For $k \geq 1$,*

$$\frac{d^k}{dx^k}x! = \int_0^\infty (\ln t)^k t^x e^{-t}\,dt \quad \textit{for } x > -1.$$

Proof. First we calculate

$$\frac{d}{dx}\int_0^\infty t^x e^{-t}\,dt = \frac{d}{dx}\int_0^\infty e^{x\ln t}e^{-t}\,dt.$$

We are allowed to exchange $\frac{d}{dx}$ and $\int_0^\infty$. If you have learned Real Analysis, you know why. If you have not learned Real Analysis, I only want you to appreciate that this is a generalized form of the *sum rule* of differentiation. Think of the integral as a sum, and of t as the summation index. The derivative of the sum is the sum of the derivatives. So

$$\frac{d}{dx}\int_0^\infty e^{x\ln t}e^{-t}dt = \int_0^\infty \frac{d}{dx}e^{x\ln t}e^{-t}dt = \int_0^\infty (\ln t)e^{x\ln t}e^{-t}dt = \int_0^\infty (\ln t)t^x e^{-t}dt.$$

We repeat this argument to obtain higher-order derivatives. Each new derivative gives a new factor of $\ln t$ in the integrand. □

Recall that F_0 isn't even twice differentiable at the integers $n \geq 0$. So there we have one aesthetic advantage of Euler's definition of $x!$ over F_0. However, it is possible to modify the idea of Section 4.1.1 to obtain an interpolant that is arbitrarily often differentiable, and there are in fact infinitely many different ways of doing that. So the existence of arbitrary many derivatives is an appealing property of Euler's interpolant, but it is not unique to Euler's interpolant.

4.2 A unique property of Euler's interpolant

4.2.1 Stirling's formula

James Stirling, Thomas Bayes, and Leonhard Euler all lived around the same time:

James Stirling:	1692–1770
Thomas Bayes:	1701–1761
Leonhard Euler:	1707–1783

(Laplace came several decades later.) Stirling became famous for the following result.

Theorem 4.5 (Stirling's formula). *As $n \to \infty$, n integer,*

$$n! \sim n^n e^{-n} \sqrt{2\pi n}.$$

The symbol "$\sim$" is read as "is asymptotic to", and means something completely precise: The ratio of the two expressions converges to 1, so

$$\lim_{n\to\infty} \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} = 1.$$

I want you to absorb this formula. First of all, for large n, it couldn't possibly be true that

$$n! \approx n^n$$

for large n. In fact, n^n is *much* greater than $n!$ for large n: There are n factors of n in n^n, but most of the n factors in $n!$ are much smaller than n. Another way of looking at it: n^n grows much faster than $n!$. In fact,

$$\frac{(n+1)!}{n!} = n+1$$

but

$$\frac{(n+1)^{n+1}}{n^n} = \frac{(n+1)^n}{n^n}(n+1) = \left(\frac{n+1}{n}\right)^n (n+1) = \left(1+\frac{1}{n}\right)^n (n+1).$$

Now we use the beautiful fact that

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e.$$

(You once learned how to prove that, when you took calculus, but I won't digress.) So

$$\frac{(n+1)^{n+1}}{n^n} \approx e\frac{(n+1)!}{n!}$$

for large n. Going from one integer to the next, n^n grows e times faster than $n!$. This suggests that

$$n! \approx n^n e^{-n}$$

would be a more promising approximation. It is, but it's not quite right. You have to multiply be $\sqrt{2\pi n}$ to get the right asymptotic behavior for $n!$. The factor of $\sqrt{2\pi n}$ is the real surprise here. How does π enter into this? Where's the circle?

Stirling's formula gives a good approximation for $n!$ even for modest n. For instance, $5! = 120$, and $5^5 e^{-5}\sqrt{10\pi} \approx 118$.

It turns out that Euler's definition of $x!$ is useful for proving Stirling's formula. I will show you that next.

Theorem 4.6 (Stirling's formula for Euler's definition of $x!$). *As $x \to \infty$, x real,*

$$x! \sim x^x e^{-x}\sqrt{2\pi x}.$$

Note that Theorem 4.6 implies Theorem 4.5. So Theorem 4.6 is a stronger result.

Proof.

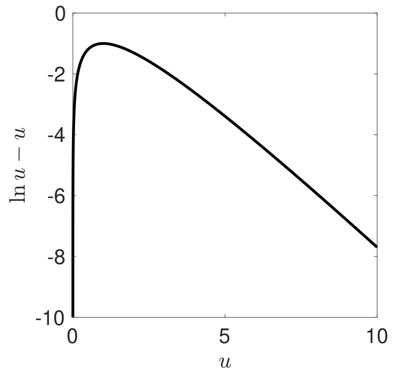
$$x! = \int_0^\infty t^x e^{-t} dt = \int_0^\infty e^{x \ln t - t}\, dt.$$

Assume $x > 0$, and substitute $t = xu$:

$$x! = x \int_0^\infty e^{x\ln(xu) - xu} du = x \int_0^\infty e^{x(\ln u - u) + x \ln x}\, du = x^x \cdot x \cdot \int_0^\infty e^{x(\ln u - u)}\, du.$$

Think about the function $u \ln u - u$. It converges to $-\infty$ as $u \to 0+$, and also to $-\infty$ as $u \to \infty$. Its maximum occurs when its derivative is zero:

$$\frac{d}{du}(u \ln u - u) = 0 \quad \Leftrightarrow \quad \ln u = 0 \quad \Leftrightarrow \quad u = 1.$$



The value of $\ln u - u$ is always negative. It is the least negative when $u = 1$; there it is -1. Now in

$$x! = x^x \cdot x \cdot \int_0^\infty e^{x(\ln u - u)}\, du,$$

the negative $\ln u - u$ is multiplied by x, and we are interested in large x. So all the values of $e^{x(\ln u - u)}$ are very small. They are the least small when $u \approx 1$. It seems plausible, therefore, that you get the correct asymptotic behavior as $x \to \infty$ is you approximate $\ln u - u$ by an expression that is accurate near $u = 1$, and very negative away from $u = 1$. This is in fact correct, but we will omit a rigorous discussion of why it is correct. (That's not so difficult, but would take us too far away from our focus in this chapter.) The idea used here is called *Laplace's method.*

We use the quadratic Taylor expansion of $\ln u - u$ near $u = 1$:

$$\ln u - u \approx -1 - \frac{1}{2}(u-1)^2 \quad \text{for } u \approx 1.$$

So

$$x! \sim x^{x+1} \int_0^\infty e^{x\left(-1-\frac{1}{2}(u-1)^2\right)}\, du = x^{x+1} e^{-x} \int_0^\infty e^{-\frac{x}{2}(u-1)^2} du.$$

Susbstitute $\frac{x}{2}(u-1)^2 = v^2$, so $v = (u-1)\sqrt{\frac{x}{2}}$:

$$x! \sim x^{x+1} e^{-x} \int_0^\infty e^{-\frac{x}{2}(u-1)^2} du =$$

$$x^{x+1} e^{-x} \int_{-\sqrt{x/2}}^\infty e^{-v^2} \sqrt{\frac{2}{x}}\, dv = x^x e^{-x} \sqrt{2x} \int_{-\sqrt{x/2}}^\infty e^{-v^2/2}\, dv. \tag{4.5}$$

As $x \to \infty$,

$$\int_{-\sqrt{x/2}}^\infty e^{-v^2/2}\, dv \to \int_{-\infty}^\infty e^{-v^2} dv, \tag{4.6}$$

and

$$\int_{-\infty}^\infty e^{-v^2} dv = \sqrt{\pi}. \tag{4.7}$$

(This, too, is a beautiful formula, and you learned it when you took multivariable calculus. I will resist reviewing its proof here. Suffice it to say that here is where the circle comes into play.)

Equations (4.5)–(4.7) imply the assertion. □

Note that this can't be the proof that Stirling gave, since Laplace lived after Stirling, and the proof uses Laplace's method. But it is a useful proof: For instance, if we used a higher-order Taylor expansion, we would get even more accurate approximation formulas for $x!$ (and in particular for $n!$).

If this is all you want from this chapter. you could stop reading here, but I would suggest that *at least* you read the theorem that the next section starts with.

4.2.2 What makes Euler's interpolant special

Theorem 4.7. *Let $F = F(x)$ be an interpolant of the factorials, defined for $x > -1$, satisfying the recursion formula. If Stirling's formula holds for $F(x)$, i.e., if if*

$$F(x) \sim x^x e^{-x}\sqrt{2\pi x} \quad \text{as } x \to \infty, \tag{4.8}$$

then $F(x) = x!$.

Proof. Suppose $F = F(x)$ is an interpolant of the integers that satisfies the recursion formula, and (4.8) holds. Let $x > -1$. Then by the recursion formula,

$$\frac{F(x)}{x!} = \frac{F(x+1)/(x+1)}{(x+1)!/(x+1)} = \frac{F(x+1)}{(x+1)!}.$$

Iterating this argument, we find

$$\frac{F(x)}{x!} = \frac{F(x+k)}{(x+k)!}.$$

for all integers $k \geq 1$.So

$$\frac{F(x)}{x!} = \frac{F(x+k)/\left((x+k)^{x+k}e^{-(x+k)}\sqrt{2\pi(x+k)}\right)}{(x+k)!/\left((x+k)^{x+k}e^{-(x+k)}\sqrt{2\pi(x+k)}\right)}.$$

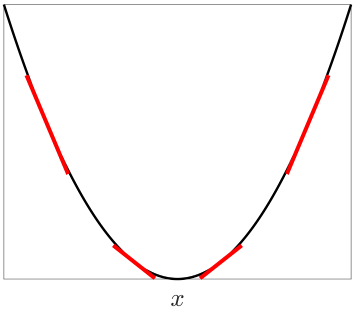
As $k \to \infty$, both numerator and denominator on the right-hand side converge to 1, since Stirling's formula holds for both F and Euler's interpolant $x!$. So $\frac{F(x)}{x!} = 1$ or $F(x) = x!$. □

Note that no continuity or differentiability assumption on F is needed.

4.3 Another unique property of Euler's interpolant

4.3.1 Convexity

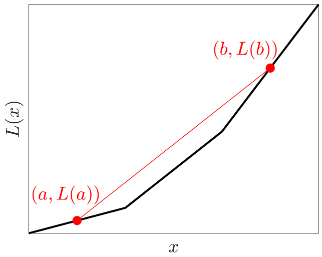
A function is "convex" if its graph is "concave-up" in the terminology used in Calculus courses. In Calculus, you connected convexity with positivity of the second derivative, as shown in the following picture.



tangent slope increases
from left to right
first derivative increases
second derivative is positive

However, to be convex, a function need not be differentiable:

Definition 4.8. *Let $I \subseteq \mathbb{R}$ be an interval. A function $L : I \to \mathbb{R}$ is called* convex *if for any pair $a, b \in I$ with $a < b$, the graph of L on the interval (a, b) does not rise above the secant line through the points $(a, L(a))$ and $(b, L(b))$. It is called* strictly convex *if the graph of L on (a, b) lies strictly below the secant line.*

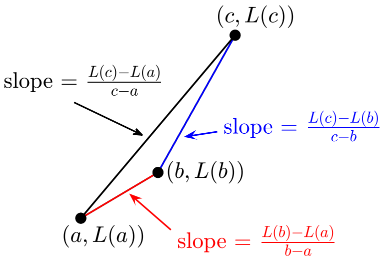


Lemma 4.9. *Let $L(x)$ be a function defined for $x \in I$, where $I \subseteq \mathbb{R}$ is an interval. Then L is convex if and only if for any $a, b, c \in I$ with $a < b < c$,*

$$\frac{L(b) - L(a)}{b - a} \leq \frac{L(c) - L(a)}{c - a} \leq \frac{L(c) - L(b)}{c - b}. \tag{4.9}$$

It is strictly convex if and only if the same holds with strict inequalities.

Proof by picture.


$(c, L(c))$
slope $= \frac{L(c)-L(a)}{c-a}$
slope $= \frac{L(c)-L(b)}{c-b}$
$(b, L(b))$
$(a, L(a))$
slope $= \frac{L(b)-L(a)}{b-a}$

□

4.3.2 Logarithmic convexity

Definition 4.10. *Let $I \subseteq \mathbb{R}$ be an interval, and $f : I \to (0, \infty)$ a function with positive real values. We call the function* (strictly) logarithmically convex *if $L(x) = \ln(f(x))$ is (strictly) convex.*

An increasing function that is strictly logarithmically convex grows faster than exponentially. For instance, $f(x) = e^{x^2}$ is strictly logarithmically convex and grows faster than exponentially, in the sense that

$$\frac{e^{x^2}}{e^{cx}} \to \infty \quad \text{as } x \to \infty$$

for any $c > 0$.

4.3.3 $x!$ is strictly logarithmically convex

Theorem 4.11. *The function $L(x) = \ln(x!)$, $x > -1$, is strictly convex.*

Don't confuse this with the statement that $x!$ is strictly convex. That's also true, and very straightforward to see.

Proof. Write $f(x) = x!$ in this proof. We have

$$L(x) = \ln(f(x)) \quad \Rightarrow \quad L'(x) = \frac{f'(x)}{f(x)} \quad \Rightarrow L''(x) = \frac{f(x)f''(x) - (f'(x))^2}{(f(x))^2}.$$

Our assertion is therefore equivalent to

$$(f'(x))^2 < f(x)f''(x) \quad \text{for } x > -1,$$

or to

$$\left|\int_0^\infty (\ln t)\, t^x\, e^{-t}\, dt\right| < \sqrt{\int_0^\infty t^x e^{-t}\, dt}\, \sqrt{\int_0^\infty (\ln t)^2 t^x e^{-t}\, dt} \quad \text{for } x > -1. \qquad (4.10)$$

If we define $g(t) = \sqrt{t^x e^{-t}}$ and $h(t) = (\ln t)\, \sqrt{t^x e^{-t}}$, (4.10) means

$$\left|\int_0^\infty g(t)h(t)\, dt\right| < \sqrt{\int_0^\infty g(t)^2\, dt}\, \sqrt{\int_0^\infty h(t)^2\, dt}. \qquad (4.11)$$

This holds by the Cauchy-Schwarz inequality; see below. The inequality is strict because $h(t)$ is not a constant multiple of $g(t)$. □

You may not know what the Cauchy-Schwarz inequality is, but when you took Linear Algebra, you probably learned a version of it. That version goes as follows. Let

$$g = \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix} \quad \text{and} \quad h = \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix}$$

be real vectors. Then

$$\left| \sum_{k=1}^{n} g_k h_k \right| \leq \sqrt{\sum_{k=1}^{n} g_k^2} \sqrt{\sum_{k=1}^{n} h_k^2}. \tag{4.12}$$

The inequality is strict unless g is a multiple of h or vice versa. You may see the analogy between (4.11) and (4.12) here. In (4.11), the summation index k has become the integration variable t, and summation has turned into integration.

If you denote by g^T the transpose of g, and by $\| \cdot \|$ the Euclidean norm, then (4.12) can be written like this:

$$\left| g^T h \right| \leq \|g\| \, \|h\|.$$

You might have learned it in this form.

4.3.4 Why we might like an interpolant of the factorials to be strictly logarithmically convex

To say that $\ln(x!)$ is strictly convex is to say that the first derivative of $\ln(x!)$ is strictly increasing. The analogous statement for the sequence $\{\ln(n!)\}_{n=0,1,2,...}$ is that its increments are strictly increasing. In fact, its increments are

$$\ln((n+1)!) - \ln(n!) = \ln(n+1),$$

which of course is a strictly increasing sequence.

4.3.5 The Bohr-Mollerup Theorem

Theorem 4.12 (Bohr-Mollerup Theorem, 1922). *Euler's interpolant is the only interpolant of the factorials that satisfies the recursion formula and is logarithmically convex.*

Proof. Let F be an interpolant of the factorials that satisfies the recursion formula and is logarithmically convex. It is suffcient to prove $F(x) = x!$ for $x \in (0,1)$; the recursion formula than implies $F(x) = x!$ for all $x > -1$.

Let $L(x) = \ln f(x)$. Then L is convex. Let $n \geq 1$. Then

$$\frac{L(n+x) - L(n)}{x} \geq \frac{L(n) - L(n-1)}{1} \tag{4.13}$$

and

$$\frac{L(n+x)-L(n)}{x} \le \frac{L(n+1)-L(n)}{1}. \tag{4.14}$$

Inequality (4.13) will yield lower bounds on $F(x)$ (one for each n), and (4.14) will yield upper bounds. We will work out what they are.

We start by relating $L(n+x)$ to $F(x)$.

$$\begin{aligned}
L(n+x) &= \ln F(n+x) \\
&= \ln\left((n+x)F(n-1+x)\right) \\
&= \ln\left((n+x)(n-1+x)F(n-2+x)\right) \\
&= \cdots \\
&= \ln\left((n+x)(n-1+x)\cdots(1+x)F(x)\right) \\
&= L(x)+\sum_{k=1}^{n}\ln(k+x). \qquad (4.15)
\end{aligned}$$

Using this, we re-write (4.13) as follows.

$$\begin{aligned}
& \frac{L(n+x)-L(n)}{x} \ge \frac{L(n)-L(n-1)}{1} \\
\Leftrightarrow\quad & \frac{L(x)+\sum_{k=1}^{n}\ln(k+x)-\ln(n!)}{x} \ge \ln(n!)-\ln((n-1)!) \\
\Leftrightarrow\quad & \frac{L(x)+\sum_{k=1}^{n}\ln(k+x)-\ln(n!)}{x} \ge \ln\frac{n!}{(n-1)!} \\
\Leftrightarrow\quad & \frac{L(x)+\sum_{k=1}^{n}\ln(k+x)-\ln(n!)}{x} \ge \ln n \\
\Leftrightarrow\quad & L(x) \ge x\ln n+\ln(n!)-\sum_{k=1}^{n}\ln(k+x)
\end{aligned}$$

Exponentiating on both sides, we obtain

$$F(x) \ge \frac{n!\; n^x}{\prod_{k=1}^{n}(k+x)}. \tag{4.16}$$

Similarly, (4.14) can be re-written as follows.

$$\begin{aligned}
& \frac{L(n+x)-L(n)}{x} \le \frac{L(n+1)-L(n)}{1} \\
\Leftrightarrow\quad & \frac{L(x)+\sum_{k=1}^{n}\ln(k+x)-\ln(n!)}{x} \le \ln((n+1)!)-\ln(n!) \\
\Leftrightarrow\quad & \frac{L(x)+\sum_{k=1}^{n}\ln(k+x)-\ln(n!)}{x} \le \ln(n+1) \\
\Leftrightarrow\quad & L(x) \le x\ln(n+1)+\ln(n!)-\sum_{k=1}^{n}\ln(k+x)
\end{aligned}$$

Exponentiating on both sides, we obtain

$$F(x) \le \frac{n!\,(n+1)^x}{\prod_{k=1}^{n}(k+x)}. \tag{4.17}$$

The ratio of the right-hand sides of (4.17) and (4.16) equals

$$\frac{(n+1)^x}{n^x} = \left(1+\frac{1}{n}\right)^x,$$

which converges to 1 as $n \to \infty$. We write

$$\frac{n!\,n^x}{\prod_{k=1}^{n}(k+x)} \sim \frac{n!\,(n+1)^x}{\prod_{k=1}^{n}(k+x)} \quad \text{as } n \to \infty.$$

The symbol $\sim$ stands for "asymptotic equivalence", meaning that the ratio of the two expressions converges to 1.

The ratio of the lower and upper bounds on $F(x)$ converges to 1 as $n \to \infty$. We conclude:

$$F(x) = \lim_{n\to\infty} \frac{n!\,n^x}{\prod_{k=1}^{n}(k+x)} \quad \text{for } x \in (0,1]. \tag{4.18}$$

(In particular, our reasoning implies that this limit exists.) Since $x!$ satisfies our assumptions, it must be given by (4.18) for $x \in (0,1)$, and therefore $F(x) = x!$ for $x \in (0,1)$. □

4.3.6 Euler's product formula as a byproduct of the proof

We have proved that

$$x! = \lim_{n\to\infty} \frac{n!\,n^x}{\prod_{k=1}^{n}(k+x)} \quad \text{for } x \in (0,1).$$

Incidentally, this formula holds for $x = 1$; that is straightforward to verify. So it holds for $x \in (0,1]$. Using the recursion formula, then for $x \in (-1,0]$,

$$\begin{aligned} x! = \frac{(x+1)!}{x+1} &= \frac{1}{x+1}\lim_{n\to\infty}\frac{n!\,n^{x+1}}{\prod_{k=1}^{n}(k+x+1)} \\ &= \lim_{n\to\infty}\frac{n!\,n^{x+1}}{\prod_{k=1}^{n+1}(k+x)} \\ &= \lim_{n\to\infty}\left(\frac{n!\,n^x}{\prod_{k=1}^{n}(k+x)}\,\frac{n}{n+x+1}\right). \end{aligned}$$

Since $n/(n+x+1)$ tends to 1 as $n \to \infty$, we conclude that

$$x! = \lim_{n\to\infty} \frac{n!\,n^x}{\prod_{k=1}^{n}(k+x)} \tag{4.19}$$

also holds for $x \in (-1, 0]$. Similarly, we can prove that this formula holds for all $x \in (1, 2]$, and so on. So it holds for all $x > -1$.

Finally, we massage the formula (4.19) a little bit to make it look nicer:

$$\begin{aligned}
x! &= \lim_{n\to\infty} \frac{n!\, n^x}{\prod_{k=1}^{n}(k+x)} \\
&= \lim_{n\to\infty} \left(n^x \prod_{k=1}^{n} \frac{k}{k+x} \right) \\
&= \lim_{n\to\infty} \left(\left(\frac{2 \cdot 3 \cdot \ldots \cdot n}{1 \cdot 2 \cdot \ldots \cdot (n-1)} \right)^x \prod_{k=1}^{n} \frac{k}{k+x} \right) \\
&= \lim_{n\to\infty} \left(\prod_{k=1}^{n-1} \left(\frac{k+1}{k} \right)^x \prod_{k=1}^{n} \frac{k}{k+x} \right) \\
&= \lim_{n\to\infty} \left(\left(\frac{n}{n+1} \right)^x \prod_{k=1}^{n} \left(\left(\frac{k+1}{k} \right)^x \frac{k}{k+x} \right) \right) = \prod_{k=1}^{\infty} \frac{\left(1+\frac{1}{k}\right)^x}{1+\frac{x}{k}}
\end{aligned}$$

So here is what this calculation proves.

Theorem 4.13. *For $x > -1$,*

$$x! = \prod_{k=1}^{\infty} \frac{\left(1+\frac{1}{k}\right)^x}{1+\frac{x}{k}}. \tag{4.20}$$

This formula comes from the proof of the Bohr-Mollerup theorem, but in fact Euler knew it already.

4.3.7 Who were Bohr and Mollerup?

Harald Bohr (1887–1951) was a distinguished Danish mathematician, a distinguished soccer player (he played on the Danish team in the olympics in 1908, they won the silver medal), and the younger brother of Niels Bohr, the Nobel-Prize-winning physicist. Johannes Mollerup (1872–1937) was a Danish mathematician as well. He wrote a highly influential textbook on mathematical analysis together with Harald Bohr.

4.4 Questions and extensions

1. **Factorials of half-integers.** Calculate

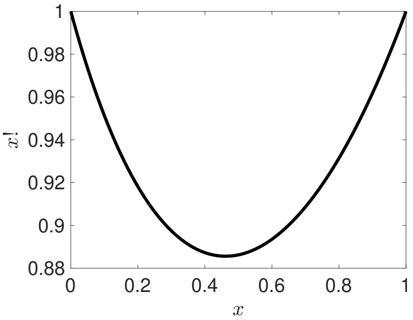
$$(0.5)! = \int_0^\infty \sqrt{t}\, e^{-t}\, dt.$$

Hint: Use $t = u^2$, and $\int_{-\infty}^{\infty} e^{-u^2} du = \sqrt{\pi}$. Compare with $F_0(0.5)$, where F_0 is the interpolant defined at the end of Section 4.1.1.

Then compute $(-0.5)!$ and $(-1.5)!$. These can all be obtained easily using the recursion formula.

2. **How big is the factorial of 1000?** Of course, 1000! is an integer. How many digits does it have?

3. **The shape of $x!$ for $0 < x < 1$.** As a function of $x \in [0, 1]$, the graph of $x!$ looks like this:



(a) Can you *prove* that it is concave-up, the derivative is negative at $x = 0$, and the derivative is positive at $x = 1$? (Hint: Don't try to prove the statements about the derivatives at 0 and at 1 directly. They follow from the facts that $0! = 1! = 0$ and the graph is concave-up.)

(b) The absolute value of the derivative of $x!$ at $x = 0$ is called the *Euler-Mascheroni constant* and appears in countless places in mathematics. It is denoted by γ. Its value is approximately 0.577. Show that

$$\gamma = \int_0^\infty e^{-t} \ln \frac{1}{t}\, dt.$$

Amazingly, it is not known whether γ is irrational. It is known, however, that *if* $\gamma = p/q$ where p and q are positive integers, then $q > 10^{244663}$ (no joke).

4. **Stirling lite.** Let $n \geq 1$ be an integer.

(a) Show (by drawing a picture) that the right Riemann sum for

$$\int_0^n \ln x\, dx$$

with $\Delta x = 1$ is larger than the integral itself. Conclude that

$$n! > n^n e^{-n}.$$

(b) Show (by drawing a picture) that the left Riemann sum for

$$\int_1^{n+1} \ln x\, dx$$

with $\Delta x = 1$ is smaller than the integral itself. Conclude that

$$n! < (n+1)^{n+1} e^{-n}.$$

(c) Explain why

$$(n+1)^{n+1} < e(n+1)n^n.$$

Combining this with (a) and (b), we obtain:

$$n^n e^{-n} < n! < e(n+1)n^n e^{-n}.$$

So up to a factor of no more than $e(n+1)$, $n!$ is $n^n e^{-n}$. Stirling did much better: The factor is about $\sqrt{2\pi n}$. But it is striking how easy it is to come within a factor of $e(n+1)$, which is smaller than the factor by which $(n+2)!$ differs from $n!$.

5. **The Cauchy-Schwarz inequality is the triangle inequality.** The Cauchy-Schwarz inequality came up in proving that $\ln(x!)$ is a strictly convex function. You learned about it in Linear Algebra. We can write it as

$$|g^T h| \le \|g\|\|h\|,$$

or as

$$g^T h \le \|g\|\|h\| \quad \text{and} \quad -g^T h \le \|g\|\|h\| \tag{4.21}$$

for vectors g and h in $\mathbb{R}^n$. This is one of the most important inequalities in all of mathematics. Another equally important inequality is the triangle inequality:

$$\|g+h\| \le \|g\| + \|h\|.$$

Suggestively, I'll write it as

$$\|g+h\| \le \|g\| + \|h\| \quad \text{and} \quad \|g-h\| \le \|g\| + \|h\|, \tag{4.22}$$

where the second inequality is just the first with h replaced by $-h$. Explain why (4.21) and (4.22) are the same statement in slightly different notation.

Chapter 5

Entropy of a probability vector

5.1 Introduction

Definition 5.1. *Let $n \geq 1$ be an integer. A vector $\rho = (\rho_1, \ldots, \rho_n)$ is called a* probability vector *if $\rho_i \in [0, 1]$ for all i, and*

$$\sum_{i=1}^{n} \rho_i = 1.$$

It's called a *probability vector* because we can talk about picking a random integer $I \in \{1, 2, \ldots, n\}$ so that

$$P(I = i) = \rho_i.$$

We say that ρ *defines a probability distribution on* $\{1, \ldots, n\}$. This means nothing other than what we just said.

Definition 5.2. *Let $n \geq 1$ be an integer. The vector*

$$u = \left(\frac{1}{n}, \ldots \frac{1}{n}\right)$$

is called the uniform probability vector.

You may have heard that the quantity

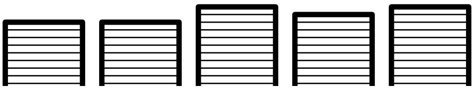
$$\sum_{i=1}^{n} \rho_i \ln \frac{1}{\rho_i} \tag{5.1}$$

is called the *entropy* of ρ. (Don't worry if you have never heard this.) Where does (5.1) come from? What is the significance of this expression? That's what this chapter is about. The notion of entropy appears in many contexts in the sciences, and alsc

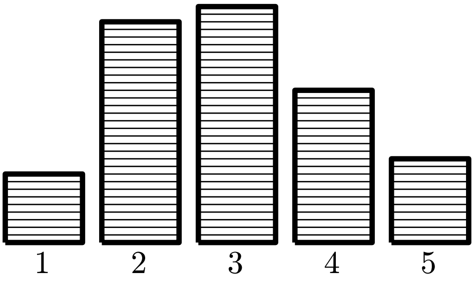
5.2 n

I have 1 somewhat obsessiv

On a go angement:



But cert



There seems to be nothing interesting here. If you want to put N building blocks into n piles of (approximately) equal height, then put approximately N/n blocks into each pile — approximately, since N/n might not be an integer. (In my example, it happens to be one: 100/5=20.)

In general, denote by k_i the number of blocks in the i-th pile, $1 \le i \le n$. So $\sum_{i=1}^n k_i = N$. We define

$$\rho_i = \frac{k_i}{N}, \quad 1 \le i \le n.$$

Then $\rho = (\rho_1, \dots, \rho_n)$ is a probability vector. How would you measure whether ρ is close to the uniform probability vector $u = \left(\frac{1}{n}, \dots, \frac{1}{n}\right)$? You might just calculate some quantity such as

$$\|\rho - u\| = \sqrt{\sum_{i=1}^n (\rho_i - u_i)^2}$$

(the *Euclidean norm* of $\rho - u$), or

$$\sum_{i=1}^n |\rho_i - u_i|,$$

(the *1-norm* of $\rho - u$), or

$$\max_i |\rho_i - u_i|$$

(the ∞-*norm* of $\rho - u$).

We will discuss an alternative to these simple ideas. This really seems unnecessary at first. It will lead us to the expression (5.1), however, and will eventually prove to be useful.

5.3 The probabilistic uniformity measure π_N

We want this discussion to be relevant to the natural world, where there is no magical hand placing building blocks but rather a natural process that might be (or might be well-modeled as) random. This motivates the following way of measuring how far ρ is from u.

Suppose that the building blocks are placed one at a time. When the j-th building block is placed ($1 \le j \le N$), it is placed into a random site $I_j \in \{1, 2, \dots, n\}$, with uniform distribution — any site is equally likely. The I_j are independent of each other. I will call the N-tuple

$$(I_1, \dots, I_N)$$

the *placement schedule*. There are n^N different placement schedules, since you get to choose one of of n possible times N times in a row. All of these placement schedules are equally likely.

Suppose as before that the end result is k_i blocks in location i, with $0 \le k_i \le N$ and $\sum_{i=1}^n k_i = N$. I will then call the n-tuple

$$(k_1, \dots, k_n)$$

the *outcome*. To determine the probability of the outcome $(k_1, \dots, k_n)$, we simply need to count how many of the equally like placement schedules produce this outcome. To generate a placement schedule leading to the outcome $(k_1, \dots, k_n)$, we start by selecting the k_1 integers j for which $I_j = 1$. There are

$$\left(\begin{array}{c} N \\ k_1 \end{array} \right) = \frac{N!}{k_1!(N-k_1)!}$$

different ways of doing that. Then we select, among the $N - k_1$ remaining integers j, the k_2 with $I_j = 2$. There are

$$\left(\begin{array}{c} N - k_1 \\ k_2 \end{array} \right) = \frac{(N-k_1)!}{k_2!(N-k_2)!}$$

different ways of doing that. And so on. The number of placement schedules resulting in the outcome $(k_1, \ldots, k_n)$ therefore equals

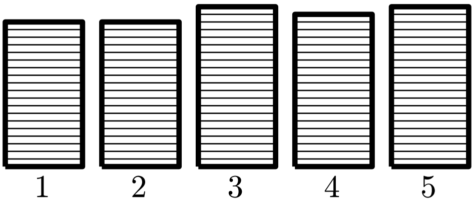
$$\frac{N!}{k_1!(N-k_1)!} \; \frac{(N-k_1)!}{k_2!(N-k_1-k_2)!} \; \frac{(N-k_1-k_2)!}{k_3!(N-k_1-k_2-k_3)!} \cdot \ldots \cdot =$$

$$N!$$

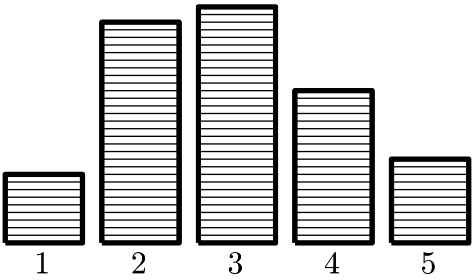
We writ

W of building blocks is e perfectly uniform listributed to:

Here π_N :



is even less likely, though only very slightly less: $\pi_{100} = 1.26 \cdot 10^{-4}$. The revoltingly non-uniform arrangement



is, however, *overwhelmingly* less likele: $\pi_{100} = 4.62 \cdot 10^{-9}$. Any one specific arrangement of building blocks is utterly unlikely to arise by pure chance, but ones that are far from uniformity are really astronomically unlikely.

Definition 5.3. *Let $\rho = (\rho_1, \ldots, \rho_n)$ be a probability vector. We call*

$$\pi_N(\rho) = \frac{N!}{(\rho_1 N)!(\rho_2 N)! \ldots (\rho_n N)!}$$

the probabilistic uniformity measure *of ρ.*

5.4 Simplifying π_N using Stirling's formula

5.4.1 Approximating π_N

We apply Stirling's formula to the factorials that appear in the formula for π_N, assuming that $\rho_i > 0$ for all i, so that $\rho_i N \to \infty$ as $N \to \infty$:

$$\pi_N \sim \frac{N^N e^{-N} \sqrt{2\pi N}}{\prod_i (\rho_i N)^{\rho_i N} e^{-\rho_i N} \sqrt{2\pi \rho_i N}} \frac{1}{n^N} \quad \text{as } N \to \infty.$$

Since $\sum_i \rho_i = 1$, this expression equals

$$\frac{\sqrt{2\pi N}}{n^N \prod_i \rho_i^{\rho_i N} \sqrt{2\pi \rho_i N}} = e^{N\left(-\ln n + \sum_i \rho_i \ln \frac{1}{\rho_i}\right)} \sqrt{\frac{(2\pi N)^{-n+1}}{\prod_i \rho_i}}.$$

(You may have to think about that for a moment, but I promise there's nothing deep here.) We summarize the result of this calculation in the following proposition.

Proposition 5.4. *Using the notation above, and assuming $\rho_i > 0$ for all i, the quantity*

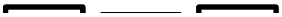
$$\tilde{\pi}_N = e^{N\left(-\ln n + \sum_i \rho_i \ln \frac{1}{\rho_i}\right)} \sqrt{\frac{(2\pi N)^{-n+1}}{\prod_i \rho_i}} \tag{5.2}$$

is asymptotic to π_N as $N \to \infty$.

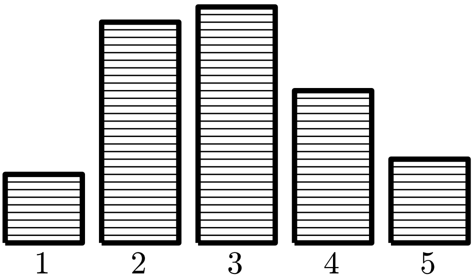
"Is asymptotic to" means the ratio, $\tilde{\pi}_N / \pi_N$, tends to 1 as $N \to \infty$.

ı for modest values of N.
ıtly, $\tilde{\pi}_N$ approximates π_N

$$_N \approx 1.39 \times 10^{-4}$$
$$_N \approx 1.42 \times 10^{-4}$$



$$_N \approx 1.26 \times 10^{-4}$$
$$_N \approx 1.28 \times 10^{-4}$$



$$\pi_N \approx 4.62 \times 10^{-9}$$
$$\tilde{\pi}_N \approx 4.74 \times 10^{-9}$$

5.4.2 $\lim_{N\to\infty} (\ln \pi_N)/N$

In the limit as $N \to \infty$, $\pi_N \sim \tilde{\pi}_N$. That means

$$\lim_{N\to\infty} \frac{\pi_N}{\tilde{\pi}_N} = 1,$$

or equivalently,

$$\lim_{N\to\infty} (\ln \pi_N - \ln \tilde{\pi}_N) = 0. \tag{5.3}$$

(A quantity converges to 1 if and only if its logarithm converges to 0.) Equation (5.3) implies, of course, that

$$\lim_{N\to\infty} \frac{\ln \pi_N - \ln \tilde{\pi}_N}{N} = 0. \tag{5.4}$$

Using the definition of $\tilde{\pi}_N$, eq. (5.2), we obtain the following observation, due to the great Austrian physicist Ludwig Boltzmann (1844–1906).

Theorem 5.5.

$$\lim_{N\to\infty} \frac{\ln \pi_N}{N} = \sum_{i=1}^{n} \rho_i \ln \left(\frac{1}{\rho_i} \right) - \ln n \tag{5.5}$$

It is tempting to think that because of Theorem 5.5, the quantity

$$e^{N\left(\sum_{i=1}^{n} \rho_i \ln\left(\frac{1}{\rho_i}\right) - \ln n\right)} \tag{5.6}$$

would be a good approximation for π_N. On second thought, however, that cannot be true: Think abut what happens when $\rho_i = \frac{1}{n}$ for all i. In that case,

$$\sum_{i=1}^{n} \rho_i \ln \left(\frac{1}{\rho_i} \right) - \ln n = \ln n - \ln n = 0,$$

so (5.6) equals 1, and that for sure isn't a good approximation for π_N, which is always much smaller than 1. Equation (5.4) is a much weaker statement than (5.3). Theorem 5.5 tells you abut $\frac{\ln \pi_N}{N}$ for large N, not about π_N.[8]

5.5 Entropy

Definition 5.6. *Let $\rho = (\rho_1, \ldots, \rho_n)$ be a probability vector. This means $0 \leq \rho_i \leq 1$ for all i, and $\sum_{i=1}^{n} \rho_i = 1$. The quantity*

$$S(\rho) = \sum_{i=1}^{n} \rho_i \ln \frac{1}{\rho_i},$$

where $\rho_i \ln \frac{1}{\rho_i}$ is taken to be 0 when $\rho_i = 0$, is called the entropy of ρ.

It is sensible to take $\rho_i \ln \frac{1}{\rho_i}$ to be 0 when $\rho_i = 0$, since

$$\lim_{x \to 0+} x \ln \frac{1}{x} = 0.$$

Nonetheless I cheated a little bit: In the preceding discussion, we really did need $\rho_i > 0$, since we applied Stirling's formula to $(\rho_i N)!$ in the limit as $N \to \infty$, but now I allowed $\rho_i = 0$ anyway. This might require a little bit of further thought, but I'll skip that thought for now.

The preceding discussion shows why the entropy should be a measure of uniformity. In fact:

[8] To understand the distinction, think about the following simple example: $e^{N+\sqrt{N}} \gg e^N$ but $\frac{\ln e^{N+\sqrt{N}})}{N} = 1 + \frac{1}{\sqrt{N}} \approx \frac{\ln e^N}{N} = 1$ for large N.

Lemma 5.7. *For any probability vector ρ,*

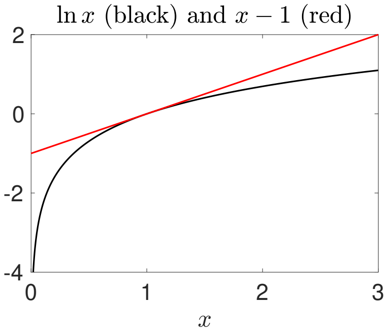
$$0 \leq S(\rho) \leq \ln n,$$

with $S(\rho) = \ln n$ if and only if $\rho_i = \frac{1}{n}$ for all i.

Proof. It is clear that $S(\rho) \geq 0$. To show that $S(\rho) \leq \ln n$, I will again assume $\rho_i > 0$ for all i, and leave it to you to convince yourself that the argument is correct even if some of the ρ_i are equal to 0. (It is not hard.) We have

$$S(\rho) = \sum_{i=1}^{n} \rho_i \ln \frac{1}{\rho_i} = \sum_{i=1}^{n} \rho_i \left(\ln \frac{1}{n\rho_i} - \ln \frac{1}{n} \right) = \left(\sum_{i=1}^{n} \rho_i \ln \frac{1}{n\rho_i} \right) + \ln n.$$

Next use that for all $x > 0$, $\ln x \leq x - 1$, with equality only for $x = 1$.



So

$$\begin{aligned}
\left(\sum_{i=1}^{n} \rho_i \ln \frac{1}{n\rho_i} \right) + \ln n &\leq \left(\sum_{i=1}^{n} \rho_i \left(\frac{1}{n\rho_i} - 1 \right) \right) + \ln n \\
&= \sum_{i=1}^{n} \frac{1}{n} - \sum_{i=1}^{n} \rho_i + \ln n \\
&= 1 - 1 + \ln n = \ln n,
\end{aligned}$$

with equality holding if and only if $\frac{1}{n\rho_i} = 1$ for all i. □

5.6 Entropic vs. Euclidean distance to uniformity

5.6.1 Definitions

If $\rho = (\rho_1, \ldots, \rho_n)$ is a probability vector, then $S(\rho)$ measures how close ρ is to uniformity — the large it is, the close is ρ to uniformity. Since the largest possible value of $S(\rho)$ is $\ln n$, we will instead consider the difference

$$\ln n - S(\rho) = \ln n - \sum_{i=1}^{n} \rho_i \ln \frac{1}{\rho_i},$$

which is something like a "distance" between $(\rho_1, \ldots, \rho_n)$ and $u = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. We have $\ln n - S(\rho) \geq 0$, and $= 0$ if and only if $\rho = u$.

Definition 5.8. *For a probability vector ρ, we call the quantity $\ln n - S(\rho)$ the* entropic distance to uniformity*.*

As we mentioned before, the distance to uniformity could be measured more naively by computing the Euclidean norm of $\rho - u$,

$$\|\rho - u\| = \sqrt{\sum_{i=1}^{n} \left(\rho_i - \frac{1}{n}\right)^2}.$$

Definition 5.9. *For a probability vector ρ, we call the quantity $\|\rho - u\|$ the* Euclidean distance to uniformity*.*

5.6.2 The entropic distance in the vicinity of uniformity

We will examine how $\ln n - S(\rho)$ can be simplified when the vector ρ is near uniformity, that is, $\rho_i \approx \frac{1}{n}$ for all i. Define $g(x) = x \ln x$. The quadratic Taylor expansion of g near a number $a > 0$ is

$$\begin{aligned} g(x) &\approx g(a) + g'(a)(x-a) + \frac{g''(a)}{2}(x-a)^2 \\ &= a \ln a + (\ln a + 1)(x-a) + \frac{1}{2a}(x-a)^2 \quad \text{for } x \approx a. \end{aligned}$$

When $a = \frac{1}{n}$, this becomes

$$g(x) \approx -\frac{\ln n}{n} + (-\ln n + 1)\left(x - \frac{1}{n}\right) + \frac{n}{2}\left(x - \frac{1}{n}\right)^2 \quad \text{for } x \approx \tfrac{1}{n}.$$

Therefore, when $\rho_i \approx \frac{1}{n}$ for all i,

$$\begin{aligned} \ln n - S(\rho) &= \ln n - \sum_{i=1}^{n} \rho_i \ln \frac{1}{\rho_i} \\ &= \ln n + \sum_{i=1}^{n} \rho_i \ln \rho_i \\ &\approx \ln n + \sum_{i=1}^{n} \left(-\frac{\ln n}{n} + (-\ln n + 1)\left(\rho_i - \frac{1}{n}\right) + \frac{n}{2}\left(\rho_i - \frac{1}{n}\right)^2\right) \\ &= \frac{n}{2} \sum_{i=1}^{n} \left(\rho_i - \frac{1}{n}\right)^2. \end{aligned}$$

We summarize this calculation in the following proposition.

Proposition 5.10 (loose version). *For probability vectors ρ near the uniform probability vector $u = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$,*

$$\ln n - S(\rho) \approx \frac{n}{2} \left\| \rho - u \right\|^2 . \tag{5.7}$$

Admittedly I was a little sloppy with my "≈" here. If that didn't bother you, then don't worry, you aren't missing anything important. If it did bother you, then you will also know how to give a more precise statement, using Taylor's theorem with remainder.

5.6.3 $\ln n - S(\rho)$ vs. $\frac{n}{2}\|\rho - u\|^2$ for $n = 2$

When $n = 2$, a probability vector ρ is of the form

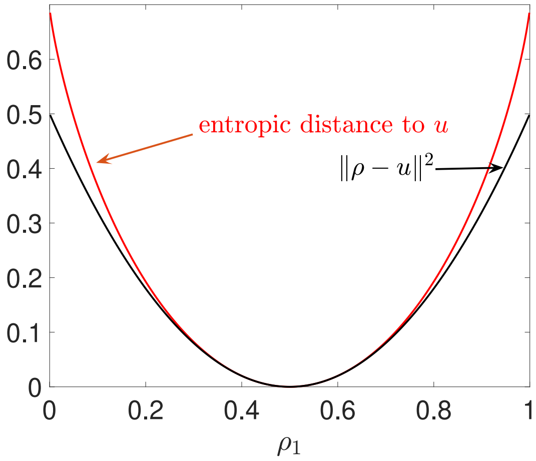
$$\rho = (\rho_1, 1 - \rho_1),$$

with $0 \leq \rho_1 \leq 1$. Then

$$\ln n - S(\rho) = \ln 2 + \rho_1 \ln \rho_1 + (1 - \rho_1) \ln(1 - \rho_1)$$

and

$$\frac{n}{2} \left\| \rho - u \right\|^2 = \left\| \rho - u \right\|^2 = \left(\rho_1 - \frac{1}{2} \right)^2 + \left(1 - \rho_1 - \frac{1}{2} \right)^2 = 2 \left(\rho_1 - \frac{1}{2} \right)^2 .$$



So for $n = 2$, the entropic distance from uniformity is nearly the same as $\|\rho - u\|^2$ for *any* probability vector ρ.

5.6.4 $\ln n - S(\rho)$ vs. $\frac{n}{2}\|\rho - u\|^2$ for larger n

To see how similar the $\ln n - S(\rho)$ and $\frac{n}{2}\|\rho - u\|^2$ are for $n > 2$, we pick a random probability vector Q as follows. First, we define $R = (R_1, \ldots, R_n)$, where the R_i are uniformly distributed random numbers in $(0, 1)$. Then we define $Q = (Q_1, \ldots, Q_n)$ with

$$Q_i = \frac{R_i}{\sum_{j=1}^n R_j}.$$

For any $s \in [0, 1]$, the vector

$$\rho_s = (1 - s)u + sQ$$

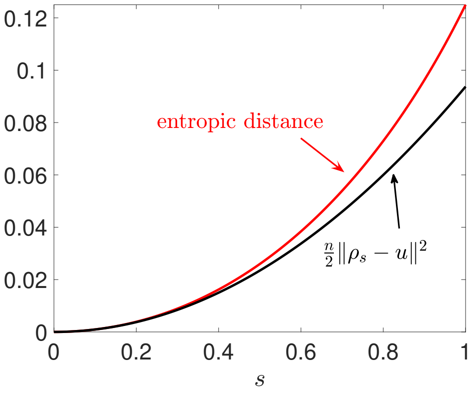
is a probability vector as well. We plot

$$\ln n - S(\rho_s)$$

and

$$\frac{n}{2}\|u - \rho_s\|^2$$

as functions of s in a single plot. For $n = 5$, for instance:



Again, the difference between $\ln n - S(\rho)$ and $\frac{n}{2}\|u - \rho\|^2$ appears to be quite minor. Of course, one would get a different picture using a different seed for the random number generator, but it appears that the discrepancy between $\ln n - S(\rho)$ and $\frac{n}{2}\|u - \rho\|^2$ is never dramatic. Furthermore, the same experiment for much larger n, say $n = 50,000$, gives very similar results.

This conclusion cannot be true in complete generality. Consider the probability vector ρ defined by $\rho_1 = 1$, $\rho_i = 0$ for $2 \leq i \leq n$. This probability vector maximizes both $\ln n - S(\rho)$ and $\frac{n}{2}\|\rho - u\|^2$, but their values are very different, especially when n is large:

$$\ln n - S(\rho) = \ln n$$

and

$$\frac{n}{2}\|\rho - u\|^2 = \frac{n}{2}.$$

I am afraid Section 5.6 remains inconclusive. Often, but not always, $\ln n - S(\rho)$ and $\frac{n}{2}\|\rho - u\|^2$ are similar quantities. They are close when $\rho \approx u$, not always close otherwise.

5.7 Information entropy

I will now describe an application of entropy in information theory. The ideas are due to Claude Shannon (1916–2001), a professor at MIT and the creator of Information Theory as a discipline.

David Tse, "How Claude Channon invented the future",
Quanta Magazine 2020

5.7.1 Information content of a message

If you tell me that an event, E, has occurred, the degree to which I am surprised depends on how likely I considered E to occur before you told me it had occurred. "Man bites dog" is far more surprising than "Dog bites man".

Suppose that p is your estimate of the probability of E *a priori* (before you are told that it has occurred). We will define a quantity called the *information content* of the news that E has occurred. (We identify information content with the degree to which you are surprised by the news.) Whatever it is, it is a quantity that depends on p. We denote it by $f(p)$.

If $p = 1$, then E is sure to occur, and then f should be 0:

$$f(1) = 0. \tag{5.8}$$

"The sun rose this morning" carries no information at all (regardless of what Laplace thought). Furthermore, it is clear from what I said initially that

$$f \text{ should be a strictly decreasing function of } p. \tag{5.9}$$

(An *a priori* strictly less likely news item carries strictly more information.) If E and F are independent events with *a priori* probabilities p and q, then the *a priority* probability that both occur is pq. The information content of the news that both have occurred, however, should be the sum of $f(p)$ and $f(q)$. So f should be defined such that

$$f(pq) = f(p) + f(q). \tag{5.10}$$

We know a function that has the properties (5.8)–(5.10):

$$f(p) = -\ln p, \qquad p \in [0, 1].$$

(If $p = 0$ and E occurs anyway, that's a miracle. Upon hearing the news that E has occurred, we are infinitely surprised.) We could scale $-\ln p$ by any positive constant and still obtain a function satisfying (5.8)–(5.10):

$$f(p) = -C \ln p, \quad p \in [0,1], \quad C > 0.$$

These are the only functions satisfying (5.8)–(5.10), at least the only differentiable ones. I won't spend time on proving that here, but see Section **??**.

For instance, C could be $\frac{1}{\ln 2}$, in which case we obtain

$$f(p) = -\frac{\ln p}{\ln 2} = -\log_2 p.$$

This was Claude Shannon's choice of C, for a reason described in the next section. We will use this choice of C from here on, but any choice of $C > 0$ would be compatible with (5.8)–(5.10).

5.7.2 Why $f(p) = -\log_2 p$ is a natural definition

Shannon thought in terms of messages encoded as sequences of 0's and 1's. The message

$$1011001 \tag{5.11}$$

carries more information than the message

$$101 \tag{5.12}$$

simply because it is longer. If each digit has the same probability of being 0 or 1, then the probability, p, of (5.11) is 2^{-7}, and its length is 7, while the probability of (5.12) is 2^{-3}, and its length is 3. In general, the length of a string of 0's and 1's is $-\log_2 p$.

5.7.3 Average content of a source of information

Suppose now that we have a source of information that may send out one of n possible messages, which we simply label $1, 2, \ldots, n$. Suppose that the probability of seeing the i-th message is ρ_i, $1 \le i \le n$. So $\rho = (\rho_1, \ldots, \rho_n)$ is a probability measure. If message i is sent out, the information content received is $-\log_2 \rho_i$. This happens with probability ρ_i. The *average* information content received is

$$\sum_{i=1}^{n} \rho_i \left(-\log_2 \rho_i\right) = \sum_{i=1}^{n} \rho_i \log_2 \frac{1}{\rho_i}. \tag{5.13}$$

The expression in (5.13) is called the *information entropy* of the source of information. Note that it is a property of the source of information, not of one particular message emenating from that source. It is precisely the entropy of the probability vector as defined earlier, except that we use $\log_2$ here instead of ln, following Shannon's convention. As I pointed out, any positive multiple of ln (that means any logarithm with any base $b > 1$) could be used in place of $\log_2$.

Think of a simple case of a weather channel that only sends out one of two messages every day: “1. It will rain tomorrow” or “2. It won’t rain tomorrow”. In a part of the world where it is guaranteed to rain every day, $\rho_1 = 1$ and $\rho_2 = 0$. Therefore in such a part of the world, the information entropy of the weather channel is zero. In a part of the world where it is just as likely to rain as not to rain, the information entropy of the weather channel is $\log_2 2 = 1$, the largest possible value.

As another example, think of a professor who assigns grades A, B, C, D, or F. If the professor gives an A to everybody, we have $\rho_1 = 1$ and $\rho_i = 0$ for $i \neq 1$. This means that the information entropy is zero — the average information content signaled by the professor’s grade is zero. To maximize the information entropy, the professor should give each of the five possible grades with the same frequency. This would have adverse consequences for the professor’s course evaluations, though.

Exercises

5.1. The entropy of a probability vector is a measure of its uniformity. You could also think of it as a measure of spread. The spread is great when ρ_i is small on the average. This line of thought results in the following alternative approach to measuring spread. Pick a random integer $I \in \{1, \ldots, n\}$ with $P(I = i) = \rho_i$. Then look at ρ_I. If that's small on the average, then ρ is spread out. The average value of ρ_I is

$$\sum_{i=1}^{n} \rho_i \cdot \rho_i.$$

The first factor ρ_i is there because the fraction of cases in which $I = i$ equals ρ_i. The second is there because when $I = i$, then $\rho_I = \rho_i$. So the spread is large when

$$Q(\rho) = \sum_{i=1}^{n} \rho_i^2$$

is small. What is the smallest that $Q(\rho)$ can get? (Not zero. Remember $\sum_{i=1}^{n} \rho_i = 1$.) When is it smallest? What is the largest that it can get? When is it largest? Does it seem like a good measure of "spread"?

Hint: The sledgehammer way of doing this uses Lagrange multipliers, but that's unnecessary and ugly.

5.2. Problem 1 suggests that spread is large when $1 - Q(\rho)$ is large. Verify that

$$1 - Q(\rho) = \sum_{i=1}^{n} \rho_i g(\rho_i)$$

with

$$g(x) = 1 - x.$$

This leads to a generalization of the idea in problem 1. You could in general define the quantity

$$S_g(\rho) = \sum_{i=1}^{n} \rho_i g(\rho_i)$$

as your measure of spread, where

$$g : (0, 1) \to \mathbb{R}$$

is a strictly decreasing function. The quantity S_g is the average value of $g(\rho_I)$ when $I \in \{1, \ldots, n\}$ is picked as in problem 1. Think about other possible choices of g. Do $g(x) = 1/x$ or $g(x) = 1/x^2$ work well? How about $g(x) = -\ln x$?

5.3. The real question here is whether there is some compelling reason to prefer $g(x) = -\ln x$ (which results in entropy) over all other possible choices. Here is one try. Suppose that you place your building blocks not in a line of n piles, but in a rectangular array of $m \times n$ piles: m rows and n columns. You choose a column index I, with column i chosen with probability ρ_i, $1 \leq i \leq n$. You independently choose a row index J, with column j chosen with probability η_j, $1 \leq j \leq m$. So both $\rho = (\rho_1, \ldots, \rho_n)$ and $\eta = (\eta_1, \ldots, \eta_m)$ are probability vectors. Now you should measure the spread of this arrangement by

$$\sum_i \sum_j \rho_i \eta_j g(\rho_i \eta_j). \tag{5.14}$$

On the other hand, you are making two independent random choices to place each building block, the choice of I and the choice of J. Therefore you might want your measure of spread to be equal to

$$\sum_i \rho_i g(\rho_i) + \sum_j \eta_j g(\eta_j). \tag{5.15}$$

Prove that this equals

$$\sum_i \sum_j \rho_i \eta_j (g(\rho_i) + g(\rho_j)). \tag{5.16}$$

If you want (5.14) to be (5.16), the natural thing to do would be to define g such that

$$g(xy) = g(x) + g(y) \tag{5.17}$$

for all x and y. This is the *law of logarithms*, and suggests that $g(x)$ should be a negative multiple of $\ln x$, for instance $-\ln x = \ln(1/x)$. One question that remains in this line of reasoning is whether (5.17) really implies that g is a constant multiple of $\ln x$. This is going to be settled in Chapter 6. But right now, prove that (5.17) implies $g(x) = C \ln x$ if you assume that g is differentiable.

5.4. Let's think about $n = \infty$. So now

$$\rho = (\rho_1, \rho_2, \ldots)$$

is an infinite sequence with $\rho_i > 0$ and

$$\sum_{i=1}^{\infty} \rho_i = 1. \tag{5.18}$$

While the ρ_i can't all be the same, so there's no such thing as a "uniform" infinite sequence, one can still view entropy as a measure of spread. Find an example for which

$$\sum_{i=1}^{\infty} \rho_i \ln \frac{1}{\rho_i} = \infty.$$

(Of course, (5.18) should hold.) It will help here to remember that

$$\sum_{i=2}^{\infty} \frac{1}{i \ln i} = \infty \quad \text{but} \quad \sum_{i=2}^{\infty} \frac{1}{i(\ln i)^2} < \infty.$$

(These were probably examples in your Calculus 2 class.)

Chapter 6

Does the sum rule imply the constant factor rule?

6.1 Linearity

By definition, linearity means that the sum rule and the constant factor rule hold. For instance, differentiation is linear because for any two differentiable functions $f(x)$ and $g(x)$,

$$\frac{d}{dx}(f(x)+g(x)) = \frac{df}{dx}(x) + \frac{dg}{dx}(x)$$

(that's the sum rule), and

$$\frac{d}{dx}(cf(x)) = c\frac{df}{dx}(x)$$

(that's the constant factor rule). In Linear Algebra, you define a mapping $L : \mathbb{R}^n \to \mathbb{R}^m$ to be linear if

$$\forall x, y \in \mathbb{R}^n \quad L(x+y) = L(x) + L(y)$$

(the sum rule), and

$$\forall x \in \mathbb{R}^n, c \in \mathbb{R} \quad L(cx) = cL(x)$$

(the constant factor rule). Our question in this chapter is whether these two rules are independent of each other, or whether one implies the other. It's quite clear that the constant factor rule can't imply the sum rule, but:

Does the sum rule imply the constant factor rule?

Can you give an example of a mapping where the sum rule holds but the constant factor rule doesn't?

6.2 The simplest case: Functions from $\mathbb{R}$ into $\mathbb{R}$

Let's make it as simple as possible. Let $f : \mathbb{R} \to \mathbb{R}$ be a function. Let's assume the sum rule:

$$f(x+y) = f(x) + f(y)$$

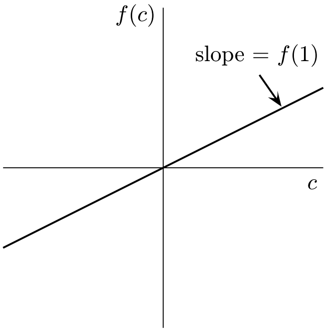
for all $x, y \in \mathbb{R}$. We say f is *additive*. Does the constant factor rule follow? That is, does it follow that

$$f(cx) = cf(x)$$

for all $c, x \in \mathbb{R}$? Notice that the constant factor rule implies, settig $x = 1$, that

$$f(c) = cf(1)$$

for all $c \in \mathbb{R}$, so f is linear in the sense that is well-familiar from high school.



Does additivity imply linearity?

6.3 Two equivalent questions

Does a function that obeys the law of exponentials have to be exponential?

Specifically, suppose $E:\ \mathbb{R} \to (0, \infty)$ satisfies

$$E(x + y) = E(x)E(y)$$

for all x and y. This is the *law of exponentials*. What can we say about E? We make the observation that $\ln E$ is an additive function. If we may conclude that $\ln E$ is linear, so $\ln E(x) = rx$ for all x, for some constant r, then indeed $E(x) = e^{rx}$, so $E(x)$ is an exponential function.

Does a function that obeys the law of logarithms have to be logarithmic?

Specifically, suppose that $L:\ (0, \infty) \to \mathbb{R}$ satisfies

$$L(st) = L(s) + L(t)$$

for all $s, t > 0$. This is called the *law of logarithms*. Then $L(e^x)$ is additive. If we may conclude that $L(e^x)$ is linear, so $L(e^x) = rx$ for all x, for some constant r, then we conclude $L(t) = r \ln t$ for all $t > 0$. That's what we will call a *logarithmic* function, using the phrase a little loosely (even $L(t) = 0$ is a logarithmic function by this terminology).

6.4 Reasons for asking these questions

There are several applications of these questions in probability theory. For instance:

Why are positive random numbers with lack of memory *exponentially distributed?*

The answer is: Because they satisfy the law of exponentials. But to complete the reasoning, we must know that the law of exponentials is *only* satisfied by exponential functions.

In Shannon's information theory, the "surprise" or "information content" of the news that an event has occurred is taken to be $\log_b \frac{1}{p}$, where p is the *a priori* probability of the event, and the basis b can be chosen to be any number > 1. (Shannon picked $b = 2$.)

Why should information content depend logarithmically on probability?

It is easy to see that f should satisfy the law of logarithms. To complete the reasoning, we must know that the logarithmic functions are the only ones satisfying the law of logarithms.

However, historically the motivation for asking whether additivity implies linearity was physical:

Why should we assume that two physical forces combine via vector addition?

This question goes back at least to an 1875 paper by Darboux. A nice exposition (and generalization) can be found in a paper by the present-day Hungarian mathematician Miklós Laczkovich, titled "On the resultant of forces", *Acta Mathematica Hungarica* 1995. The argument explained there shows that a simple and plausible set of axioms for the combination of forces implies that the correct description must be vector addition, *provided* that additivity implies linearity.

However, I think the main reason for studying these questions is aesthetic. Linearity is one of the simplest notions in all of mathematics. I want to understand it fully.

6.5 Cauchy's 1821 result

If $f(x+y) = f(x) + f(y)$ for all x and y, then, setting $y = x$, $f(x) = 2f(x)$, and therefore $f(3x) = f(2x+x) = f(2x) + f(x) = 3f(x)$, and so on. So $f(nx) = nf(x)$ for all $n \in \mathbb{N}$. But also $f(0+0) = f(0) + f(0)$, and therefore $f(0) = 0$, and $0 = f(0) = f(x + (-x)) = f(x) + f(-x)$, therefore $f(-x) = -f(x)$. Further, if n is a negative integer, $f(nx) = -f((-n)x) = -(-n)f(x) = nf(x)$. In summary, additivity implies the constant factor rule for all integer factors c.

If c is a rational number, so $c = p/q$ with $p \in \mathbb{Z}$ and $q \in \mathbb{N}$, then

$$qf\left(\frac{p}{q}x\right) = f\left(q \cdot \frac{p}{q}x\right) = f(px) = pf(x),$$

and therefore $f(cx) = cf(x)$. We have therefore proved:

Lemma 6.1. *If f is additive, then $f(cx) = cf(x)$ for all $c \in \mathbb{Q}$, $x \in \mathbb{R}$.*

What if c is irrational? The idea is to approximate c by rational numbers. There exists a sequence $\{c_n\}$ of rational numbers with $c_n \to c$, and therefore

$$f(cx) = f\left(\lim_{n\to\infty} c_n\, x\right) = \lim_{n\to\infty} f(c_n x) = \lim_{n\to\infty} c_n f(x) = cf(x).$$

(The second equality sign is justified because f is continuous.) So altogether, we conclude:

Theorem 6.2 (Cauchy, 1821). *If $f : \mathbb{R} \to \mathbb{R}$ is continuous and additive, then f satisfies the constant factor rule, i.e., f is linear.*

6.6 Bases of vector spaces

The only remaining question is whether continuity is actually a necessary assumption here. This question remained unresolved for more than 80 years, until it was settled in 1905 by Georg Hamel (1877–1954). The main theorem of Hamel's 1905 paper appears at first sight to have nothing to do with our question:

Theorem 6.3 (Hamel, 1905). *Every vector space has a basis.*

You learned in Linear Algebra what the words mean, but I'll review briefly. A *vector space* is a set of objects that can be added and multiplied by scalars so that all the laws of arithmetic familiar from $\mathbb{R}^n$ are valid. The scalars can be complex numbers, real numbers, rational numbers, or more generally elements of a *field* F. We then refer to V as a *vector space over the field* F. (If you don't know what a field is, that's not relevant to our discussion here, as long as you know that F could be $\mathbb{R}$ or $\mathbb{Q}$.)

A *basis* of a vector space is a collection $(b_i)_{i\in I}$ of elements of V with the property that every $v \in V$ has a *unique* representation in the form

$$v = \sum_{i\in I} c_i b_i$$

with $c_i \in F$, and such that *at most finitely many c_i* are non-zero (and therefore there is no question about whether and how the sum is defined).

For example, the vector space $\mathbb{R}^2$ over the field $\mathbb{R}$ has the basis

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

which is also called its *canonical* basis. The infinite sequences $(x_1, x_2, x_3, x_4, \ldots)$ of real numbers also form a vector space over $\mathbb{R}$. The sequences

$$(1,0,0,0,\ldots), \quad (0,1,0,0,\ldots), \quad (0,0,1,0,\ldots), \quad \text{etc.}$$

do *not* form a basis of this vector space. Can you see why not?

You may think you learned in Linear Algebra that any vector space has a basis. You may have been told, but you likely did not learn the proof. It's not a straightforward result by any means. In fact, its proof relies on the axiom of choice:

Axiom of choice. *Let I be a non-empty set, and for every $i \in I$, let S_i be a non-empty set. Then there exists a set*

$$\{x_i \ : \ i \in I\}$$

with $x_i \in S_i$ for all I.

This seems quite obvious, but it has been proved to be independent of the other axioms underlying set theory. A large part of mathematics relies on the axiom of choice. Not only does Hamel's proof rely on it, but his result is equivalent to it:

Theorem 6.4. *The statement that every vector space has a basis is equivalent to the axiom of choice.*

Even though the axiom of choice is widely used, it is known to be equivalent to statements that sound downright crazy. The most famous of these may be the following.

Theorem 6.5 (Banach-Tarski paradox). *The axiom of choice is equivalent to the statement that one can cut a solid ball of radius 1 in three-dimensional space into finitely many pieces (five are enough), and re-assemble them, using translations and rotations only, into two solid balls, each of radius 1.*

You will think that's impossible, because the total volume of the pieces is either $\frac{4}{3}\pi$ or $\frac{8}{3}\pi$, but not both. But these are very complicated pieces, too complicated to ascribe to them a *volume* in any meaningful way. They are what one calls *not measurable.*

All that notwithstanding, mathematics often relies on the axiom of choice, and therefore the statement that every vector space has a basis is considered true.

6.7 How Hamel's 1905 result answers the question about additivity and linearity

Think of the vector space $V = \mathbb{R}$ over the field $F = \mathbb{Q}$. It has a basis. So there is a collection $(b_i)_{i \in I}$ of real numbers with the property that every real number x has exactly one representation in the form

$$x = \sum_i c_i b_i,$$

where the sum is finite and the c_i are rational. Don't ask me to give you an example of such a basis:

Theorem 6.6. *The statement that the vector space $\mathbb{R}$ over the field $\mathbb{Q}$ has a basis is equivalent to the axiom of choice.*

Equipped with all of that, we can answer the remaining question about Cauchy's 1821 result. This question, in fact, was the aim and motivation of Hamel's 1905 paper.

Theorem 6.7 (Hamel, 1905). *Let $(b_i)_{i\in I}$ be a basis of $\mathbb{R}$ over $\mathbb{Q}$. Then an additive function $f : \mathbb{R} \to \mathbb{R}$ is obtained by defining the $f(b_i)$ any way we wish, then setting*

$$f\left(\sum_{i\in I} c_i b_i\right) = \sum_{i\in I} c_i f(b_i) \tag{6.1}$$

for any choice of rational c_i, at most finitely many of which are non-zero. All additive functions are obtained in this way.

Notice that we would have to define $f(b_i) = rb_i$ for a fixed constant $r \in \mathbb{R}$ to get a *linear* function. It takes a very special choice of the $f(b_i)$ for f to become linear. Most additive functions are non-linear.

Proof. It is clear that (6.1) must hold if f is additive, since additivity implies the constant factor rule for rational factors. Conversely, assume now that (6.1) holds. Let

$$x = \sum_i c_i b_i \quad \text{and} \quad y = \sum_i d_i b_i$$

be real numbers, where the c_i and d_i are rational, and at most finitely many of them are non-zero. Then

$$f(x+y) = f\left(\sum_i (c_i + d_i) b_i\right) = \sum_i (c_i + d_i) f(b_i) = \sum_i c_i f(b_i) + \sum_i d_i f(b_i) =$$

$$f\left(\sum_i c_i b_i\right) + f\left(\sum_i d_i b_i\right) = f(x) + f(y).$$

We used (6.1) three times here. I am sure you can see where. □

Definition 6.8. *An additive function that is not linear, and therefore (by Cauchy's 1821 theorem) not continuous, is called* wild.

Read the following theorem only if you know what "Lebesgue-measurable" means.

Theorem 6.9. *Wild additive functions are never Lebesgue-measurable. Therefore, in Cauchy's 1821 result, "continuous" can be replaced by "Lebesgue-measurable".*

Theorem 6.10. *The statement that wild additive functions exist is equivalent to the axiom of choice.*

You can make them disappear altogether by refusing to accept the axiom of choice. That also removes the disturbing Banach-Tarski paradox. Sadly, it removes large parts of mathematics.

6.8 The graph of a wild additive function is dense

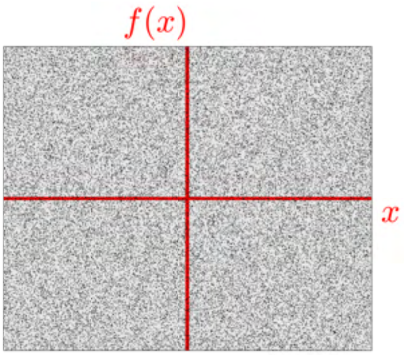
To emphasize the point that wild additive functions are very strange constructs, we prove the following theorem.

Theorem 6.11. *Let $f : \mathbb{R} \to \mathbb{R}$ be a wild additive function. Then the graph of f, i.e., the set*

$$\{(x, f(x)) \ : \ x \in \mathbb{R}\},$$

is dense in $\mathbb{R}^2$.

This means that for every point $(x_*, y_*) \in \mathbb{R}^2$, there is a sequence of points (x_n, y_n) with $y_n = f(x_n)$ and $x_n \to x_*$, $y_n \to y_*$. The graph of such a function might be drawn like this:

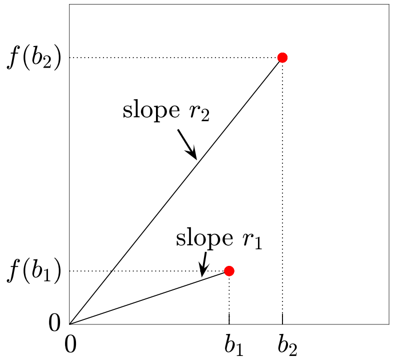


(The gray stuff is the graph.)

Proof. Let $(b_i)_{i\in I}$ be a basis of the vector space $\mathbb{R}$ over the field $\mathbb{Q}$, and let b_1 and b_2 be two of the b_i with

$$f(b_1) = r_1 b_1, \quad f(b_2) = r_2 b_2, \qquad r_1, r_2 \in \mathbb{R}, \qquad r_1 \neq r_2.$$

Because f is not linear, b_1 and b_2 exist.



Since $(b_1, f(b_1))$ and $(b_2, f(b_2))$ are linearly independent, every $(x_*, y_*) \in \mathbb{R}$ can be written as a linear combination

$$(x_*, y_*) = c_*(b_1, f(b_1)) + d_*(b_2, f(b_2)).$$

Pick a sequence of rational numbers $\{c_n\}$ with $c_n \to c_ast$, and a sequence of rational numbers $\{d_n\}$ with $d_n \to d_*$. Then

$$c_n(b_1, f(b_1)) + d_n(b_2, f(b_2)) \to (x_*, y_*),$$

and

$$c_n(b_1, f(b_1)) + d_n(b_2, f(b_2)) = (c_n b_1 + d_n b_2, f(c_n b_1 + d_n b_2))$$

lies on the graph of f. □

6.9 Briefly back to exponentials and logarithms

Additive functions that aren't linear are truly wild: Not Lebesgue-measurable (if you know what that means), and their graph is dense. We pointed out earlier that the question about additivity and linearity is equivalent to questions about exponentials and logarithms. I'll just very briefly say here that the preceding discussion immediately implies that functions that satisfy the law of exponentials are exponentials, or otherwise they aren't Lebesgue-measurable and their graph is dense. The same holds for logarithms.

Exercises

6.1. Can you give an example of an infinite-dimensional vector space for which you can easily and explicitly describe a basis?

6.2. Let $T > 0$ be a random time. (Think of the time it takes between your arrival at the green line platform and the departure of the train.) We say that T *lacks memory* or *does not age* if for all $t > 0$ and $s > 0$,

$$P(T > t + s \mid T > s) = P(T > t).$$

To make sure that the conditional probability is well-defined, assuming $P(T > s) > 0$ for all $s > 0$. Prove that there exists a $\lambda > 0$ so that $P(T > t) = e^{-\lambda t}$ for all $t > 0$. Do you need any assumption?

6.3. Most random quantities $T > 0$ do not have the *lack of memory* of *lack of aging* property described in Problem 2. In fact, the inequalities can go both ways here. Explain the following examples. The life span of your refrigerator, in years, probably satisfies $P(T > 15 \mid T > 10) < P(T > 5)$. The life span of a human, in years, certainly satisfies $P(T > 120 \mid T > 70) < P(T > 50)$, but quite possibly also $P(T > 2 \mid T > 1) > P(T > 1)$. The time that a tenured professor stays at the same university, in years, might well satisfy $P(T > 12 \mid T > 10) > P(T > 2)$. The time you must wait for the subway, in minutes, likely satisfies $P(T > 45 \mid T > 30) > P(T > 15)$.

6.4. In Problem 3 of Chapter 5, we wondered whether a decreasing function $g(x)$, $x \in (0, 1)$, with

$$g(xy) = g(x) + g(y)$$

for all x and y in $(0, 1)$ must automatically be a constant multiple of a logarithm. Explain why the answer is yes, and you do *not* need to impose *any* additional assumption whatsoever.

6.5. We discussed functions $f :\ \mathbb{R} \to \mathbb{R}$ in this chapter. But in Linear Algebra, the focus is on mappings

$$L :\ \mathbb{R}^n \to \mathbb{R}^m.$$

For such mappings, is it true that the sum rule implies linearity? Does the reasoning in this chapter extend?

6.6. Let V and W be infinite-dimensional real vector spaces, equipped with norms.[9] Let

$$L :\ V \to W$$

be an additive mapping. In the spirit of Cauchy's 1821 result, which simple extra assumption on L guarantees that L is linear? "Continuity" isn't a good answer anymore, since linear maps between infinite dimensional vector spaces don't need to be continuous.

[9] If you don't know what "norm" means, you may want to skip this question.

6.7. Suppose that the "utility" you derive from x gadgets is $U(x)$. (Economists talk about "utility".) Suppose $U(x) > 0$ for all $x > 0$, and suppose

$$U(x+y) = U(x) + U(y) \quad \text{for } x, y > 0,$$

Can you deduce that $U(x) = cx$ for some c? Notice that you know the additivity condition only for positive x and y. Does it matter? Do you need any other assumption about U?

6.8. The *intensity* I of a sound has a physical definition. It is the rate of energy transfer by the sound wave per unit area, but we don't need to worry about that here. It is believed that doubling I is perceived by people as an increase in loudness by a fixed amount. (Whatever that really means.) So if we denote by $L(I)$ the perceived loudness,

$$L(2I) = L(I) + c \tag{6.2}$$

for some constant c. For a given $c > 0$, what are all the functions

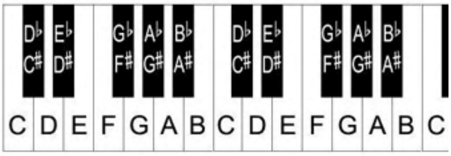
$$L : \ (0, \infty) \to \mathbb{R}$$

that satisfy (6.2)?

Chapter 7

Why do we divide the octave into twelve steps?

Here is what a piano keyboard looks like:



The octave (from C to C, for instance) is divided into 12 half steps. Go **here** to listen to what they sound like. The question in this chapter is why 12? Is that cultural, or is there physics behind that?

7.1 Musical tones and Fourier expansions

7.1.1 Musical tones

We perceive rapid oscillations in air pressure through our ears. We hear oscillations between approximately 20 Hz and 20,000 Hz. (Hz = Hertz = 1/second.) By contrast, a cat can hear between 50 Hz and 85,000 Hz. Medical ultrasound, which we do not hear, is in the millions of Hz.

To understand musical tones, one must therefore think about oscillatory functions.

7.1.2 Oscillatory functions, periods, and frequencies

We will think about functions

$$f: \mathbb{R} \to \mathbb{R}$$

here, and always denote the independent variable by t, thinking of it as time.

Definition 7.1. *Let $T > 0$. A function*

$$f: \mathbb{R} \to \mathbb{R}$$

has period T *if*

$$f(t+T) = f(t) \quad \textit{for all } t. \tag{7.1}$$

We say that f has minimal period T *if* (7.1) *holds, and there is no smaller positive T for which* (7.1) *also holds. If the minimal period of f is T, then the* frequency of f *is*

$$\alpha = \frac{1}{T}.$$

I find it useful to distinguish between "f has period T" and "f has minimal period T", even though this is not common terminology. (When people say "f has period T" they may or may not mean "f has minimal period T".) By definition, the frequency α is the reciprocal of the minimal period; I don't refer to it as the "maximal frequency".

7.1.3 Cosine oscillations

Let $T > 0$, $a \geq 0$, and $\theta \in [0, 1)$. The period of the function

$$a \cos\left(2\pi\left(\frac{t}{T} + \theta\right)\right) \tag{7.2}$$

equals T. Formula (7.2) gives us, for a fixed $T > 0$, a two-parameter family of periodic functions with period T: One for each $a \geq 0$ and $\theta \in [0, 1)$. We call a the *amplitude*, and θ the *phase shift*. If $a > 0$, then T is the minimal frequency.

We could allow $\theta \in \mathbb{R}$, but nothing would be gained, since (7.2) remains the same if an integer is added to or subtracted from θ. We could also allow $a < 0$, but again nothing would be gained, since the sign of (7.2) changes when θ is raised or lowered by $1/2$. We could also replace cos by sin in (7.2), but again nothing would be gained, since

$$a \sin\left(2\pi\left(\frac{t}{T} + \theta\right)\right) = a \cos\left(2\pi\left(\frac{t}{T} + \theta - \frac{1}{4}\right)\right)$$

Starting with (7.2), we can construct other functions with period T:

$$a \cos\left(2\pi k\left(\frac{t}{T} + \theta\right)\right),$$

where $k > 0$ is an integer, has minimal period T/k, and therefore also has period T. Finally we can take linear combinations of functions of this form:

$$\sum_k a_k \cos\left(2\pi k \left(\frac{t}{T} + \theta\right)\right).$$

7.1.4 Fourier series

Theorem 7.2 (Joseph Fourier, early 1800s). *Let $T > 0$. If*

$$f:\ \mathbb{R} \to \mathbb{R}$$

is a function with period T, then f can be expanded in the form

$$f(t) = \sum_{k=0}^{\infty} a_k \cos\left(2\pi k \left(\frac{t}{T} + \theta_k\right)\right) \tag{7.3}$$

where $a_0 \in \mathbb{R}$, $a_k \geq 0$ for $k \geq 1$, $\theta_k \in [0, 1)$ for all k. The a_k with $k \geq 0$ and the θ_k with $k \geq 1$ are uniquely determined. (Obviously the value of θ_0 doesn't matter, and can be chosen any way you like.)

Note that (7.3) can be written in terms of the frequency $\alpha = 1/T$ as

$$f(t) = \sum_{k=0}^{\infty} a_k \cos(2\pi k(\alpha t + \theta_k)).$$

Fourier's expansion of periodic functions into trigonometric functions was initially controversial. However, in 1841 (after Fourier had died), a paper appeared in the *Cambridge Mathematical Journal*, under the pseudonym "P. Q. R.", titled "On Fourier's expansions of functions in trigonometric series", defending Fourier's theory. P. Q. R. was 17-year-old William Thomson (1824–1907). In the 1850s, Thomson analyzed the physics of the first transatlantic cable. In 1891, Thomson became Lord Kelvin, recognized as one of the greatest physicists of his time. In the late 1940s, Kelvin's cable theory became a central component of the Hodgkin-Huxley theory of nerve cells, which underlies most theoretical and computational neuroscience nowadays.

7.1.5 Fundamental tone and overtones

Think of pressure oscillating periodically. In (7.3), the mean is a_0. We are only interested in fluctuations around the mean, so we will leave out the term with $k = 0$, and consider

$$\sum_{k=1}^{\infty} a_k \cos\left(2\pi k \left(\frac{t}{T} + \theta_k\right)\right)$$

with $a_k \geq 0$ and $\theta_k \in [0, 1)$. The function

$$\cos\left(2\pi k\left(\frac{t}{T} + \theta_k\right)\right)$$

has minimal period T/k. So the Fourier series represents the tone as a weighted combination of tones with periods T, $T/2$, $T/3$, and so on. Writing $\alpha = 1/T$, the frequencies are α, 2α, 3α, ... The tone that corresponds to

$$a_1 \cos\left(2\pi\left(\frac{t}{T} + \theta_1\right)\right)$$

is called the *fundamental tone*, and the tones corresponding to

$$a_k \cos\left(2\pi k\left(\frac{t}{T} + \theta_k\right)\right)$$

with $k \geq 2$ are called the *overtones*.

7.1.6 Pitch and timbre

When you play a note on a musical instrument, typically you hear a mixture of a fundamental tone at frequency α, and overtones at frequencies 2α, 3α, and so on. The fundamental tone determines what we call the *pitch*. Higher frequency corresponds to higher pitch. The overtones determine what we call the *timbre*, the character of the sound. A trumpet and a violin differ in the overtone mixtures that they generate.

7.1.7 Intervals

Go **here** and click on the lowest C, then the E above it. (These are called C_3 and E_3 in music.) The frequency of C_3 is 131 Hz (this is, to be precise, the frequency of the fundamental tone), and that of E_3 is 165 Hz. Now play the A above that (that's called A_3), and the $C^{\#}$ above it (that's called $C_4^{\#}$). You may agree that you are hearing the "same interval". What you are hearing, however, is 220 Hz (that's A_3) and 277 Hz (that's $C_4^{\#}$). The frequency differences aren't the same:

$$165 - 131 \neq 277 - 220.$$

In what sense are the intervals "the same"? The answer is that the ratios are the same (up to rounding-caused errors):

$$\frac{165}{131} \approx \frac{277}{220}.$$

We hear two intervals as "the same" when the frequency *ratios* are the same. You can play "Twinkle-twinkle little star" beginning on two different keys, and what you will hear is "the same melody", but all the frequency differences between subsequent notes are different. It's "the same melody" because the frequency *ratios* between subsequent notes are the same, or in other words, the differences between the *logarithms* of frequencies. We hear frequency ratios, not frequency differences. The reason lies in the physiology of the inner ear. I will not try to study that here.

7.1.8 The intervals between overtones

The frequency of the first overtone is twice that of the fundamental tone, so the frequency ratio is 2 : 1. The interval that corresponds to that ratio is called an *octave.* The word has to do with *eight.* To see why, go **here**, and play the white keys, starting at the lowest key, counting "1" on the lowest key, which is a C, then "2" on the next key, which is a D, and so on. When you get to "8", you are back to a note called C. That's an octave higher.

The frequency of the second overtone is 3 times higher than that of the fundamental tome, so between the first and second overtone, there is a frequency ratio of 3:2. Go **here** and play the C at the left end, and the first G to its right. That's a 3:2 frequency ratio. It is called a "fifth". The reason is the same as before — count white keys from the leftmost C to the G. At "5" you arrive on G.

The next overtone interval, between the second and third overtones, is characterized by a frequency ratio of 4:3. That's the interval because the low C and the next-higher F. It's called a *fourth.* The next interval has frequency ratio 5:4, and it's the interval between low C and the next-higher E, called a *major third.* The next interval has frequency ratio 6:5, and it's the interval between low C and $E^\flat$. (So for the first time, you need to use a black key.) That's called a *minor third.*

The intervals between overtones become smaller and smaller, and beyond the minor third, not all of them have names anymore. It should be said that what you hear on the piano are not *exactly* the frequency ratios 3:2, 4:3, 5:4, 6:5. If they were exactly those ratios, one would call the interval the *pure fifth*, the *pure fourth*, the *pure major third*, and the *pure minor third.* The frequency ratio of 9:8 is called the *pure major second*; it is the interval between C and the D immediately to its right. The frequency ratio of 16:15 is the *pure minor second*, the interval between C and the $C^\#$ immediately to its right. For reasons to be explained later, you cannot play any pure interval on a modern piano.

7.2 Playing two tones together

If we identify a tone with a Fourier series

$$\sum_{k=1}^{\infty} a_k \cos\left(2k\pi(\alpha t + \theta_k)\right).$$

then playing two tones together gives rise to a sum of the form

$$\sum_{k=1}^{\infty} a_k \cos\left(2k\pi(\alpha t + \theta_k)\right) + \sum_{k=1}^{\infty} b_k \cos\left(2k\pi(\beta t + \eta_k)\right).$$

This is too complicated, so I am going to focus only on the sum of the fundamental tones, and ignore the phase shifts and the amplitudes:

$$\cos(2\pi\alpha t) + \cos(2\pi\beta t).$$

We assume of course $\alpha \neq \beta$, and might then as well assume $\alpha < \beta$.

Many questions are left unanswered here: What are the effects of amplitudes, phase shifts, and overtones?

7.2.1 Period of the sum of two pure cosine oscillations

Proposition 7.3. *Let $0 < \alpha < \beta$. The sum*

$$\cos(2\pi\alpha t) + \cos(2\pi\beta t). \tag{7.4}$$

is periodic if and only if the ratio $\frac{\beta}{\alpha}$ is rational. If

$$\frac{\beta}{\alpha} = \frac{q}{p}$$

with $0 < p < q$, p and q relatively prime, then the frequency of (7.4) is α/p.

Proof. The sum (7.4) has period T if

$$\cos(2\pi\alpha(t+T)) + \cos(2\pi\beta(t+T)) = \cos(2\pi\alpha t) + \cos(2\pi\beta t) \tag{7.5}$$

for all t . Let $k \geq 1$ be a positive integer multiple of 4. Differentiate (7.5) k times:

$$(2\pi\alpha)^k \cos(2\pi\alpha(t+T)) + (2\pi\beta)^k \cos(2\pi\beta(t+T)) =$$

$$(2\pi\alpha)^k \cos(2\pi\alpha t) + (2\pi\beta)^k \cos(2\pi\beta t).$$

In the limit as $k \to \infty$, only the terms involving β survive, since $\beta > \alpha$, so $\cos(2\pi\beta(t+T)) = \cos(2\pi\beta t)$ for all t. We conclude that (7.5) is equivalent to

$$\cos(2\pi\alpha(t+T)) = \cos(2\pi\alpha t) \quad \text{and} \quad \cos(2\pi\beta(t+T)) = \cos(2\pi\beta t)$$

for all t. This means

$$\alpha T = p \quad \text{and} \quad \beta T = q \tag{7.6}$$

for integers p and q with $0 < p < q$. Therefore

$$\frac{\alpha}{\beta} = \frac{p}{q} \tag{7.7}$$

is rational. Conversely, (7.7) implies (7.6) with

$$T = \frac{p}{\alpha} = \frac{q}{\beta}.$$

When p and q are relatively prime, then p and q is minimal, therefore T is minimal, and therefore its reciprocal, α/p, is the frequency of (7.4). □

7.2.2 An experiment about pitch

Go here. Play a tone at 131 Hz. (In music, that's called C$_3$.) That's the first overtone when the fundamental tone is 65.5 Hz (called C$_2$). But we aren't playing the fundamental tone here. Add now (using a different window of the browser) the next overtone of 65.5 Hz, which is $3 \cdot 65.5\,\text{Hz} = 196.5\,\text{Hz} = 393/2\,\text{Hz}$.

Which pitch do you hear? If you are like me, you believe to hear the pitch drop by an octave. The experiment doesn't work on everybody, and it may not work on you. But I very distinctly hear the pitch drop down by an octave — even though a *higher* note was added, not a lower one.

Why might this happen? The answer is clear from Proposition 7.3: The frequency of

$$\cos(2\pi\alpha t) + \cos\left(2\pi\,\frac{3}{2}\alpha\,t\right)$$

is not α, but $\alpha/2$.

7.2.3 Waveforms of the intervals of the overtone sequence

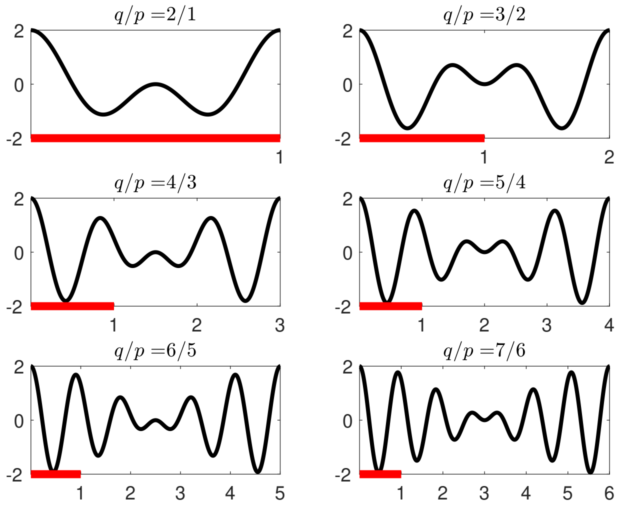
I will assume now that $\alpha = 1$. This is just a matter of choosing the time unit. If α were 440 Hz, then it is 1 if your time unit is $\frac{1}{440}$ seconds. Therefore to understand the graph of

$$\cos(2\pi\alpha t) + \cos\left(2\pi\,\frac{q}{p}\,\alpha t\right),$$

we really just have to consider

$$\cos(2\pi t) + \cos\left(2\pi\,\frac{q}{p}\,t\right). \tag{7.8}$$

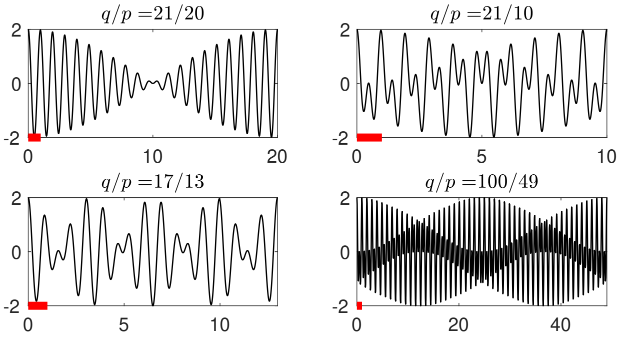
We'll first plot this function for $\frac{q}{p} = \frac{2}{1}, \frac{3}{2}, \frac{4}{3}, \frac{5}{4}, \frac{6}{5}, \frac{7}{6}$. These are the intervals called the octave and the (pure) fifth, fourth, major third, minor third, and the (much less well-known) "subminor third".



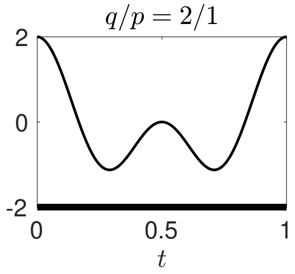
The red bar always indicates 1, the period of the lower of the two notes that are being combined. Going up in the overtone sequence, the periods become longer and more complex.

7.2.4 Waveforms of other intervals

A few other interesting examples are shown in the next plot.



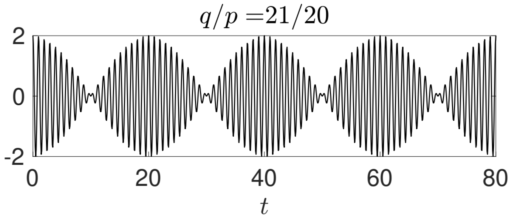
In the 21:20 interval, you see what is called the *beats* phenomenon — periodic fluctuations in amplitude. You can see the beginnings of this phenomenon much earlier in the overtone interval sequence already, strictly speaking even in the octave:



In the 21:10 and 100:49 intervals, you see a different kind of beat phenomenon.

7.2.5 Beats arising from the interaction of nearly equal frequencies

Here is a plot of the 21:20 interval again, but over several periods:



To understand what is going on here, assume $0 < \alpha < \beta$. Then

$$\begin{aligned} & \cos(2\pi\alpha t) + \cos(2\pi\beta t) \\ = \quad & \cos\left(2\pi\left(\frac{\beta+\alpha}{2} - \frac{\beta-\alpha}{2}\right)t\right) + \cos\left(2\pi\left(\frac{\beta+\alpha}{2} + \frac{\beta-\alpha}{2}\right)t\right) \end{aligned} \tag{7.9}$$

Using $\cos(x \pm y) = \cos x \cos y \mp \sin x \sin y$, (7.9) equals

$$2\cos\left(2\pi\left(\frac{\beta+\alpha}{2}\right)t\right)\cos\left(2\pi\left(\frac{\beta-\alpha}{2}\right)t\right) \tag{7.10}$$

The two cosine factors in (7.10) have the minimal periods

$$\frac{2}{\beta+\alpha} \quad \text{and} \quad \frac{2}{\beta-\alpha}.$$

Therefore (7.10) can be viewed as an oscillation at the "fast" frequency

$$\frac{\beta+\alpha}{2}$$

modulated in an oscillatory manner with the "slow" frquency

$$\frac{\beta-\alpha}{2}.$$

The duration of a beat is in fact not $\frac{2}{\beta-\alpha}$ but $\frac{1}{\beta-\alpha}$, since it is not the period of

$$\cos\left(2\pi\left(\frac{\beta-\alpha}{2}\right)t\right)$$

but the time between two instances in which this function equals zero.

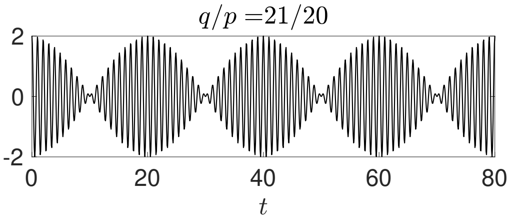
We arrive at the following result.

Proposition 7.4. *Let p and q be natural numbers, relatively prime, with $q > p$. The sum*

$$\cos(2\pi t) + \cos\left(2\pi\,\frac{q}{p}\,t\right)$$

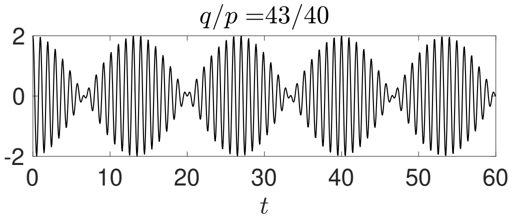
has the exact period p. However, for $p \approx q$, it approximately oscillates at frequency 1, with modulation into "beats" of duration $\frac{p}{q-p}$.

In the picture at the start of this section:



we have beats of duration 20, and oscillations within each beat at frequency approximately 1. The overall period of the oscillation is 20, precisely the duration of a beat. To hear the 21:20 interval, go **here** and play a note at 440 Hz, and another at 462 Hz. Then gradually reduce 462 to 444; you will hear the beats slow down.

Another example:



Here we have beats of duration $\frac{40}{3}$, while the period is 40 (3 beats). You can see that the period is 3 beats if you look very carefully at what is happening in between the beats. The frequency within each beat is, again, approximately 1.

7.3 Consonance and dissonance

7.3.1 The octave, fifth, fourth, major third, minor third

We will begin by considering the 2:1, 3:2, 4:3, 5:4, and 6:5 intervals — the octave, and the pure fifth, fourth, major third, and minor third. Go **here** to listen what they sound like. All of them sound "harmonious", or "consonant". Compare that with a "pure major second" (a 9:8 frequency ratio, for instance 440 Hz and 495 Hz). You may hear what people mean when they say that the major second is "dissonant", whereas the fifth, fourth, major third, and minor third are "consonant". Whether these words have a precise meaning is another matter; it isn't very clear to me.

7.3.2 Later intervals in the overtone sequence

I mentioned the major second already, the 9:8 frequency ratio. The ratio 16:15 is called the *minor second*. Try that one out **here**, for instance using one tone at 450 and one at 480 Hz. Again, it sounds "dissonant", and this time you hear the beats associated with two almost equal frequencies. Earlier intervals in the overtone sequence sound "consonant", later ones sound "dissonant".

7.3.3 What is dissonance?

It seems like a reasonable guess that "dissonance" means a complicated wave form, and that is typically associated with large p. (Remember we consider the summation of tones with frequencies $\alpha < \beta$, assuming $\frac{\alpha}{\beta} = \frac{p}{q}$, where $p, q \in \mathbb{N}$, $p < q$, and p and q are relatively prime.)

This isn't quite right. A 7:3 frequency ratio (so $p = 3$) sounds somewhat dissonant, whereas 100:47 (so $p = 47$) sounds quite consonant, because it is sufficiently close to 100:50=1:2, the octave. So perhaps a frequency ratio of $p : q$ sounds dissonant if it cannot be approximated well by a ratio $p_0 : q_0$ where $0 < p_0 < q_0$ are integers and p_0 is small.

7.3.4 Dissonance in the Indonesian classical tradition

Go **here** to hear gamelan music, the traditional music of Indonesia. Gamelan has ancient roots, but is still alive in Indonesia today. All intervals in gamelan music, except for the octave, are not the ones used in European-derived music, and they sound "dissonant" to people who grew up with European-derived music. Here is the gamelan group at Tufts University:

Tufts University Department of Music

7.3.5 The emancipation of the dissonance in the European classical tradition

In European classical music, dissonance became increasingly acceptable over the centuries. Composers as early as J. S. Bach made extensive use of it. In the late 19th century more and more "dissonant" sounds appeared in European classical music, culminating in the early twentieth century with the Arnold Schönberg, who did away with the distinction between consonance and dissonance altogether. Go **here** for one of his "12-tone" compositions. (I enjoy Schönberg's music, and am particularly fond of opus 25, which the link will take you to.)

He called it the "emancipation of the dissonance". Here he is, after having fled from Nazi-ruled Europe to Los Angeles.

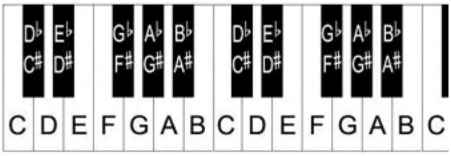
by George Platt Lynes, 1947, Metropolitan Museum of Art
Arnold Schoenberg, Hollywood

7.4 Equal temperament tuning

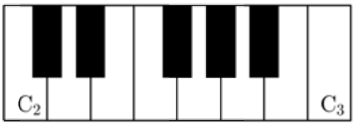
7.4.1 n-tone equal temperament (n-TET)

In order to be able to play “the same” melody starting on any key of the piano, the piano is nowadays tuned so that the frequency ratio between two subsequent notes is always the same. This is called *equal temperament tuning.* In earlier centuries, European classical music used different tuning systems, which I won’t discuss here.[10]

Almost all modern western music of all genres divides the octave into 12 equal steps. The frequency ratio between C and $C^{\#} = D^{\flat}$ is $2^{1/12} : 1$, so is the frequency ratio between $C^{\#} = D^{\flat}$ and D, and so on.

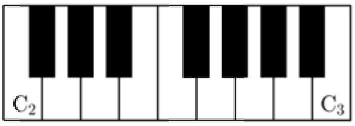


But one could divide the octave into a different number, n, of equal steps. This is called n-tone equal temperament tuning, or briefly n-TET. The traditional keyboard



12 tones per octave

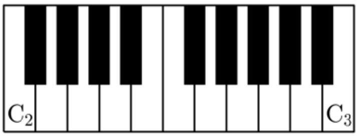
could be replaced by



15 tones per octave

or perhaps by

[10] For instance, J. S. Bach’s *Well-Tempered Clavier* was intended for a keyboard instrument that did *not* have equal temperament tuning. Unlike earlier tuning systems, the tuning system that Bach had in mind did allow playing in all keys. The collection of pieces in the *Well-Tempered Clavier* demonstrated that. But different keys sounded (slightly) different, whereas they sound the same in the modern equal temperament system.



19 tones per octave

for instance.

I will first explain why there is no n for which n-TET allows you to play "pure" overtone intervals.

7.4.2 No pure overtone interval other than an octave is ever playable in any equal temperament tuning

Proposition 7.5. *Let $n \geq 3$ and $j \geq 2$ be integers, and let k be an integer with $1 \leq k \leq n$. Then*

$$\frac{j+1}{j} \neq 2^{k/n}. \tag{7.11}$$

Note that the $2^{k/n}$ are the frequency ratio that appear on a keyboard in which the octave is divided into n equal (by frequency ratio) steps, and the $\frac{j+1}{j}$ are the frequency ratios between overtones. For $j = 2, 3, 4, 5$, we get the pure fifth, fourth, major third, and minor third, the most important "consonant" intervals in the European-derived music tradition.

Proof. If (7.11) were to hold, then

$$(j+1)^n = 2^k j^n. \tag{7.12}$$

If j is odd, then the left-hand side of (7.12) has at least n factors of 2 in it, and the right-hand side has $k < n$ factors of 2 in it. If j is even, then the left-hand side of (7.12) has no factors of 2 in it, whereas the right-hand side has at least $n+k$ such factors. So j can be neither even nor odd, and this contradiction proves the assertion. □

7.4.3 If we care about consonance, $n = 12$ is good, but $n = 19$ may be better

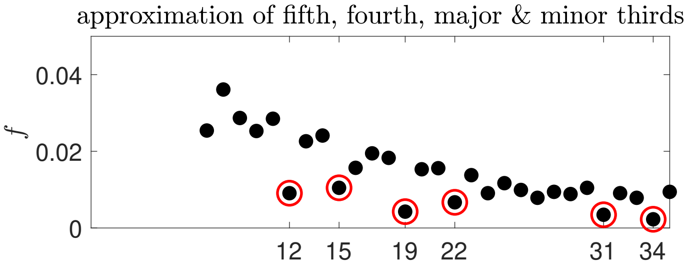
For $\gamma \in (1, 2)$, I define

$$\phi(\gamma, n) = \min\left\{\left|\frac{\gamma}{2^{k/n}} - 1\right| \ : \ 1 \leq k \leq n-1\right\}$$

So $\phi(\gamma, n)$ measures how well a frequency ratio of γ is represented in n-TET. Then I define

$$f(n) = \max\left\{\phi\left(\frac{j+1}{j}, n\right) \ : \ 2 \leq j \leq 5\right\}.$$

So $f(n)$ measures how well the pure fifth, fourth, major third, and minor third are represented in n-TET. Now let's plot f as a function of n:

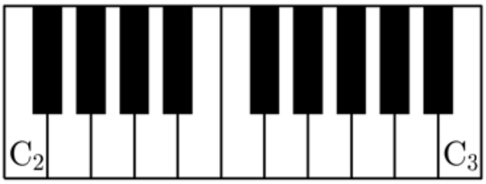


If you like fifths, fourths, major and minor thirds, then small $f(n)$ is good. Therefore $n = 12$ is the smallest exceptionally good value. If you want to do better, you have to go all the way to $n = 19$. The following table shows more detail about the comparison between $n = 12$ and $n = 19$.

	pure	12-TET	19-TET
fifth	1.5	$2^{7/12} = 1.4983$	$2^{11/19} = 1.4938$
fourth	1.3333	$2^{5/12} = 1.3348$	$2^{8/19} = 1.3389$
major third	1.25	$2^{4/12} = 1.2599$	$2^{6/19} = 1.2447$
minor third	1.2	$2^{3/12} = 1.1892$	$2^{5/19} = 1.2001$

So 12-TET is actually better than 19-TET for pure fifths and pure fourths, but worse for the major thirds, and much worse for the minor thirds.

Go **here** for music in 19-TET.



7.4.4 The emancipation of the dissonance makes other n possible

The choice $n = 12$ is motivated by wanting to represent the "consonant" fifth, fourth, major and minor thirds as well as possible in equal temperament, while not using an excessively large n.

If that's not important, then you can use other values of n. Gamelan uses two scales. The older one is called *slendro*. It divides the octave into 5 steps. Some authors claim that slendro is close to 5-TET, others dispute this. The second, newer gamelan scale is called *pelog*. It divides the octave into seven unequal steps. Some authors have claimed that the intervals in pelog are approximately ones (though not all) that would appear on an instrument tuned in 9-TET. This, too, is controversial.

People have experimented with many other values of n. Click **here** for a 22-TET piano piece which I particularly enjoyed. **Here** you can see a pianist playing a 31-TET piano piece on a microtonal keyboard.

Exercises

7.1. (no math) The note that is called A4 in music has a frequency of 440 Hz. Some people claim that it's better for your health to hear music where the A4 is tuned to 432 Hz. Do you believe it? Check online whether there is any sensible-sounding evidence for this claim. (That, of course, is largely a matter of opinion.)

7.2. (easy) Let $\theta \in [0,1)$. Let

$$f(t) = \cos(6\pi(t+\theta)).$$

What is the frequency of f?

7.3. (easy) Here is a simple example of a Fourier "series" with only three terms:

$$f(t) = a_0 + a_1 \cos\left(2\pi\left(\frac{t}{T} + \theta_1\right)\right) + a_2 \cos\left(4\pi\left(\frac{t}{T} + \theta_2\right)\right)$$

The period is $T > 0$. Explain why

$$a_0 = \frac{1}{T}\int_0^T f(t)dt.$$

7.4. (easy after doing the preceding exercise) Explain why always

$$a_0 = \frac{1}{T}\int_0^T f(t)dt$$

in (7.3). Integrate both sides of (7.3) to do this. Don't worry about why it's legal to integrate the series term by term.

7.5. (easy) Let $\alpha > 0$ and $\theta \in [0,1)$. Let

$$f(t) = \cos(\alpha\pi(t+\theta)).$$

Write $f(t)$ in the form

$$f(t) = c\cos(\alpha\pi t) + d\sin(\alpha\pi t).$$

(You do need to remember here that $\cos(x+y) = \cos x \cos y - \sin x \sin y$.)

7.6. (still fairly easy) Suppose you simultaneously play a note at 440 Hz (the note that's called A4 in music), and one at 441 Hz (the A4 that the Boston Symphony Orchestra tunes to). You will get beats. How many beats per second? Hints: You need to remember eq. (7.10) here, which is

$$\cos(2\pi\alpha t) + \cos(2\pi\beta t) = 2\cos\left(2\pi\left(\frac{\beta+\alpha}{2}\right)t\right)\cos\left(2\pi\left(\frac{\beta-\alpha}{2}\right)t\right).$$

The second cosine factor is the slow one that shapes the beats. A beat ends and the next one starts when this factor is zero.

7.7. (medium) The Fourier expansion (7.3) is usually written like this:

$$f(t) = a_0 + \sum_{k=1}^{\infty} \left(\alpha_k \cos\left(2\pi k \frac{t}{T}\right) + \beta_k \sin\left(2\pi k \frac{t}{T}\right)\right),$$

for constants a_0 and α_k, β_k, $k \geq 1$, without any sign constraints. Explain why this is precisely equivalent to (7.3).

7.8. (medium) Remember eq. (7.10):

$$\cos(2\pi\alpha t) + \cos(2\pi\beta t) = 2\cos\left(2\pi\left(\frac{\beta+\alpha}{2}\right)t\right)\cos\left(2\pi\left(\frac{\beta-\alpha}{2}\right)t\right).$$

What happens if you introduce a phase shift into one of the two oscillations that are added together here? So suppose $\theta \in [0, 1)$. What is the formula for

$$\cos(2\pi(\alpha t + \theta)) + \cos(2\pi\beta t)$$

analogous to (7.10)?

7.9. (medium) Let's think about the octave: A fundamental tone, and a tone at twice the frequency. Let's model playing the two tones together by studying the sum

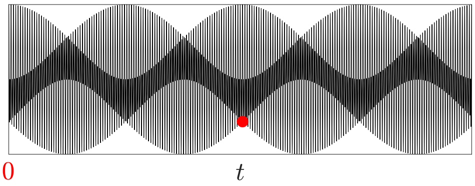
$$f(t) = \cos(2\pi t) + \cos(4\pi t).$$

(This is a tone at frequency 1 and a tone at frequency 2. If you measure frequency in Hz, then 1 and 2 are not audible, but perhaps we are measuring frequency in a different unit, so that these become audible tones.) The maximum value of $f(t)$ is 2, but what is its minimum value? Hint: It is useful to remember that $\cos(2x) = \cos^2 x - \sin^2 x = \cos^2 x - (1 - \cos^2 x) = 2\cos^2 x - 1$.

7.10. (hard, but it helps to do the preceding problem first) This is about the "other kind" of beats, the kind that you see when you play something that's almost, but not quite, an octave. Below is the graph of the function

$$\cos(2\pi t) + \cos\left(2\pi \, \frac{200}{101} \, t\right).$$

$q/p = 200/101$



Find the approximate coordinates of the red point. You don't need plotting software. All you need is to think about it.

Can you hear these beats? Try it using **this** web page.

Chapter 8

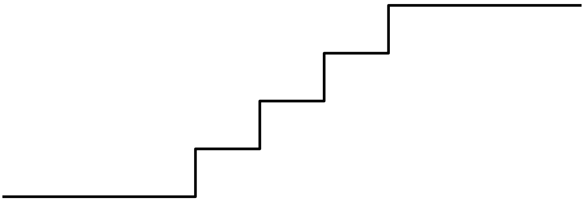
Thinking backward

This chapter is inspired by a beautiful one-page essay by Nick Trefethen, which you can find **here**. The puzzle in Section 8.1.4 comes from David Acheson's book **The Spirit of Mathematics**, which has many wonderful and entertaining examples of elementary and not-quite-so-elementary mathematics. The matrix multiplication example comes from Trefethen's article. Trefethen also mentions backpropagation in the training of neural networks as an example of "starting at the end". Additional examples in this chapter were suggested to me by Eduard Harabetian.

8.1 Puzzles

8.1.1 The number of ways in which you can climb a staircase

Suppose you want to climb a staircase with n steps. Here is the case $n = 4$:



You can take single steps, or double steps (two steps at a time), and you can mix them up. How many different ways of climbing the staircase are there? We denote the number of ways of doing this by A_n. It is quite clear that $A_1 = 1$ and $A_2 = 2$. (When $n = 2$, you can either take two single steps, or one double step.) For $n = 3$, we can easily list all possible schedules explicitly:

$$(1,1,1), \quad (1,2), \quad (2,1).$$

So $A_3 = 3$. This means you either take three single steps, or a single step followed by a double one, or a double step followed by a single one. For $n = 4$, it is still easy to list all possibilities:

$$(1,1,1,1), \quad (1,1,2), \quad (1,2,1), \quad (2,1,1), \quad (2,2).$$

So $A_4 = 5$.

Let's calculate a general formula for A_n. Let k denote the number of times you take double steps: $0 \leq k \leq n/2$. Once we have settled on the value of k, we know that we will take $n-k$ steps in total. We will now have to decide which of those $n-k$ steps are to be double steps, and there are exactly

$$\binom{n-k}{k}$$

was of doing that. So the total number of ways in which we can climb the staircase equals

$$A_n = \sum_{0 \leq k \leq n/2} \binom{n-k}{k}. \tag{8.1}$$

For instance,

$$A_4 = \binom{4}{0} + \binom{3}{1} + \binom{2}{2} = 1 + 3 + 1 = 5,$$

in agreement with what we found by just listing all options for $n = 4$. Or

$$A_5 = \binom{5}{0} + \binom{4}{1} + \binom{3}{2} = 1 + 4 + 3 = 8.$$

We could end here. But eq. (8.1) is not insightful. We can obtain a much more transparent formula by starting at the end. At the very end, you either take a single step, or a double step. How many ways of climbing the staircase ending with a single step are there? Well, first you have to climb $n-1$ steps, and then comes the single step that finishes it off. So the answer is A_{n-1}. Similarly, how many ways of climbing the staircase ending with a double step are there? First you have to climb $n-2$ steps, and then comes the double step that finishes it off. So the answer is A_{n-2}. We conclude:

$$A_n = A_{n-1} + A_{n-2}.$$

Since clearly

$$A_1 = 1, \quad A_2 = 2,$$

we have

$$A_3 = A_1 + A_3 = 3, \quad A_4 = A_2 + A_3 = 5, \quad A_5 = A_3 + A_4 = 8, \quad A_6 = A_4 + A_5 = 13,$$

and so on. These are called the *Fibonacci numbers.* To be precise, the Fibonacci sequence $\{F\}_{n=1,2,\ldots}$ starts with two 1's:

$$F_1 = 1, \quad F_2 = 1,$$

and then follows the recursion

$$F_n = F_{n-1} + F_{n-2}, \tag{8.2}$$

and consequently

$$A_n = F_{n+1}. \tag{8.3}$$

The main point here, and the theme of this chapter, is that it is sometimes useful to think from end to beginning, instead of thinking from beginning to end. But while we are here, I'll say a bit more about the Fibonacci numbers. They are named after an Italian mathematician who lived around the year 1200, and whose name was not Fibonacci. His name was Leonardo; Italians didn't have last names at the time. Later (apparently centuries after his death) he became known as Fibonacci, which stood for "son of the Bonacci family". The Fibonacci numbers, however, had already been known to ancient Indian mathematicians more than 2000 years ago. They were discovered in the context of computing the number of rhythmic patterns that could be formed with one- and two-syllable words. In other words, the stair climbing puzzle, stated in slightly different words, led to the discovery of the Fibonacci numbers.

You can use the recursion formula (8.2) to compute F_n for any n. But there is a quicker way of doing it. First look for sequences $\{F_n\}_n$ that satisfy (8.2), regardless of the condition that $F_1 = F_2 = 1$. Let's try to find sequences of the form

$$F_n = r^n$$

for some number r. Inserting this into (8.2), we find that

$$r^n = r^{n-1} + r^{n-2}$$

for $n \geq 3$, or

$$r^2 = r + 1.$$

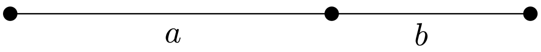
This is a quadratic equation, which has the solutions

$$\frac{1 \pm \sqrt{5}}{2}.$$

The number

$$1 + \sqrt{5}$$

is also kno



$$\frac{a+b}{a} = \frac{a}{b} = \varphi$$

The number

$$\psi = \frac{1-\sqrt{5}}{2} = -0.61803\ldots$$

is also called the *conjugate* of the golden ratio. Notice that $\phi + \psi = 1$, so

$$\psi = 1 - \phi \quad \text{and} \quad \phi = 1 - \psi.$$

Since ϕ and ψ are the solutions of

$$r^2 = r + 1,$$

or equivalently of

$$1 - r = -\frac{1}{r},$$

we also have

$$\psi = -\frac{1}{\phi} \quad \text{and} \quad \phi = -\frac{1}{\psi}.$$

So we now have two sequences that satisfy the recursion formula (8.2):

$$F_n = \phi^n \quad \text{and} \quad F_n = \psi^n.$$

We immediately get infinitely many others:

$$F_n = c\phi^n + d\psi^n,$$

where c and d are constants. We can set c and d so that F_1 and F_2 are anything we like. For instance, if we want $F_1 = F_2 = 1$, we just need to ensure that

$$c\phi + d\psi = 1, \quad c\phi^2 + d\psi^2 = 1.$$

These are two linear equations for c and d. In matrix-vector form:

$$\begin{bmatrix} \phi & \psi \\ \phi^2 & \psi^2 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The matrix is non-singular because $\phi\psi^2 - \psi\phi^2 = \phi\psi(\psi - \phi) \neq 0$. We have

$$\begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} \phi & \psi \\ \phi^2 & \psi^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\phi\psi(\psi-\phi)} \begin{bmatrix} \psi^2 & -\psi \\ -\phi^2 & \phi \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} =$$

$$\frac{1}{\phi\psi(\psi-\phi)} \begin{bmatrix} \psi^2 - \psi \\ -\phi^2 + \phi \end{bmatrix} = \frac{1}{\phi\psi(\psi-\phi)} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{1}{\phi - \psi} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

So

$$c = \frac{1}{\sqrt{5}}, \quad d = -\frac{1}{\sqrt{5}}.$$

In conclusion:

Theorem 8.1 (Euler, 1765). *The Fibonacci numbers are*

$$F_n = \frac{1}{\sqrt{5}} \left(\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right) = \frac{\phi^n - \psi^n}{\sqrt{5}}. \tag{8.4}$$

Since $\psi = -0.618\ldots$, *this implies*

$$F_n = \left[\frac{\phi^n}{\sqrt{5}} \right],$$

where the brackets mean rounding to the nearest integer.

For instance, if the staircase has 533 steps (as does the staircase to the top of the South Tower of the Cologne Cathedral in Germany), then the number of different ways of climbing up, taking one or two steps at a time,

$$A_{533} = F_{534} = \frac{1}{\sqrt{5}} \left(\left(\frac{1+\sqrt{5}}{2} \right)^{534} - \left(\frac{1-\sqrt{5}}{2} \right)^{534} \right) =$$

$$177793095480941658714766079178479431478443211152680070609378957 94\ldots$$

$$\ldots0313896094016507582005031756220276694802823751 2 \approx 1.78 \cdot 10^{111}.$$

8.1.2 Probability of an odd number of heads in n coin tosses

Suppose you toss a biased coin n times. Suppose the probability of getting heads on a given toss is $p \in (0,1)$. How likely are you to obtain an odd number of heads?

Let's answer the question for three tosses first. The ways in which you get an odd number of heads are:

$$HHH, \quad HTT, \quad THT, \quad TTH.$$

The likelihood of getting one of these is

$$p^3 + 3p(1-p)^2. \tag{8.5}$$

That's the formula for $n = 3$, but what is the formula for a general n?

We could answer this by brute force. Denote by P_n the probability of getting an odd number of heads in n tosses. We want to compute a formula for P_n. Let k be an odd integer with $1 \leq k \leq n$. The probability that a prescribed set of k tosses out of n will yield heads, and the others will yield tails, is

$$p^k(1-p)^{n-k}.$$

Therefore the probability of getting exact k heads on n tosses is

$$\left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k},$$

where the factor of $\left(\begin{array}{c} n \\ k \end{array} \right)$ is present because it equals the number of different ways in which we can single out exactly k out of n tosses. Therefore

$$P_n = \sum_{k \text{ odd}} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k},$$

or

$$P_n = \sum_{0 \le j < n/2} \left(\begin{array}{c} n \\ 2j+1 \end{array} \right) p^{2j+1} (1-p)^{n-2j-1}. \tag{8.6}$$

For instance,

$$P_2 = \left(\begin{array}{c} 2 \\ 1 \end{array} \right) p(1-p) = 2p(1-p)$$

and

$$P_3 = \left(\begin{array}{c} 3 \\ 1 \end{array} \right) p(1-p)^2 + \left(\begin{array}{c} 3 \\ 3 \end{array} \right) p^3(1-p)^0 = p^3 + 3p(1-p)^2.$$

Again, a much better formula than (8.6) can be obtained by starting at the end. You can get an odd number of heads either by getting an even number of heads on the first $n-1$ tosses, followed by a head, or by getting an odd number of heads on the first $n-1$ tosses, followed by a tail. If P_n is the probability of getting an odd number of heads on n tosses, therefore,

$$P_n = (1 - P_{n-1})p + P_{n-1}(1-p).$$

On the right-hand side, $1 - P_{n-1}$ is the probability of getting an even number of heads on the first $n-1$ tosses, and the factor p is the probability of getting heads on the last toss. Similarly, P_{n-1} is the probability of getting an odd number of heads on the first $n-1$ tosses, and the factor $1-p$ is the probability of getting tails on the last toss. Simplifying a bit, we get

$$P_n = (1-2p)P_{n-1} + p. \tag{8.7}$$

To calculate P_n, first think about sequences $\{P_n\}$ solving (8.7) in general. Is there a solution that is constant, so $P_n = c$ for a fixed number c, for all n? We would have to have

$$c = (1-2p)c + p,$$

or

$$c = \frac{1}{2}.$$

Now let

$$Q_n = P_n - \frac{1}{2},$$

the difference between P_n and the constant solution $\frac{1}{2}$. From (8.7), we obtain

$$Q_n = (1-2p)Q_{n-1}.$$

This implies, by induction,

$$Q_n = (1-2p)^{n-1} Q_1,$$

or

$$P_n = \frac{1}{2} + (1-2p)^{n-1}\left(P_1 - \frac{1}{2}\right).$$

Now P_1 is the probability of getting an odd number of heads on one toss, and that is p. We therefore arrive at

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

The probability of getting an odd number of heads on n tosses equals

$$P_n = \frac{1-(1-2p)^n}{2}. \tag{8.8}$$

if p is the probability of heads on a single toss.

You will agree that (8.6) is a truly terrible way of writing the simple formula (8.8).

8.1.3 Randomness extractors

Notice that (8.8) is very close to $\frac{1}{2}$ if $|1-2p| \ll 1$ and n is large. If $p = 0.4$ (corresponding to a heavily biased coin), we have

$$P_3 = \frac{1-0.2^3}{2} \approx 0.496.$$

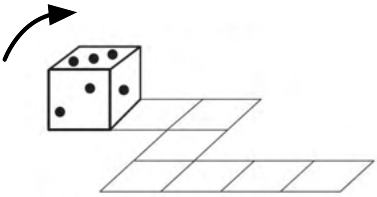
If $p = 0.2$,

$$P_{10} = \frac{1-0.6^{10}}{2} \approx 0.497.$$

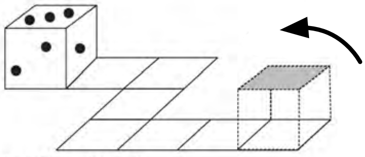
This is a method for simulating an unbiased coin toss with a biased coin. We toss the biased coin many times, then determine whether the number of heads is even or odd — both have approximately the probability 1/2. A method for simulating an unbiased coin with a biased one is called a *randomness extractor*. Randomness extractors are used in cryptography when it is crucial to get a random bit that is very close to unbiased.

8.1.4 Rolling a die along a track

You have a die and a litte track. (The picture comes from the book by Acheson mentioned earlier.)



The question is: If you roll the die through the track, which number will be on top at the end? It's not easy to do in your head, until you have the idea of starting at the end. You picture the die on the last field of the track, and picture whichever side is on top as being painted grey.



When you get to the start of the track, where will the grey side be? This is easy to figure out in your head. It will be opposite to the face with one eye, so it will be the face with six eyes. Therefore the answer to the original question is "six".

8.2 Computing matrix products

Trefethen's essay points out the following analogy. Suppose you wanted to compute a matrix product

$$A_n A_{n-1} \ldots A_2 A_1. \tag{8.9}$$

Assume that the matrices are not necessarily of the same dimensions, but of course their dimensions should be such that the matrix product is well-defined.

In some instances, working from right to left can require far less, or far more computational effort than working from left to right. For instance, computing a matrix product of the form

$$\begin{bmatrix} \ast & \ast & \ast & \ast & \ast \end{bmatrix} \begin{bmatrix} \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \end{bmatrix} \begin{bmatrix} \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \end{bmatrix} \begin{bmatrix} \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \\ \ast & \ast & \ast & \ast & \ast \end{bmatrix}$$

(the product is a 1 × 5-matrix) takes 275 multiplications and 220 additions if you work from right to left, but it takes only 75 multiplications and 60 additions if you work from left to right.

The matrix product (8.9) represents the composition of n linear maps. The first linear map that gets applied is the one represented by A_1, the second is A_2, and so on. So working from right to left should be viewed as "starting at the beginning" here, and working from left to right should be viewed as "starting at the end". Starting at the end is more efficient here.

Notice that starting at the end (working from left to right) can also be viewed as computing the transpose of the desired matrix product, that is, the adjoint of the mapping represented by the matrix product.

8.3 Backpropagation in the abstract

Abstractly, neural networks solve the following kind of problem. We are given test data points (x_i, y_i), $1 \leq i \leq n$. We think of the x_i as "inputs" and the y_i as "outputs". Think of each x_i and each y_i as a real vector. The x's and the y's don't need to have the same length. We are looking for a mapping $F = F(x)$ so that

$$F(x_i) \approx y_i \quad \text{for all } i.$$

If F is taken to be of parameterized form,

$$F = F(x; \eta),$$

where one should think of η as a high-dimensional vector, then the task is to find a parameter vector η for which

$$F(x_i; \eta) \approx y_i \quad \text{for all } i.$$

This amounts to an optimization problem. For instance, we could try to choose η to minimize

$$\sum_{i=1}^{n} \left\| F(x_i; \eta) - y_i \right\|^2.$$

Solving the minimization problem by the method of steepest descent will require, in every step, evaluation of the first derivatives of $F(x_i; \eta)$ with respect to the components of the vector η. We denote the Jacobi matrix[11] of F with respect to η by

$$\frac{\partial F}{\partial \eta}.$$

Since η will be large, it is important to do this calculation efficiently.

In a neural network, F is typically defined in "layers". For example, in a network with four layers, evaluation of $F(x)$ proceeds in four steps:

$$\begin{aligned} x \mapsto P(x;\alpha) \mapsto Q(P(x;\alpha);\beta) \mapsto \\ R(Q(P(x;\alpha);\beta);\gamma) \mapsto S(R(Q(P(x;\alpha);\beta);\gamma);\delta). \end{aligned} \tag{8.10}$$

Here $P = P(x; \alpha)$ is the mapping describing the transformation from the input layer into the first layer, dependent on the parameter vector α; $Q = Q(P; \beta)$ is the

[11] The j-th row of the Jacobi matrix of F with respect to η contains the first derivatives of the j-th component of F with respect to all components of η.

mapping describing the transformation from the first layer into the second layer, depending on the parameter vector β; and so on. We have

$$\eta = (\alpha, \beta, \gamma, \delta).$$

To compute the Jacobi matrix $\frac{\partial F}{\partial \eta}$, we have to compute $\frac{\partial F}{\partial \alpha}$, $\frac{\partial F}{\partial \beta}$, $\frac{\partial F}{\partial \gamma}$, and $\frac{\partial F}{\partial \delta}$. Using the chain rule,

$$\begin{aligned}
\frac{\partial F}{\partial \alpha} &= \frac{\partial S}{\partial R}\frac{\partial R}{\partial Q}\frac{\partial Q}{\partial P}\frac{\partial P}{\partial \alpha}, && (8.11)\\
\frac{\partial F}{\partial \beta} &= \frac{\partial S}{\partial R}\frac{\partial R}{\partial Q}\frac{\partial Q}{\partial \beta}, && (8.12)\\
\frac{\partial F}{\partial \gamma} &= \frac{\partial S}{\partial R}\frac{\partial R}{\partial \beta}, && (8.13)\\
\frac{\partial F}{\partial \delta} &= \frac{\partial S}{\partial \delta}. && (8.14)
\end{aligned}$$

It is most natural and efficient to evaluate (8.10) "forward", meaning from left to right, but (8.11)–(8.14) "backward", from bottom to top.

8.4 Mortgages

Suppose you borrow \$400,000 at an interest rate of 6.5%. Now you make monthly payments to the bank which lent you the money: P_1 dollars after one month, P_2 dollars after two months, and so on. The bank will typically prescribe that the P_i can't be smaller than some positive P, which would be sufficient to pay off your debt in 30 or 15 (depending on which kind of mortgage you get) years. But you can pay more if you want to.

8.4.1 Forward and backward thinking are almost the same here

After n months, the bank considers the \$400,000 that you received worth

$$\$400,000 \cdot q^n,$$

where

$$q = 1 + \frac{0.065}{12}.$$

(This is called *monthly compounding.*) At the same time, your payments $P_1, P_2, \ldots,$ P_n are worth

$$\sum_{i=1}^{n} q^{n-i} P_i$$

dollars. You are done with your payments as soon as

$$\sum_{i=1}^{n} q^{n-i} P_i = 400,000\, q^n. \qquad (8.15)$$

Instead of reasoning with the *future* value of your payments, you can value with their *past* values, namely, their values on the day on which you got the loan. In fact, this leads to (8.15), just with both sides divided by q^n:

$$\sum_{i=1}^{n} q^{-i} P_i = 400,000. \tag{8.16}$$

The i-th payment is worth $q^{-i}P_i$ at the time of loan initiation.

In this example, there is really no essential difference between "thinking forward" and "thinking backward". Nonetheless, I find the "backward" thinking a little more intuitive, perhaps because we are trying to bring the total payment amount to the fixed loan amount of \$400,000 instead of chasing the moving target of \$400,00$\cdot q^n$.

8.4.2 Mostly interest initially, mostly principal later

People say that mortgage payments are "mostly interest earlier, mostly principal later". Is this bank policy, or mathematics? The answer is, it is mathematics. To see why, suppose that all P_i are equal to the same value P. (This is the usual arrangement.) For a 30-year mortgage loan of \$4000,000, one must then choose P such that

$$\sum_{i=1}^{360} q^{-i} P = 400,000$$

or

$$P = \frac{400,000}{\sum_{i=1}^{360} q^{-i}}. \tag{8.17}$$

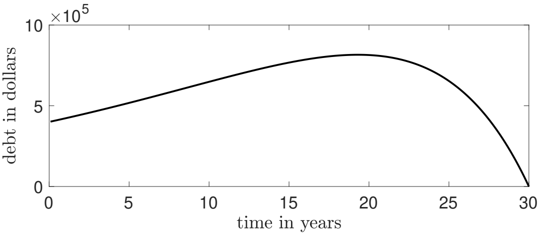
(This is the formula that a "mortgage calculator" would evaluate.) If you wanted to pay off your debt after $n < 360$ months, how much would you owe? The answer is Q with

$$\sum_{i=1}^{n} q^{-i} P + q^{-n} Q = 400,000.$$

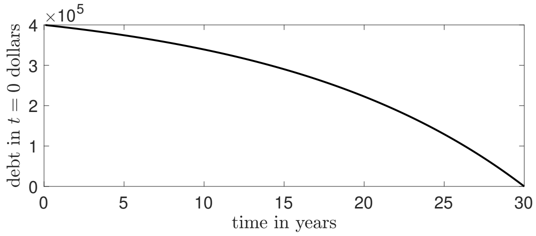
Solving for Q and using (8.17), we obtain

$$Q = q^n \left(1 - \frac{\sum_{i=1}^{n} q^{-i}}{\sum_{i=1}^{360} q^{-i}} \right) 400,000.$$

With $q = 1 + 0.065/12$, the graph of Q as a function of n looks as follows.



After making payments for 20 years, you will owe about \$800,000, twice as much as your original loan amount! If we measure your debt in *dollars at the time of loan initiation*, then of course each payment does bring down your debt:



The curve is concave-down: The decline is slow initially, and accelerates later.

Backward thinking, that is, thinking in terms of dollars at the time of loan initiation, results in a monotonically declining debt, whereas forward thinking, thinking in terms of present-day dollars, does not. This is another sense in which backward thinking is a little more intuitive here.

8.4.3 Should I apply early payments to principal?

I asked the Google AI. This is what it told me:

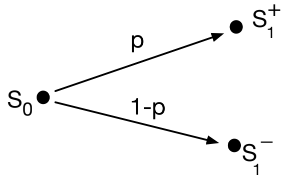
> *To apply early payments to your loan principal, contact your lender and specify that you want the extra payment to reduce your outstanding loan balance, rather than being applied to future payments.*

It goes on to recommend that you do this. The recommendation makes little sense to me. When you make an extra payment of Q dollars after n months, the bank really has no choice to make. In dollars at the time of loan initiation, the payment should reduce your debt exactly by $q^{-n}Q$. The bank *might* allow you to skip some future monthly payments because of your early payment of Q dollars, that would be conceivable at least. (They typically do not do that.) So perhaps you could interpret the request “Apply my early payment to principal” as really meaning “Do not allow me to skip future monthly payments”. But why would you explicitly ask the bank to restrict your future options?

8.5 Stock option pricing using the no-arbitrage principle on a tree

8.5.1 One-period tree

Suppose that today’s value of a stock were S_0, and you knew with certainty that tomorrow, it will be S_1^+ with probability p, and $S_1^- < S_1^+$ with probability $1 - p$. The following diagram symbolically describes our assumptions.



All there is to know about the *statistics* of the stock price is known to us. This does not mean that we actually can predict tomorrow's stock price, only that we know what the possibilities are, and how likely they are.

Suppose I offered you an *option* on this stock — the right to sell at a certain prescribed price tomorrow (a *put option*), or the right to buy at a certain prescribed price tomorrow (a *call option*), for instance. We needn't be specific about what kind of option we are talking about, but we assume that we will know its value tomorrow, after we find out whether the stock price will be S_1^+ or S_1^-. If the stock price goes to S_1^+, the value of the option will be V_1^+, and if the stock price goes to S_1^-, the value of the option will be V_1^-.

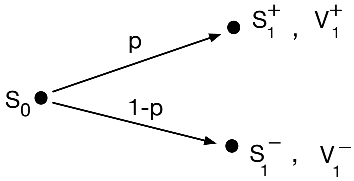
For instance, suppose we were talking about a *European put option*. Such an option gives its holder the right to sell the stock at a specified price, let's call it S, at a specified time in the future, let's say tomorrow. If the value of the stock goes to S_1^+ tomorrow, the value of the option tomorrow will become

$$V_1^+ = \max(S - S_1^+, 0).$$

If the value of the stock goes to S_1^-, the value of the option will be

$$V_1^- = \max(S - S_1^-, 0).$$

It doesn't have to be this example. All that matters is that S_1^+ determines V_1^+, and S_1^- determines V_1^-. Let's add V_1^+ and V_1^- to our diagram of things we know:



What should be the price of such an option today? There are two big surprises here. First, there is an objective answer to this question. If we agree on what S_1^+ and S_1^- are, and if tomorrow's value of the option that we are pricing will be determined once we know whether the stock price goes to S_1^+ or to S_1^-, then we know how much the option should cost today. The second big surprise is that the answer is independent of p. Here is the argument that leads to these conclusions.

Suppose that I purchased the option, but at the same time also bought a fraction c of a stock. (Let's assume that one can buy fractions of this stock. Perhaps that becomes more plausible if you think of buying a large number N of options, and cN stocks.) If the stock price goes to S_1^+, the value of my portfolio will be

$$V_1^+ + cS_1^+.$$

If the stock price goes to S_1^-, the value of my portfolio will be

$$V_1^- + cS_1^-.$$

I could choose c to eliminate all uncertainty:

$$V_1^+ + cS_1^+ = V_1^- + cS_1^-.$$

Solving for c,

$$c = -\frac{V_1^+ - V_1^-}{S_1^+ - S_1^-}.$$

For the simple put option discussed earlier, this isn't negative, but negative c is not a problem: To buy a negative amount of stock simply means to sell stock.

If I choose this value of c, then I have no risk, no uncertainty. The value of my portfolio tomorrow will be

$$V_1^+ + cS_1^+ = V_1^- + cS_1^-. \tag{8.18}$$

The value of my portfolio today is

$$V_0 + cS_0,$$

where as before S_0 is today's stock price, and V_0 is today's option price.

Assume for simplicity that there is no risk-free return to be had by putting money into the bank. [12] Then the price of my risk-free portfolio shouldn't change between today and tomorrow. If my portfolio were guaranteed to rise from today to tomorrow, everybody would buy the same portfolio many times over, and the prices would adjust as a result. If my portfolio were guaranteed to lose, everybody would sell the same porfolio many times over, again causing a price adjustment. This is called the *no arbitrage principle.* An *arbitrage opportunity* is an opportunity for a risk-free profit. In a market that is sufficiently transparent and responds sufficiently fast, there should be no arbitrage opportunities.

We conclude:

$$V_0 + cS_0 = V_1^+ + cS_1^+,$$

and therefore

$$\begin{aligned} V_0 &= V_1^+ + c(S_1^+ - S_0) \\ &= V_1^+ - \frac{V_1^+ - V_1^-}{S_1^+ - S_1^-}(S_1^+ - S_0) \\ &= \frac{S_0 - S_1^-}{S_1^+ - S_1^-} V_1^+ + \frac{S_1^+ - S_0}{S_1^+ - S_1^-} V_1^-. \end{aligned}$$

In summary:

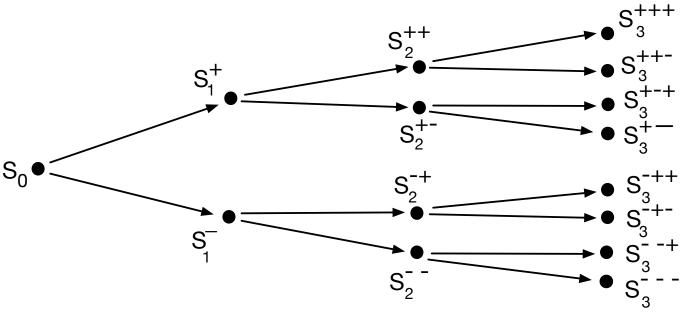
[12] If there were a risk-free rate of return, the discussion could easily be adjusted to take that into account, but the formulas would become a little less clean.

Proposition 8.2. *Assume that the risk-free rate of return is zero. Given that tomorrow's stock and option prices will either be S_1^+ and V_1^+, or S_1^- and V_1^-, and given that today's stock price is S_0, the only value for today's option price that is compatible with the no-arbitrage principle is*

$$V_0 = \frac{S_0 - S_1^-}{S_1^+ - S_1^-} V_1^+ + \frac{S_1^+ - S_0}{S_1^+ - S_1^-} V_1^- . \tag{8.19}$$

8.5.2 Multi-period tree

Now suppose that we knew the possible changes in the stock price not only over one day, but over multiple days. For three days, that might look like this:



Consider the price of an option that can only be exercised on day 3. (A *European option* can only be exercised on the end date, whereas an *American option* can be exercised on or before the end date.) Assume that we will know the option price on day 3 once we know the stock price on day 3. Then we can compute the option price on day 2, using (8.19). Therefore we can compute the option price on day 1, again using (8.19). Finally we can compute the option price on day 0, again using (8.19).

It is another example of "starting at the end"!

Exercises

8.1. (easy) For a staircase of five steps, can you figure out how many ways of climbing, taking one or two steps at a time, there will be *without* starting at the end? Verify that you get it right by comparing with the sixth Fibonacci number.

8.2. (easy) Fibonacci's recursion formula is

$$F_n = F_{n-1} + F_{n-2}.$$

To understand it, the key is to have the idea of looking for sequences in the form

$$F_n = r^n.$$

Suppose the recursion formula were

$$F_n = F_{n-1} - F_{n-2}. \tag{8.20}$$

Are there solutions in the form $F_n = r^n$, $r \in \mathbb{R}$?

8.3. (easy) Suppose that $\{F_n\}$ is a sequence that solves the recursion relation (8.20). Suppose $F_1 = F_2 = 1$. Compute F_n for $n = 3, 4, \ldots, 15$.

8.4. (easy) Now let's put the minus sign in front of F_{n-1}, not F_{n-2}:

$$F_n = -F_{n-1} + F_{n-2}. \tag{8.21}$$

Does this have solutions in the from $F_n = r^n$, $r \in \mathbb{R}$? Can you find an explicit expression for the sequence that satisfies (8.21) and $F_1 = F_2 = 1$?

8.5. (easy if you use Section 8.1.2) Mr. Smith has five children. How likely is Mr. Smith to have an even number of boys? What would be the answer if he had six children?

8.6. (easy) For the case of five children, do Problem 5 *without* using Section 8.1.2.

8.7. (easy if you use Section 8.1.2) You have a coin that gives heads with probability 0.47. You toss it four times. How likely is it that the number of heads is odd?

8.8. (easy) Suppose you have a coin that gives heads with some probability $p \in (0, 1)$. Toss the coin twice. If you get the same result twice, then toss it twice again. If you get two different results, accept the first result and discard the second. Show that this will yield heads with probability 1/2 exactly, regardless of what p is.

This idea is called the *von Neumann randomness extractor*, after John von Neumann, one of the great mathematicians of the 20th century.

8.9. (easy) The von Neumann randomness extractor (see Problem 8) requires a random number of coin tosses, since you have to do pairs of coin tosses until you get a pair with two different results. Explain why the probability of getting two different results on a pair of coin tosses equals

$$2p(1 - p).$$

8.10. (easy if you have learned some calculus-based probability) For the von Neumann randomness extractor (see Problem 8), what is the expected number of coin tosses needed? The answer will be smaller when p is close to $1/2$, larger when p is further from $1/2$. However, show that it is never smaller than 4.

8.11. (medium) For the case of six children, do Problem 5 *without* using Section 8.1.2.

8.12. (medium) To evaluate the matrix-vector product

$$\begin{bmatrix} * & * & * & * \end{bmatrix} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}$$

it is best to proceed from left to right. For

$$\begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{bmatrix} * \\ * \\ * \\ * \end{bmatrix}$$

it is best to proceed from right to left. How about the following product?

$$\begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{bmatrix} * \\ * \\ * \\ * \end{bmatrix} \begin{bmatrix} * & * & * & * \end{bmatrix} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}$$

What is the most efficient order in which to evaluate this product? (The answer might neither be "left to right" nor "right to left".)

8.13. (hard) Suppose that today's stock price were \$100, and tomorrow's stock price were \$110 with probability 0.999999, but \$99 dollars with probability 0.000001. Think about an option that allows me to sell the stock tomorrow at \$100. Having such an option will protect me against the very unlikely event of losing a dollar. If the stock goes up to \$110, the option will be worthless tomorrow, but if it goes down to \$99, the option will be worth one dollar. Formula (8.19) tells me that this option should today have the price

$$V_0 = \frac{10}{11}$$

dollars, or about 91 cents. Something seems seriously wrong here. I am supposed to pay 91 cents today to protect myself against a loss of one dollar tomorrow, which only has a likelihood of one in a million? I don't think I'd do that.

Can you understand why nothing is wrong with (8.19) in spite of this disturbing example?

8.14. (requires both coding and thinking) Think about the recursion formula

$$F_n = F_{n-1} \pm F_{n-2},$$

where the sign is chosen at random, with each of + and − equally likely. The resulting sequence is called the *random Fibonacci theorem.* Write code demonstrating that

$$\lim_{n\to\infty} |F_n|^{1/n} \approx 1.13 \quad \text{with probability 1.}$$

The fact that there exists some value in $[0, \infty)$ equal to this limit with probability 1 follows from a 1960 theorem by Harry Kesten and Hillel Furstenberg about random matrix products. The value of the limit was shown to be 1.1319882487943... by Divakar Viswanath in 1999.

Can you prove mathematically that

$$|F_n|^{1/n} \leq \phi$$

(the golden ratio) for all n?

Random Fibonacci sequences are just about the simplest random dynamical systems that there are. It is clear that random dynamical systems are hugely important. Almost anything that changes with time obeys deterministic laws but also should be thought of as having“random” aspects, even if those may simply be those aspects that we can’t model in detail because we don’t understand them.

Chapter 9

Brouwer's fixed point theorem and the Borsuk-Ulam theorem

9.1 Brouwer's fixed point theorem

9.1.1 Statement of the theorem

Brouwer's fixed point theorem is about systems of nonlinear equations. For instance, let's think about this one:

$$\sin(x^2+y) - \cos(z+x) - 2x = 0, \tag{9.1}$$
$$x^2+y^4 - \cos(xz) + 4y = 0, \tag{9.2}$$
$$\sin(z^2+y)\cos^2(4x+y) - z = 0. \tag{9.3}$$

Totally crazy, of course, and who really cares whether this system has a solution? Amazingly, though, Brouwer tells you that yes, this system, opaque though it is, does have a solution. The theorem applies to countless other nonlinear systems of equations, many of which *are* important.

A system of three equations in three unknowns can always be written as

$$F(x,y,z) = 0, \tag{9.4}$$
$$G(x,y,z) = 0, \tag{9.5}$$
$$H(x,y,z) = 0. \tag{9.6}$$

Alternatively, any such system can be written in the form

$$x = f(x,y,z), \tag{9.7}$$
$$y = g(x,y,z), \tag{9.8}$$
$$z = h(x,y,z). \tag{9.9}$$

For instance, we could take $f(x,y,z) = F(x,y,z) + x$, $g(x,y,z) = G(x,y,z) + y$, and $h(x,y,z) = H(x,y,z) + z$. That's just one possibility, and there are infinitely many other ways of re-writing (9.4)–(9.6) in the form (9.7)–(9.9).

For instance, we can write (9.1)–(9.3) as

$$\begin{aligned} x &= \frac{\sin(x^2+y)-\cos(z+x)}{2}, & (9.10)\\ y &= \frac{\cos(xz)-x^2-y^4}{4}, & (9.11)\\ z &= \sin(z^2+y)\cos^2(4x+y). & (9.12) \end{aligned}$$

In general, Brouwer's theorem concerns systems of n equations in the n real unknowns $x_1, x_2, \ldots, x_n$. We write

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

We assume that a function

$$F: \mathbb{R}^n \to \mathbb{R}^n$$

is given, and consider the system

$$x = F(x). \qquad (9.13)$$

More explicitly:

$$\begin{aligned} x_1 &= F_1(x_1, x_2, \ldots, x_n),\\ x_2 &= F_2(x_1, x_2, \ldots, x_n),\\ &\vdots\\ x_n &= F_n(x_1, x_2, \ldots, x_n). \end{aligned}$$

Equation (9.13) is said to be *in fixed point form.* We are looking for an $x \in \mathbb{R}^n$ that is kept *fixed* by F. To state Brouwer's theorem about such equations, we need some terminology. For an integer $n \geq 1$, B^n is the unit ball in $\mathbb{R}^n$:

$$B^n = \{x \;:\; \|x\| \leq 1\}.$$

(The notation $\|x\|$ means the euclidean length of the vector x.) A set $K \subseteq \mathbb{R}^n$ is called *topologically equivalent to* B^n if there exists a continuous, bijective (one-to-one and onto) mapping

$$\varphi: K \to B^n$$

with a continuous inverse φ^{-1}.

Theorem 9.1 (Brouwer, 1911). *Let $K \subseteq \mathbb{R}^n$ be topologically equivalent to B^n. Let*

$$F : K \to K$$

be a continuous mapping. Then there exists at least one $x \in K$ with

$$x = F(x).$$

For example, K might be the region in space occupied by your morning coffee. We conclude that you can stir your coffee any way you like, it will always be true that after it comes back to rest, at least one molecule comes to rest in the precisely same place that it was in before you starting stirring.

Brouwer's fixed point theorem is named after L. E. J. Brouwer (1881–1966), who is famous for many great results in topology and analysis, and also for founding *intuitionism*, one of several competing philosophies of the foundations of mathematics.[13]

9.1.2 $K = B^n$ is all we need

The general theorem follows if we know that the theorem holds for $K = B^n$. Namely, suppose that we know the theorem for $K = B^n$, and suppose that now $K \subseteq \mathbb{R}^n$ is some other set, topologically equivalent to B^n. Suppose

$$\varphi : \ K \to B^n$$

is a continuous bijection with a continuous inverse. Then the function

$$u \mapsto \varphi(F(\varphi^{-1}(u)))$$

is a continuous mapping

$$B^n \to B^n.$$

[13] Brouwer is also infamous for anti-semitic comments, which persuaded David Hilbert to fire him from the editorial board of the journal *Mathematische Annalen*, of which Hilbert was the editor-in-chief.

Brouwer's theorem therefore tells us that there exists a $u \in R^n$ with

$$u = \varphi(F(\varphi^{-1}(u))).$$

This implies

$$\varphi^{-1}(u) = F(\varphi^{-1}(u)),$$

so $x = \varphi^{-1}(u) \in K$ satisfies

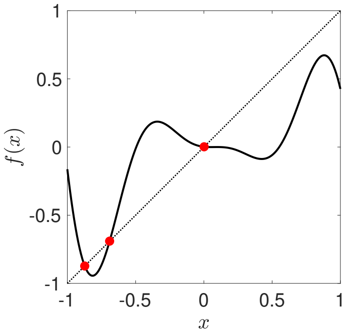
$$x = F(x).$$

9.1.3 $n = 1$

For $n = 1$, Brouwer's theorem says that for any continuous function

$$f: \ [-1, 1] \to [-1, 1],$$

there exists an $x \in [-1, 1]$ with $x = f(x)$.



The equation $x = f(x)$ means that the point (x, x) lies both on the graph, and on the 45-degree line. In the picture, there are three such points, so three solutions of $x = f(x)$. There has to be one at least: The graph cannot get from the left edge to the right edge without getting across the 45-degree line. It's the intermediate value theorem, applied to $f(x) - x$. In fact, Brouwer's fixed point theorem for $n = 1$ is an equivalent statement of the intermediate value theorem.

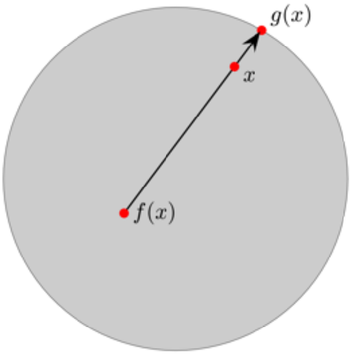
9.1.4 $n \geq 2$, outline

Suppose that

$$f: \ B^n \to B^n$$

is continuous and has no fixed point. We can then define $g(x)$ to be the point at which the ray starting at $f(x)$ and pointing in the direction of x meets the boundary of B^n, which is the sphere

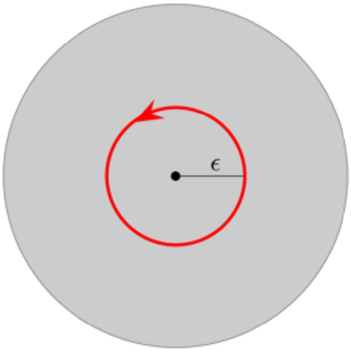
$$S^{n-1} = \{x \in \mathbb{R}^n \ : \ \|x\| = 1\}.$$



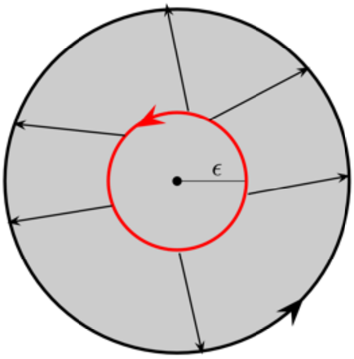
The function $g: B^n \to S^{n-1}$ is continuous with $g(x) = x$ for $x \in S^{n-1}$. To complete the proof, we prove that no such function exists. Notice that this is utterly plausible — we can't push all points in B^n to the boundary in such a way that the points on the boundary stay wher they are, and no "tear" occurs in the interior.

9.1.5 $n = 2$, precise argument using winding numbers

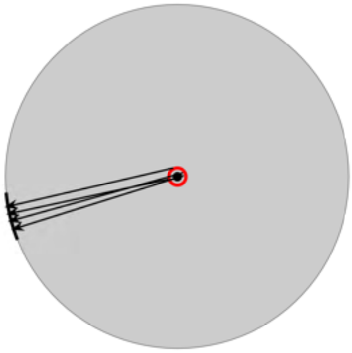
Think about the circle with radius $\epsilon \in (0, 1]$ centered around the origin in $\mathbb{R}^2$, oriented counter-c



Picture what the image of this circle under g would look like. It is a closed curve in the plane, and it is a subset of S^1.



First consider $\epsilon = 1$. In that case, the red circle has radius 1, and is its own image under g. It winds around the origin once in the counter-clockwise direction, so we say that its *winding number with respect to the origin* equals 1. For very small ϵ, it is also easy to understand what happens:

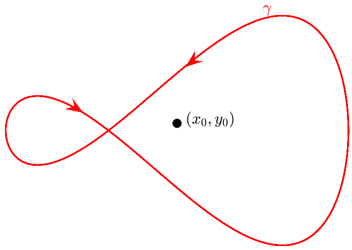


Since g is continuous, it must map all points on the small circle with radius ϵ to points near the image of the origin under g, when ϵ is small enough. So the closed curve that's the image of the small red circle actually does *not* loop around the origin. We say that its *winding number with respect to the origin* is 0.

But the winding number must clearly depend continuously on ϵ, and it is always an integer. It cannot jump from 1 to 0 as ϵ is reduced from 1 to 0. This contradiction proves our assertion.

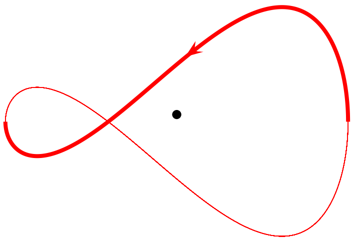
9.1.6 More on winding numbers

Let (x_0, y_0) be a point in the plane, and think about a closed curve γ in the plane that does not run through the point (x_0, y_0). ("Closed" means that it starts and ends at the same plac

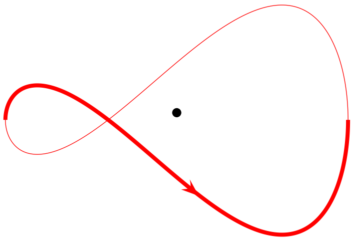


Definition 9.2. *The* winding number *of* γ *with respect to* (x_0, y_0) *is the number of times that the curve winds around* (x_0, y_0) *in the counter-clockwise direction. We denote it by* $\text{wind}(\gamma; x_0, y_0)$.

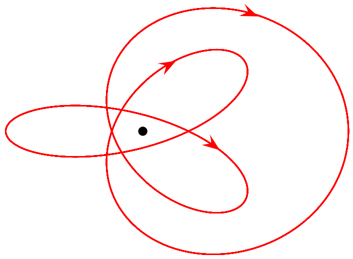
Even though we have not said exactly what this means, you will probably agree that in the example given above, $\text{wind}(\gamma; x_0, y_0) = 1$. In case you don't see that, here is the first half of the counter-clockwise path around (x_0, y_0):



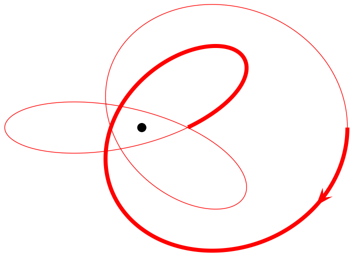
And here is the second half:

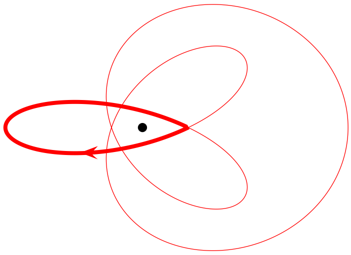


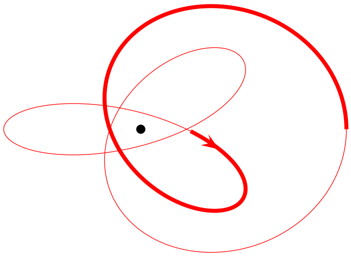
You may also see that in the example below, $\text{wind}(\gamma; x_0, y_0) = -3$, meaning that the curve winds around (x_0, y_0) in the *clockwise* direction three times.



In case you don't see that, I am going to show you the three pieces of the curve corresponding to the three times that the curve winds around (x_0, y_0) in the clockwise direction:







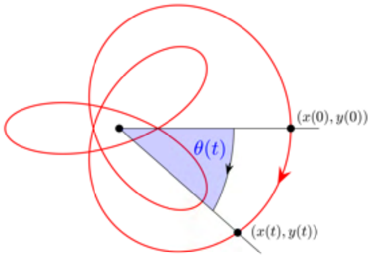
These examples may give you a clear understanding of what a winding number is. I won't be completely precise about its definition. However, here is a sketch. Let $T > 0$, and let $(x(t), y(t))$, $0 \leq t \leq T$, be a parametrization of the curve, with

$$(x(T), y(T)) = (x(0), y(0))$$

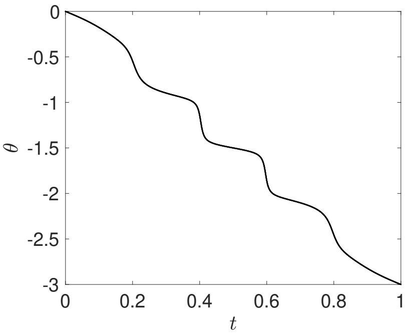
and

$$(x(t), y(t)) \neq (x(0), y(0)) \quad \text{for } 0 < t < T.$$

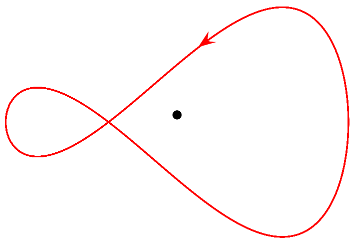
Let $\theta = \theta(t)$ be the angle indicated in the following figure.



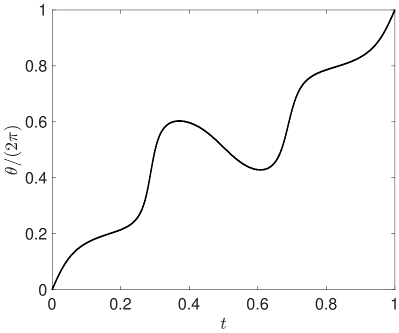
In this example, $\theta(t) < 0$ because the rotation is clockwise. Plotting $\theta(t)/(2\pi)$ as a function of $t \in [0, T]$, we get this:



Of course, θ is only defined up to adding or subtracting integer multiples of 2π. However, if you want $\theta(t)$ to be a *continuous* function, then the fact that $\theta(k)/(2\pi)$ decreases by 3 during one trip through the curve (in the direction of the arrows) is inescapable. Let's return to our first example.

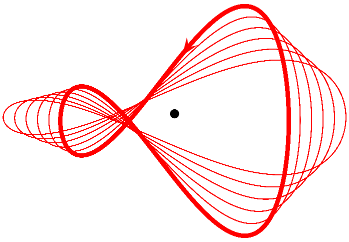


Here the plot of $\theta/(2\pi)$ as a function of t looks as follows.



If we start in a different point of the curve, or add an integer multiple of 2π to θ, the plot will be shifted up or down, but the fact that $\theta(t)/(2\pi)$ rises by 1 during one round trip is inescapable as long as we define $\theta(t)$ so that it is continuous. This is why $\text{wind}(\gamma; x_0, y_0) = 1$ here.

Now suppose we slightly, continuously, deformed the curve γ, for instance like this:



The winding number will then also change continuously. But the winding number is an integer! An integer-value function cannot be continuous unless it is constant. We conclude:

Theorem 9.3. *The winding number* $\text{wind}(\gamma; x_0, y_0)$ *does not change when the curve* γ *is deformed continuously, as long as none of the deformed curves passes through* (x_0, y_0).

We therefore call the winding number a topological invariant.

9.1.7 Applications of Brouwer's fixed point theorem

Brouwer's fixed point theorem is a rather theoretical tool. It assures us of the *existence* of (at least one) solution of an equation of the form $x = F(x)$. It does not help us find a solution. However, any nonlinear system of equations can be written in the from $x = F(x)$, and therefore potentially be analyzed using Brouwer's theorem.

Existence of economic equilibria:

Brouwer's theorem and other fixed point theorems are used in theoretical economics. As a greatly simplified example, imagine an economy in which only three goods are being traded, for simplicity. Suppose their current prices are p_1, p_2, and p_3. We assume that $p_1, p_2, p_3 \geq 0$. (So there are no "goods" that people will pay money to get away from.) We will assume

$$p_1 + p_2 + p_3 = 1. \tag{9.14}$$

This assumption reflects the fact that what matters is no the number on the price tag, but the price *relative* to the prices of other goods. Therefore we might as well, for theoretical purposes, normalize so that (9.14) holds.

Given the current price structure, there will be more or less demand for the three goods, and as a result, prices will adjust — goods in high demand will become more expensive, and ones in low demand will become less expensive. We will assume that p_i is adjusted as follows:

$$p_i \to \tilde{p}_i = \frac{\max(p_i + Z_i(p_1, p_2, p_3), 0)}{\sum_{k=1}^{3} \max(p_k + Z_k(p_1, p_2, p_3), 0)}, \tag{9.15}$$

where Z_1, Z_2, Z_3 are continuous functions of p_1, p_2, p_3. The denominator in (9.15) normalizes the $\tilde{p}_i$ so that their sum is 1. We assume that the functions Z_k are chosen such that the denominator in (9.15) is not zero; if it were zero, all prices would be adjusted to zero, so there would be no scarcity in the economy.

Note that because of (9.14), p_3 can be viewed as a function of p_1 and p_2:

$$p_3 = 1 - p_1 - p_2,$$

and similarly

$$\tilde{p}_3 = 1 - \tilde{p}_1 - \tilde{p}_2.$$

The *price vector* $p = (p_1, p_2)$ and the *adjusted price vector* $\tilde{p} = (\tilde{p}_1, \tilde{p}_2)$ lie in the triangle defined by

$$T = \{(x_1, x_2) \ : \ 0 \le x_1, x_2 \le 1, \quad x_1 + x_2 \le 1\}.$$

We denote the right-hand side of (9.15) by $F(p_1, p_2) \in T$.

Is there an *equilibrium* price vector, one that will lead to no adjustments? This question is now equivalent to asking whether there is a price vector p with

$$p = F(p).$$

Brouwer's theorem answers the question: There is such a price vector. To read more about the (much more complex) story, google "Arrow-Debreu model".

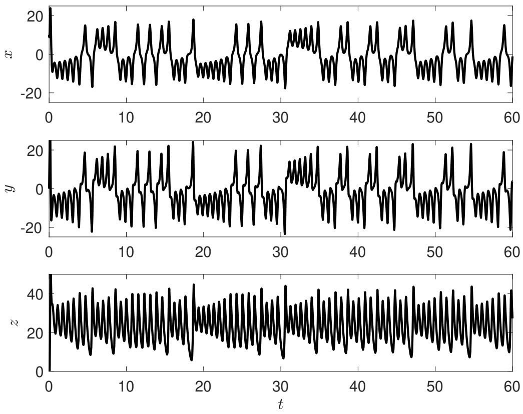
There are many other applications of fixed point theorems in theoretical economics.

Existence of periodic solutions of differential equations:

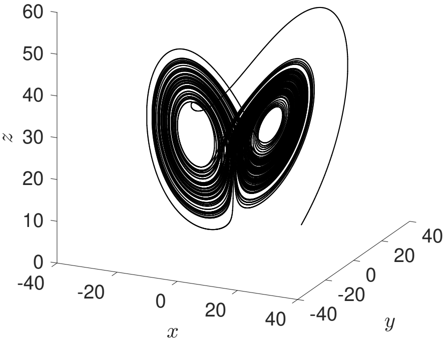
The Lorenz system is a system of three ordinary differential equations with chaotic solutions. The system is

$$\begin{aligned}\frac{dx}{dt} &= 10(y - x),\\ \frac{dy}{dt} &= -y + 28x - xz,\\ \frac{dz}{dt} &= -\frac{8}{3}z + xy.\end{aligned}$$

Edward Lorenz, a meteorologist at MIT, came upon these equations as a very much simplified model of convection rolls in the atmosphere. Here is an example of a solution:



Go **here** to see ; point in three-dimensional spa



The behavior of x, y, and z as functions of t is erratic, highly sensitive to perturbations in the initial values of x, y, and z, and aperiodic. In fact, it has been proved that there are no stable periodic solutions. ("Stable" means that the solution would return to the same periodic pattern if slightly perturbed away from it.) There are, however, unstable periodic solutions, and they play a great role in the theory of the Lorenz equations.

Brouwer's fixed point theorem allows us to see that there must be (unstable) periodic solutions. Consider a solution that starts at

$$x(0) = 0, \quad y(0) = y_0, \quad z(0) = z_0 \tag{9.16}$$

with

$$(y_0, z_0) \in K = [-40, 40] \times [10, 60].$$

Inspection of the numerical solutions shows that such a solution must return, at some later time, to a point with $x = 0$ and $(y, z) \in K$. Let's suppose that this happens for the first time at time $t_1 > 0$, and let

$$y_1 = y(t_1), \quad z_1 = z(t_1),$$

so

$$(y_1, z_1) \in K.$$

Denote the mapping

$$(y_0, z_0) \mapsto (y_1, z_1)$$

by F. It is a continuous mapping $K \to K$. Therefore it has a fixed point, by Brower's theorem. The solution $(x(t), y(t), z(t))$ with (9.16) must be a periodic solution.

Much stronger results are known. The Lorenz equations have infinitely many unstable periodic solutions. The point of my discussion here is just to illustrate that Brouwer's fixed point theorem can sometimes be used to analyze the behavior of ordinary differential equations, for example to prove the existence of periodic solutions.

Existence of positive eigenvalues:

Suppose that A is a real $n \times n$-matrix. An *eigenvector* is a non-zero vector x with

$$Ax = \lambda x$$

for some number λ. We call x an *eigenvector*, and λ and *eigenvalue*. You learned this when you learned linear algebra.

Suppose all entries in A are positive. For instance, if $n = 2$, an example would be

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}.$$

Are then all eigenvalues of A positive? The answer is no:

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} = (-1) \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

so -1 is an eigenvalue of A. However, in this example, also

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

So 3 is an eigenvalue. At least there does exist a positive eigenvalue.

Is this always true? If the entries of A are all positive, must there be a positive eigenvalue? It turns out that the answer is yes. In fact, there is an eigenvector x in which all entries are ≥ 0; the associated eigenvalue must then be positive. Here is a way of seeing that using Brouwer's fixed point theorem.

We are trying to find a non-zero vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

with $x_i \geq 0$ for all i so that Ax and x point in the same direction. This means

$$\frac{Ax}{\|Ax\|} = x.$$

There we have a fixed point equation! I will now leave it to you to fill in the rest; see Problem 9.7

9.2 The Borsuk-Ulam theorem

9.2.1 Statement of the theorem

For any integer $n \geq 1$, we set

$$S^n = \left\{ x \in \mathbb{R}^{n+1} \ : \ \|x\| = 1 \right\}.$$

This is an n-dimensional sphere with radius 1, embedded in $\mathbb{R}^{n+1}$, centered at the origin.

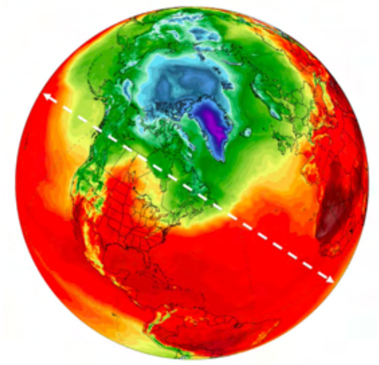
Theorem 9.4 (Borsuk, 1933). *If*

$$F: \ S^n \to \mathbb{R}^n$$

is a continuous mapping, then there exists an $x \in S^n$ with

$$F(x) = F(-x).$$

For example, if $n = 2$, the sphere S^2 is the earth, and the components of F are pressure and temperature (both of which are plausibly continuous functions), the theorem tells us that there is a pair of antipodal points on earth in which both temperature and pressure are the same, and therefore the weather is the same, at least if you buy that temperature and pressure determine the weather.



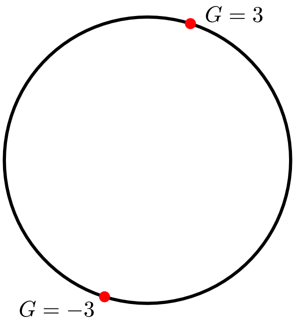
The theorem is named after the Polish mathematician Karol Borsuk (1905–1982). Borsuk credited Sanisław Ulam (1909–1984) for having suggested the result.[14]

9.2.2 $n = 1$

Suppose F is a continuous function on the circle S^1. Write

$$G(x) = F(x) - F(-x) \quad \text{for } x \in S^1.$$

Then G is an odd function, meaning $G(-x) = -G(x)$. For instance, if $G(x) = 3$ for some x, then $G(-x) = -3$.



Somewhere in between, G must be zero. It's again an equivalent statement of the intermediate value theorem.

If you think of the circle as the equator, and of F as the temperature, this result states that there must always be a pair of antipodal points on the equator in which the temperatures are the same.

[14] Ulam's uncle, Michał Ulam, was in the habit of gambling in Monte Carlo. This is how the Monte Carlo method, invented by Sanisław Ulam, got its name.

9.2.3 $n \geq 2$, outline

Now suppose that

$$F: \ S^n \to \mathbb{R}^n$$

is continuous. Define

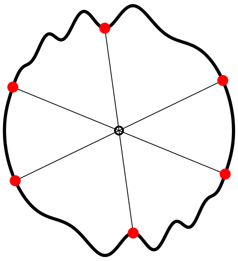
$$G(x) = F(x) - F(-x).$$

We want to show that $G(x) = 0$ for some x.

To see this, suppose that $G(x) \neq 0$ for all x. Consider the "equator" of S^n, which is

$$C_0 = \left\{(x_1, \ldots, x_n, 0) \ : \ x_1^2 + \ldots + x_n^2 = 1\right\},$$

and think about $G(C_0)$. We should picture this as some $(n-1)$-dimensional "surface" S_0 in $\mathbb{R}^n$, w



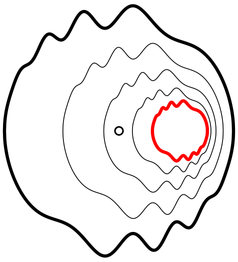
S_0 *wraps around* the origin. (This is the main idea that needs to be made more precise.) It does not *contain* the origin, since we are assuming $G(x) \neq 0$ for all $x \in S^n$. Now gradually move C_0 upwards:

$$C_\epsilon = \left\{(x_1, \ldots, x_n, \epsilon) \ : \ x_1^2 + \ldots + x_n^2 = 1 - \epsilon^2\right\} \subseteq S^n$$

for $\epsilon \in [0, 1]$. Define

$$S_\epsilon = G(C_\epsilon).$$

None of these "surfaces" contains the origin, but as ϵ rises from 0 to 1, they first "enclose" the origin, and then, when ϵ is near 1, they don't, since then all points in S_ϵ are near $G(0, .$



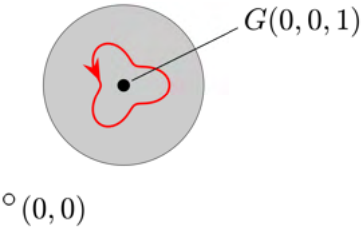
Somehow the S_ϵ passed over 0 as ϵ rose, but 0 was never part of any S_ϵ. This is impossible, and this contradiction completes the proof.

To make this argument precise, we have to discuss more precisely what we mean by saying S_0 *wraps around* the origin. For $n = 2$, this is done using winding numbers. For $n > 2$, it can be done using the notion of the *degree* of a mapping, but we won't do that here.

9.2.4 $n = 2$, precise argument using winding numbers

For $n = 2$, C_0 can be thought of as a circle with radius 1 in the (x_1, x_2)-plane (disregarding the third coordinate, which is 0). We think of it as oriented counterclockwise. Its image under G is the circle itself. It has winding number 1 with respect to the origin. Similarly, C_ϵ can be thought of as a circle with radius $\sqrt{1-\epsilon^2}$

in th ϵ). We think of
it as $(0, 0, 1)$ and not
cont of the curve C_ϵ
is en th respect to the
origi



This is not possible, since the winding number is a topological invariant.

9.2.5 The ham sandwich theorem

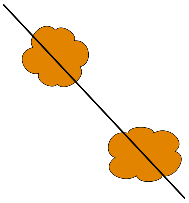
Theorem 9.5 (Ham sandwich theorem, Hugo Steinhaus and Stefan Banach, 1938). *Let $A_1, \ldots, A_n$ be measurable sets of finite volume in $\mathbb{R}^n$. Then there exists a hyperplane $H \subseteq \mathbb{R}^n$ such that for each i, half of A_i (by volume) lies on each side of H.*

This is called the *ham sandwich theorem* because you could think of $n = 3$, and think of A_1 as bread, A_2 as ham, and A_3 as cheese. It then says that no matter how disorderly you were in arranging the three ingredients, you can make a single straight cut that will divide all three of them in half.

Hannah Fry, on Numberphile

As silly as this sounds, it turns out to have interesting applications.

The two-dimensional case is also known as the *pancake theorem*: Any two pancakes can be bisected with a single cut.



Proof. A hyperplane in $\mathbb{R}^n$ is given by an equation of the form

$$u \cdot x = c,$$

or more explicitly

$$u_1 x_1 + \ldots + u_n x_n = c,$$

where $u \in S^{n-1}$ (so $u \in \mathbb{R}^n$ and $\|u\| = 1$) and $c \in \mathbb{R}$. The hyperplane is perpendicular to u. The complement of the hyperplane has two connected components, $\{u \cdot x > c\}$ and $\{u \cdot x < c\}$. If $A \subseteq \mathbb{R}^n$, the hyperplane separates

$$A^+_{u,c} = \{x \in A \ : \ u \cdot x > c\}$$

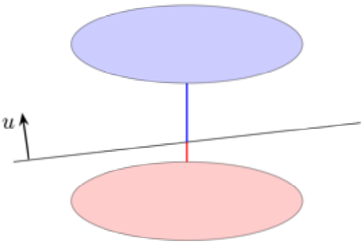
from

$$A^-_{u,c} = \{x \in A \ : \ u \cdot x < c\}\,.$$

Given $u \in S^{n-1}$, the volume of $A^+ u, c$ continuously decreases from $\text{vol}(A)$ to 0 as c varies from $-\infty$ to $+\infty$. By the intermediate value theorem, there exists a number c so that

$$\text{vol}(A^-_{u,c}) = \frac{1}{2}\text{vol}(A). \tag{9.17}$$

In extreme cases, there should actually be a range of values of c satisfying this equation, for example ⌉



Here A consists of two elliptical regions connected by a single line. Any line with the property that one of the two elliptical regions lies on one side, the other on the other, bisects A. It is, however, always true that the set of all c with (9.17) is a closed interval. We denote the midpoint of this interval by $\gamma_A(u)$. So

$$u \cdot x = \gamma_A(u)$$

bisects the region A. We note that

$$u \cdot x = c \quad \Leftrightarrow \quad -u \cdot x = -c$$

and therefore

$$\gamma_A(-u) = -\gamma_A(u).$$

Our goal is to prove that there exists an $u \in S^{n-1}$ such that

$$\mathrm{vol}\left(A_{i,u,\gamma_{A_n}(u)}\right) = \frac{1}{2}\mathrm{vol}(A_i) \tag{9.18}$$

for all i with $1 \leq i \leq n-1$. In other words, a $u \in S^{n-1}$ so that a hyperplane perpendicular to u that bisects A_n also bisects $A_1, \ldots, A_{n-1}$. We can also write (9.18) as

$$\mathrm{vol}\left(A_{i,u,\gamma_{A_n}(u)}\right) - \frac{1}{2}\mathrm{vol}(A_i) = 0.$$

Define

$$G: \ S^{n-1} \to \mathbb{R}^{n-1}$$

by

$$G(u) = \left(\mathrm{vol}\left(A_{i,u,\gamma_{A_n}(u)}\right) - \frac{1}{2}\mathrm{vol}(A_i)\right)_{1 \leq i \leq n-1}.$$

We want to prove that $G(u) = 0$ for some $u \in S^{n-1}$.

We note that G is continuous, and for all $u \in S^{n-1}$,

$$\begin{aligned}
G(-u) &= \left(\mathrm{vol}\left(A_{i,-u,\gamma_{A_n}(-u)}\right) - \frac{1}{2}\mathrm{vol}(A_i)\right)_{1 \leq i \leq n-1} \\
&= \left(\mathrm{vol}\left(A_{i,-u,-\gamma_{A_n}(u)}\right) - \frac{1}{2}\mathrm{vol}(A_i)\right)_{1 \leq i \leq n-1} \\
&= \left(\mathrm{vol}(A_i) - \mathrm{vol}\left(A_{i,u,\gamma_{A_n}(u)}\right) - \frac{1}{2}\mathrm{vol}(A_i)\right)_{1 \leq i \leq n-1} \\
&= \left(\frac{1}{2}\mathrm{vol}(A_i) - \mathrm{vol}\left(A_{i,u,\gamma_{A_n}(u)}\right)\right)_{1 \leq i \leq n-1} = -G(u).
\end{aligned}$$

The assertion that $G(u) = 0$ for some $u \in S^{n-1}$ now follows from the Borsuk-Ulam theorem. □

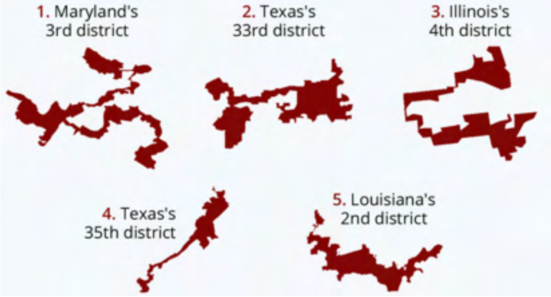
It is possible to deduce the Borsuk-Ulam theorem from the ham sandwich theorem as well, but I won't discuss that direction here. So the ham sandwich theorem is an alternative way of stating the Borsuk-Ulam theorem.

9.2.6 Partisan gerrymandering

Imagine a state in which part A gets 60% of the vote, and party B gets 40%. If the congressional districts are drawn such that in each district, party A gets 60%

of the vote, then party A gets *all* seats in the House of Representatives for the state. Deliberately drawing district boundaries in a way that favors one party over the other is called *partisan gerrymandering.* Astonishingly, it is legal in the United States, according to the 2019 *Rucho vs. Common Cause* decision by the United States Supreme Court. (The vote was 5:4.)

It is often said that gerrymandered districts can be recognized from their complicated shapes:


1. Maryland's 3rd district
2. Texas's 33rd district
3. Illinois's 4th district
4. Texas's 35th district
5. Louisiana's 2nd district

The *Polsky-Popper test*, for example, considers a district suspect if the ratio

$$\frac{4\pi A}{P^2}$$

is small, where A denotes the area, and P the length of the perimeter. Note that for a circular district of radius r, we would have $A = \pi r^2$ and $P = 2\pi r$, and therefore $\frac{4\pi A}{P^2} = 1$. For any other shape, $\frac{4\pi A}{P^2} < 1$.

The pancake theorem casts doubt on this idea. Partisan gerrymandering does not require complicated district boundaries. In fact, if you think of areas dominated by party A as one pancake, and areas dominated by party B as another, you can find a single straight cut that bisects both. Therefore you could create two districts so that in each of them, A would get 60% of the vote. If you wanted three districts, you could cut one of the two districts in half, again with a single straight cut, again in such a way that A would get 60% of the vote in both of the two resulting districts. And so on. Party A can win every single district, with very simple straight district boundaries.

Exercises

9.1. (easy) **A simple application of Brouwer's theorem.** Explain why the following system of equation must have a solution.

$$\begin{aligned} x &= \sin(x^2 - y^4), \\ y &= \frac{\sin(x+y) - \cos(x^2 - \sin(y))}{2}. \end{aligned}$$

(Don't work hard. If you have to work hard, you haven't understood.)

9.2. (easy but creative) **Further simple applications of Brouwer's theorem.** Can you come up with nicer, more surprising, more interesting examples of systems of nonlinear equations that Brouwer's theorem applies to?

9.3. (easy) **Trying to create a counterexample to Brouwer's theorem.** Let me try to create a counterexample to Brouwers' theorem. I will try to define a continuous function

$$\varphi : \; B^2 \to B^2$$

without any fixed points. I will make it a composition of three functions. First, I apply a rotation by 90 degrees:

$$\varphi_1(x, y) = (-y, x).$$

That already has almost no fixed point. Well, it has one, at $(0, 0)$. But then I'll shrink the disk:

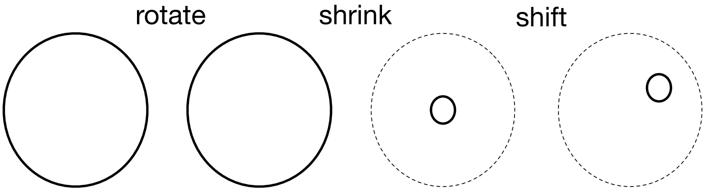
$$\varphi_2(\phi_1(x, y)) = \left(-\frac{y}{10}, \frac{x}{10}\right).$$

So the composition maps the unit disk into a disk of radius $1/10$ centered at the origin. The only fixed point is still $(0, 0)$. But now I'll get rid of that one, by shifting the small disk with radius $1/10$ away from the origin:

$$\varphi_3(x, y) = (x + 1/2, y + 1/2).$$

So

$$\varphi(x, y) = \varphi_3(\varphi_2(\varphi_1(x, y))).$$



Brouwer tells us that we can't escape. There's still a fixed point. Can you find it?

9.4. (medium) **Uniqueness of fixed points.** Prove in Problem 3 that there is *exactly* one fixed point. Hint: If (x_1, y_1) and (x_2, y_2) are fixed points, think about the distance between them.

9.5. (easy) **Infinitely many antipodal points with equal temperature.** Let $f : S^2 \to \mathbb{R}$ be a continuous function. Prove that there are *infinitely many* $x \in S^2$ with $f(x) = f(-x)$. In meteorological terms, there are infinitely many pairs of antipodal points on earths where the temperatures are the same. (Hint: The argument we gave for the equator applies to any great circle on earth. However, be a little careful: Great circles intersect.)

9.6. (easy) **Review of eigenvectors.** In Problem 7, you'll show that a matrix whose entries are all positive has an eigenvector with components ≥ 0, associated with a positive eigenvalue. This preliminary problem is intended to remind you of eigenvalues and eigenvectors.

Verify that

$$\left[\begin{array}{cc} 1 & 2 \\ 2 & 1 \end{array}\right]\left[\begin{array}{c} 1 \\ 1 \end{array}\right] = 3\left[\begin{array}{c} 1 \\ 1 \end{array}\right].$$

So the matrix $\left[\begin{array}{cc} 1 & 2 \\ 2 & 1 \end{array}\right]$ has the eigenvector $\left[\begin{array}{c} 1 \\ 1 \end{array}\right]$. The matrix has all positive entries, and there is an eigenvector in which all components are ≥ 0. (In this example, they are even strictly greater than 0.) Show that this matrix also has an eigenvector in which one entry is positive, and the other is negative. Don't try to recall any method for computing eigenvectors. Just recall what "eigenvector" means, then figure it out by thinking.

9.7. (medium) **Non-negative eigenvectors of positive matrices.** Suppose that A is an $n \times n$-matrix:

$$A = [a_{ij}]_{1 \leq i,j \leq n} \, .$$

Assume that all a_{ij} are (strictly) positive. Prove that there exists an eigenvector of A with non-negative components, associated with a positive eigenvalue. That is, there exists a non-zero vector

$$x = [x_i]_{1 \leq i \leq n} \in \mathbb{R}^n$$

with $x_i \geq 0$ for all i such that

$$Ax = \lambda x$$

for some real number $\lambda > 0$.

Hint: Let

$$P = \left\{ x \in S^{n-1} \ : \ x_i \geq 0 \text{ for all } i \right\}.$$

For $x \in P$, define

$$f(x) = \frac{Ax}{\|Ax\|}.$$

(Why can $\|Ax\|$ not be zero?) Explain why f has a fixed point by Brouwer's theorem, and why this proves the assertion.

Hint within the hint: The set $P \subseteq \mathbb{R}^n$ is not topologically equivalent to B^n. But for $x \in P$, we have

$$x_n = \sqrt{1 - \sum_{i=1}^{n-1} x_i^2},$$

so f can be interpreted as a function from

$$B = \left\{ (x_1, \ldots, x_{n-1}) \in \mathbb{R}^{n-1} \ : \ x_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{n-1} x_i^2 \leq 1 \right\}$$

into itself, and this set is topologically equivalent to B^{n-1}.

9.8. (hard, have to understand some real analysis here) **Non-negative eigenvectors of non-negative matrices.** Suppose that in the previous problem, all a_{ij} are ≥ 0, but not necessarily > 0. Prove that there is still an eigenvector $x = [x_i]_{1 \leq i \leq n}$ with $x_i \geq 0$ for all i. Hint: Replace a_{ij} by $a_{ij} + 1/k$, where $k \in \mathbb{N}$. Then, by the previous problem, there exists an eigenvector with non-negative components. Let's call it $x^{(k)}$. The sequence of vectors $x^{(k)}$ does not necessarily have a limit, but it has a subsequence that has a limit.

9.9. (medium) **A variation on the proof of the Borsuk-Ulam theorem.** (thanks to Nicholas Cummings). Use winding numbers to prove the following statement. If f and g are continuous functions defined on

$$\left\{(x, y) \ : \ x^2 + y^2 \leq 1\right\}$$

with

$$f(x, y) = -f(-x, -y) \ \text{ and } \ g(x, y) = -g(-x, -y) \quad \text{for } x^2 + y^2 = 1,$$

then there exists an (x, y) with $f(x, y) = g(x, y) = 0$. Explain why this implies the Borsuk-Ulam theorem.

Chapter 10

Why you cannot smoothly comb a hedgehog

10.1 The hedgehog theorem

A hedgehog looks like this:

By Gibe, CC BY-SA 3.0, via Wikimedia Commons

However, when it feels threatened, it rolls itself into a ball:

By SumandaMaritz. CC BY-SA 3.0, via Wikimedia Commons.

The theorem of this chapter says that once it's rolled up into a ball, you cannot comb it. That is, you cannot get all its spikes to lie flat in continuously varying directions. Or, mathematically:

Theorem 10.1 (Hedgehog Theorem, Henri Poincaré 1885). *Let V be a continuous vector field on S^2. That is, for every $x \in S^2$, let $V(x) \in \mathbb{R}^3$ be a vector that is tangential to S^2 at x. Then $V(x) = 0$ for some $x \in S^2$.*

Since the wind blows parallel to the earth's surface, this also implies that at any moment, there is at least one point on earth where there is no wind. Go **here** to see that there are usually many such points on earth. The theorem is called the *hairy ball theorem* in English, but I prefer the German *Igelsatz*, which means *hedgehog theorem*. One of its proofs uses the notion of *rotation number*, which I'll explain next.

10.2 Rotation numbers

10.2.1 Rotation number of a closed C^1 curve in the plane with respect to a C^0 vector field

Suppose that

$$\gamma(t) = (x(t), y(t)), \quad 0 \le t \le T,$$

defines a curve in the plane with

$$\gamma(0) = \gamma(T).$$

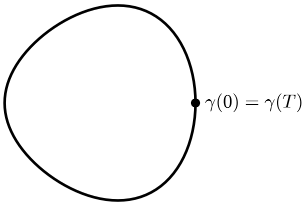
We assume that $x(t)$ and $y(t)$ are continuously differentiable functions, that is, that $\gamma'(t) = (x'(t), y'(t))$ is well-defined and continuous for $0 \le t \le T$. This is what C^1 stands for in the section title; in general, C^1-functions are differentiable functions with a continuous derivative. Think of $\gamma(t)$ as the location of a particle at time t. Then $\gamma'(t)$ is the velocity of the particle. We assume

$$\gamma'(t) \neq 0 \quad \text{for all } t \in [0, T]$$

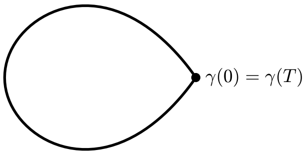
and

$$\gamma'(0) = \gamma'(T)$$

So our curve coul



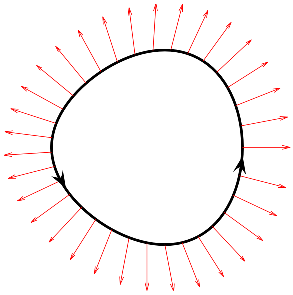
but not like this:



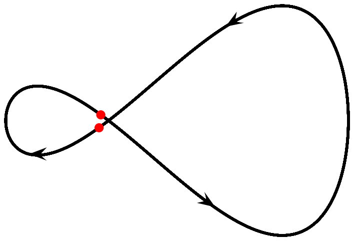
For every point (x, y) on the curve γ, assume that

$$(u(x, y), v(x, y))$$

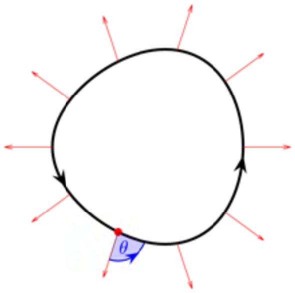
is a vector, which we think of as an arrow attached to the point (x, y).



We assume that the dependence of (u, v) on (x, y) is continuous. The symbol C^0 in the section heading stands for "continuous". Notice that the assumption is that $u(x, y)$ and $v(x, y)$ depend continuously on (x, y), not only that $u(x(t), y(t))$ and $v(x(t), y(t))$ depend continuously on t. In the following picture, for instance, since the two red dots are near each other, the values of u and v in those points should be close to each other, even though the values of t corresponding to these two points are very different.



For each $t \in [0, T]$ we now let θ be the angle by which you have to rotate the vector (u, v) attached to $(x(t), y(t))$ in the counter-clockwise direction to turn it into the direction of $\gamma'(t)$.



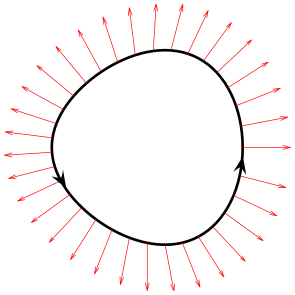
Of course, θ is only defined up to adding integer multiples of 2π. However, if we want θ to change *continuously* as we move once around the curve, then $\theta(T)-\theta(0)$ is well-defined. Since we assume $(x'(0), y'(0)) = (x'(T), y'(T))$, $\theta(T)-\theta(0)$ must be an integer multiple of 2π.

Definition 10.2. *The integer*

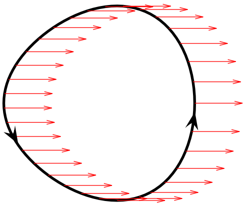
$$\frac{\theta(T)-\theta(0)}{2\pi}$$

is called the rotation number *of the curve with respect to the given vector field on the curve.*

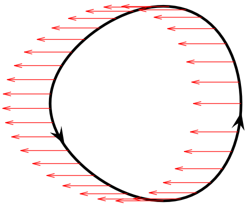
For instance, here the rotation number is zero:



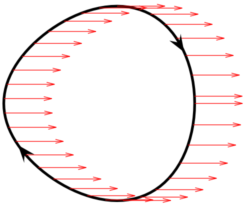
And here it is



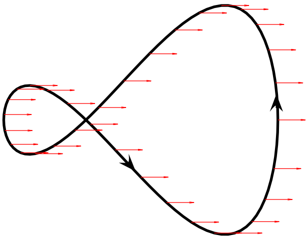
Reversing all the directions of the arrows does not change the rotation number, it is still 1:



More generally, rotating all the vectors by the same angle θ_0 just replaces θ by $\theta - \theta_0$, and therefore does not affect $\theta(T) - \theta(0)$ at all. On the other hand, if we reverse the direction of the curve, the rotation number changes sign. For instance, here the rotation nu

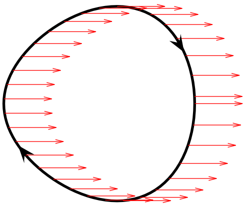


And in the following pi

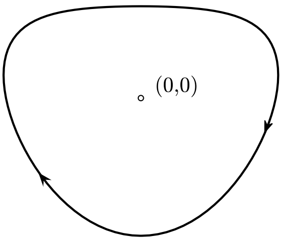


10.2.2 Rotation number of a closed C^1 curve in the plane with respect to a constant vector field

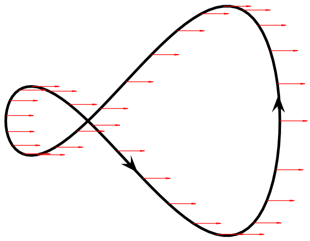
When all vectors (u, v) are the same, they might as well be all equal to $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, since rotation of all vectors (u, v) by the same angle does not change the rotation number. Then the rotation number of a curve is the winding number of the curve $\gamma'(t) = (x'(t), y'(t))$ xample:



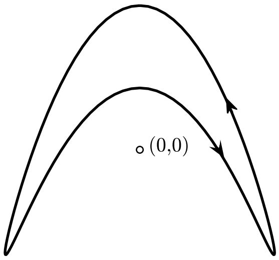
the curve $\gamma'(t) = (x'(t), y'(t))$ looks as follows.



For this example:



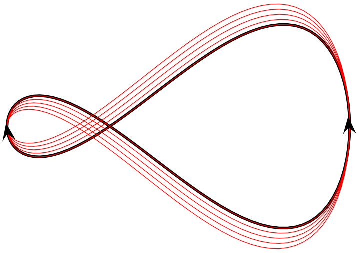
the curve $\gamma'(t) = (x'(t), y'(t))$ looks like this:



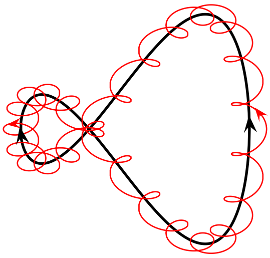
Another way of saying the same thing: We track how $\gamma'(t)$ changes as we move around the curve once, and count how often it rotates in the counter-clockwise direction. That number is the rotation number with respect to a constant vector field.

10.2.3 Invariance under smooth deformations

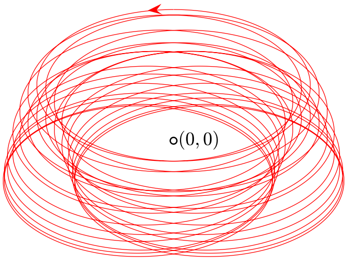
Now suppose the curve γ is continuously deformed, for example like this:



If the vectors (u, v) are defined on all the deformed curves, depend on location continuously, and are never zero, then the rotation number must change continuously, and therefore — since it is an integer — not at all. There is one caveat. The following perturbatio [illegible] esn't count as small:



The red curve does not have the same rotation number as the black one with respect to the vectors $(u, v) = (1, 0)$. The assumption that we need here is that not only γ, but also the velocity vector γ' changes gradually, continuously. In fact, for the crazy wiggly red curve shown above, $\gamma'(t) = (x'(t), y'(t))$ as a curve in the plane looks like this:



Its winding number with respect to $(0, 0)$ is not zero.

Theorem 10.3 (invariance of rotation numbers under smooth deformation). *If the closed curve γ is deformed continuously in such a way that γ' also changes continuously, and the vectors (u, v) are defined and non-zero on all deformed curves as well and depend continuously on location, then the rotation number of the curve with respect to the vector field remains the same.*

10.2.4 Rotation number of a closed C^1 curve on the unit sphere with respect to a C^0 vector field

Consider now a curve γ on the sphere S^2:

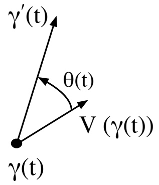
$$\gamma(t) = (x(t), y(t), z(t)), \quad 0 \le t \le T,$$

and

$$x(t)^2 + y(t)^2 + z(t)^2 = 1 \quad \text{for all } t \in [0, T].$$

Assume that $x(t)$, $y(t)$, and $z(t)$ are C^1, that $\gamma(0) = \gamma(T)$, and that $\gamma'(0) = \gamma'(T)$.

Suppose that $V = V(x) \in \mathbb{R}^3$ is defined for all x that lie on the curve γ, tangential to the sphere, continuously dependent on x, and $V(x) \neq 0$ for all x that lie on the curve. For each t, we have a picture like this in the tangent plane to the sphere at $\gamma(t)$:



The angle θ is still defined up to additive integer multiples of 2π. We can therefore define the rotation number of γ with respect to V precisely as before.

10.3 Proof of the hedgehog theorem

Suppose now that a continuous tangential vector field $V(x)$, $x \in S^2$, is nonzero everywhere on the sphere. Consider the curves

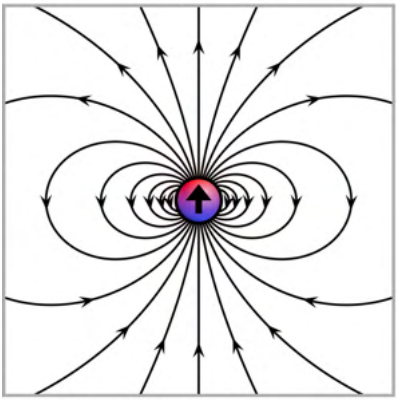
$$\gamma_h(t) = (\cos t, \sin t, h), \quad -1 \leq h \leq 1.$$

For $h \approx 1$, the curve encircles the point $(0, 0, 1)$ in the counter-clockwise direction, and the vector field is approximately $V(0, 0, 1)$ everywhere on the curve. This implies that the rotation number of γ_h with respect to V is $+1$ for $h \approx 1$. For $h \approx -1$, the curve encircles the point $(0, 0, -1)$ in the clockwise direction, and the vector field is approximately $V(0, 0, -1)$ everywhere on the curve. This implies that the rotation number of γ_h with respect to V is -1 for $h \approx -1$. But this contradicts Theorem 10.3, and this contradiction proves the hedgehog theorem.

10.4 Various other combing tasks

10.4.1 Combing the hedgehog so only one spike sticks out

Is it possible for a continuous vector field V on S^2 to be zero in only one point? The answer is yes. Here is an example. At the "north pole" (0,0,1), define the vector field to be the tangent vectors of the following curves:

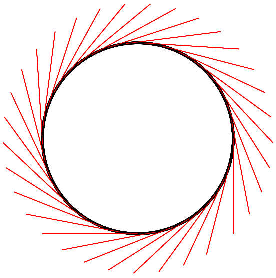


By Geek3, CC BY-SA 4.0, via Wikimedia Commons

(These are the electrical field lines of a dipole.) It is clear now that this vector field can be extended to all of S^2 in such a way that the north pole remains the only place where $V = 0$.

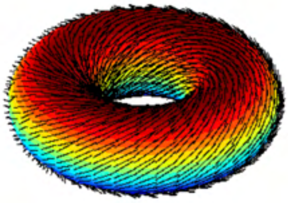
10.4.2 Combing a circle

Of course we can comb a hairy circle:



10.4.3 Combing a doughnut

There are multiple ways to comb a hairy doughnut.

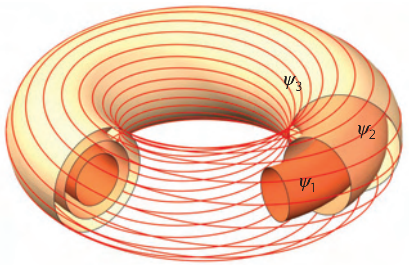


So the doughnut (torus) does not have the problem that the hedgehog (sphere) has.

10.5 Applications

10.5.1 Fusion reactors

Fusion reactors are built in the shape of a torus, not of a sphere. The reason is that a very hot plasma (10 times hotter than the core of the sun) has to be confined to the reactor container. This is done by magnets, and the magnetic field lines must be everywhere non-zero and parallel to the container walls. By the hedgehog theorem, this rules out a spherical design. A torus, on the other hand, is possible. The tokamak reactor, which is one of the leading candidates for a long-term solution of the world's energy problem, is torus-shaped.

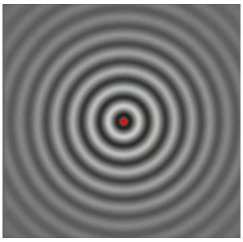


Fasoli et al, Naturę Physics 2016

The red lines are magnetic field lines. A hairy doughnut with its hair aligned with these field lines would be combed perfectly.

10.5.2 Isotropic antennas

An *isotropic antenna* would be an antenna that radiates radio waves of the same intensity in all directions. (Real-life antennas have preferred directions.)



Chetvorno, via Wikimedia Commons

If there were such an antenna, it would have, on a sphere surrounding the antenna, electrical field lines that are everywhere tangential to the sphere, and continuously varying. This is impossible by the hedgehog theorem. We conclude:

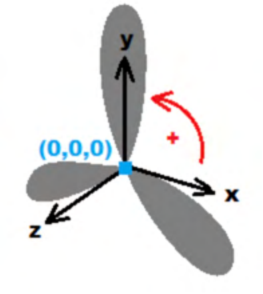
Isotropic antennas are not possible.

10.5.3 Rotating a three-dimensional image on the computer

Suppose that

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

is a unit vector. We want to rotate an image around the axis in the direction of z.



The rotated image is easy to compute if we can calculate unit vectors

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

so that x, y, and z are orthogonal to each other. In fact, finding x would be enough: Then we can define $y = x \times z$ (where $\times$ is the cross product that you heard about in Calculus 3). So here is our task:

Given a unit vector $z \in \mathbb{R}^3$, find a unit vector $x \in \mathbb{R}^3$ perpendicular to z.

Here is a way of stating the hedgehog theorem that makes clear why it is relevant to this task.

Theorem 10.4. *There is no* continuous *function*

$$F: \ S^2 \to S^2$$

so that $F(z)$ is perpendiular to z for all $z \in S^2$.

So any method for finding $x \perp z$, given z, will need to be a little ugly, involving cases. For instance:

$$x = \frac{1}{z_2^2 + z_3^2} \begin{bmatrix} 0 \\ -z_3 \\ z_2 \end{bmatrix}$$

if $z_1 \neq 1$ and therefore $z_2^2 + z_3^2 > 0$, and

$$x = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

if $z_1 = 1$.

Exercises

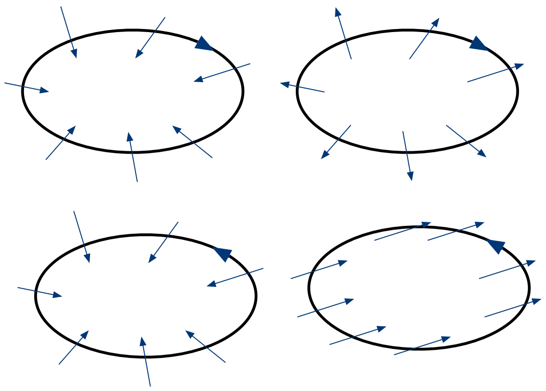
10.1. (easy) **Index of a vector field with respect to a curve in the plane.** Bewlow is a list of three somewhat similar notions. We have talked about the first two already, but not about the third.

(1) The *winding number* of a closed continuous curve with respect to a point not on the curve. [We used this to prove Brouwer's fixed point theorem in two dimensions, and the Borsuk-Ulam theorem in two dimensions.]

(2) The *rotation number* of a closed continuously differentiable curve with respect to a continuous vector field that is nowhere zero on the curve. [We used this to prove the hedgehog theorem]

(3) The *index* of a closed continuous curve with respect to a continuous vector field that is nowhere zero on the curve.

The index is computed by counting how many full rotations the vector undergoes in the counter-clockwise direction as you go around the curve once. In the following examples, what are the indices?

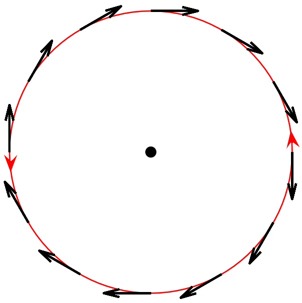


10.2. (medium) **Index of an odd vector field on S^1.** Suppose that you are given a continuously varying vector field (u, v) defined on S^1 that is *odd*, meaning

$$(u(-x,-y), v(-x,-y)) = -(u(x,y), v(x,y)),$$

and everywhere non-zero. Explain why the index of the unit circle, traversed once counter-clockwise, must be an odd integer.

10.3. (easy) **Index of a tangential vector field on S^1.** In the following picture, what is the index?



Does it matter whether we reverse the direction in which the circle is traversed and/or the direction of the vectors?

10.4. (medium) **Index of an isolated zero of a continuous vector field in the plane.** Suppose that

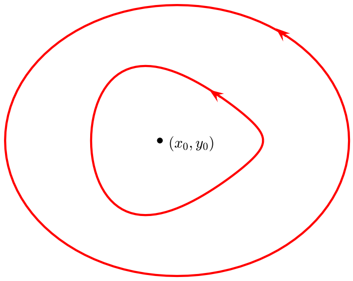
$$(u(x,y), v(x,y))$$

is a continuous vector field in the plane. A point (x_0, y_0) is called a *zero* of the vector field if

$$(u(x_0,y_0), v(x_0,y_0)) = (0,0).$$

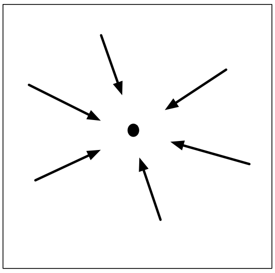
It is called an *isolated zero* if there exists some radius $r > 0$ so that there is no other zero within distance r of (x_0, y_0).

Now consider two continuous closed curves that wind around (x_0, y_0) once in the counter-clockwise direction. Assume that both have the property that there is no zero of the vector field on the curve, nor in the region enclosed by the curve.

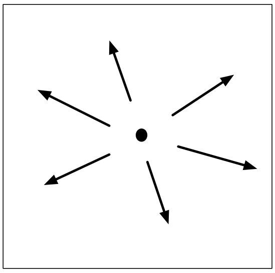


Explain why then the indices of the two curves with respect to the given vector field must be the same. This common index, the index of all continuous closed curves winding around (x_0, y_0) once in the counter-clockwise direction, and around no other zero of the vector field, is called the *index of the isolated zero* (x_0, y_0).

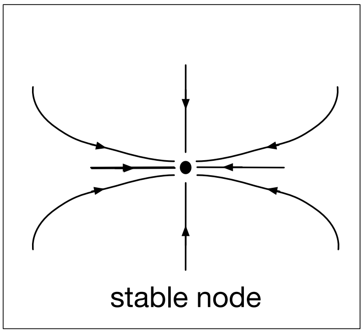
10.5. (easy) **Indices of sinks, sources, nodes, and saddles.** Here are examples of vector fields. (In the lower two examples, I plotted curves following the direction of the vector fields, instead of the vectors themselves, because it is easier to visualize the vector field that way.) You may remember these examples if you took a class on ordinary differential equations.



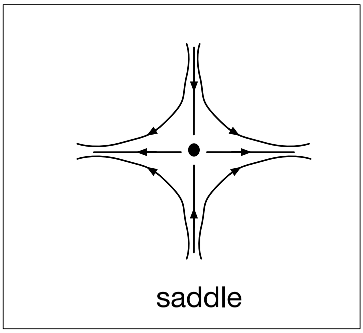
sink



source



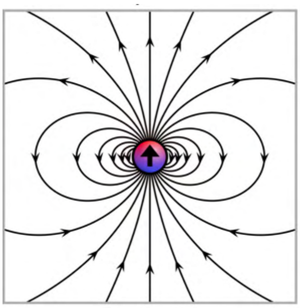
stable node



saddle

In each case, determine the index of the isolated zero at the center.

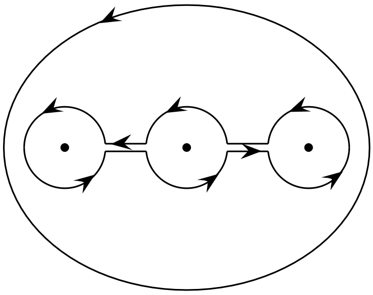
10.6. (easy) **Index of a dipole.** What is the index here?



By Geek3, CC BY-SA 4.0, via Wikimedia Commons

10.7. (medium) **An isolated zero of a vector field with index zero.** Give an example of an isolated zero of a vector field with index zero.

10.8. (medium) **Index of a curve surrounding several isolated zeros.** Given a continuous vector field in the plane, a curve that winds once around finitely many isolated zeros in the counter-clockwise direction has index equal to the sum of the indices of the zeros. Explain why this has to be true using the following suggestive picture.



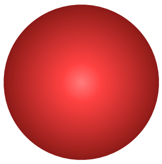
10.9. (medium) **Borsuk-Ulam theorem via indices.** Combine Exercises 10.8 and 10.2 to prove the statement of Exercise 9.9 (and therefore, the Borsuk-Ulam theorem) using the notion of index rather than the notion of winding number.

10.10. (medium, requires some understanding of differential equations) **Periodic and equilibrium solutions of two-by-two systems of ODEs.** Explain, using Problem 10.8: If f and g are continuously differentiable[15] functions $\mathbb{R}^2 \to \mathbb{R}$, and $(x(t), y(t))$ is a periodic solution of

$$\begin{aligned}\frac{dx}{dt} &= f(x,y),\\ \frac{dy}{dt} &= g(x,y),\end{aligned}$$

then the system has at least one equilibrium solution; that is, there is a point (x_0, y_0) with $f(x_0, y_0) = g(x_0, y_0) = 0$.

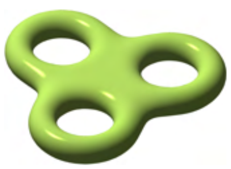
10.11. (medium) **Euler charactreristic.** Let's think about bounded smooth surfaces in $\mathbb{R}^3$ without boundary. (This is unnecessarily restrictive, but let's just focus on that simple case.) Here are some examples:



By Jahobr, CC0
via Wikimedia Commons

Oleg Alexandrov, Public Domain
via Wikimedia Commons

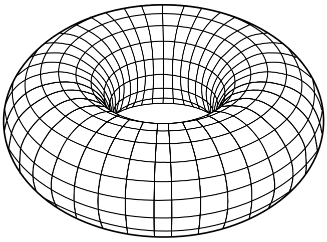
Oleg Alexandrov, Public Domain
via Wikimedia Commons



Oleg Alexandrov, Public Domain
via Wikimedia Commons

Any such surface has an *Euler characteristic*, which is the number, v, of vertices minus the number, e, of edges plus the number, f, of faces for any polygonal net covering the surface. Here is an example of such a net for a torus:

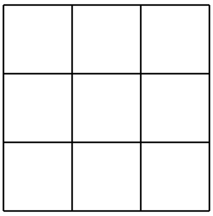
[15] I make this assumption to avoid issues concerning non-uniqueness of solutions.



By GYassineMrabetTalk, CC BY-SA 3.0,
via Wikimedia

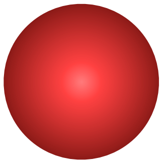
The number $v - e + f$ does not depend on which net you look at. (I am not explaining why that's true here.) In Section 1.5.1 we proved that for the sphere, the Euler characteristic is 2.

Can you see why for the torus, the Euler characteristic is zero? Hint: Start with the following



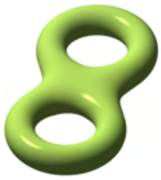
What is $v-e+f$ here? If you now bend this figure and glue the right edge to the left edge, you get a grid of rectangles covering a cylinder. Since you glued the right edge to the left, you lost three edges and four vertices. If you then bend the cylinder and glue the two edges together, you get a torus covered by a net of rectangles. What is $v - e + f$ now?

For examples such as

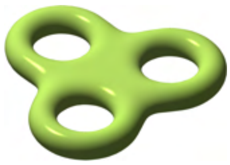


By Jahobr, CC0
via Wikimedia Commons

Oleg Alexandrov, Public Domain
via Wikimedia Commons



Oleg Alexandrov, Public Domain
via Wikimedia Commons



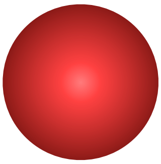
Oleg Alexandrov, Public Domain
via Wikimedia Commons

the number of "holes" is called the "genus". So the sphere has genus 0, the torus has genus 1, and the other two surfaces shown have genus 2 and 3. It can be shown that

$$v - e + f = 2 - 2g$$

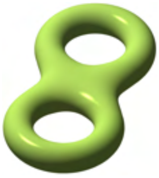
where g is the genus. Verify that this agrees with what we know about the sphere and the torus.

10.12. (easy) **Poincaré-Hopf theorem.** Suppose a continuous tangential vector field on one of the surfaces

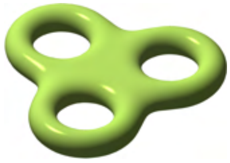


By Jahobr, CC0
via Wikimedia Commons

Oleg Alexandrov, Public Domain
via Wikimedia Commons



Oleg Alexandrov, Public Domain
via Wikimedia Commons



Oleg Alexandrov, Public Domain
via Wikimedia Commons

has only finitely many zeros. The *index* can be defined for a zero of a continuous tangential vector field on one of these surfaces just as it is defined in the plane. The *Poincaré-Hopf theorem* then states that the sum of all indices of the zeros equals the Euler characteristic. Explain why this again implies the hedgehog theorem, and why among the four surfaces above, the torus is the only one on which there exists a continuous tangential vector field without any zero.

This is why a fusion reactor can be a torus, but neither a sphere, nor a double or triple torus.

Chapter 11

Infinitely many different kinds of infinity

11.1 Sets of equal size

11.1.1 Definition

Definition 11.1. *If A and B are sets, we say that A and B are* of equal size *or* of equal cardinality*, in symbols*

$$|A| = |B|,$$

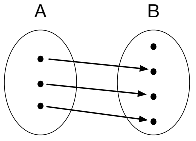
if there exists a bijection

$$\varphi: \ A \to B.$$

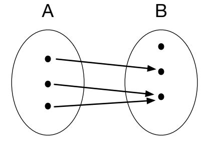
A "bijection" is a mapping that is both one-to-one (injective) and onto (surjective). "One-to-one" means

$$x \neq y \Rightarrow \varphi(x) \neq \varphi(y).$$

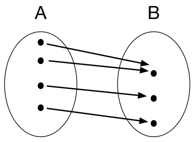
The images of two different elements are different. "Onto" means that for every $y \in B$, there is some $x \in B$ so that $\varphi(x) = y$.



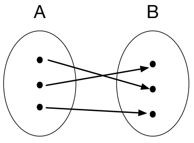
one-to-one, but not onto



neither one-to-one nor onto



onto but not one-to-one



a bijection

When the sets are finite, "of equal size" of course means exactly what you would have thought it should mean. They have equally many elements. But when the sets are infinite, Definition 11.1 still applies and leads to interesting conclusions.

11.1.2 Countable sets

Definition 11.2. *A set is called* countably infinite *if it is of the same size as the set $\mathbb{N}$ of natural numbers. A set is called* countable *if it is either finite or countably infinite, but not empty.*

Lemma 11.3. *A set A is countable if and only if there exists a sequence $x_1, x_2, x_3, \ldots$ such that*

$$\{x_1, x_2, x_3 \ldots\} = A.$$

I leave it to you to prove this lemma to yourself. So A is countable if and only if you can construct a sequence that sweeps through all of A.

Proposition 11.4. *The set $\mathbb{Z}$ of integers (positive or negative) is countably infinite.*

Proof.

$$\mathbb{Z} = \{0, -1, 1, -2, 2, -3, 3, -4, 4, \ldots\}.$$

So there you have it, a sequence that sweeps through all of $\mathbb{Z}$. □

Naively, you might have thought that $\mathbb{Z}$ is somehow "twice as big" as $\mathbb{N}$: When you list $\mathbb{N}$,

$$1, 2, 3, \ldots,$$

you are leaving out the negative integers (and zero). But that's misleading. The fact that you can create a sequence that *does not* run through all integers doesn't mean that you can't create one that does. In fact, we showed that you *can* create one that does.

Proposition 11.5. *The set of all* rational *numbers in $[0, 1)$ is countably infinite.*

Proof. A rational number is a fraction of two integers. The ones in $[0, 1)$ are

$$0, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \ldots$$

You will see some repetition in the sequence that I created: $\frac{2}{4}$ is the same as $\frac{1}{2}$, and if you go a little bit further down the line, you will find $\frac{2}{6}$, which is also $\frac{1}{3}$, etc. But Definition 11.3 said nothing about repetitions being prohibited. If you don't like repetitions, you can simply erase numbers that have already appeared from the list, keeping only the first appearance of each number on the list. □

Lemma 11.6. *Suppose that $A_1, A_2, A_3, \ldots$ are countable sets. Then the union*

$$A = \bigcup_{j \in \mathbb{N}} A_j$$

is countable.

Proof. Let $A_j = \{a_{1,j}, a_{2,j}, a_{3,j}, \ldots\}$. Then the elements of A are

$$a_{11}, \quad a_{21}, \quad a_{12}, \quad a_{31}, \quad a_{22}, \quad a_{13}, \quad a_{41}, \quad a_{32}, \quad a_{23}, \quad a_{14}, \quad a_{51}, \quad a_{42}, \quad \ldots$$

There you have it: A sequence that runs through all of A. I have written the argument as if all the A_j were infinite, but it is clear that there is no significant complication if any of them are finite. □

You can picture the argument as follows. Imagine the a_{ij} in a kind of infinite rectangular array, like this:

$$\begin{array}{cccccc}
a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & \cdots \\
a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & \cdots \\
a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & \cdots \\
a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & \cdots \\
a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

Then put them all into a single row, one diagonal at a time. First a_{11}:

$$\begin{array}{cccccc}
a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & \cdots \\
a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & \cdots \\
a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & \cdots \\
a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & \cdots \\
a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

Then a_{21} and a_{12}:

$$\begin{array}{cccccc}
\mathsf{a}_{11} & \mathsf{a}_{12} & \mathsf{a}_{13} & \mathsf{a}_{14} & \mathsf{a}_{15} & \cdots \\
\mathsf{a}_{21} & \mathsf{a}_{22} & \mathsf{a}_{23} & \mathsf{a}_{24} & \mathsf{a}_{25} & \cdots \\
\mathsf{a}_{31} & \mathsf{a}_{32} & \mathsf{a}_{33} & \mathsf{a}_{34} & \mathsf{a}_{35} & \cdots \\
\mathsf{a}_{41} & \mathsf{a}_{42} & \mathsf{a}_{43} & \mathsf{a}_{44} & \mathsf{a}_{45} & \cdots \\
\mathsf{a}_{51} & \mathsf{a}_{52} & \mathsf{a}_{53} & \mathsf{a}_{54} & \mathsf{a}_{55} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

Then a_{31}, a_{22}, and a_{13}:

$$\begin{array}{cccccc}
\mathsf{a}_{11} & \mathsf{a}_{12} & \mathsf{a}_{13} & \mathsf{a}_{14} & \mathsf{a}_{15} & \cdots \\
\mathsf{a}_{21} & \mathsf{a}_{22} & \mathsf{a}_{23} & \mathsf{a}_{24} & \mathsf{a}_{25} & \cdots \\
\mathsf{a}_{31} & \mathsf{a}_{32} & \mathsf{a}_{33} & \mathsf{a}_{34} & \mathsf{a}_{35} & \cdots \\
\mathsf{a}_{41} & \mathsf{a}_{42} & \mathsf{a}_{43} & \mathsf{a}_{44} & \mathsf{a}_{45} & \cdots \\
\mathsf{a}_{51} & \mathsf{a}_{52} & \mathsf{a}_{53} & \mathsf{a}_{54} & \mathsf{a}_{55} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

And so on.

We could not do the same thing row by row (we would never get done with even the first row!), nor coulumn by column. But this does not disprove the statement that A is countable. A failed attempt at proving that A is countable does not prove that A is not countable.

By the same reasoning, it is true that

$$\bigcup_{j \in J} A_j$$

is countable if the A_j are countable non-empty sets, and J is a countable set (not necessarily $\mathbb{N}$).

Countable unions of countable sets are countable.

Proposition 11.7. *The set $\mathbb{Q}$ of all rational numbers is countable.*

Proof.

$$\mathbb{Q} = \bigcup_{j \in \mathbb{Z}} \left(j + (\mathbb{Q} \cap [0,1)) \right).$$

The notation $j+(\mathbb{Q}\cap[0,1))$ stands for the set of all numbers of the form $j+x$, where x is a rational number in $[0,1)$. Since $\mathbb{Q}\cap[0,1)$ is countable, so is $j+(\mathbb{Q}\cap[0,1))$. Therefore $\mathbb{Q}$ is a countable union (recall $\mathbb{Z}$ is countable) of countable sets, so it is countable. □

This is striking. You would think that surely there are "far more" rational numbers than natural nubmers. But that's not correct. You can write all rational numbers in a single sequence.

11.2 Sets of unequal size

11.2.1 Cantor's first proof of the uncountability of $\mathbb{R}$

Theorem 11.8 (Georg Cantor, 1874). *The set $\mathbb{R}$ of real numbers is not countable.*

Proof. Some of you will now expect me to recount Cantor's famous "diagonal argument", which you may have learned in a previous class. I will instead give you Cantor's first proof, presented in 1874, which did not use the diagonal argument at all.

In his 1874 paper, Cantor proved the following statement. If $\{x_n\}_{n=1,2,3,...}$ is a sequence of real numbers and a, b are real numbers with $a<b$, then there exists a real number in (a,b) that is not x_n for any n. It's a stronger statement than the one demonstrated by the diagonal argument.

So let's assume that every number in (a,b) appears in the sequence $\{x_n\}$. Let a_1 be the earliest of the x_n (the one with smallest n) for which $a<x_n<b$. Let b_1 be the earliest of the x_n for which $a_1<x_n<b$. Let a_2 be the earliest of the x_n for which $a_1<x_n<b_1$. Let b_2 be the earliest of the x_n for which $a_2<x_n<b_2$. And so on. We construct in this way sequences

$$a<a_1<a_2<a_3<\dots$$

and

$$b>b_1>b_2>b_3>\dots$$

so that all the a_i are strictly smaller than all the b_i.

Now we need a lemma, the proof of which will be postponed. The lemma is:

$$x_n\notin(a_n,b_n).$$

Let's first see how that completes our proof. Since the a_i are increasing and bounded above, they have a limit. Call that limit x:

$$x=\lim_{i\to\infty}a_i.$$

Similarly, the b_i are decreasing and bounded below, so they have a limit y:

$$y=\lim_{i\to\infty}b_i.$$

We have, for all i:

$$a_i < x \leq y < b_i.$$

So $x \in (a_i, b_i)$ for all i. But $x = x_n$ for some n, by our initial assumption, so $x \notin (a_n, b_n)$ by the yet-to-be-proved lemma. This contradiction proves the assertion. □

Lemma 11.9. *$x_n \notin (a_n, b_n)$ for all n in the preceding proof.*

Proof. We will use induction. First, we have to show $x_1 \notin (a_1, b_1)$. If x_1 were in (a_1, b_1), it would be the earliest member of the sequence in (a, b), and therefore $a_1 = x_1$, not $a_1 < x_1$.

Next, assume that $n \geq 2$, and $x_k \notin (a_k, b_k)$ is already known for $1 \leq k \leq n-1$. If x_n were in (a_n, b_n), it would be the earliest member of the sequence in (a_{n-1}, b_{n-1}), and therefore $a_n = x_n$, not $a_n < x_n$. □

Georg Cantor lived from 1845 to 1918. He was the first to realize that not all infinite sets are "of equal size". His ideas were highly controversial initially, among both mathematicians and philosophers. Their objections were not at all nonsensical. Many decades earlier, Gauss had argued that "infinity" is not something that "exists", but that it is merely a convenient way of speaking about limits. Leopold Kronecker called Cantor a "corrupter of youth". Henri Poincaré, one of the two greatest mathematicians of Cantor's time, said that the discussion of "infinite sets", as if they were objects that "exist", simply was not mathematics.

However, David Hilbert, the other of the two greatest mathematicians of the time, took Cantor's side, and that side prevailed eventually. Nowadays the distinction between countable and uncountable infinite sets is fundamental to many branches of mathematics, for instance to analysis and therefore to probability theory and statistics.

The attacks on his ideas weighed heavily on Georg Cantor. He had a severe bouts of depression, leading to repeated stays in hospitals and sanatoriums. He died of a heart attack in 1918, during the First World War, in poverty.

11.2.2 Inequalities among cardinalities of sets

"Cardinality" is the word that set theorists use for "size".

Definition 11.10. *If A and B are non-empty sets, then we say that B is at least as large as A, in symbols*

$$|A| \leq |B|,$$

if there exists a mapping

$$\varphi: \; A \to B$$

that is one-to-one (injective).

Theorem 11.11. *Let A and B be non-empty sets. There exists an injective mapping*

$$\varphi: \; A \to B$$

if and only if there exists a surjective mapping

$$\psi: \; B \to A.$$

Proof. First, assume that there exists an injective mapping

$$\varphi: \; A \to B.$$

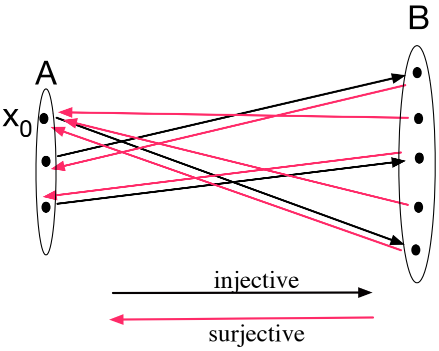
Let $x_0 \in A$, and define, for $y \in B$,

$$\psi(y) = \begin{cases} x \in A \text{ with } \varphi(x) = y & \text{if such an } x \text{ exists} \\ x_0 & \text{otherwise} \end{cases}$$

Then

$$\psi: \; B \to A$$

is surjective.



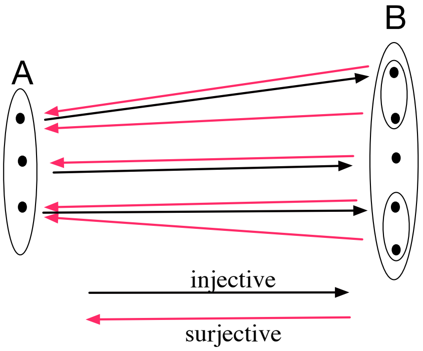
Now assume conversely that

$$\psi : \ B \to A$$

is surjective. For every $x \in A$, there is a set $S_x \subseteq B$ with $\psi(y) = x$ for all $y \in S_x$. Define $\varphi(x)$ to be one of the elements of S_x. Then

$$\varphi : \ A \to B$$

is injective.



□

The second half of this proof relied on the idea that we can select from every S_x one of its elements. In fact, the statement that the existence of a surjective map $\psi : \ B \to A$ implies the existence of an injective map $\varphi : \ A \to B$ is equivalent to the axiom of choice.

11.2.3 The Cantor-Schröder-Bernstein theorem

This theorem has three names. Georg Cantor for stating it without proof in 1895, Ernst Schröder for proving it in 1896, and Felix Bernstein for giving an alternative proof in 1897, at age 19, when he was a student in Cantor's seminar.

Theorem 11.12 (Cantor-Schröder-Bernstein Theorem). *Let A and B be non-empty sets. If $|A| \leq |B|$ and $|B| \leq |A|$, then $|A| = |B|$. That is, A and B are of the same size; there exists a bijection $\varphi : \ A \to B$.*

Proof. Suppose

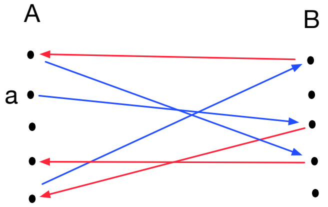
$$f : \ A \to B$$

and

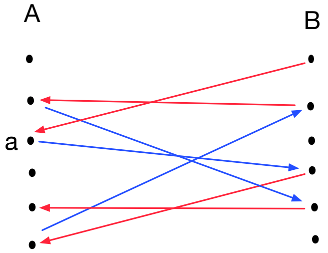
$$g : \ B \to A$$

are injections. Given $a \in A$, consider the sequence

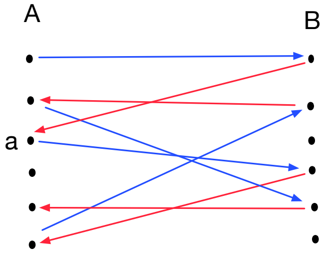
$$a, f(a), g(f(a)), f(g(f(a)), \ldots$$



The sequence might eventually return to a, but before returning to a, it cannot visit any element of A or B more than once. We extend the sequence to the left, first with $g^{-1}(a)$ (if a lies in the range of g):



then with $f^{-1}(g^{-1}(a)$ (if $g^{-1}(a)$ lies in the range of f):



and so on. On the left, the sequence ends either with a point in A, or with a point in B, or never.

Every point in A is in exactly one sequence of this sort, and so is every point in B. It is therefore enough to show that there is a bijection between the points of one such sequence in A, and the points of the sequence in B.

If the sequence ends with a point in A, then f is a bijection between the points of the sequence in A, and those in B. If it ends in a point in B, then g is a bijection between the points of the sequence in A, and those in B. If it never ends, either f or g will work. □

This proof was given by the Hungarian mathematician Julius König in 1906.

11.2.4 Strict inequalities among cardinalities of sets

Definition 11.13. *Let A and B be non-empty sets. We say that A is strictly smaller than B, in symbols*

$$|A| < |B|,$$

if there is an injective mapping from A into B, but no injective mapping from B into A.

11.2.5 The continuum hypothesis

We now know

$$|\mathbb{N}| = |\mathbb{Z}| = |\mathbb{Q}| < |\mathbb{R}|.$$

Continuum hypothesis: *There exists no set A with*

$$|\mathbb{N}| < |A| < |\mathbb{R}|.$$

Georg Cantor thought this was true, but couldn't prove it. On David Hilbert's famous list of open mathematical problems, published in 1900, proving or disproving the continuum hypothesis was problem 1. In 1940, Kurt Gödel (1906–1978) proved that standard set theory does not allow a proof that the continuum hypothesis is false. In 1963, Paul Cohen (1934–2007) proved that standard set theory does not allow a proof that the continuum hypothesis is true. He won the 1966 Fields Medal for this result. In combination, this yields the following theorem.

Theorem 11.14 (Kurt Gödel and Paul Cohen). *The continuum hypothesis is independent of the standard axioms of set theory (the "Zermelo-Frankel axioms" and the axiom of choice).*

Whether the continuum hypothesis is true or false is *undecidable.*

11.3 Infinitely many different kinds of infinity

11.3.1 The power set of A is greater than A

Definition 11.15. *If A is a set, then the* power set *of A is the set of all subsets of A. It is denoted by $P(A)$.*

For instance, if

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

then

$$P(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

Theorem 11.16. *If A is a set, than $|A| < |P(A)|$.*

Proof. Suppose that

$$f: \ A \to P(A)$$

were a bijection. Form the set

$$S = \{x \in A \ : \ x \notin f(x)\}.$$

Why does this make sense? Because $f(x) \in P(A)$, so $f(x)$ is a subset of A. We can ask whether x belongs to it. If no, then $x \in S$, but if yes, then $x \notin S$.

Note that $S \in P(A)$. Since f is a bijection, there is an $a \in A$ with $f(a) = S$. Is a an element of $f(a)$? If yes, then $a \notin S$. If no, then $a \in S$. But now remember that $f(a)$ is S. therefore what we just proved is

$$a \in S \ \Rightarrow \ a \notin S$$

and

$$a \notin S \ \Rightarrow \ a \in S.$$

This contradition proves that there cannot be a bijection $f: \ A \to P(A)$.

Of course, there is an *injective* map $f: \ A \to P(A)$. Just define $f(a) = \{a\}$.
□

11.3.2 How many different kinds of infinity are there?

Theorem 11.16 shows that there are infinitely many different kinds of infinity:

$$|\mathbb{N}| < |P(\mathbb{N})| < |P(P(\mathbb{N}))| < |P(P(P(\mathbb{N})))| < \dots$$

How many different cardinalities are there?

Many years ago, I asked a logician friend this question. He looked at me as if he had just spotted a mouse in the room, and said with an air of definite authority: "This question makes no sense." I left it there, because the question was not *that* urgent to me. Many years later, I will return to the question, and explain why the question indeed "makes no sense", and what that means.

We would plausibly formalize the question as follows. Let S be a set of sets with the following property. For every set A, there is at exactly one set $\tilde{A} \in S$ so that $|A| = |\tilde{A}|$. What is the cardinality of S? Instead of answering this question, we will argue that there cannot be any such set S.

Suppose S were indeed a set of the desired kind. Define

$$U = \bigcup \{A : A \in S\}.$$

Let $M = P(U)$. For any set $A \in S$, we have:

$$|A| \leq |U| < P(U) = M.$$

So there is no set in S with the same cardinality as M. This contradicts the definition of S, showing that there is no such set S, and therefore no way of making sense of the question "How many different kinds of infinity are there?"

Exercises

11.1. (easy) Show that the set of pairs of natural numbers is countable.

11.2. (easy) Show that the set of pairs of integers (positive, negative, or zero) is countable.

11.3. (easy after exercise 11.2) Show by induction that for any n, the set of n-tuples of natural numbers is countable.

11.4. (easy) Let $n \in \mathbb{N}$. We call a number $x \in \mathbb{R}$ *algebraic of degree* n if there exist integers

$$a_0, a_1, \ldots, a_{n-1}, a_n$$

with $a_n \neq 0$ so that

$$a_n x^n + a_{n-1} x^{-n-1} + \ldots + a_1 x + a_0 = 0.$$

and the same does not hold for any smaller n. Show that the algebraic numbers of degree 1 are the rational numbers.

11.5. (easy) Show that in exercise 11.4, the a_i can equivalently be taken to be *rational* numbers; that would not change the set of algebraic numbers of degree n.

11.6. (easy) Show that $\sqrt{2}$ is algebraic of degree 2.

11.7. (medium) Show that $\sqrt{\sqrt{2} + \sqrt{3}}$ is algebraic of degree at most 8.

11.8. (easy after exercise 11.3) For $n \in \mathbb{N}$, let A_n be the set of algebraic numbers of degree n. Show that A_n is countable.

11.9. (easy after exercise 11.8) A number $x \in \mathbb{R}$ is called *algebraic* if there exists an $n \in \mathbb{N}$ so that x is algebraic of degree n. Let A denote the set of all algebraic numbers. Show that A is countable. So A is a subset of $\mathbb{R}$ that includes many irrational numbers, yet it is still countable.

11.10. (easy after exercise 11.9) A number $x \in \mathbb{R}$ is called *transcendental* if it is not algebraic. Show that the set of all transcendental numbers is uncountable. So in this sense, on overwhelming majority of real numbers are transcendental.

11.11. Fun facts about algebraic and transcendental numbers:

(a) (too hard to serve as an exercise here) π and e are transcendental.

(b) (easy) If x is algebraic, then $-x$ is algebraic, and if also $x \neq 0$, then $1/x$ is algebraic.

(c) (hard) If x and y are algebraic, so are $x + y$ and xy.

(d) (easy using part (c)) Let x be algebraic and y transcendental. Then $x + y$ is transcendental, and if also $x \neq 0$, then xy is transcendental.

(e) (easy) if x^2 is algebraic, so is x.

(f) (easy using part (c)) If x is algebraic, so is x^2.

(g) (easy using parts (e) and (f)) x is transcendental if and only if x^2 is transcendental.

(e) (easy) If x and y are transcendental, then $x + y$ need not be transcendental.

(f) (open questions) Is $e + \pi$ transcendental? Is $e\pi$ transcendental? We don't know, and in fact we do not even know whether $e + \pi$ or $e\pi$ are *irrational.*[16]

11.12. (easy) Find a bijection $(0, 1) \to \mathbb{R}$, to prove that $(0, 1)$ and $\mathbb{R}$ have the same cardinality.

11.13. (easy) Find an injection $(0, 1) \to [0, 1]$, and find an injection $[0, 1] \to (0, 1)$. Therefore, by the Cantor-Schröder-Bernstein theorem, $[0, 1]$ and $(0, 1)$ have the same cardinality.

11.14. (medium) Explicitly construct a bijection $\varphi : [0, 1] \to (0, 1)$. Hint: First construct a bijection

$$\varphi_1 : [0, 1] - \mathbb{Q} \to (0, 1) - \mathbb{Q}.$$

That's easy, since the irrational numbers in $[0, 1]$ are precisely the irrational numbers in $(0, 1)$. Then construct a bijection

$$\varphi_2 : [0, 1] \cap \mathbb{Q} \to (0, 1) \cap \mathbb{Q}.$$

Then put the two together.

11.15. (easy) Let S be the set of all sequences $d_1 d_2 d_3 \ldots$ with $d_i \in \{0, 1\}$ for all i, $d_i = 0$ for infinitely many i, and $d_i \neq 0$ for at least one i. Construct a bijection $S \to (0, 1)$. (Hint: Identifiy the sequence $d_1 d_2 d_3 \ldots$ with the binary expansion $0.d_1 d_2 d_3 \ldots$.)

11.16. (easy after exercise 11.15) Show that there exists a bijective mapping from $(0, 1)$ into the square $(0, 1) \times (0, 1)$.

11.17. (easy after exercises 11.16 and 11.12) Show that $\mathbb{R}$ and $\mathbb{R}^2$ have the same cardinality.

11.18. (medium after 11.15) Let S be the set of all sequence of numbers in $(0, 1)$. Show $|S| = |\mathbb{R}|$.

11.19. (easy after 11.18 and 11.12) Show that the set of sequences of real numbers has the same cardinality as $\mathbb{R}$.

11.20. (medium after exercise **??**.19) Show that the set of all continuous functions

$$f : \mathbb{R} \to \mathbb{R}$$

has the same cardinality as $\mathbb{R}$.

11.21. (medium) Show that the set of all functions

$$f : \mathbb{R} \to \mathbb{R}$$

[16] What *is* known, however, is that not *both* of them can be algebraic.

has greater cardinality than $\mathbb{R}$. (Hint: Even the functions with values 0 and 1 only have greater cardinality than $\mathbb{R}$. Why?)

Chapter 12

Mathematics of single-winner election systems

12.1 What is this subject about?

Imagine you wanted to elect a single winner from a field of n candidates. Assume that N people will vote. How should you do this?

The question comes up in many contexts. Perhaps you and four friends want to decide whether to eat Indian, Chinese, or Thai food. Here $n = 3$ and $N = 5$. Or a city might want to elect a mayor, and there might be $n = 5$ candidates and $N = 400,000$ voters. Or the Mathematics Department might want to decide whether to focus its next faculty search on Numerical Analysis, Dynamics, or Number Theory. Here $n = 3$, and N is the number of department members who are eligible to vote on such questions.[17]

There are many other kinds of elections — elect two out of n candidates, or elect one representative per district, and so on. We will focus exclusively on the simplest case: N voters, n candidates, and exactly one is to be elected.

You could ask each voter to name their favorite candidate, and declare the one named most often the winner. If one candidate is named by 21 voters, one by 19 voters, and three others by 20 voters each, then the one with 21 votes wins. This is called *plurality voting.* It doesn't seem obvious that this is the best thing you can do.

We will assume instead that we ask each voter to *rank* all candidates, and for simplicity we will assume *strict* and *complete* rankings. For instance, when there are $n = 4$ candidates, A, B, C, and D, and $N = 5$ voters (think of people voting on which restaurant to eat at), the outcome of the election might be described by the following table.

C	B	A	C	D
A	A	B	A	B
B	C	C	B	A
D	D	D	D	C

(12.1)

[17] We actually do use fancy single-winner election systems to decide such questions.

Here the first voter ranks the candidates C≻A≻B≻D. (The symbol ≻ stands for "ranks higher than".) The second voter ranks them B≻A≻C≻D. And so on. We call this sort of table a *preference schedule with* $n = 4$ *candidates and* $N = 5$ *voters.*

Of course, in practice we might be asking too much if we asked each voter in the Presidential election to strictly rank all candidates. (In Massachusetts in 2024, there were six candidates for U.S. President, some of whom you might never even have heard of.) The systems that we'll discuss here can easily be adjusted to allow for rankings with ties, and for incomplete rankings.

Back to the preference schedule

C	B	A	C	D
A	A	B	A	B
B	C	C	B	A
D	D	D	D	C

.

If we only paid attention to each voter's favorite candidate, then C would get two votes, and the others would get one each. Therefore C would win. But note that actually, three out of five voters would prefer A over C, so if we declared A the winner instead of C, a majority of voters would be happier.

Definition 12.1. *A* single-winner election system *is a mapping that assigns to each preference schedule with* N *voters and* n *candidates a single winner.*

An election system, according to this definition, selects a single winner for any preference schedule. Even when $n = N = 3$ and the preference schedule is

T =

A	B	C
B	C	A
C	A	B

,

so all three candidates do perfectly equally well with the voters, there must be a single winner. We assume that the election system breaks ties. In practice, it is almost always necessary to do that. The answer to the question "Which restaurant should we eat at?" can't be "Indian and Chinese are tied", unless we are prepared to eat two dinners in a row, and there usually can't be two mayors at the same time either.

We will assume that ties are resolved in alphabetical order, with preference given to candidates with labels that come earlier in the alphabet. If the labels are assigned to candidates *at random* independently of the election outcome, this is equivalent to resolving ties with fair coin tosses.

12.2 Methods

12.2.1 Plurality voting

Plurality voting is the simplest single-winner election system, and the one most often used in the United States. Whoever is ranked first most often wins the election.

It is therefore unnecessary to ask the voters for their complete rankings, only their top choice matters. For the preference schedule

C	B	A	C	D
A	A	B	A	B
B	C	C	B	A
D	D	D	D	C

,

C wins. For the preference schedule

A	B	C
B	C	A
C	A	B

,

A wins. (Remember that we resolve ties alphabetically.)

12.2.2 Instant runoff

In *instant runoff voting*, the first step is to eliminate the candidate with the *smallest* number of first-place rankings. If there are several candidates who have the minimal number of first-place rankings, we eliminate the one who is last in the alphabet. Elimination of a candidate leaves gaps in the preference schedule, which we fill by moving candidates below the gap up by one notch.

For example:

C	B	A	C	D
A	A	B	A	B
B	C	C	B	A
D	D	D	D	C

→

C	B	A	C	B
A	A	B	A	A
B	C	C	B	C

→

C	B	B	C	B
B	C	C	B	C

→

B	B	B	B	B

So here B wins. (This is the same preference schedule which we considered before. C wins by plurality voting.)

This single-winner election system, arguably the second-simplest possible, is called *instant runoff*. It is also often referred to as *ranked choice voting*. That's a misnomer. There are many different single-winner election systems that are based on ranking the choices.

12.2.3 Borda count

Borda count is a single-winner election system named after the engineer Jean-Charles de Borda (1733–1799). The scheme is to give each candidate n points for a first-place vote, $n-1$ points for a second-place vote, and so on. Whoever gets the most points wins. Ties are resolved alphabetically as before.

Again suppose that this is the preference schedule:

C	B	A	C	D
A	A	B	A	B
B	C	C	B	A
D	D	D	D	C

/

Here A gets 15 Borda points, B gets 14 points, C gets 13 points, and D gets 8 points, so A wins.

12.2.4 Pairwise comparison

We imagine a competition among the candidates called the *pairwise comparison tournament*. For any pair of candidates X and Y, we decide, based on the preference schedule, whether X would beat Y if they were the only two candidates on the ballot, or vice versa. If X would beat Y, X gets a point. If Y would beat X, Y gets a point. If there would be a tie (this is only possible if N is even), both X and Y get half a point. Whoever gets the most points wins, with ties resolved alphabetically as usual.

For the preference schedule (12.1), A beats B, C, and D in one-on-one contest, and therefore gets 3 points in the pairwise comparison tournament. This implies that A wins — none of the other candidates win *all* one-on-one contest, since they all lose against A.

12.2.5 Dictatorship

For a fixed k with $1 \leq k \leq N$, dictatorship of the k-th voter is a single-winner election system: The winner is simply the candidate placed first by voter k. It may not be widely popular, but it is probably the voting systems in the majority of all countries in the world.

12.3 Arguably desirable fairness criteria

12.3.1 Pareto-efficiency

The criterion of Pareto-efficiency is absurdly weak:

Pareto-efficiency criterion. *We say that an election methods is* Pareto-efficient *if it has the property that X will win if X ranks first on every single ballot.*

Each of the winner election systems we have mentioned satisfies this criterion.

12.3.2 Majority-fairness

Majority criterion. *We say that an election method satisfies the* majority criterion *or is* majority-fair *if it is guaranteed to make a majority candidate the winner. A* majority candidate *is one placed first by more than half the voters.*

There may not be a majority candidate. Even presidential elections in the U.S., at the state level, are often examples. The two leading candidates might get 46% and 49% of the vote, with the rest going to third-party candidates. In such a case, the majority criterion does not tell us what *should* happen.

Proposition 12.2. *Borda count violates the majority criterion.*

Proof. Think about the following preference schedule.

A	A	A	B	B
B	B	B	C	C
C	C	C	A	A

Here A has a majority of first-place votes. However, A gets 11 Borda points, and B gets 12. So by Borda count, B wins. □

You can easily convince yourself that the other systems that we have discussed satisfy the majority criterion, except for dictatorship.

12.3.3 Condorcet-fairness

When there is no majority candidate, there may still be a "generalized majority candidate" in the sense of the following definition.

Definition 12.3. *For a given preference schedule with n candidates and N voters, candidate X is called a* Condorcet candidate *if X gets $n-1$ points in the pairwise comparison contest. i.e., if X would beat every other candidate in one-on-one competition.*

Clearly there can be at most one Condorcet candidate. For instance, for the preference schedule

C	B	A	C	D
A	A	B	A	B
B	C	C	B	A
D	D	D	D	C

,

A is the Condorcet candidate.

Definition 12.4. *We say that an election method satisfies the* Condorcet criterion *or is* Condorcet-fair *if it is guaranteed to make the (if there is one) Condorcet candidate the winner.*

Condorcet-fairness implies majority-fairness, but the reverse is not the case.

Proposition 12.5. *Plurality voting, instant runoff, and Borda count are not Condorcet-fair. Pairwise comparison, on the other hand, is.*

Proof. For preference schedule (12.1), the winner by plurality voting is C, the winner by instant runoff is B, but the Condorcet candidate is A. Therefore plurality voting and instant runoff are not Condorcet-fair. Borda count selects the Condorcet candidate in the example (12.1), but is in general not even majority-fair, therefore certainly not Condorcet-fair. (A majority candidate is a Condorcet candidate, therefore Condorcet-fairness implies majority-fairness.) □

The notion of Condorcet-fairness is named after Nicolas de Condorcet (1743–1794), a philosopher, political scientist, and mathematician, and friend of Thomas Jefferson. He was killed during the French Revolution, which lasted for 10 years, from 1789 to 1799.

12.3.4 Weak monotonicity

Moving a candidate X upwards on a single ballot, while leaving all else unchanged, is called a *ballot change favorable to X.*

Definition 12.6. *We call an election method* weakly monotonic *if a ballot change favorable to X cannot turn X from the winner into a loser.*

Proposition 12.7. *Among the methods that we have discussed, the only one which is not weakly monotonic is instant runoff.*

Proof. It is easy to verify that plurality voting, Borda count, pairwise comparison, and dictatorship are weakly monotonic. To see that instant runoff voting is not weakly monotonic, consider the preference schedule

C	C	C	C	A	A	A	B	B	B
B	B	B	B	C	B	B	C	C	C
A	A	A	A	B	C	C	A	A	A

.

Under instant runoff, here is what happens: B gets eliminated, because B comes after A in the alphabet. Then C wins. But suppose that voter 5 changed their mind and rankced C above A:

C	C	C	C	C	A	A	B	B	B
B	B	B	B	A	B	B	C	C	C
A	A	A	A	B	C	C	A	A	A

.

Now A gets eliminated, and therefore B ties with C. Since B comes earlier in the alphabet than C, now B wins.

This example relies heavily on our tie-breaking conventions, and that's how it must be if a single voter's change of mind is to change the outcome. I leave it to you to construct an example *not* involving any ties in which *several* voters make a change favorable to C, and as a result C loses when before C would have won. □

12.4 Dubious fairness criteria

12.4.1 Strong monotonicity

Definition 12.8. *We call a voting method* strongly monotonic *if a change in single ballot resulting in a new ballot in which every candidate who ranked below X prior to the change still ranks below X cannot turn X from the winner into a loser.*

Notice that the definition says nothing about how the change in the ballot affects the candidates ranking above X in the original ballot. They may be scrambled. But if everybody who was below X before the change is still below X after the change, and X was the winner before the change, then X should win after the change.

This may or may not sound convincing to you. To me, it only sounds very mildly convincing. This is why this fairness criterion comes under the heading "dubious fairness criteria". None of the methods that we have discussed, except for dictatorship, is strongly monotonic. For instance, let's show that plurality voting isn't strongly monotonic. If the preference schedule is

A	A	B	C	C
B	B	C	A	D
C	C	A	D	A
D	D	D	B	B

,

then A wins by the plurality method, using our alphabetic tie breaking convention. But if the preference schedule is

A	A	C	C	C
B	B	B	A	D
C	C	A	D	A
D	D	D	B	B

,

C wins. Only the third ballot was changed, and on that ballot, only D ranked below A before the change, and D still ranks below A after the change.

12.4.2 Strategy-proofness

Voters who really would like the Green Party candidate to be President often vote for the Democrat, realizing that their vote for their true choice might cause a candidate to be elected whom they like even less than the Democrat. Similarly, voters who really would like the Libertarian Party candidate to win might vote for the Republican instead. This sort of behavior will be called *strategic voting* here.

In some instances, strategic voting can in fact be successful. To illustrate this, we again return to

C	B	A	C	D
A	A	B	A	B
B	C	C	B	A
D	D	D	D	C

.

By plurality voting, C wins. But if the fifth voter (corresponding to the right-most column) were to swap their first two choices, the preference schedule would become

C	B	A	C	B
A	A	B	A	D
B	C	C	B	A
D	D	D	D	C

.

Now B and C have equally many first-place votes, and since ties are resolved alphabetically, now B wins. This is an outcome that the fifth voter prefers to C's victory. Remember that

C	B	A	C	D
A	A	B	A	B
B	C	C	B	A
D	D	D	D	C

reflects the fifth voter's true rankings.

Definition 12.9. *We call an election method* strategy-proof *if it is* never *possible for a voter to cast a ballot not reflecting the voter's true preference, and thereby affect the outcome in a way that the voter likes.*

There is no question that strategy-proofness would be desirable. However, no voting method we have discussed up to this point is strategy-proof, except for dictatorship.

12.5 Impossibility of a perfect system

12.5.1 Strong monotonicity implies dictatorship

Social choice theory has several "dictatorship theorems", saying that the only election system with certain seemingly desirable properties is dictatorship. I will prove the following example first.

Theorem 12.10 (Muller-Satterthwaite theorem, 1977). *Suppose $n \geq 3$. The only Pareto-efficient and strongly monotonic election method is dictatorship.*

The charm of the proof is that it really involves nothing that you didn't know in fifth grade, and yet it is not easy. Before we proof the Muller-Satterthaite theorem, we state and prove the following lemma.

Lemma 12.11. *Assume a strongly monotonic, Pareto-efficient election method. If X is ranked above Y by every single voter, then Y is not the winner.*

Proof. Suppose this were not the case, so there were a preference schedule in which X ranks above Y in each ballot, yet Y is the winner. By strong monotonicity, Y would still be the winner if X were moved to first place on every single ballot, without disturbing rankings of candidates below Y. But since X would then be in first place on every single ballot, X would have to be the winner by Pareto-efficiency, so Y would *not* be the winner anymore. □

Now we are ready for the proof of the Muller-Satterthaite theorem.

Proof. Assume a strongly monotonic, Pareto-efficient election method. Pick an arbitrary candidate, let's say A. (The argument can be done with any other candidate in place of A, and this point will be important later.) Let us think about a situation in which we know what must happen: Suppose A is placed first by every voter. By Pareto-efficiency, A is then the winner. We indicate this as follows:

1	...	$k-1$	k	$k+1$	...	N
A	...	A	A	A	...	A
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.

$$\rightarrow \quad A\ . \tag{12.2}$$

Each column refers to a single voter. The numbers at the top of the columns label voters.

Of course, A is extremely popular in (12.2) and *should* be the winner! Every voter places A first. However, we will now change the preference schedule gradually, making A look worse and worse, yet still ensuring that A must remain the winner. In the end, A will be utterly unpopular, yet still the winner, and the only possible explanation will be that the winner election system is dictatorship.

To make A look less popular, let's bring a specific competitor into play, say candidate B. As long as A is in first place on each ballot, A wins, since the method is Pareto-efficient. For instance:

1	...	$k-1$	k	$k+1$	...	N
A	...	A	A	A	...	A
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
B	...	B	B	B	...	B

$$\rightarrow \quad \text{A} . \tag{12.3}$$

Now let's make B look better, by moving B up on the first ballot. A remains the winner as long as B does not reach the top position, by strong monotonicity (and also by Pareto-efficiency). Even when B reaches the second position in the first ballot, A is still the winner:

1	2	...	$k-1$	k	$k+1$	...	N
A	A	...	A	A	A	...	A
B	.	...	.	.	.	...	.
.	.	...	.	.	.	...	.
.	.	...	.	.	.	...	.
.	.	...	.	.	.	...	.
.	B	...	B	B	B	...	B

$$\rightarrow \quad \text{A} . \tag{12.4}$$

The moment, however, when B rises above A on the first ballot, it is unclear what will happen:

1	2	...	$k-1$	k	$k+1$	...	N
B	A	...	A	A	A	...	A
A	.	...	.	.	.	...	.
.	.	...	.	.	.	...	.
.	.	...	.	.	.	...	.
.	.	...	.	.	.	...	.
.	B	...	B	B	B	...	B

$$\rightarrow \quad ? \tag{12.5}$$

You may be tempted to say "A is still in first place on all ballots except the first, and in second place on the first ballot, so A should still be the winner." This will of course be true for most single-winner election systems, but not for all: If the system is dictatorship of the first voter, then going from (12.4) to (12.5) changes the winner from A to B.

The winner in (12.5) is, in any case, either A or B. To see this, suppose that some third candidate C were the winner in (12.5). Then C would have to be the winner even in (12.4), by strong monotonicity. So we have concluded:

1	2	...	$k-1$	k	$k+1$	...	N
B	A	...	A	A	A	...	A
A	.	...	.	.	.	...	.
.	.	...	.	.	.	...	.
.	.	...	.	.	.	...	.
.	.	...	.	.	.	...	.
.	B	...	B	B	B	...	B

→ A or B . (12.6)

Suppose the winner is still A. Then we now move B upward on the second ballot. As long as B does not reach the top in the second ballot, A remains the winner, by strong monotonicity. When B reaches the top, and A falls to second place on the second ballot, either A remains the winner, or B becomes the winner, by the argument we just gave. If A remains the winner, we proceed to the third ballot and let B rise to the top there. We continue in this way until, at some point, the winner changes from A to B when B is moved to the top of one of the ballot. This *must* happen eventually, for if we move B into the top position in *all* ballots, then B is the winner by Pareto-efficiency. Let's say that B becomes the winner the moment when B is moved into the top position in the k-th ballot. So

1	...	$k-1$	k	$k+1$	...	N
B	...	B	A	A	...	A
A	...	A	B	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	B	...	B

→ A , (12.7)

but

1	...	$k-1$	k	$k+1$	...	N
B	...	B	B	A	...	A
A	...	A	A	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	B	...	B

→ B . (12.8)

In this example, therefore, voter k plays a pivotal role: As long as voter k places A above B, the overall winner is A, but as soon as voter k places B above A, the overall winner becomes B. Not much seems surprising so far. You might think, for instance, that k is probably approximately $N/2$ — the winner changes from A to B as soon as B, not A, has the majority.

However, we can make these examples much stranger. To start with, we will prove that in (12.7), we can move A to the bottom of ballots 1 through $k-1$, and to second-to-last place in columns $k+1$ through N, and A will still be the winner overall:

1	...	$k-1$	k	$k+1$	...	N
B	...	B	A	.	...	.
.	...	.	B	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	A	...	A
A	...	A	.	B	...	B

$\rightarrow$ A . (12.9)

To see this, you have to make two observations. First, the winner in (12.9) is certainly *not* B. For if the winner were B in (12.9), then the winner would have to be B in (12.7), by strong monotonicity. Second, the winner in (12.9) cannot be some third candidate C either. For if

1	...	$k-1$	k	$k+1$	...	N
B	...	B	A	.	...	.
.	...	.	B	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	A	...	A
A	...	A	.	B	...	B

$\rightarrow$ C , (12.10)

then

1	...	$k-1$	k	$k+1$	...	N
B	...	B	B	.	...	.
.	...	.	A	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	A	...	A
A	...	A	.	B	...	B

$\rightarrow$ C (12.11)

by strong monotonicity, but (12.8) implies

1	...	$k-1$	k	$k+1$	...	N
B	...	B	B	.	...	.
.	...	.	A	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	A	...	A
A	...	A	.	B	...	B

$\rightarrow$ B , (12.12)

again by strong monotonicity.

We have now proved (12.9), and that is certainly a rather surprising conclusion — A is at or near the bottom on all ballots except the k-th, yet A is still the winner!

It would become truly astonishing if we could move A all the way to the bottom in ballots $k+1$ through N. However, it is not clear that we can do that:

1	...	$k-1$	k	$k+1$	...	N
B	...	B	A	.	...	.
.	...	.	B	.	...	.
.	...	.	.	.	...	.
.	...	.	.	.	...	.
.	...	.	.	B	...	B
A	...	A	.	A	...	A

→ ? (12.13)

One thing is clear, though: In (12.13), the winner cannot be anybody other than A or B. For if a third candidate C were the winner, then C would have to be the winner even in (12.9), by strong monotonicity. So we would like to find an argument that rules out that B is the winner in (12.13).

We do this by introducing a third candidate C. (This is where the argument needs the assumption that there are at least three candidates.) We insert C into (12.9) in a way that, by strong monotonicity, does not alter the fact that A is the winner:

1	...	$k-1$	k	$k+1$	...	N
C	...	C	A	.	...	.
B	...	B	C	.	...	.
.	...	.	B	.	...	.
.	...	.	.	C	...	C
.	...	.	.	A	...	A
A	...	A	.	B	...	B

→ A . (12.14)

If in this preference schedule, we swap the positions of A and B in the $k+1$-st through N-th ballots, the winner is still A or B, as argued earlier, but now the winner *cannot be B* any more, since C is ahead of B on every single ballot. So

1	...	$k-1$	k	$k+1$	...	N
C	...	C	A	.	...	.
B	...	B	C	.	...	.
.	...	.	B	.	...	.
.	...	.	.	C	...	C
.	...	.	.	B	...	B
A	...	A	.	A	...	A

→ A . (12.15)

So now we have an example of a preference schedule where A is ranked *last* by every voter except voter number k, who ranks A first, and A wins anyway!

Any preference schedule in which voter k ranks A first can be obtained from (12.15) by first re-ordering (if necessary) the candidates other than A, while keeping A in its position, then moving A upwards in ballots 1 through $k-1$ and $k+1$ through N (if necessary). By strong monotonicity, this does not turn A from a winner into a loser. So whenever voter k ranks A first, A is the winner. We say that voter k

is the *A-dictator*. An *A-dictator* is a voter who does not necessarily determine the outcome of the election in all cases, but *if* the voter chooses to rank A first, then A wins.

Now recall that A was chosen arbitrarily at the beginning. So there is also a B-dictator, and there is a C-dictator, etc. Could they be different voters? The answer is no. For instance, if the A-dictator were to put A first, and the B-dictator put B first, then both A and B would have to be winners, and that cannot be true. This proves that there is a *single* dictator whose top choice is the winner of the election. □

12.5.2 Strategy-proofness implies dictatorship

Theorem 12.12 (Gibbard-Satterthwaite Theorem, 1973/1975). *Let $n \geq 3$. If a single-winner election method is Pareto-efficient and strategy-proof, it is dictatorial.*

Proof. Let the election method be Pareto-efficient and strategy-proof. We will show that it is then strongly monotonic, and therefore the Muller-Satterthwaite theorem implies our assertion.

Suppose that X is the election winner for a given preference schedule. Now we change the i-th ballot in such a way that everybody ranking below X originally still ranks below X after the change. After the change, Y is the winner. We need to prove that X=Y.

We denote the original i-th ballot by B_i, and the changed i-th ballot by B_i'. Suppose that B_i reflected the i-th voter's honest opinion. Then Y cannot rank above X in B_i, since otherwise the change in the i-th ballot would constitute successful strategic voting. Therefore Y cannot rank above X in B_i' either, since the change was such that everybody ranking below X originally still ranks below X after the change.

But now assume that B_i' reflected the i-th voter's honest opinion. If X ranked above Y in B_i', then changing from B_i' to B_i would constitute successful strategic voting for voter i. We conclude that X cannot rank above Y in B_i'.

So in B_i', Y cannot rank above X, and X cannot rank above Y. This implies X=Y. □

12.6 Which system should one use in practice?

This is of course in the eye of the beholder. I think that Condorcet-fairness is a compelling requirement. This would rule out all systems discussed so far, except for pairwise comparison. However, pairwise comparison is not practical, because it will often lead to ties, even in elections with millions of voters. The total number of points distributed in the pairwise comparison tournament is $n(n-1)/2$. This is why ties are fairly likely for small n, even when N is large.

There are elegant Condorcet-fair systems, for instance the beatpath method. However, they are complicated, and therefore almost certainly not politically feasible. My opinion is that one should simply use instant runoff, but with one modification: *If* there is a Condorcet candidate, declare that candidate the winner. If not, then use instant runoff.

12.7 Probabilistic analysis

Instead of asking whether a given system *can* violate a given fairness criterion, such as the Condorcet criterion, one might ask how *likely* such violations are to occur in a randomly chosen election outcome. To make this precise, one must first say what one means by a "randomly chosen election outcome".

The relevant distribution, the distribution from which election outcomes should be drawn here, would be the distribution of actual, real-world election outcomes. Perhaps one can determine that distribution, with some accuracy, through empirical studies, but I can't. This is why I will consider the following way of drawing "random election outcomes", which strikes me as somewhat plausible at least.

12.7.1 Randomly placed candidates with a continuum of voters

I assume that anybody's views can be characterized by a real number x, measuring how far to the left or to the right their views are. Left-wing voters or candidates have negative x, right-wing ones have positive x. In recognition of the fact that hard left-wing and hard right-wing views often resemble each other strikingly, we should perhaps think of x as a point on a circle, but we'll think of it as a point on the real axis, called here the "left-right axis". We assume that the electorate is so large that it can be thought of as a continuum, and that $f = f(x)$ is the density characterizing the distribution of views in the electorate. For simplicity (and not very realistically), I will assume $f(x) > 0$ for all $x \in \mathbb{R}$. So the fraction of voters with views x between a and b is

$$\int_a^b f(x)\,dx.$$

In particular,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

We will further assume that 0 is the political center, so exactly half the electorate is to the left of 0:

$$\int_{-\infty}^{0} f(x)\,dx = \frac{1}{2}. \tag{12.16}$$

This is a matter of convention.

I further assume that n candidates are placed on the left-right axis at random, independently of each other, with views that have density $g = g(x)$. We denote the positions of these candidates by $X_1, \ldots, X_n$. I assume that voters rank candidates in the order of their proximity to the candidate positions. For example, the fraction of voters who place candidate 2 first would be, assuming $n \geq 3$,

$$\int_{(X_1+X_2)/2}^{(X_2+X_3)/2} f(x)\,dx.$$

For any choice of $X_1, \ldots, X_n$, we can now determine what the outcome of the election would be using any of the systems we have discussed. The probability of any ties is zero in this model, because the voter views and the candidate views of probability densities.

Proposition 12.13. *Under the assumptions made here, a candidate whose position is closest to 0 is the Condorcet candidate. In particular, the probability that there is no Condorcet candidate is zero.*

Proof. Consider two candidates positioned at X and Y, with X closer to 0 than Y. Assume without loss of generality that X < 0. If Y<X, then the fraction of the vote won by X in a one-on-one contest with X would be

$$\int_{\frac{X+Y}{2}}^{\infty} f(x)dx > \frac{1}{2},$$

so X would beat Y. If Y > X, then in fact Y > |X| since X is closest to 0. Therefore the fraction of the vote won by X in a one-on-one contest with X would be

$$\int_{-\infty}^{\frac{X+Y}{2}} f(x)dx > \frac{1}{2},$$

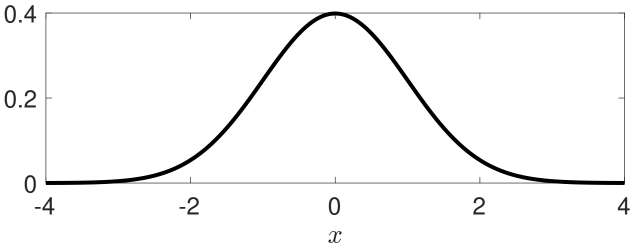
so again X would beat Y. So in a one-on-one contest, the candidate closer to 0 always wins. □

12.7.2 The probability that the Condorcet candidate will lose

We can now draw random candidate positions, and compute whether the Condorcet candidate would win under plurality voting for instance, or under instant runoff. For a very simple example, assume that $f(x)$ and $g(x)$ are standard Gaussian densities. This means

$$f(x) = g(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

(This is the Gaussian or normal density with mean zero and standard deviation 1.)



We consider elections with $n = 3$ candidates. We draw three candidate positions at random with this density. I did that, and got (rounding to two significant digits) -1.3, 0.32, and 0.86. The fraction of the vote won by the candidate positions at -1.3 is

$$\int_{-\infty}^{(-1.3+0.32)/2} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = \int_{-\infty}^{-0.49} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx.$$

Substitute $u = x/\sqrt{2}$:

$$\int_{-\infty}^{-0.49/\sqrt{2}} \frac{e^{-u^2}}{\sqrt{\pi}}\, du = \frac{1}{2} + \int_{0}^{-0.49/\sqrt{2}} \frac{e^{-u^2}}{\sqrt{\pi}}\, du = \frac{1+\operatorname{erf}(-0.49/\sqrt{2})}{2},$$

where

$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$$

is the function known as the *error function.* (It is a built-in function in Matlab because it is so important in calculations involving Gaussian distributions.) We have

$$\frac{1+\operatorname{erf}(-0.49/\sqrt{2})}{2} \approx 0.31$$

Similarly, the candidate positioned at 0.32 gets a fration of the vote equal to

$$\frac{-\operatorname{erf}(-0.49/\sqrt{2})+\operatorname{erf}(0.59/\sqrt{2})}{2} \approx 0.41,$$

and the candidate positioned at 0.86 gets a fraction of

$$\frac{1-\operatorname{erf}(0.59/\sqrt{2})}{2} \approx 0.28.$$

So here the Condorcet candidate, i.e., the one closest to 0, namely the one positioned at 0.32, wins with about 41% of the vote. Repeating this 100 million times on the computer, we find:

Under our assumptions, the probability that plurality voting will elect somebody who isn't the Condorcet candidate is about 29.33%.

We can do a similar calculation for instant runoff. Again, with 100 million computer simulations, we find with good confidence:

Under our assumptions, the probability that instant runoff will elect somebody who isn't the Condorcet candidate is about 14.72%.

The probability of a "Condorcet failure", a Condorcet candidate who loses the election, is approximately half of what it is with plurality voting.[18]

[18] The calculations are accurate enough to say with some confidence that it isn't *exactly* half.

Exercises

12.1. (easy) Can you construct a preference schedule with $n = 3$ and $N = 3$ where there is a majority candidate, but Borda count would lead to the election of one of the other candidates?

12.2. (easy) Can you think of any reason why we would want to consider a single winner election system that is not Pareto-efficient?

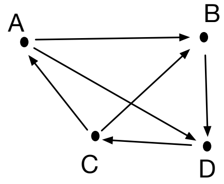
12.3. (easy) Prove that plurality voting, Borda count, pairwise comparison, and dictatorship are weakly monotonic.

12.4. (medium) Prove directly (without citing the Muller-Satterthwaite theorem) that none of the election systems in Exercise 12.3 are strongly monotonic, except for dictatorship.

12.5. (medium) Let's say that an election system satisfies the *independence of irrelevant comparisons* (IIC) criterion if a winner X remains the winner if candidates above X are permuted but stay above X, or candidates below X are permuted but stay below X, on one of the ballots. Prove that a system is strongly monotonic if and only if it is weakly monotonic and satisfies the IIC criterion.

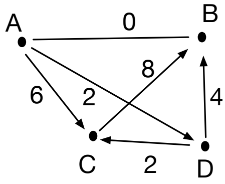
12.6. (easy) Construct a preference schedule with three candidates A, B, and C, so that a majority of voters prefers A to B, a majority prefers B to C, and a majority prefers C to A. This is called a *Condorcet cycle.*

12.7. (easy) We can visualize the pairwise comparison election system as follows. We make a dot on the page for each candidate, and draw an arrow from candidate X to candidate Y if there are more voters who prefer X to Y than there are voters who prefer Y to X. We draw an arrow from candidate Y to X if there are more voters who prefer Y to X than there are voters who prefer X to Y. We draw no arrow between X and Y if exactly half the voters prefer X to Y, and the other half prefers Y to X.



Explain: X is a Condorcet candidate if and only if there are arrows pointing from X to all other candidates in the pairwise comparison graph.

12.8. (easy) Now attach *weights* to the arrows in the pairwise comparison graph, indicating the *margin* by which the candidate winning the one-on-one contest would win. If K voters prefer X to Y, and L voters prefer Y to X, with $K > L$, then there is an arrow pointing from X to Y with weight $K-L$. If the outcome of a one-on-one contest between X and Y would be a tie, there is no arrow between X and Y, but we now insert a link (without a direction) and give it weight 0.



Prove that either all weights are even, or all weights are odd.

12.9. (medium) Assume N is odd, so that there are no ties in pairwise comparisons. Assume there is no Condorcet candidate. Prove that every candidate lies on a Condorcet cycle. (See Exercise 12.6 for the definition of *Condorcet cycle.*)

12.10. (hard) Suppose we are given *any* weighted pairwise comparison graph in which all weights are of the same parity. (Either all even, or all odd.) Prove that there is a preference schedule that gives rise to the given pairwise comparison graph. Furthermore, prove that such a preference schedule can be constructed with $N \leq 1 + \sum_j w_j$ voters, where the w_j are the given weights.

Hint: First think about the case when all w_j are even. Start with the "empty preference schedule" — there are no voters. Then notice that any given link can be strengthened by 2 without affecting any other links, simply by adding two columns to the existing preference schedule. For instance, to strengthen the link from A to B by 2, assuming there are five candidates A, B, C, D, E, you would add

$$\boxed{\begin{array}{c}\text{A}\\\text{B}\\\text{C}\\\text{D}\\\text{E}\end{array}} \quad \text{and} \quad \boxed{\begin{array}{c}\text{C}\\\text{D}\\\text{E}\\\text{A}\\\text{B}\end{array}}\ .$$

12.11. (easy after you do Exercise 12.10 or simply believe the result) Define an election system as follows. Draw the *weighted pairwise comparison graph* and assign to each candidate X the sum of the weights of the outgoing arrows starting at X. Whoever gets most points wins. Is this Condorcet-fair?

12.12. (easy) Prove that in Exercise 12.10, you couldn't do with fewer than $\max_j w_j$ voters.

12.13. (easy but creative) Can you think of other ways of formalizing the question "How likely is a Condorcet candidate to lose the election?"

Chapter 13

Review

We revisit each of our twelve topics again for a few minutes, adding something new each time.

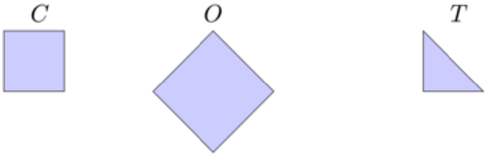
13.1 Hypercube architecture in parallel computing

Mathematics is appealing to me only if there is a link with the world. It doesn't have to be a link that you can use for practical purposes. But in my mind, mathematics should paint an abstract picture of the world that we can clearly understand, in order to better understand the world. That's what makes it interesting. It's not "solving challenging puzzles" for its own sake. Life poses enough challenges. I don't need to manufacture my own.

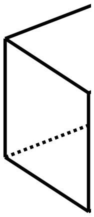
Therefore the Platonic solids (regular polytopes) in dimensions higher than three are of little interest to me, at least at first sight. The fact that there are six of them in four dimensions, and only three in all higher dimensions, is a mildly amusing curiosity, but nothing more than that, to me. The three Platonic solids that exist in all dimensions are

$$C = [0,1]^d, \qquad O = \left\{ x \in \mathbb{R}^d \;:\; \|x\|_1 \leq 1 \right\}, \qquad T = O \cap C,$$

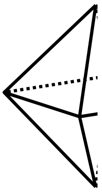
where $\|x\| = \sum_{i=1}^{d} |x_i|$. We call C the *d-dimensional hypercube*, O the *d-dimensional hyperoctahedron*, and T the *d-dimensional hypertetrahedron*.



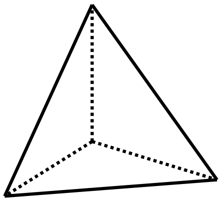
For $d = 3$, we get the ordinary cube, octahedron, and tetrahedron.



C



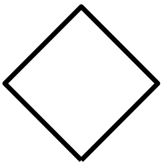
T



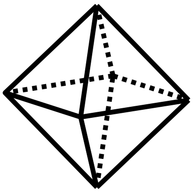
The hypercube in d dimensions has $N = 2^d$ vertices. Two vertices are connected by an edge if and only if their coordinate vectors differ in only one entry, so each vertex has $d = \log_2 N$ neighboring vertices. The total number of edges is $Nd/2 = N(\log_2 N)/2$. We can walk from any vertex to any other vertex, along edges, in at most $d = \log_2 N$ steps. We say that the network of vertices has *diameter* $\log_2 N$.

The hyperoctahedron O is the unit ball in the 1-norm in $\mathbb{R}^d$. It is the convex hall of the vectors $\pm e_i$, $1 \leq i \leq d$, where e_i is the i-th canonical basis vector of $\mathbb{R}^d$. The $\pm e_i$ are its vertices, so it has $N = 2d$ vertices.Each vertex has an *antipodal* vertex (e_i and $-e_i$). There is an edge between any two vertices that aren't antipodal. Each vertex is therefore connected to $N - 2$ edges. The total number of edges is $\frac{N-2}{2} N$. The diameter of the network of vertices is 2.

$d = 2$

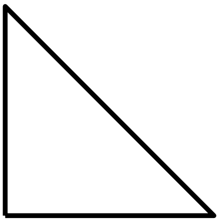


$d = 3$

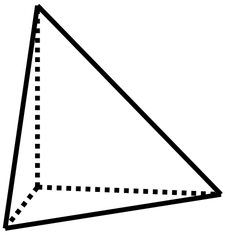


The hypertetrahedron T is the part of the hyperoctahedron O that lies in the positive hyperoctant. It has $N = d + 1$ vertices. There is an edge between any two vertices. The total number of edges is $\frac{N-1}{2} N$. The diameter of the network of vertices is 1.

$d = 2$

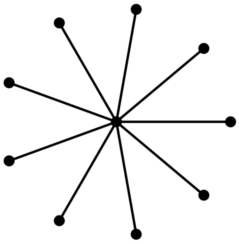


$d = 3$



A few decades ago, a real application of hypercubes emerged, making the hypercube more interesting to me: The processors of parallel computers were connected like the vertices of a hypercube, with a wire between two processors if and only if there is an edge between the two associated vertices in the hypercube. The hypercube network makes a reasonable compromise between diameter ($\log_2 N$) and the total number of edges ($\frac{\log_2 N}{2} N$). The hperoctahedron and hypertetrahedron networks have far smaller diameters at the expense of far greater numbers of edges.

Both the diameter and the number of edges can be smaller than in the hypercube network, for ins



Here the toal number of edges is $N-1$, and the diameter is 2. Of course, the nodes are no longer all the same, and the central one is very heavily connected.

The hypercube architecture has become less popular in recent years. It appears that one reason is that it is hard to maintain the hypercube structure when you want to increase the number of processors. You have to *double* the number, you can't increase it incrementally. Another disadvantage appears to be that it is difficult to connect nodes on a (two-dimensional) chip in a hypercube fashion.

My new MacBook Air has an M4 chip, which has 10 CPU cores, 10 GPU cores, and a 16-core Neural Engine. If you don't know exactly what that means: neither to I. But as far as I know there is no hypercube architecture in the M4 chip.

13.2 The Monty Hall puzzle

When I present Martin Gardner's puzzle in its original form (without naming one of the children), invariably somebody says "Isn't that the Monty Hall puzzle?" It isn't. The two have in common that both involve conditional probability, and in both cases, the surprising answer is 2/3. But that's where the similarity ends.

The Monty Hall puzzle goes like this: Monty Hall (1921–2017), host of "Let's Make a Deal", shows you three doors on stage. You know that there is a car behind one of them, and there are goats behind the other two. You can choose a door, and get what's behind it. The tacit assumption is that you'd rather have a car than a goat. (I'd much rather have a goat, I have a car already.) Monty asks you to guess which door has the car. You point to a door. Monty opens one of the two doors that you did not point to, and there is a goat behind that one. Then Monty asks "Do you stick with your initial guess, or do you change your guess to the other still-closed door?"

What should you do? Stick with your guess, or switch? Or maybe it doesn't matter? The answer is simple. Let's call the door that you picked "door 1". There are two *a priori* possibilities: The car is behind door 1 (probability 1/3), or it is not (probability 2/3). Sticking with your original choice wins if the car is behind door 1, and switching wins if the car is between doors 2 or 3. So if you switch, your chance of winning the car rises from 1/3 to 2/3.

Do these puzzles satisfy my requirement of being connected to the world? The Monty Hall puzzle doesn't really, but conditional probability is crucial in many contexts, as discussed in Chapter 2.

13.3 Frequentists vs. Bayesians

13.3.1 Is it mathematics or philosophy?

I think of probability as the frequency with which something happens. That's the *frequentist* point of view. Of course, it has to be taken with a grain of salt. When my smart speaker tells me that there is a 14% chance of rain tomorrow, it can't possibly mean that among all tomorrows, 14 out of 100 have rain. There's only one tomorrow. I tend to vaguely think of the statement as meaning "Among otherwise somewhat similar days in the past, 14% have been rainy."

But I am not a statistician, and listening to statisticians, I sometimes feel that I should adopt the Bayesian way of thinking. I don't even know exactly what that means. Is it a matter of philosophy, or does it affect the math that we do?

The standard way of making probability precise seems to have a frequentist flavor. Underlying everything is the *sample space* Ω. It doesn't really matter for the mathematics how we think of that intuitively, but the common way of thinking of it is to consider Ω the "set of all outcomes of a random experiment". We assume that we are given a probability measure P on Ω (before we even start doing any math), and what would that mean, intuitively, if $P(E)$ is *not* the fraction of instances in which the outcome lies in E?

For the longest time, I thought of Ω not as the set of all possible outcomes of a random experiment, but as the "set of all (pertinent) experiments, past and future". So for instance when you think about coin tosses, Ω would not be $\{H, T\}$, as people usually say, but it would be the set of all coin tosses in the history of humanity. It is vaguer than $\Omega = \{H, T\}$ (for instance, does it count when a coin falls off a shelf accidentally?), but the math does not change.

I'll walk through a standard simple problem in *Bayesian* probability theory and see whether it could naturally be interpreted with a *frequentist* mindset. Suppose you see somebody toss a possibly biased coin 10 times, and get 7 heads. How likely is heads on the 11th toss?

This, in essence, is Laplace's sunrise problem. I'll review very briefly how the analysis goes. We assumed that the probability of getting heads is $B \in (0, 1)$, uniformly distributed. Then we proved (using Bayes' formula) that the density of B *given* that 10 tosses yielded 7 heads equals

$$f(\beta) = 1320\beta^7(1-\beta)^3. \tag{13.1}$$

(Compare eq. (3.3).) The probability of getting heads on the 11th toss is therefore

$$\int_0^1 \beta f(\beta) d\beta = \frac{2}{3}. \tag{13.2}$$

I can easily interpret everything we just did in a frequentist mindset. Fate draws a random number $B \in (0,1)$. Then Fate makes a coin with probability B of getting heads. Then Fate tosses the coin 11 times. We do this very many times, but discard all runs except for the ones where Fate gets 7 heads on the first 10 tosses. Among these, how *frequent* is it for Fate to get heads on toss number 11? The answer is: Fate gets heads on the 11th toss two times out of three.

13.3.2 The dependence on a priori assumptions

The formula (13.1), and therefore (13.2), depends on the assumption that B is chosen uniformly, reflecting that we know nothing about B. In the long run, this does not matter: If you see $7n$ heads in $10n$ tosses, then the answer will converge to the unsurprising value of 7/10 as $n \to \infty$ no matter which initial (unconditional) density f_0 of B you assume, as long as $f_0(\beta) > 0$ for all $\beta \in (0,1)$. The dependence on f_0 in the answer washes out as n increases, but as it washes out, the answer becomes the obvious one.

For example, if you look at an item you want to buy, and it has 7 good reviews and 0 bad reviews, then Laplace's theory tells you that you have a 1/8 chance of a bad experience if you buy the item. (This ignore the obvious objection that you shouldn't trust 7 reviews; they might have been written by seller and their family.) But if the item is made by a company that you have a positive impression of, Laplace's answer is no longer valid, since it no longer makes any sense to assume B to be uniformly distributed *a priori*.

For small n, the answer you compute is only as good as your *a priori* assumption. For large n, your answer is the unsurprising one. I wonder whether there are values of n in between, where the dependence on the *a priori* assumption has washed out, but the answer is still significantly different from the obvious one.

13.3.3 Forward vs. backward

The typical "frequentist" question is:

> Given the characteristics of some stochastic process, how likely is it to see certain outcomes.?

The typical "Bayesian" question is:

> Given that we see certain outcomes, what are the characteristics of the stochastic process?

So while it doesn't seem to me that there is a distinction between "frequentist" and "Bayesian" probability theory, the "Bayesian" questions differ from the "frequentist" ones, and they are roughly speaking inverses of each other.

13.4 Euler's gamma function

13.4.1 Review of Euler's definition of $x!$

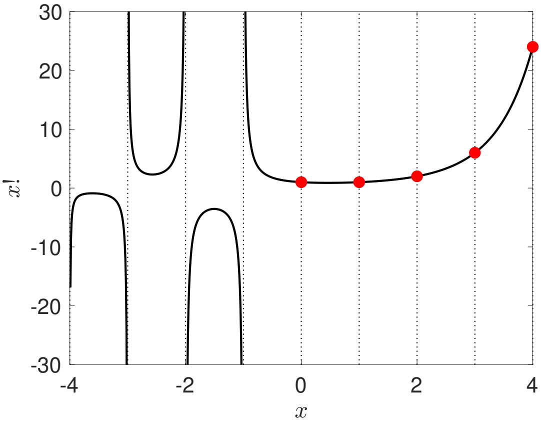
Euler's definition of the factorial of a real number

$$x! = \int_0^\infty t^x e^{-t}\, dt.$$

This converges for $x > -1$, not for $x \le -1$. However, for $x < -1$, x not integer, we can define $x!$ using the recursion formula

$$x! = \frac{(x+1)!}{x+1}$$

repeatedly. The resulting function looks like this:



The red dots indicate $k!$, $k = 0, 1, 2, 3, 4$.

13.4.2 Why I failed to answer my main question in this chapter

I tried to think about the question whether anything is very special about this particular interpolant of the factorials. I gave you two answers: If you want the recursion formula to hold for all x, and you want Stirling's formula to hold for all x, then Euler's definition is the only possible one. Or if you want the recursion formula to hold for all x, and you want $\ln x!$ to be a convex function, then Euler's definition is the only possible one. Neither answer is convincing. These are pretty properties, but why should they be important?

I am still looking to a good answer to the question "What is special about Euler's interpolation of the factorials?"

13.4.3 $z!$ for complex z

We can actually define $z!$ for a complex number with real part > -1:

$$z! = \int_0^\infty t^z e^{-t}\, dt.$$

For instance,

$$i! = \int_0^\infty t^i e^{-t}\,dt = \int_0^\infty (\cos\ln t + i\sin\ln t)\,e^{-t}\,dt \approx 0.498 - 0.155i.$$

(Whether this means anything at all is a different question.) Once we have defined $z!$ for $\mathrm{Re}(z) > -1$, we can use the recursion formula

$$z! = \frac{(z+1)!}{z+1}$$

to define it for z with $\mathrm{Re}(z) \in (-2,-1]$, *except* for $z=-1$. Then again we can use the recursion formula to define $z!$ for $\mathrm{Re}(z) \in (-3,-2]$, except for $z=-2$. And so on. So $z!$ becomes a function defined on

$$\mathbb{C} - \{-1,-2,-3,\ldots\}\,.$$

It is differentiable as a function of the complex variable z, so it is analytic everywhere except for poles at $-1,-2,-3,\ldots$.

13.4.4 Euler's gamma function

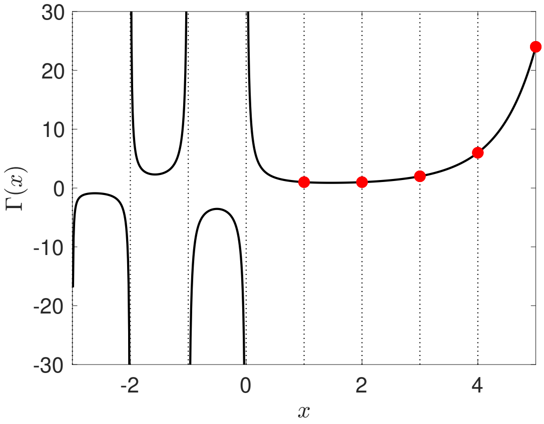
The standard notation is to shift $z!$ by one unit:

$$\Gamma(z) = (z-1)!.$$

This is called *Euler's gamma function.* So $\Gamma(z)$ is an analytic function defined on

$$\mathbb{C} - \{0,-1,-2,\ldots\}\,.$$

The graph is simply shifted to the right by one unit:



13.5 Entropy

13.5.1 Sensible ways of measuring spread

Let $n \in \mathbb{N}$, and let $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$ be a probability vector. We assume $\rho_i > 0$ for all i, and

$$\sum_{i=1}^{n} \rho_i = 1. \tag{13.3}$$

Such a vector is called a *positive probability vector.*

Exercise 5.2 suggested that formulas of the form

$$S_g(\rho) = \sum_{i=1}^{n} \rho_i g(\rho_i) \tag{13.4}$$

could in general be sensible ways of measuring the "spread" of ρ, assuming

$$g: \ (0,1) \to (0,\infty)$$

is a decreasing function. We will write $g = g(r)$, $0 < r < 1$.[19]

To understand the motivation for (13.4), think of ρ as describing mass distributed over n sites. The fraction of the mass that sits at the i-th site equals ρ_i. Then $S_g(\rho)$ is large if most of the mass is located in places where g is large — that is, most of the mass is located in places where there isn't a lot of mass. This means that the mass is widely spread.

However, we also saw in Exercise 5.2 that not every decreasing function $g: \ (0,1) \to (0,\infty)$ yields a reasonable measure of spread. For instance, $g(\rho_i) = \frac{1}{\rho_i}$ yields $S_g(\rho) = n$ for all ρ. And $g(\rho_i) = \frac{1}{\rho_i^2}$ yields $S_g(\rho) = \sum_{i=1}^{n} \frac{1}{\rho_i}$, which becomes arbitrarily large as one particular ρ_i tends to 1 and all the others tend to zero — so that in fact seems to be a measure of *concentration*, not of *spread.*

Which are the reasonable choices of g? We should surely require that $S_g(\rho)$ is maximal for the uniform probability vector

$$\rho = u = \left(\frac{1}{n}, \ldots \frac{1}{n}\right).$$

Note that

$$S_g(u) = g\left(\frac{1}{n}\right).$$

Lemma 13.1. *If $g: (0,1) \to (0,\infty)$ is strictly decreasing and differentiable, and $rg(r)$ is strictly concave-down, then $S_g(\rho) < S_g(u)$ for all positive probability vectors $\rho \neq u$.*

[19] It would seem more sensible to denote the independent variable of g by ρ, but ρ is already taken in this section — it denotes the probability vector $(\rho_1, \ldots, \rho_n)$.

Proof. Write

$$h(r) = rg(r).$$

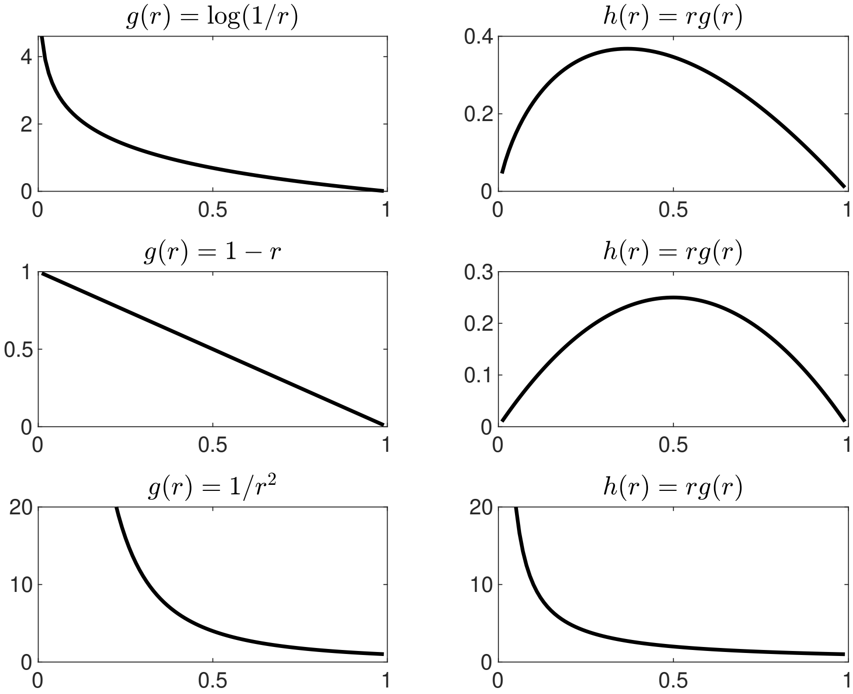
Let ρ be a positive probability vector of length n with $\rho \neq u$. Then

$$S_g(\rho) = \sum_{i=1}^{n} h(\rho_i) < \sum_{i=1}^{n} \left[h\left(\frac{1}{n}\right) + h'\left(\frac{1}{n}\right)\left(\rho_i - \frac{1}{n}\right)\right] = \sum_{i=1}^{n} h\left(\frac{1}{n}\right) = S_g(u)$$

because h is strictly concave-down. □

13.5.2 Three choices of g

The following figure shows, in the left column, three different choices of $g(r)$, and in the right column, the corresponding functions $h(r) = rg(r)$.



The first row in the figure corresponds to Boltzmann's choice. The second row yields the entropy

$$S_g(\rho) = \sum_{i=1}^{n} \rho_i(1 - \rho_i) = 1 - \|\rho\|^2.$$

The third row corresponds to $g(r) = 1/r^2$, which does not give rise to a reasonable notion of entropy, as noted earlier.

13.5.3 Tsallis entropy

Constantino Tsallis, a Brazilian physicist, proposed to use

$$g(r) = 1 - r^{q-1}, \tag{13.5}$$

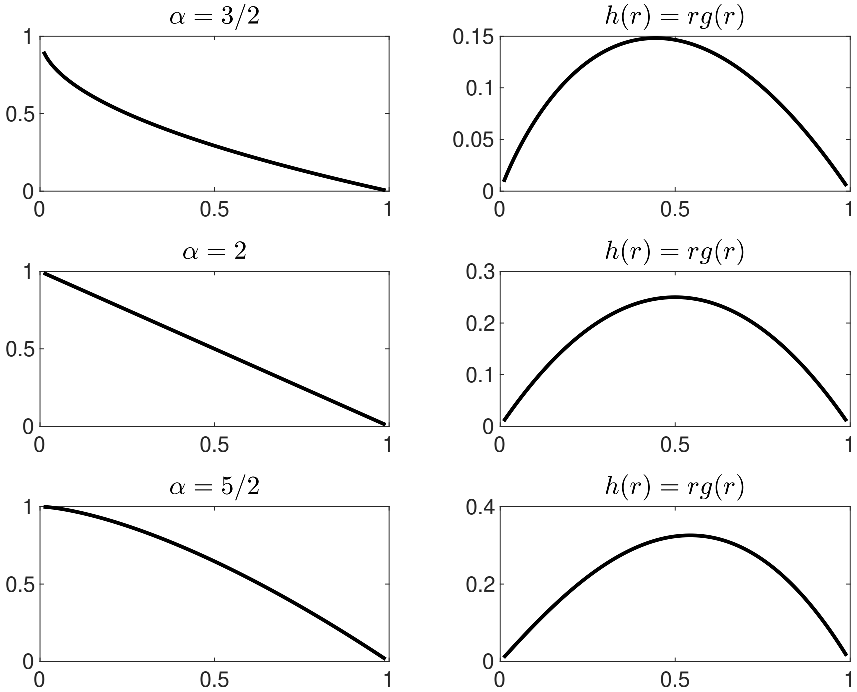
where $q > 1$ is called the *entropic index.* This yields

$$S_g(\rho) = \sum_{i=1}^{n} (\rho_i - \rho_i^q) = 1 - \|\rho\|_q^q.$$

The function

$$h(r) = r - r^q$$

is strictly concave-down for any $q > 1$. Therefore $S_g(\rho)$ is a sensible definition of entropy for any index $q > 1$.

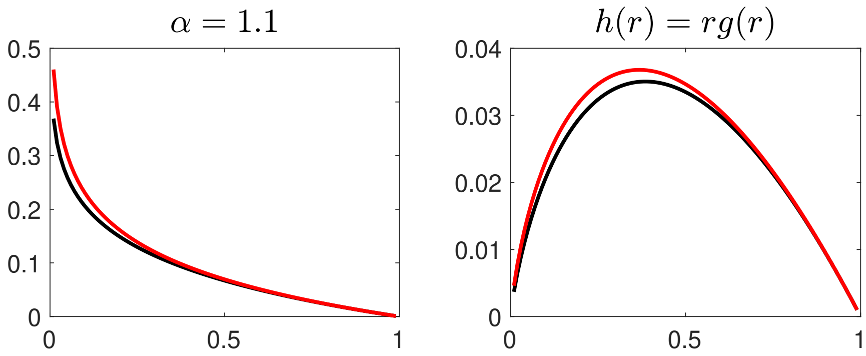


13.5.4 The Boltzmann limit of Tsallis entropy

The local linear approximation of $g(r) = 1 - r^{q-1}$ around $q = 1$ is

$$(q-1)\ln\frac{1}{r}.$$

Here for instance is the case of $q = 1.1$, together with $(q-1)\ln\frac{1}{r}$ (in red):



So up to a factor of $q-1$, the Tsallis entropy becomes the Boltzmann entropy in the limit as $q \to 1+$. In fact, Tsallis scaled his definition like this:

$$g(r) = \frac{1 - r^{q-1}}{q-1}.$$

Then the Tsallis entropy becomes the Boltzmann entropy as $q \to 1+$.

13.6 How physical forces combine

13.6.1 The question

Consider two physical forces, represented by vectors $F \in \mathbb{R}^3$ and $G \in \mathbb{R}^3$. When both forces act on the same object, what should be the vector representing their combined effect? We learn in high school that the answer is $F + G$ — we add the two vectors. But why should that be true? This question has a long history. Here are some references:

D'Alembert, Histoire de l'Académie des Sciences, Paris 1769
Darboux, Bulletin des Sciences Mathématiques 1875
Hamel, Mathematische Annalen 1905
Miklós Laczkovich, Acta Mathematica Hungarica 1995

We will *not* assume that the appropriate way of combining F and G is vector addition, but simply assume that they are combined in *some* way, which we denote by $F \oplus G$. We will formulate axioms that the operation $\oplus$ should satisfy, and then investigate whether those axioms force us to define $\oplus = +$.

13.6.2 Axioms

Axioms for force addition.

1. *The commutative and associative laws:*

   (a) $F \oplus G = G \oplus F$ *for all* $F, G \in \mathbb{R}^3$

   (b) $(F \oplus G) \oplus H = F \oplus (G \oplus H)$ *for all* $F, G, H \in \mathbb{R}^3$

2. *The combination of forces is rotation and reflection invariant:* $(QF) \oplus (QG) = F \oplus G$ *if* $F, G \in \mathbb{R}^3$ *and* Q *is either a rotation, or a reflection.*

3. *One-dimensional forces add:*

$$F \oplus cF = (1 + c)F \quad \textit{for } c \in \mathbb{R},\ F \in \mathbb{R}^3$$

The first two are pretty straightforward, but why should the third hold? I have no compelling answer to that.

13.6.3 Additive bijections from $\mathbb{R}$ into $\mathbb{R}$

An additive function from $\mathbb{R}$ into $\mathbb{R}$ is of the form

$$f\left(\sum_i c_i b_i\right) = \sum_i c_i r_i b_i, \tag{13.6}$$

where the b_i are a basis of $\mathbb{R}$ over $\mathbb{Q}$, the c_i are rational coefficients, the r_i are real numbers, and the sums are finite. Any of these functions is automatically odd. Which ones are bijections?

Lemma 13.2. *Either one of the following conditions implies that* (13.6) *is bijective.*

(a) *All* r_i *are the same, and their common value is not zero.*

(b) *All* r_i *are rational and non-zero.*

(c) *For all* i,

$$r_i = \frac{b_{\sigma_i}}{b_i},$$

where σ *is a permutation of the indices* i*; that is,* $\sigma : I \to I$ *is a bijection, where* I *is the set of indices* i.

Proof. (a) means $f(x) = rx$ for some non-zero r, so of course f is then bijective. To prove (b) and (c), we must show that for any

$$y = \sum_i d_i b_i$$

(with $d_i \in \mathbb{Q}$ and only finitely many of them non-zero), there is exactly one

$$x = \sum_i c_i b_i$$

(with $c_i \in \mathbb{Q}$ and only finitely many of them non-zero) so that

$$\sum_i c_i r_i b_i = \sum_i d_i b_i.$$

Under the condition (b), this simply means $c_i = d_i/r_i$, which will be rational since r_i is rational. Under the condition (c), we need

$$\sum_i c_i b_{\sigma_i} = \sum_i d_i b_i = \sum_i d_{\sigma_i} b_{\sigma_i},$$

which means $c_i = d_{\sigma_i}$. □

13.6.4 All the operations that satisfy the axioms

Theorem 13.3. *(a) Let*

$$f : \mathbb{R} \to \mathbb{R}$$

be additive and bijective. Define a mapping

$$\varphi_f : \mathbb{R}^3 \to \mathbb{R}^3$$

by

$$\varphi_f(F) = \begin{cases} f(\|F\|) \frac{F}{\|F\|} & \text{if } F \neq 0, \\ 0 & \text{if } F = 0. \end{cases}$$

For $F, G \in \mathbb{R}^3$, define

$$F \oplus G = \varphi_{f^{-1}} \left(\varphi_f(F) + \varphi_f(G) \right).$$

Then $\oplus$ satisfies the axioms stated in Section 13.6.2.

(b) Any operation $\oplus$ satisfying the axioms of Section 13.6.2 is of this form.

I learned this theorem from a paper by Miklós Laczkovich, titled "On the resultant of forces", *Acta Mathematica Hungarica* 1995. There isn't enough space here to think about its proof, although the proof is not overly complicated. I will, however, think about its consequences.

13.6.5 If f is measurable, then $\oplus$ is $+$

We know that f, if measurable, must be linear:

$$f(x) = rx$$

for some $r \in \mathbb{R}$. Since f is also a bijection, $r \neq 0$. Therefore

$$\varphi_f(F) = rF,$$

and

$$F \oplus G = \varphi_{f^{-1}}\left(\varphi_f(F) + \varphi_f(G)\right) = \frac{1}{r}(rF + rG) = F + G.$$

13.7 Emancipation of the dissonance

13.7.1 A minute of music

Click **here** for a great early example of Schönberg's 12-tone music, the piano suite opus 25. Listen to number 2, titled Gavotte. It is only about a minute long.

13.7.2 10-tone equal temperament

How does this fit into our discussion of 12 tones in the octave? The choice of 12 (rather than, say, 10, or any other number you like) makes mathematical sense only if you have a preference for consonance over dissonance. European classical music gradually abandoned that preference over the centuries, starting very early with J. S. Bach if not earlier. (It appears that the Gamelan music tradition never had it in the first place.) Click **here** for a piece of music in 10-tone equal temperament.

13.8 Thinking backwards to prove limit statements

It occurred to me that the proofs of limit statements are often examples of "thinking backwards". This is what I will explain here.

13.8.1 $\Rightarrow$ and $\Leftarrow$ between inequalities

Useful observation. If $\tilde{a} \leq a$ then

$$a < b \quad \Rightarrow \quad \tilde{a} < b.$$

If $\tilde{b} \geq b$ then

$$a < b \quad \Rightarrow \quad a < \tilde{b}.$$

If $\tilde{a} \geq a$ then

$$a < b \quad \Leftarrow \quad \tilde{a} < b$$

If $\tilde{b} \leq b$ then

$$a < b \quad \Leftarrow \quad a < \tilde{b}.$$

For short:

1. *If you make the smaller side smaller or the larger side larger, you get an implication $\Rightarrow$.*

2. *If you make the smaller side larger or the larger side smaller, you get a reverse implication $\Leftarrow$.*

Similar statements hold with $<$ replaced by $>$ or by $\leq$ or by $\geq$. When you think about it, these statements are all obvious. Nonetheless, it's good to keep in mind.

13.8.2 Proving limit statements

You probably know this:

Definition 13.4. *If $\{a_n\}_{n\geq 1}$ is a sequence of real numbers, and $a \in \mathbb{R}$, we say that*

$$\lim_{n\to\infty} a_n = a$$

if for every $\epsilon > 0$ there is an $R \in \mathbb{R}$ so that $|a_n - a| < \epsilon$ for all n with $n > R$. With quantifiers:

$$\forall \epsilon > 0\ \exists R \in \mathbb{R}\ \forall n \in \mathbb{N} \quad n > R \Rightarrow |a_n - a| < \epsilon$$

As a simple example, let's prove that

$$\lim_{n\to\infty} \frac{2n+3}{n+7} = 2,$$

using the definition. Let $\epsilon > 0$ we start at the end:

$$\begin{aligned} & |a_n - a| < \epsilon \\ \Leftrightarrow \quad & \left|\frac{2n+3}{n+7} - 2\right| < \epsilon \end{aligned}$$

$$\begin{aligned}
\Leftrightarrow &\quad \left|\frac{-11}{n+7}\right| < \epsilon \\
\Leftrightarrow &\quad \frac{11}{n+7} < \epsilon \\
\Leftrightarrow &\quad n > \frac{11}{\epsilon} - 7
\end{aligned}$$

You will notice that we started at the end. We started with $|a_n - a| > \epsilon$. It would have been harder to see that $R = \frac{11}{\epsilon} - 7$ works from the start.

For a slightly harder example, let's prove that

$$\lim_{n\to\infty} [(n+1)^p - n^p] = 0 \quad \text{if } 0 < p < 1.$$

Here the trick is to observe that the graph of x^p, $x > 0$, is concave-down when $0 < p < 1$. This implies that

$$\frac{(n+1)^p - n^p}{1},$$

the slope of the secant between $x = n$ and $x = n + 1$, is smaller than pn^{p-1}, the slope of the tangent at $x = n$. Therefore

$$\begin{aligned}
&\quad |(n+1)^p - n^p| < \epsilon \\
\Leftrightarrow &\quad (n+1)^p - n^p < \epsilon \\
\Leftarrow &\quad pn^{p-1} < \epsilon \\
\Leftrightarrow &\quad \frac{p}{\epsilon} < n^{1-p} \\
\Leftrightarrow &\quad n > \left(\frac{p}{\epsilon}\right)^{1/(1-p)}.
\end{aligned}$$

13.9 Nash equilibria and Brouwer's fixed point theorem

13.9.1 Two-player games

Suppose two people are playing a game. The first player has m possible strategies, and the second has n strategies. If the first player chooses strategy i and the second chooses strategy j, then the payout to the first player is $a_{ij} \geq 0$, and the payout to the second is $b_{ij} \geq 0$. It is no restriction of generality to assume the payouts to be non-negative: We could charge each of the two players the same entrance fee first, so that any payout smaller than the entrance fee would amount to a negative "payout".

So the payouts are characterized by the $m \times n$-matrices

$$A = [a_{ij}]_{1\leq i\leq m, 1\leq j\leq n} \quad \text{and} \quad B = [b_{ij}]_{1\leq i\leq m, 1\leq j\leq n}.$$

13.9.2 Mixed strategies

The strategies $i = 1, \ldots, m$ (for the first player) and $j = 1, \ldots, n$ (for the second player) are also called *pure* strategies. A player might not settle on one of the possible pure strategies, but pick one of them at random. For instance, the first player might pick an $i \in \{1, \ldots, m\}$ at random, with the probability of choosing strategy i equal to p_i. The probability vector

$$p = \begin{bmatrix} p_1 \\ p_2 \\ \cdots \\ p_m \end{bmatrix} \in \mathbb{R}^m$$

(thought of as a column vector here) is called a *mixed strategy* for the first player. Think of a soccer player and a goalie during a penalty kick. The player has a choice between aiming for the left corner of the goal or for the right corner. The player will want to make the choice between those two options random, so as not to be predictable. Similarly, a mixed strategy of the second player is a probability vector

$$q = \begin{bmatrix} q_1 \\ q_2 \\ \cdots \\ q_n \end{bmatrix} \in \mathbb{R}^n.$$

The second player chooses the j-th pure strategy with probability q_j.

13.9.3 Definition of Nash equlibria

Assuming that the players make independent choices, the *expected* (mean) payout for the first player is $p^T Aq$, and for the second player it is $p^T Bq$.

Definition 13.5. *Using the notation above, the pair (p, q) is a* Nash equilibrium *if*

$$p^T Aq \geq \tilde{p}^T Aq$$

and

$$p^T Bq \geq p^T B\tilde{q}$$

for all probability vectors $\tilde{p} \in \mathbb{R}^m$ and $\tilde{q} \in \mathbb{R}^n$.

So (p, q) is a Nash equilibrium if p is optimal for the first player, given that the second player adopts the mixed strategy q, and vice versa. Here "optimal" means "expectation-maximizing". One could also think about reducing variance, for the risk-averse among us, but that's not what we'll do here.

13.9.4 Existence of Nash equlibria

Theorem 13.6 (John Nash, 1950). *With the notation introduced above, there exist probability vectors $p^* \in \mathbb{R}^m$ and $q^* \in \mathbb{R}^n$ so that (p^*, q^*) is a Nash equilibrium.*

Proof. Suppose p is any mixed strategy for player 1. We compare it with the i-th pure strategy. We see the i-th pure strategy as a special case of a mixed strategy, identifying it with the canonical basis vector $e_i \in \mathbb{R}^m$. If the first player uses e_i as their strategy, their payout changes from from $p^T Aq$ to $e_i^T Aq$. If this is a great improvement, then it makes sense for the first player to adopt a strategy that makes choosing the pure strategy i much more likely. This motivates replacing the p_i by

$$P_i = p_i + \max\left(e_i^T Aq - p^T Aq, 0\right), \quad 1 \le i \le m.$$

The sum of the P_i, defined in this way, would not be 1 any more. Therefore we define instead:

$$P_i = \frac{p_i + \max\left(e_i^T Aq - p^T Aq, 0\right)}{1 + \sum_{k=1}^m \max\left(e_k^T Aq - p^T Aq, 0\right)}, \quad 1 \le i \le m.$$

Similarly, we define

$$Q_j = \frac{q_j + \max\left(p^T Ae_j - p^T Aq, 0\right)}{1 + \sum_{\ell=1}^n \max\left(p^T Ae_\ell - p^T Aq, 0\right)}, \quad 1 \le j \le n.$$

The vectors $P \in \mathbb{R}^m$ and $Q \in \mathbb{R}^n$ are then probability vectors. We denote the mapping from (p, q) to (P, Q) by F:

$$F: \ (p, q) \mapsto (P, Q).$$

We note that F is continuous.

Since p is a probability vector, it can be specified by merely specifying the probabilities $p_1, \ldots, p_{m-1}$ with $p_i \ge 0$ and $\sum_{i=1}^{m-1} p_i \le 1$. Then

$$p_m = 1 - \sum_{i=1}^{m-1} p_i.$$

Analogous statements hold for q, P, and Q. We define

$$\Delta = \Bigg\{ (p_1, \ldots, p_{m-1}, q_1, \ldots, q_{n-1}) \in \mathbb{R}^{m+n-2} \ : \\ p_i \ge 0 \text{ for } 1 \le i \le m-1, \quad \sum_{i=1}^{m-1} p_i \le 1, \\ q_j \ge 0 \text{ for } 1 \le j \le n-1, \quad \sum_{j=1}^{n-1} q_j \le 1 \Bigg\}.$$

The mapping F can also be seen as a continuous mapping $F: \ \Delta \to \Delta$. By Brouwer's theorem, it has a fixed point. Thus there is a pair (p_*, q_*) so that the corresponding pair (P_*, Q_*) equals (p_*, q_*). We will now show that (p_*, q_*) is a Nash equilibrium.

The equation $P^* = p^*$ is equivalent to

$$\forall i \in \{1, \dots, m\}$$

$$\max(e_i^T A q^* - (p^*)^T A q^*, 0) = p_{*,i} \sum_{k=1}^{m} \max\left(e_k^T A, q^* - (p^*)^T A q^*, 0 \right). \qquad (13.7)$$

This is equivalent to

$$\exists c \geq 0 \quad \forall i \in \{1, \dots, m\} \quad \max\left(e_i^T A q^* - (p^*)^T A q^*, 0 \right) = c\, p_i^*. \qquad (13.8)$$

It is obvious that (13.7) implies (13.8), with

$$c = \sum_{k=1}^{m} \max\left(e_k^T A, q^* - (p^*)^T A q^*, 0 \right).$$

To see that (13.8) also implies (13.7), sum (13.8) over i.

We will show that $c = 0$. Suppose $c > 0$. Then for all those $i \in \{1, \dots, m\}$ for which $p_i^* > 0$, the left-hand side of the equation in (13.8) would have to be positive, so

$$e_i^T A q^* > (p^*)^T A q^*.$$

Therefore:

$$(p^*)^T A q^* = \left(\sum_{i=1}^{m} p_i^* e_i \right)^T (A q^*) = \sum_{i=1}^{m} p_i^* \left(e_i^T A q^* \right) > \sum_{i=1}^{m} p_i^* \left(p^* A q^* \right) = p^* A q^*,$$

a contradiction.

So $c = 0$, and therefore $(p^*)^T A q^* \geq e_i^T A q^*$ for all i. For any probability vector $\tilde{p}$, then

$$(\tilde{p})^T A q^* = \left(\sum_{i=1}^{m} \tilde{p}_i e_i \right)^T A q^* = \sum_{i=1}^{m} \tilde{p}_i \left(e_i^T A q^* \right) \leq \sum_{i=1}^{m} \tilde{p}_i \left((p^*)^T A q^* \right) = (p^*)^T A q^*.$$

Similarly, for any probability vector $\tilde{q} \in \mathbb{R}^n$,

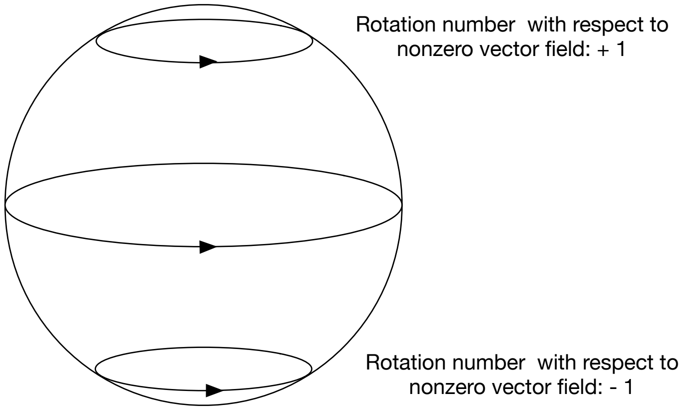
$$(p^*)^T B \tilde{q} \leq (p^*)^T B q^*.$$

This proves that (p^*, q^*) is a Nash equilibrium. □

This proof appeared (for the case of arbitrarily many players) in John Nash's 1951 paper in the *Annals of Mathematics*, which was based on his Ph. D. thesis at Princeton University. Less than a decade letter, he was found to suffer from schizophrenia. For his contributions to game theory, he received the Nobel Prize in Economics in 1994. For his contributions to the theory of partial differential equations (during a few years following his Ph. D.), he received the 2015 Abel Prize, the equivalent of a Nobel Prize in Mathematics, awarded annually by the King of Norway. On his way back to Princeton from Oslo, a few days after having received the Abel Prize, Nash and his wife died in a car accident on the New Jersey turnpike in May 2015.

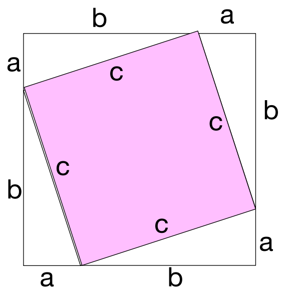
13.10 Single-picture proofs

13.10.1 Hedgehog theorem


Rotation number with respect to nonzero vector field: + 1
Rotation number with respect to nonzero vector field: - 1

13.10.2 Pythagorean theorem

Single-picture proofs are appealing. A particularly famous example:



$$4 \cdot \tfrac{ab}{2} + c^2 = (a+b)^2 \;\Leftrightarrow\; a^2 + b^2 = c^2$$

13.10.3 Cauchy-Schwarz inequality

Proofs that can be grasped in "one shot" of algebra are almost equally appealing. Here for instance is the proof that the triangle inequality and the Cauchy-Schwarz inequality are the same statement (compare Exercise 4.5):

$$\begin{array}{ll} & \|x+y\| \le \|x\| + \|y\| \\ \Leftrightarrow & \|x+y\|^2 \le \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 \\ \Leftrightarrow & \|x\|^2 + 2x\cdot y + \|y\|^2 \le \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 \\ \Leftrightarrow & x\cdot y \le \|x\|\,\|y\| \end{array}$$

13.10.4 Cramer's rule

Theorem 13.7 (Cramer's rule). *Let $A \in \mathbb{R}^{n\times n}$ be invertible, and $b \in \mathbb{R}^n$. Then the solution of the system $Ax = b$ is given by*

$$x_i = \frac{\det A_i}{\det A}, \quad 1 \le i \le n,$$

where A_i is the matrix obtained by replacing the i-th column in A by b.

Proof. The following (almost) "one-shot proof" is due to Richard Ehrenborg, *Mathematics Magazine* 2004. We know that it is possible to use "elementary row operations" to convert

$$Ax = b$$

into an equivalent system of the form

$$Ix = y, \tag{13.9}$$

which then of course has the solution $x = y$. There are three kinds of elementary row operations: (i) Add one equation to another, (ii) multiply an equation with a nonzero scalar, (iii) exchange two equations. Notice that Cramer's rule clearly holds for (13.9). Therefore it is sufficient to prove that none of these three operations alters

$$\frac{\det A_i}{\det A}. \tag{13.10}$$

Adding one row to another changes neither $\det A_i$ nor $\det A$. Multiplying a row by a non-zero constant multiplies both $\det A_i$ and $\det A$ by that constant, and therefore leaves the ratio (13.10) unchanged. Exchanging two rows changes the signs of both $\det A_i$ and $\det A$, and therefore leaves the ratio (13.10) unchanged. □

Cramer's rule is an inefficient way of computing the solution of $Ax = b$.

13.11 The Cantor set

13.11.1 Definition

Cantor defined a set with interesting properties which has since played a role in many different contexts in mathematics. The definition is as follows. First, define

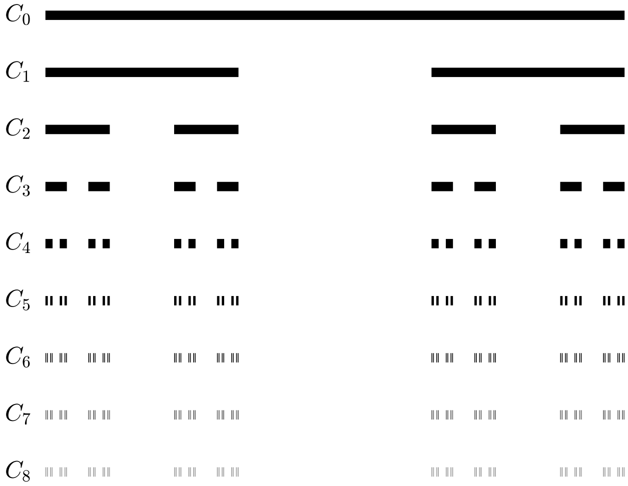
$$C_0 = [0, 1].$$

Then define C_1 to be the set obtained by removing the open middle third from C_0:

$$C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right].$$

In general, C_n $n \in \mathbb{N}$, will be a disjoint union of 2^n closed intervals, and C_{n+1} is obtained from C_n by removing the open middle thirds from each of the intervals forming C_n. For instance,

$$C_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{1}{3}\right] \cup \left[\frac{2}{3}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right].$$



The Cantor set is the intersection of all the C_n:

$$C = \bigcap_{n \geq 0} C_n.$$

The C_n, being finite unions of closed intervals, are closed sets. (This means that the boundary points of C_n belong to C_n.) Then C, being the intersection of closed sets, is a closed set. It is certainly not empty. For instance, $0 \in C$ and $1 \in C$. Also, $1/3 \in C$ and $2/3 \in C$, etc.

13.11.2 Measure zero

Clearly C must be a very "sparse" subset of $[0, 1]$. Even C_8 just looks like a few specks of dust on the page, and C is just a tiny fraction of C_8! The Cantor set C is a set of *measure zero*. This means that for any $\epsilon > 0$, it is possible to find a union of countably many intervals of total length $\leq \epsilon$ that contains C. In fact, if we choose n so large that $\left(\frac{2}{3}\right)^n \leq \epsilon$, then C_n is such a union of intervals.

13.11.3 $|C| = |\mathbb{R}|$

Proposition 13.8. *The cardinality of the Cantor set is the cardinality of the entire real line.*

Proof. Every point in the Cantor set has a unique "address" that is an infinite sequence of L's (for *left*) and R's (for *right*). For instance,

$$LLRRLRR\ldots$$

would mean that the number is in the left part of C_1 (so in $\left[0, \frac{1}{3}\right]$), and in the left part of that (so in $\left[0, \frac{1}{9}\right]$, but in the right part of that (so in $\left[\frac{2}{27}, \frac{1}{9}\right]$), and so on. The set of all infinite sequences of L's and R has the same cardinality of the set of all infinite sequences of 0's and 1's. An infinite sequence of 0's and 1's can be thought of as a binary representation of a number in $[0, 1]$. For instance,

$$01100101110\ldots$$

is identified with the binary representation

$$0.01100101110\ldots = \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^6} + \frac{1}{2^8} + \frac{1}{2^9} + \frac{1}{2^{10}} + \ldots$$

There is one catch, having to do with infinite tails of 1's: The binary representations

$$0.010111111111\ldots \quad \text{(all following digits are 1)}$$

and

$$0.0110000000000\ldots \quad \text{(all following digits are 0)}$$

denote the exactly same number. There are, however, only countably many sequences of 0's and 1's that end with an infinite trail of 1's. (Can you see why this is true?) Therefore the cardinality of C is the cardinality of $[0, 1] \cup \mathbb{N}$, but that's the same as the cardinality of $[0, 1]$. (Can you define a bijection between $[0, 1]$ and $[0, 1] \cup \mathbb{N}$?) The cardinality of $[0, 1]$ is the same as the cardinality of $\mathbb{R}$. □

13.11.4 Cantor set and infinite sequences of coin tosses

If you toss a coin infinitely many times (this of course is an idealized way of thinking about *very many* coin tosses), the outcome can be thought of as an infinite sequences of heads and tails:

$$HHTTTHHTH\ldots$$

This in turn can be identified with an infinite sequence of L's and R's, in other words, with an element of C. When thinking about an infinite sequence of coin tosses, the correct *sample space* (set of possible outcomes of the experiment) is, or can be identified with, the Cantor set C.

13.12 Smith-fair voting methods

13.12.1 Dominating sets of candidates

Definition 13.9. *Given a preference schedule, a* dominating set *of candidates is a non-empty set $\mathcal{D}$ with the property that every candidate inside $\mathcal{D}$ would beat every candidate outside $\mathcal{D}$ in head-to-head comparison.*

Notice that there is always a dominating set, namely the set of all candidates. A candidate X is a Condorcet candidate if and only if {X} is dominating.

Lemma 13.10. *If $\mathcal{D}$ and $\mathcal{E}$ are dominating sets of candidates, then $\mathcal{D} \subseteq \mathcal{E}$ or $\mathcal{E} \subseteq \mathcal{D}$.*

Proof. Suppose not, so there are candidates X and Y with $\mathrm{X} \in \mathcal{D}-\mathcal{E}$ and $\mathrm{Y} \in \mathcal{E}-\mathcal{D}$. Then in head-to-head comparison, X beats Y because $\mathrm{X} \in \mathcal{D}$ and $\mathrm{Y} \notin \mathcal{D}$, but also Y beats X because $\mathrm{Y} \in \mathcal{E}$ and $\mathrm{X} \notin \mathcal{E}$. This is a contradiction, proving the assertion. □

13.12.2 The Smith set

Definition 13.11. *The smallest dominating set, i.e., the intersection of all dominating sets (which are, by Lemma 13.10, nested in each other) is called the* Smith set*.*

The Smith set is named after John Howard Smith, an emeritus professor at Boston College. He is mostly a number theorist. Intuitively, the Smith set is the set of "strong" candidates. We will also refer to the members of the Smith set as the *Smith candidates.*

13.12.3 Smith-fair methods

Definition 13.12. *A single-winner election method is called* Smith-fair *if the winner is in the Smith set in all cases.*

Clearly, a Smith-fair method must be Condorcet fair. Among the methods we have studied, therefore, the only one that has a chance at being Smith-fair is pairwise comparison. In fact, it is.

Proposition 13.13. *The method of pairwise comparison is Smith-fair.*

Proof. Suppose there are k Smith candidates, and $n - k$ non-Smith candidate. In the pairwise comparison tournament, Smith candidates get at least $n - k$ points, and non-Smith candidates get at most $n - k - 1$ points. □

13.12.4 A priori Smith-fair methods

Definition 13.14. *A singler-winner election method is called* a priori Smith-fair *if non-Smith candidates never affect the outcome of the election.*

In other words, a method is *a priori Smith-fair* if the determination of the winner could begin by removing all non-Smith candidates from the preference schedules, without changing the outcome.

Proposition 13.15. *Pairwise comparison is a priori Smith-fair.*

Proof. Suppose there are k Smith candidates, and $n - k$ non-Smith candidates. The non-Smith candidates will never win. Each Smith candidate beats each non-Smith candidate in pairwise comparison. Therefore the presence of the non-Smith candidates simply adds $n - k$ pairwise comparison points to each Smith candidate. This cannot affect which of the Smith candidates wins. □

Are there any methods that are Smith-fair, but not *a priori* Smith-fair? The answer is yes. For instance: Declare as the winner the Smith candidate with the most Borda points, even if there is a non-Smith candidate with more Borda points. This is obviously Smith fair, but you can find examples demonstrating that the presence of non-Smith candidates can affect the winner.

13.12.5 The winner selection method that I would recommend

I said earlier that we should idetermine whether there is a Condorcet candidate, and if so, declare that Condorcet candidate the winner, otherwise use instant runoff. One might call this method *a priori Condorcet-fair instant runoff*. A slight refinement is *a priori Smith-fair instant runoff*: Remove all non-Smith candidates from the ballots before carrying out the instant runoff. I think this is the best among those methods that are simple enough to use in practice.

13.13 The final exam

There is only one question on the final exam for this course:

Ten years from now, will you still read about mathematics?

If yes, I pass, but if no, I fail. Mathematics, just like music, art, novels, poems, and walks in the woods, is good for the soul. If edication in mathematics, music, art, literature is to have any meaning for people who don't occupy themselves with these fields professionally, it must be through lifelong non-professional interests.[20]

In no sense is an amateur interest in mathematics inferior to an anxiety- or ambition-driven professional research interest. There are many books, videos, and publications that can help fuel such an interest. Here are some examples:

- "The Spirit of Mathematics" by David Acheson. (I happened to find this book recently, and I find it excellent.)
- "Sync" by Steven Strogatz. (A book on synchronization phenomena.)
- "How not to be wrong" by Jordan Ellenberg. (This book has a lot about conditional probability. Ellenberg has another book that also looks very nice titled "Shape". I have only briefly browsed through it, though.)
- The "Mathematical Gazette". (A British journal with beautiful articles for broad audiences.)
- The "Mathematics Magazine". (A U.S. journal similar to the Gazette.)
- 3Blue1Brown. (The best mathematics videos that I have seen.)

Send me your suggestions! I will read or watch them, and add them to my list.

[20] I do not believe the claim that studying some mathematics for a few years makes you a better thinker somehow.

Appendix A

Answers to some of the questions

Section 1.6.4: Denote the numbers of pentagonal, square, and triangular faces by p, s, and t, respectively. Each pentagon has five square neighbors, but each square has two pentagonal neighbors. Therefore

$$s = \frac{5p}{2}.$$

Each pentagon touches five triangles at the edges, but each triangle is touched by three pentagons. Therefore

$$t = \frac{5p}{3}.$$

These two equations imply that p is divisible by 2 and 3, so by 6. It's clearly greater than 6 and smaller than 18, so it must be 12. Therefore $s = 30$, $t = 20$, and the total number of faces is $f = 62$.

Every vertex belongs to exactly one pentagon, but each pentagon has five vertices, therefore $p = \frac{v}{5}$, or $v = 5p = 60$. In each vertex, four edges come together, but each edge belongs to two vertices, so $e = \frac{4v}{2} = 2v$, or $e = 120$. So in summary,

$$(v, e, f) = (60, 120, 62).$$

Section 1.6.5: Suppose there were an Archimedean solid with four or more different kinds of faces. All vertices look identical. So among the faces that come together in a vertex, all kinds that appear in the solid must be represented. Let's say that regular polygons with $k < \ell < m < n$ vertices are among them, with $k \geq 3$, and therefore $\ell \geq 4$, $m \geq 5$, $n \geq 6$. The interior angles of these faces are

$$\geq \left(\frac{3-2}{3} + \frac{4-2}{4} + \frac{5-2}{5} + \frac{6-2}{6} \right) \pi = \frac{21}{10}\pi > 2\pi.$$

But the sum of the interior angles of the faces coming together in a vertex must be smaller than 2π.

Section 1.6.6: (a) You obtain the cells by setting one of the four coordinates to -1 or to 1, so $c = 8$.

(b) Think for instance about the cell

$$\{(x, y, z, -1) \ : \ -1 \le x, y, z \le 1\}.$$

It is a three-dimensional cube, so it has six faces. One of the faces is

$$\{(x, y, -1, -1) \ : \ -1 \le x, y \le 1\}.$$

That's also a face of the cell

$$\{(x, y, -1, u) \ : \ -1 \le x, y, u \le 1\}.$$

Each face is shared by two cells. Therefore

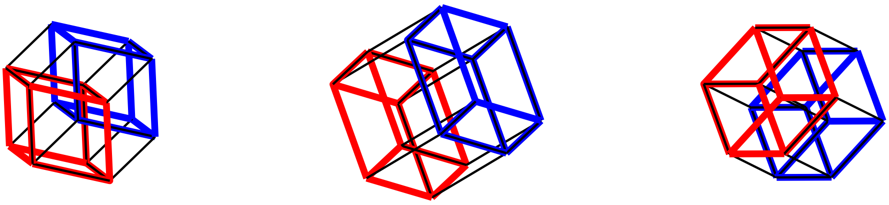
$$f = \frac{6c}{2} = 3c = 24.$$

(c) Each face has four edges. For instance,

$$\{(x, -1, -1, -1) \ : \ -1 \le x \le 1\}.$$

is an edge. It belongs to three faces:

$$\{(x, y, -1, -1) \ : \ -1 \le x, y \le 1\}, \ \ \{(x, -1, z, -1) \ : \ -1 \le a \le 1\}. \ \text{ and}$$



Section 2.6.2: Out of 10,000 women, 100 have breast cancer, and 90 of those will test positive. But also, of the 9,900 women without breast cancer, 9 percent, so 891,

will test positive. Altogether, in 10,000 women, 891+90=981 will test positive, and of those, 90 have breast cancer. Since 90/981 ≈ 0.092, among the offered answers the one that comes closest is “one in ten”.

Section 2.6.4: Observing the situation gave you the information that Mr. Smith has a boy, no more and no less. So here the answer is in fact $\frac{1}{3}$.

Section 2.6.5: The probability that a two-boy family will send in an Adam is

$$\frac{1}{2}p + \frac{1}{2}p(1-p).$$

The factors of $\frac{1}{2}$ correspond to the likelihood that the first or second boy is selected for the draft. The probability that the first boy is named Adam is p, and the probability that the second is named Adam is $p(1-p)$.

The probability that a family with an older boy and a younger girl will send in an Adam is p (namely, the probability that the boy is named Adam). The probability that a family with an older girl and a younger boy will send in an Adam is p as well.

Therefore the probability that a family sending in an Adam is a two-boy family is

$$\frac{\frac{1}{2}p + \frac{1}{2}p(1-p)}{\frac{1}{2}p + \frac{1}{2}p(1-p) + 2p} = \frac{p - \frac{p^2}{2}}{3p - \frac{p^2}{2}} = \frac{1 - \frac{p}{2}}{3 - \frac{p}{2}}. \tag{A.1}$$

Since p is usually close to 0, this is close to $\frac{1}{3}$, but it actually is a little less than $\frac{1}{3}$.

Section 2.6.6: A two-boy family will send in their Adam (their first-born) with probability $\frac{1}{2}$. A one-boy family will surely send in their Adam (their only boy). Therefore the probability that a family sending in an Adam is a two-boy family is now

$$\frac{\frac{1}{2}}{\frac{1}{2} + 1 + 1} = \frac{1}{5}.$$

You get this answer also if you set $p = 1$ in (A.1).

Section 3.6.2: We just have to set $n = s = 1$ in our general formula:

$$f(\beta) = 2\beta.$$

Section 3.6.3: We just have to set $n = 2$, $s = 1$ in our general formula:

$$f(\beta) = 6\beta(1-\beta).$$

Section 3.6.4: Let E be the event that Fate gets tails on one toss. We want to compute

$$P(B \le x \mid E) = P(E \mid B \le x)\,\frac{P(B \le x)}{P(E)}.$$

The right-hand side is evaluated based on the assumption that B has the density 2β. Then

$$P(B \leq x) = \int_0^x (2\beta)\, d\beta = x^2,$$

$$P(E) = \int_0^1 (1-\beta) \cdot 2\beta\, d\beta = \frac{1}{3}.$$

The calculation of $P(E \,|\, B \leq x)$ is a little tricky. The density of B is assumed to be 2β. The density of B, *given* that $B \leq x$, is *proportional* to 2β, but we must scale so that the integral from 0 to x comes out to be 1:

$$C \int_0^x (2\beta)\, d\beta = 1$$

means

$$C = \frac{1}{x^2}.$$

So

$$P(E \mid B \leq x) = \int_0^x (1-\beta) \frac{2\beta}{x^2} d\beta.$$

Altogether now, we have

$$P(B \leq x \mid E) = \int_0^x (1-\beta) \frac{2\beta}{x^2} d\beta \; \frac{x^2}{1/3} = \int_0^x 6\beta(1-\beta)\, d\beta.$$

So the conditional density indeed comes out to be $6\beta(1-\beta)$, precisely the same as in Section 3.6.3.

Section 3.6.5: Again we let E be the event that fate gets head on every single one of n tosses. We want to compute

$$P(B \leq x \mid E).$$

By Bayes' formula,

$$P(B \leq x \mid E) = P(E \mid B \leq x)\; \frac{P(B \leq x)}{P(E)}. \tag{A.2}$$

These probabilities, however, are now to be evaluated based on the belief that $B \in [0, 1/2]$ is uniformly distributed. This means that

$$P(B \leq x) = \begin{cases} 2x & \text{if } 0 < x \leq 1/2, \\ 1 & \text{if } x > 1/2, \end{cases}$$

and

$$P(E) = \int_0^{1/2} \beta^n \cdot 2\, d\beta = \frac{1}{(n+1)2^n}.$$

(The integrand β^n is there because that's the probability of n heads in a row if β is the probability of getting heads on one toss. It's $2d\beta$ because that's the probability of landing in an interval $I \subseteq (0, 1/2]$ of length $d\beta$.) Finally, we have to compute

$$P(E \mid B \leq x).$$

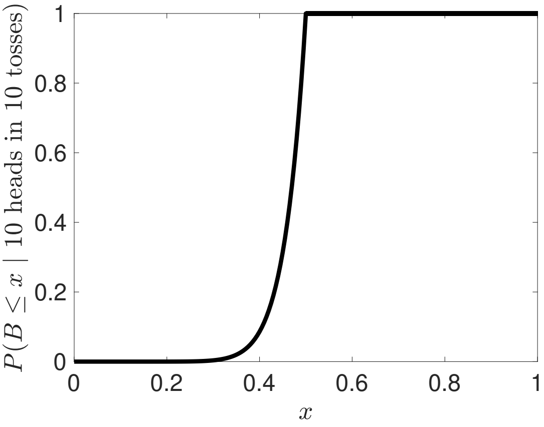
Remember we are assuming that B is uniformly distributed in $(0, 1/2]$. But now we only consider instances in which $B \leq x$. If $x \in (1/2, 1)$, then of course $P(E \mid B \leq x) = P(E)$. (If we believe that B is uniformly distributed in $(0, 1/2]$, then $B \leq 3/4$ is no news, for instance.) If $x \in (0, 1/2]$, then

$$P(E \mid B \leq x) = \int_0^x \beta^n \frac{1}{x} d\beta = \frac{x^n}{n+1}.$$

Putting all of these pieces together, and using (A.2), we obtain:

$$P(B \leq x \mid E) = \begin{cases} (2x)^{n+1} & \text{if } x \leq 1/2, \\ 1 & \text{if } x > 1/2. \end{cases}$$

So for instance, for $n = 10$, $P(B \leq x\, E)$ looks like this:



So after seeing 10 heads in 10 tosses, you will be pretty certain that the probability of heads on one toss is close to 1/2 — but you will never acknowledge that it might be bigger than 1/2. That's because you started out *certain* that it was $\leq 1/2$. To be able to learn the value of B from the evidence, you must start out open-minded, namely, you must start out with an assumed distribution according to which all values in $(0, 1)$ *possible* at least.

Chapter 4

1.

$$\int_0^\infty \sqrt{t}\; e^{-t}\, dt = \int_0^\infty u e^{-u^2}\, 2u du = \int_0^\infty u \cdot 2u e^{-u^2}\, du =$$

$$\left[-ue^{-u^2}\right]_0^\infty + \int_0^\infty e^{-u^2}\,du = \int_0^\infty e^{-u^2}\,du = \frac{1}{2}\int_{-\infty}^\infty e^{-u^2}\,du = \frac{\sqrt{\pi}}{2} \approx 0.886.$$

On the other hand,

$$F_0(0.5) = 1 - \frac{0.5(1-0.5)}{2} = 0.875.$$

They are similar, but slightly different.

Now we can also calculate

$$(-0.5)! = \frac{(0.5)!}{0.5} = \sqrt{\pi}$$

and

$$(-1.5)! = \frac{(-0.5)!}{-0.5} = -2\sqrt{\pi}.$$

That's a surprise: $(-1.5)!$ is negative!

2. Let $x = 1000$. Then

$$x! \approx x^x e^{-x}\sqrt{2\pi x}$$

and therefore

$$\log_{10}(x!) \approx x\log_{10}(x) + \log_{10}(e^{-x}) + \frac{1}{2}\log_{10}(2\pi x).$$

Now you have to know (or understand) that $\log_{10}(z) = \frac{\ln z}{\ln 10}$. So

$$\log_{10}(x!) \approx \log_{10}(x) - \frac{x}{\ln 10} + \frac{1}{2}\log_{10}(2\pi x).$$

Setting $x = 1000$, this yields 2567.60. So there are 2567 digits before the decimal point. (This assumes that Stirling's formula is accurate enough, but even for 5! it gives an answer that's accurate within an error smaller than 2%.)

3. We know that $F(0) = 1$, $F(1) = 1$. We also know that

$$F''(x) = \int_0^\infty (\ln t)^2\, t^x e^{-t}\,dt > 0$$

for all x.This means that F' is a strictly increasing function. Therefore, if $F'(0)$ were ≥ 0, then $F'(x) > 0$ for all $x > 0$, and $F(1)$ couldn't be the same as $F(0)$. Similarly, if $F'(0)$ were ≤ 0, then $F'(x) < 0$ for all $x < 1$, and $F(0)$ couldn't be the same as $F(1)$.

This proves $F'(0) < 0 < F'(1)$. But $F'(1) = 1 + F'(0)$ by (4.2), since F is a differentiable functions satisfying the recursion relation. If $F'(0)$ where ≤ -1, then $F'(1)$ would ≤ 0. Therefore

$$-1 < F'(0) < 0,$$

and with $F'(1) = 1 + F'(0)$, that implies

$$0 < F'(1) < 1.$$

4. The right Riemann sum with $\Delta x = 1$ is

$$\ln 1 + \ln 2 + \ldots + \ln n.$$

Here $\ln 1 = 0$ approximates $\int_0^1 \ln x dx < 0$. Similarly, $\ln 2$ approximates $\int_1^2 \ln x\,dx$, and since $\ln x < \ln 2$ for $1 < x < 2$, this integral is smaller than $\ln 2$. In general, for $1 \le k \le n$, $\ln k$ approximates $\int_{k-1}^k \ln x\,dx$, and since $\ln x < \ln k$ for $k-1 < x < k$, we conclude $\int_{k-1}^k \ln x\,dx < \ln k$. So

$$\sum_{k=1}^n \ln k > \int_0^n \ln x\,dx = n\ln n - n.$$

This means

$$\ln(n!) > n\ln n - n,$$

and therefore

$$n! > e^{n\ln n - n} = n^n e^{-n}.$$

The upper bound on $n!$ follows similarly. Finally,

$$(n+1)^{n+1} = \frac{(n+1)^{n+1}}{n^n}\,n^n = \left(\frac{n+1}{n}\right)^n (n+1)\,n^n =$$

$$\left(1+\frac{1}{n}\right)^n (n+1)\,n^n = e^{n\ln\left(1+\frac{1}{n}\right)}\,(n+1)\,n^n.$$

Now use $\ln(1+x) < x$ for all $x > -1$. (This is true because $y = x$ defines the tangent line to the graph of $\ln(1+x)$ at $x = 0$, and the graph is concave-down.) So we conclude

$$e^{n\ln\left(1+\frac{1}{n}\right)}\,(n+1)\,n^n < e^{n\frac{1}{n}}\,(n+1)\,n^n = e\,(n+1)\,n^n.$$

5.

$$\|g+h\| \le \|g\| + \|h\|$$

means

$$\|g+h\|^2 \le \|g\|^2 + \|h\|^2 + 2\|g\|\|h\|,$$

or

$$(g+h)^T(g+h) \le \|g\|^2 + \|h\|^2 + 2\|g\|\|h\|,$$

or

$$\|g\|^2 + 2g^Th + \|h\|^2 \le \|g\|^2 + \|h\|^2 + 2\|g\|\|h\|,$$

or

$$g^Th \le \|g\|\|h\|.$$

Similarly,

$$\|g-h\| \le \|g\| + \|h\|$$

means

$$-g^T h \leq \|g\|\|h\|.$$

Exercise 5.1. The intelligent way of finding a lower bound is to observe that

$$\sum_{i=1}^{n} \rho_i^2 = \sum_{i=1}^{n} \left(\rho_i^2 - 2a\rho_i + a^2\right) + 2a - a^2$$

for any constant a, since $\sum_{i=1}^{n} \rho_i = 1$. So minimizing $Q(\rho)$ is equivalent to minimizing

$$\sum_{i=1}^{n} \left(\rho_i^2 - 2a\rho_i + a^2\right) = \sum_{i=1}^{n} (\rho_i - a)^2.$$

You'd think the answer would be $\rho_i = a$ for all i, but that wouldn't satisfy the constraint $\sum_{i=1}^{n} \rho_i = 1$. Unless, of course, $a = 1/n$, so choose $a = 1/n$. The vector ρ that minimizes $Q(\rho)$ is the vector ρ that minimizes

$$\sum_{i=1}^{n} \left(\rho_i - \frac{1}{n}\right)^2,$$

and that's obviously $\rho_i = \frac{1}{n}$ for all i. The upper bound is even easier. Since $\rho_i \in [0,1]$ for all i, we have $\rho_i^2 \leq \rho_i$, and therefore $Q(\rho) \leq 1$. Equality holds only if $\rho_i^2 = \rho_i$ for all i, which is the case exactly if $\rho_i = 1$ for one i, and $\rho_i = 0$ for all others. So $Q(\rho)$ does seem to be a good measure of spread: It is minimal for maximal spread, and maximal for minimal spread.

Exercise 5.2.

$$1 - Q(\rho) = 1 - \sum_{i=1}^{n} \rho_i^2 = \sum_{i=1}^{n} \left(\rho_i - \rho_i^2\right) = \sum_{i=1}^{n} \rho_i(1 - \rho_i).$$

Does $g(x) = 1/x$ work? Then

$$S_g(\rho) = \sum_{i=1}^{n} \rho_i \cdot \frac{1}{\rho_i} = n$$

(assuming $\rho_i > 0$ for all i). So $S_g(\rho)$ is the same for all ρ, and surely does not measure spread. How about $g(x) = 1/x^2$? Then

$$S_g(\rho) = \sum_{i=1}^{n} \rho_i \cdot \frac{1}{\rho_i^2} = \sum_{i=1}^{n} \frac{1}{\rho_i}.$$

This can get arbitrarily large, of one of the ρ_i is close to 1 and therefore the others are close to zero. So this is a measure of *concentration.* In fact it is *smallest* for maximal spread, namely for $\rho_i = \frac{1}{n}$ for all i. How about $g(x) = -\ln x$? That gives the usual entropy: $\sum_{i=1}^{n} \rho_i \ln \frac{1}{\rho_i}$.

Exercise 5.3.

$$\sum_i \rho_i g(\rho_i) + \sum_j \eta_j g(\eta_j) = \sum_j \eta_j \sum_i \rho_i g(\rho_i) + \sum_i \rho_i \sum_j \eta_j g(\eta_j) =$$

$$\sum_i \sum_j \rho_i \eta_j (g(\rho_i) + g(\eta_j)).$$

Now assume

$$g(xy) = g(x) + g(y)$$

for all x, y. Notice that $g(1 \cdot 1) = g(1) + g(1)$ then, therefore $g(1) = 0$. If g is differentiable then $g(xy) = g(x) + g(y)$ implies, by differentiating with respect to u:

$$xg'(xy) = g'(y)$$

and therefore, setting $y = 1$,

$$xg'(x) = g'(1)$$

or

$$g'(x) = \frac{g'(1)}{x}.$$

Integrate this equation:

$$g(x) = g'(1) \ln x + c$$

for some constant c. But since $g(1) = 0$, we must have $c = 0$. Therefore

$$g(x) = C \ln x$$

with $C = g'(1)$.

Exercise 5.4. Let $\rho_1 = 0$ and

$$\rho_i = \frac{C}{i(\ln i)^2} \quad \text{for } i \geq 2,$$

where $C > 0$ is chosen to make the sum of the ρ_i equal to 1:

$$C = \left(\sum_{i=2}^{\infty} \frac{1}{i(\ln i)^2} \right)^{-1}.$$

Then

$$\sum_{i=1}^{\infty} \rho_i \ln \frac{1}{\rho_i} = \sum_{i=2}^{\infty} \rho_i \ln \frac{1}{\rho_i}$$

(remember that we consider $\rho_i \ln \frac{1}{\rho_i}$ to be 0 when $\rho_i = 0$). Now

$$\sum_{i=2}^{\infty} \rho_i \ln \frac{1}{\rho_i} = \sum_{i=2}^{\infty} \frac{C}{i(\ln i)^2} \ln \left(\frac{i(\ln i)^2}{C} \right) = \sum_{i=2}^{\infty} \frac{C}{i(\ln i)^2} \left(\ln i + 2 \ln \ln i - \ln C \right).$$

This is infinite because $\sum_{i=2}^{\infty} \frac{1}{i \ln i} = \infty$.

Exercise 6.1. The sequences $(x_1, x_2, x_3, \ldots)$ with $x_i \in \mathbb{R}$ and $x_i \neq 0$ only for at most finitely many i form a vector space, and for that vector space, , the sequences

$$(1,0,0,0,\ldots), \quad (0,1,0,0,\ldots), \quad (0,0,1,0,\ldots), \quad \text{etc.}$$

do form a basis.

Exercise 6.2. Define

$$E(t) = P(T > t), \quad t > 0.$$

This is a decreasing function that satisfies the law of exponentials:

$$\begin{aligned} & P(T > t+s \mid T > s) = P(T > t) \\ \Leftrightarrow \quad & \frac{P(T > t+s) \text{ and } P(T > s)}{P(T > s)} = P(T > t) \\ \Leftrightarrow \quad & \frac{P(T > t+s}{P(T > s)} = P(T > t) \\ \Leftrightarrow \quad & P(T > t+s) = P(T > t)P(T > s) \\ \Leftrightarrow \quad & E(t+s) = E(t)E(s). \end{aligned}$$

We are tempted to say "It must therefore be an exponential". A flaw in that reasoning: $E(t)$ is not defined for all t, only for $t > 0$. However, if we define $E(0) = 1$ and $E(-t) = \frac{1}{E(t)}$ for $t < 0$, it is easy to verify that $E(t)$ satisfies the law of exponentials for all t. Now we can in fact conclude that $E(t)$ must be an exponential, provided that its graph is not dense. But $E(t)$ is decreasing, therefore its graph cannot be dense.

Exercise 6.3. Your refrigerator is less likely to last another 5 years when it's already 10 years old. A 70-year-old human is less likely to live to 120 than a newborn is to see their 50th birthday. A child is less likely to die in their second year than in their first. (Thankfully, both are unlikely.) The tenured professor who has stayed put for 10 years already has probably settled down and will stay for many more years. The newly tenured professor, on the other hand, may still be itching to move. The probability that you must wait for the subway longer than 15 minutes is probably small. (Well okay, maybe not. I am thinking of a well-functioning public transportation system.) However, if you have been waiting for 30 minutes already, probably something really went wrong, and you'll wait for a long time before the next train comes.

Exercise 6.4. If $g(xy) = g(x) + g(y)$ for all x and y in $(0,1)$, then the function $h(s) = g(e^{-s})$, $s > 0$, is additive. We are tempted to say that therefore it is linear, but there is one flaw in the reasoning: $h(s)$ is only defined for $s > 0$. However, define $h(0) = 0$ and $h(-s) = -h(s)$ for $s > 0$, and it is easy to verify that the extended function is additive for all $s \in \mathbb{R}$. Therefore h is linear, unless the graph of h is dense. But the graph of h can't be dense, since g is decreasing (this makes

h increasing). Therefore h is in fact linear, $h(s) = Cs$ for some constant C, and therefore $g(x) = -C \ln x = C \ln \frac{1}{x}$ without any assumption on g.

Exercise 6.5. It does apply. If the mapping is measurable and obeys the sum rule, it obeys the constant factor rule.

Exercise 6.6. It suffices to assume is that for a fixed $v \in V$, the mapping from $\mathbb{R}$ into W defined by $c \mapsto L(cv)$ is continuous.

This isn't an overly strong assumption. If V and W are any real vector spaces with norms, and $L: V \to W$ is linear, then for any fixed $v \in V$, the mapping from $\mathbb{R}$ into W defined by $c \mapsto L(cv)$ is continuous, even though L need not be continuous.

Exercise 7.3. Integrating $f(t)$ from 0 to T, we obtain

$$a_0 T + 0 + 0.$$

(The integral of $\cos\left(2\pi\left(\frac{t}{T} + \theta_1\right)\right)$ over a full period is 0, and so is the integral of $\cos\left(4\pi\left(\frac{t}{T} + \theta_2\right)\right)$ over a full period.) This implies the assertion.

Exercise 7.5.

$$f(t) = \cos(\alpha\pi t)\cos(\alpha\pi\theta) - \sin(\alpha\pi t)\sin(\alpha\pi\theta).$$

So $c = \cos(\alpha\pi\theta)$ and $d = -\sin(\alpha\pi\theta)$.

Exercise 7.6. The period of

$$\cos\left(2\pi\left(\frac{\beta-\alpha}{2}\right)t\right)$$

is $\frac{2}{\beta-\alpha}$. The time between two subsequent zeros of this function is therefore $\frac{1}{\beta-\alpha}$. This means that the frequency of beats (the number of beats per unit time) is $\beta - \alpha$.

Exercise 7.8.

$$\begin{aligned} & \cos(2\pi(\alpha t + \theta)) + \cos(2\pi\beta t) \\ = \quad & \cos(\gamma + \delta) + \cos(\gamma - \delta) \end{aligned} \tag{A.3}$$

where γ and δ are chosen such that

$$\gamma + \delta = 2\pi(\alpha t + \theta),$$

$$\gamma - \delta = 2\pi\beta t.$$

This means

$$\gamma = \pi(\alpha t + \beta t + \theta),$$

$$\delta = \pi(\alpha t - \beta t + \theta).$$

Continuing with (A.3), we get, using the addition formula for the cosine,

$$\begin{aligned}
& 2\cos\gamma\cos\delta \\
= \; & 2\cos\left(\pi(\alpha t+\beta t+\theta)\right)\cos\left(\pi(\alpha t-\beta t+\theta)\right) \\
= \; & 2\cos\left(2\pi\left(\frac{\alpha+\beta}{2}t+\frac{\theta}{2}\right)\right)\cos\left(2\pi\left(\frac{\alpha-\beta}{2}t+\frac{\theta}{2}\right)\right) \\
= \; & 2\cos\left(2\pi\left(\frac{\alpha+\beta}{2}t+\frac{\theta}{2}\right)\right)\cos\left(2\pi\left(\frac{\beta-\alpha}{2}t-\frac{\theta}{2}\right)\right).
\end{aligned}$$

Exercise 7.9.

$$\cos(2\pi t)+\cos(4\pi t)=\cos(2\pi t)+2\cos^2(2\pi t)-1.$$

Write $x=\cos(2\pi t)$. Then

$$\cos(2\pi t)+\cos(4\pi t)=2x^2+x-1.$$

Write $f(x)=2x^2+x-1$, $-1\le x\le 1$. This function has one critical point in the interior of $[-1,1]$, at $x=-1/4$, and there the value is $-9/8$. At $x=-1$, the value is 0, and at $x=1$, the value is 2. So the minimum value of f on $[-1,1]$ is $-9/8$.

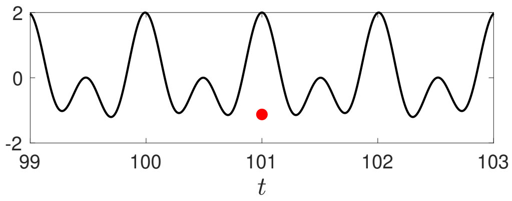
Exercise 7.10. We have

$$\cos(2\pi t)+\cos\left(2\pi\frac{200}{101}t\right)=\cos(2\pi t)+\cos\left(4\pi t-2\pi\frac{2}{101}t\right).$$

It's almost an octave, but the second cosine, the higher note, has a gradually shifting phase. Only at times when $\frac{2}{101}t$ is an integer is it in phase with $\cos(4\pi t)$. That means, it's in phase with $\cos(4\pi t)$ at integer multiples of $\frac{101}{2}$. The t-coordinate of a red point is the time at which the second cosine is in phase with $\cos(4\pi t)$ for the second time, so it must be $t=101$. Around times when the second cosine is in phase with $\cos(4\pi t)$, the minimum of the sum is about $-9/8$. So the red point should be

$$(101,-9/8).$$

That's exactly what it is, and here is a closeup:



Exercise 9.4. Think about two points (x_1,y_1) and (x_2,y_2). Rotation by 90 degrees does not change their distance. However, multiplying by $\frac{1}{10}$ reduces their distance

by a factor of $\frac{1}{10}$. Shifting does not alter the distance. So the distance between $\varphi(x_1, y_1)$ and $\varphi(x_2, y_2)$ must be one tenth of the distance between (x_1, y_1) and (x_2, y_2). But if (x_1, y_1) and (x_2, y_2) are fixed points, then $\varphi(x_1, y_1) = (x_1, y_1)$ and $\varphi(x_2, y_2) = (x_2, y_2)$. Therefore the distance between (x_1, y_1) and (x_2, y_2) must be zero, so the two fixed points are identical.

Exercise 9.5. If in any point x on the equator, the temperature is the same as at $-x$, we are done. Otherwise there is a point x_0 on the equator so that the temperature at $-x_0$ is different from the temperature at x_0. Rotate the equator around the axis through x_0 and $-x_0$. Any rotation angle between 0 and 90 degrees corresponds to a great circle, and these great circles only have the points x_0 and $-x_0$ in common. But each of the great circles must contain an antipodal pair of points in which the temperatures are the same, and that pair cannot be x_0 and $-x_0$. Therefore there are infinitely many such pairs.

Exercise 9.7. If all x_i are ≥ 0, and $\|x\| = 1$, then at least one x_i must be strictly positive. Therefore Ax must have strictly positive entries, so $Ax \neq 0$ and therefore $\|Ax\| > 0$. This explains by $f(x)$ is defined for all $x \in P$. Further, if $x \in P$, then all entries in Ax are ≥ 0, and therefore $f(x) \in P$. Now consider the mapping

$$L: \ B \to P$$

defined by

$$L(x_1, \ldots, x_{n-1}) = \left(x_1, \ldots, x_{n-1}, \sqrt{1 - \sum_{i=1}^{n-1} x_i^2} \right).$$

Then the composition

$$L^{-1} \circ f \circ L$$

is a continuous mapping $B \to B$. By Brouwer's theorem, it has a fixed point. So there exists an $s \in B$ such that

$$L^{-1} \circ f \circ L(s) = s,$$

or

$$f(L(s)) = L(s).$$

Writing $x = L(s)$, we have

$$f(x) = x.$$

This means

$$\frac{Ax}{\|Ax\|} = x,$$

so

$$Ax = \|Ax\| x,$$

so x is an eigenvector of A associated with the eigenvalue $\|Ax\|$.

Exercise 9.8. Since the $x^{(k)}$ form a bounded sequence in $\mathbb{R}^n$, there is a sequence

$$0 < k_1 < k_2 < k_3 < \dots$$

of integers such that

$$x = \lim_{j \to \infty} x^{(k_j)}$$

exists. The entries in x are still ≥ 0. Furthermore,

$$Ax = \lim_{j \to \infty} Ax^{(k_j)} = \lim_{j \to \infty} \|Ax^{(k_j)}\| x^{(k_j)} = \|Ax\| x.$$

So x is an eigenvector of A with eigenvalue $\|Ax\|$.

Exercise 9.9. Suppose that there were no point where both f and g are zero. Consider the curve

$$(f(\cos t, \sin t), g(\cos t, \sin t))$$

in the plane. At $t = 0$, it is at

$$(f_0, g_0) = (f(1,0), g(1,0)).$$

At $t = \pi$, it is at

$$(f_1, g_1) = (f(-1,0), g(-1,0)) = (-f_0, -g_0).$$

So it is at the point opposite to the starting point. As t rises from π to 2π, the curve returns to its starting point (f_0, g_0), along a path that is the reflection of the path from (f_0, g_0) to (f_1, g_1) across the origin:

$$(f(\cos t, \sin t), g(\cos t, \sin t)) = (-f(-\cos t, -\sin t), -g(-\cos t, -\sin t)) =$$

$$(-f(\cos(t-\pi), \sin(t-\pi)), -g(\cos(t-\pi), \sin(t-\pi))).$$

Overall, the curve winds around the origin a nonzero number of times. Now deform the original curve gradually:

$$(f(r \cos t, r \sin t), g(r \cos t, r \sin t))$$

where r gradually decreases from 1 to 0. None of the deformed curves pass through the origin. Therefore the winding number cannot change. On the other hand, for r near 0, the curve remains very close to

$$(f(0,0), g(0,0)) \neq (0,0),$$

and therefore its winding number must be zero. This contradiction proves the assertion.

Exercise 10.1. left upper: -1, right upper: -1, left lower: 1, right lower: 0.

Exercise 10.2. Going halfway around the circle, (u, v) rotates from its original value to the negative of that, so altogether it rotates by $2\pi k + \pi$ for some integer k.

But completing the circle, it rotates by the same amount again, because the vector field is odd. Therefore it rotates by $4\pi k + 2\pi$, a and the index is $2k+1$, an odd integer.

Exercise 10.3. The index is 1. If we reversed the directions of the vector field, the index would not change. If we traveled along the circle clockwise, the index would become -1.

Exercise 10.7. Continuous transformation of one curve into the other changes the index continuously, and since the index is always an integer, it doesn't change at all.

Exercise 10.5. The sink, source, and stable node have index 1. The saddle has index -1.

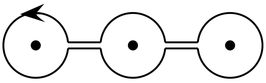
Exercise 10.6. The dipole has index 2.

Exercise 10.7. Define

$$u(x,y) = x^2 + y^2, \quad v(x,y) = 0.$$

The or

Exerc
inuously,
so they



is the sum of the three indices of the zeros.

Exercise 10.9. Suppose we were given two functions $f = f(x,y)$ and $g = g(x,y)$ defined for $(x,y) \in B^2$, with $f(-x,-y) = -f(x,y)$ and $g(-x,-y) = -g(x,y)$ for $(x,y) \in S^1$ and $(f(x,y), g(x,y)) \neq (0,0)$ for all $(x,y) \in B^2$. Then the index of S^1 (traversed once counter-clockwise) with respect to the vector field (f,g) is non-zero by Exercise 10.2. But now consider the circle of radius r, traversed counter-clockwise. Its index with respect to the vector field (f,g) must be independent of r, and yet it is non-zero for $r=1$, and zero for $r=0$. This contradiction implies that (f,g) must have a zero somewhere, and this implies the Borsuk-Ulam theorem; compare Exercise 9.9.

Exercise 10.10. A periodic solution corresponds to a closed loop in the phase plane. Now traverse this loop counter-clockwise — even if that means traveling backwards in time. The vector field $(f(x,y), g(x,y))$ is tangential to the curve everywhere, and therefore the index of the curve, traversed once counter-clockwise, is $+1$. This implies that the sum of the indices of zeros of (f,g) enclosed by the curve is $+1$, and in particular there are zeros of (f,g).

Exercise 10.11. The grid covering the square has

$$v = 16$$

vertices,

$$e = 24$$

edges, and

$$f = 9$$

faces. So

$$v - e + f = 1.$$

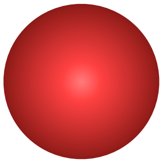
When you glue the right edge to the left, you lose 4 vertices and 3 edges, so now

$$v - e + f = 0.$$

When you then glue the upper edge ot the lower, you lose 3 vertices and 3 edges, so still

$$v - e + f = 0.$$

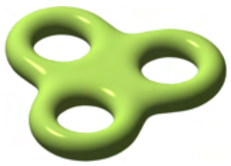
The surfaces



By Jahobr, CC0
via Wikimedia Commons

Oleg Alexandrov, Public Domain
via Wikimedia Commons

Oleg Alexandrov, Public Domain
via Wikimedia Commons



Oleg Alexandrov, Public Domain
via Wikimedia Commons

have genus 0, 1, 2, 3 and therefore Euler characteristic 2, 0, -2, -4. (We knew that the sphere has Euler characteristic 2, and now also that the torus has Euler characteristic 0.)

Exercise 10.12. The Poincaré-Hopf theorem implies that the sum of the indices of the zeros of a continuous tangential vector field on the sphere equals 2. In particular, there must be at least one zero. (If there is only one, it must have index 2, like a dipole.) The double and triple torus have non-zero Euler characteristics, and therefore any continuous tangential vector field on those surfaces must have at least one zero.