Mathematics

Dan Davison

September 27, 2018

Contents

Chapter 1

Foundations

1.0.1 Triangle inequalities

Theorem. *Let $a, b \in \mathbb{R}$ with $a \neq b$ and $a, b \neq 0$. Using $+$, $-$ and $|\cdot|$ we can generate the following 4 real numbers:*

$$-\big(|a|+|b|\big) \quad < \quad -\big||a|-|b|\big| \quad < \quad 0 \quad < \quad \big||a|-|b|\big| \quad < \quad |a|+|b|.$$

- *$a+b$ and $a-b$ can equal any of them.*
- *$|a+b|$ and $|a-b|$ can equal either of the two positive numbers.*
- *$|a|-|b|$ can equal either of the two "inner" numbers.*

If we allow $a = b$ with $a \neq 0, b \neq 0$ then

$$-\big(|a|+|b|\big) \quad < \quad -\big||a|-|b|\big| \quad \leq \quad 0 \quad \leq \quad \big||a|-|b|\big| \quad < \quad |a|+|b|.$$

If we allow $a = 0$ and $b = 0$ with $a \neq b$ then

$$-\big(|a|+|b|\big) \quad \leq \quad -\big||a|-|b|\big| \quad < \quad 0 \quad < \quad \big||a|-|b|\big| \quad \leq \quad |a|+|b|;$$

If we allow $a = b$ including $a = b = 0$ then

$$-\big(|a|+|b|\big) \quad \leq \quad -\big||a|-|b|\big| \quad \leq \quad 0 \quad \leq \quad \big||a|-|b|\big| \quad \leq \quad |a|+|b|;$$

1.0.2 The quadratic formula

Theorem. *The roots of $ax^2+bx+c=0$ are $x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$.*

Proof.

$$\begin{aligned} x^2 + \frac{b}{a}x + \frac{c}{a} &= 0 \\ \left(x+\frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} &= 0 \qquad \text{"completing the square"} \\ x &= -\frac{b}{2a} \pm \sqrt{\frac{b^2}{4a^2} - \frac{4ac}{4a^2}} \\ &= \frac{-b \pm \sqrt{b^2-4ac}}{2a}. \end{aligned}$$

□

1.0.3 Geometric series

Theorem. $a_n := \sum_{k=0}^{n} r^n = \frac{1-r^{n+1}}{1-r}$.

Therefore if $r < 1$ then $\lim_{n\to\infty} a_n = \frac{1}{1-r}$.

Proof.

$$\begin{aligned} a_n &= \sum_{k=0}^{n} r^n = 1 + r + r^2 + \ldots + r^n \\ a_n - r a_n &= 1 - r^{n+1} \\ a_n &= \frac{1 - r^{n+1}}{1 - r} \end{aligned}$$

□

Remark. Note that $a_{n+1} = 1 + r a_n$.

1.0.4 Partial fractions

TODO

1.0.5 Even and odd functions

Definition. *A function (over an additive group?) is even if and only if $f(-x) = f(x)$ for all x.*

A function (over an additive group?) is odd if and only if $f(-x) = -f(x)$ for all x.

Functions can be neither even nor odd.

Claim. *A polynomial $p(x)$ is even if an only if it has only even powers of x.*

A polynomial $p(x)$ is odd if an only if it has only odd powers of x.

1.0.6 $\sqrt{2}$ is irrational

Theorem. *$\sqrt{2}$ is irrational.*

Proof. Suppose $\sqrt{2} \in \mathbb{Q}$. Then $\sqrt{2}$ can be written as $\frac{a}{b}$ where $a, b \in \mathbb{Z}$ have no common factor (aka coprime, aka mutually prime).

Then $2 = \frac{a^2}{b^2}$, so a^2 is even.

Therefore a is even.

Let $a = 2c$. Then $b^2 = \frac{4c^2}{2} = 2c^2$, so b^2 is even.

Therefeore both a and b are even, which is a contradiction.

Therefore $\sqrt{2} \notin \mathbb{Q}$. □

Remark. It remains to be proved that $\sqrt{2}$ exists in $\mathbb{R}$. See 5.1.6.

1.0.7 Misc

0 Revision

You should check that you recall the following.

0.1 The Greek Alphabet

A	α	alpha	N	ν	nu
B	β	beta	Ξ	ξ	xi
Γ	γ	gamma	O	o	omicron
Δ	δ	delta	Π	π	pi
E	ϵ	epsilon	P	ρ	rho
Z	ζ	zeta	Σ	σ	sigma
H	η	eta	T	τ	tau
Θ	θ	theta	Υ	υ	upsilon
I	ι	iota	Φ	ϕ	phi
K	κ	kappa	X	χ	chi
Λ	λ	lambda	Ψ	ψ	psi
M	μ	mu	Ω	ω	omega

There are also typographic variations of epsilon (i.e. ε), phi (i.e. φ), and rho (i.e. ϱ).

0.2 Sums and Elementary Transcendental Functions

0.2.1 The sum of a geometric progression

$$\sum_{k=0}^{n-1} \omega^k = \frac{1-\omega^n}{1-\omega}\,. \tag{0.1}$$

0.2.2 The binomial theorem

The binomial theorem for the expansion of powers of sums states that for a non-negative integer n,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k\,, \tag{0.2a}$$

where the *binomial coefficients* are given by

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}\,. \tag{0.2b}$$

0.2.3 The exponential function

One way to define the exponential function, $\exp(x)$, is by the series

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}\,. \tag{0.3a}$$

From this definition one can deduce (after a little bit of work) that the exponential function has the following properties

$$\begin{aligned} \exp(0) &= 1\,, && (0.3b)\\ \exp(1) &= e \approx 2.71828183\,, && (0.3c)\\ \exp(x+y) &= \exp(x)\exp(y)\,, && (0.3d)\\ \exp(-x) &= \frac{1}{\exp(x)}\,. && (0.3e) \end{aligned}$$

Exercise. Show that if x is integer or rational then

$$e^x = \exp(x)\,. \tag{0.4a}$$

If x is irrational we define e^x to be $\exp(x)$, i.e.

$$e^x \equiv \exp(x)\,. \tag{0.4b}$$

0.2.4 The logarithm

For a real number $x > 0$, the logarithm of x, i.e. $\log x$ (or $\ln x$ if you really want), is defined as the unique solution y of the equation

$$\exp(y) = x\,. \tag{0.5a}$$

It has the following properties

$$\begin{aligned} \log(1) &= 0\,, && (0.5b)\\ \log(e) &= 1\,, && (0.5c)\\ \log(\exp(x)) &= x\,, && (0.5d)\\ \log(xy) &= \log(x) + \log(y)\,, && (0.5e)\\ \log(y) &= -\log\left(\frac{1}{y}\right)\,. && (0.5f) \end{aligned}$$

Exercise. Show that if x is integer or rational then

$$\log(y^x) = x\log(y)\,. \tag{0.6a}$$

If x is irrational we define $\log(y^x)$ to be $x\log(y)$, i.e.

$$y^x = \exp(x\log(y))\,. \tag{0.6b}$$

0.2.5 The cosine and sine functions

The cosine and sine functions are defined by the series

$$\cos(x) = \sum_{n=0}^{\infty} \frac{(-)^n x^{2n}}{2n!}\,, \tag{0.7a}$$

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-)^n x^{2n+1}}{(2n+1)!}\,. \tag{0.7b}$$

0.2.6 Certain trigonometric identities

You should recall the following

$$\begin{aligned} \sin(x \pm y) &= \sin(x)\cos(y) \pm \cos(x)\sin(y)\,, && (0.8a)\\ \cos(x \pm y) &= \cos(x)\cos(y) \mp \sin(x)\sin(y)\,, && (0.8b)\\ \tan(x \pm y) &= \frac{\tan(x) \pm \tan(y)}{1 \mp \tan(x)\tan(y)}\,, && (0.8c)\\ \cos(x) + \cos(y) &= 2\cos\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right)\,, && (0.8d)\\ \sin(x) + \sin(y) &= 2\sin\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right)\,, && (0.8e)\\ \cos(x) - \cos(y) &= -2\sin\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right)\,, && (0.8f)\\ \sin(x) - \sin(y) &= 2\cos\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right)\,. && (0.8g) \end{aligned}$$

0.2.7 The cosine rule

Let ABC be a triangle. Let the lengths of the sides opposite vertices A, B and C be a, b and c respectively. Further suppose that the angles subtended at A, B and C are α, β and γ respectively. Then the cosine rule (also known as the cosine formula or law of cosines) states that

$$\begin{aligned} a^2 &= b^2 + c^2 - 2bc\cos\alpha\,, && (0.9a)\\ b^2 &= a^2 + c^2 - 2ac\cos\beta\,, && (0.9b)\\ c^2 &= a^2 + b^2 - 2ab\cos\gamma\,. && (0.9c) \end{aligned}$$

Exercise: draw the figure (if it's not there).

0.3 Elemetary Geometry

0.3.1 The equation of a line

In 2D Cartesian co-ordinates, (x, y), the equation of a line with slope m which passes through (x_0, y_0) is given by

$$y - y_0 = m(x - x_0)\,. \tag{0.10a}$$

In parametric form the equation of this line is given by

$$x = x_0 + a\,t\,, \quad y = y_0 + a\,m\,t\,, \tag{0.10b}$$

where t is the parametric variable and a is an arbitrary real number.

0.3.2 The equation of a circle

In 2D Cartesian co-ordinates, (x, y), the equation of a circle of radius r and centre (p, q) is given by

$$(x-p)^2 + (y-q)^2 = r^2\,. \tag{0.11}$$

0.3.3 Plane polar co-ordinates (r, θ)

In plane polar co-ordinates the co-ordinates of a point are given in terms of a radial distance, r, from the origin and a polar angle, θ, where $0 \leqslant r < \infty$ and $0 \leqslant \theta < 2\pi$. In terms of 2D Cartesian co-ordinates, (x, y),

$$x = r\cos\theta\,, \quad y = r\sin\theta\,. \tag{0.12a}$$

From inverting (0.12a) it follows that

$$\begin{aligned} r &= \sqrt{x^2+y^2}\,, && (0.12b)\\ \theta &= \arctan\left(\frac{y}{x}\right)\,, && (0.12c) \end{aligned}$$

where the choice of arctan should be such that $0 < \theta < \pi$ if $y > 0$, $\pi < \theta < 2\pi$ if $y < 0$, $\theta = 0$ if $x > 0$ and $y = 0$, and $\theta = \pi$ if $x < 0$ and $y = 0$.

Exercise: draw the figure (if it's not there).

Remark: sometimes ρ and/or ϕ are used in place of r and/or θ respectively.

0.4 Complex Numbers

All of you should have the equivalent of a *Further Mathematics* AS-level, and hence should have encountered complex numbers before. The following is 'revision', just in case you have not!

0.4.1 Real numbers

The real numbers are denoted by $\mathbb{R}$ and consist of:

integers,	denoted by $\mathbb{Z}$,	$\ldots -3,\ -2,\ -1,\ 0,\ 1,\ 2,\ \ldots$
rationals,	denoted by $\mathbb{Q}$,	p/q where $p,\ q$ are integers $(q \neq 0)$
irrationals,		the rest of the reals, e.g. $\sqrt{2},\ e,\ \pi,\ \pi^2$.

We sometimes visualise real numbers as lying on a line (e.g. between any two distinct points on a line there is another point, and between any two distinct real numbers there is always another real number).

0.4.2 i and the general solution of a quadratic equation

Consider the quadratic equation

$$\alpha z^2 + \beta z + \gamma = 0 \quad : \quad \alpha,\, \beta,\, \gamma \in \mathbb{R} \quad , \alpha \neq 0\,,$$

where $\in$ means 'belongs to'. This has two roots

$$z_1 = -\frac{\beta + \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha} \quad \text{and} \quad z_2 = -\frac{\beta - \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}\,. \tag{0.13}$$

If $\beta^2 \geqslant 4\alpha\gamma$ then the roots are real (there is a repeated root if $\beta^2 = 4\alpha\gamma$). If $\beta^2 < 4\alpha\gamma$ then the square root is not equal to any real number. In order that we can *always* solve a quadratic equation, we introduce

$$i = \sqrt{-1}\,. \tag{0.14}$$

Remark: note that i is sometimes denoted by j by engineers (and MATLAB).

If $\beta^2 < 4\alpha\gamma$, (0.13) can now be rewritten

$$z_1 = -\frac{\beta}{2\alpha} + i\frac{\sqrt{4\alpha\gamma - \beta^2}}{2\alpha} \quad \text{and} \quad z_2 = -\frac{\beta}{2\alpha} - i\frac{\sqrt{4\alpha\gamma - \beta^2}}{2\alpha}\,, \tag{0.15}$$

where the square roots are now real [numbers]. Subject to us being happy with the introduction and existence of i, we can now *always* solve a quadratic equation.

0.4.3 Complex numbers (by algebra)

Complex numbers are denoted by $\mathbb{C}$. We define a complex number, say z, to be a number with the form

$$z = a + ib, \quad \text{where} \quad a,\, b \in \mathbb{R}, \tag{0.16}$$

where $i = \sqrt{-1}$ (see (0.14)). We say that $z \in \mathbb{C}$.

For $z = a + ib$, we sometimes write

$$\begin{aligned} a = \mathrm{Re}\,(z) &: \text{ the real part of } z,\\ b = \mathrm{Im}\,(z) &: \text{ the imaginary part of } z. \end{aligned}$$

Remarks.

(i) $\mathbb{C}$ contains all real numbers since if $a \in \mathbb{R}$ then $a + i.0 \in \mathbb{C}$.

(ii) A complex number $0 + i.b$ is said to be *pure imaginary*.

(iii) Extending the number system from real ($\mathbb{R}$) to complex ($\mathbb{C}$) allows a number of important generalisations, e.g. it is now possible to always to solve a quadratic equation (see §0.4.2), and it makes solving certain differential equations much easier.

(iv) Complex numbers were first used by Tartaglia (1500-1557) and Cardano (1501-1576). The terms *real* and *imaginary* were first introduced by Descartes (1596-1650).

Theorem 0.1. *The representation of a complex number z in terms of its real and imaginary parts is unique.*

Proof. Assume $\exists\ a, b, c, d \in \mathbb{R}$ such that

$$z = a + ib = c + id.$$

Then $a - c = i\,(d - b)$, and so $(a - c)^2 = -\,(d - b)^2$. But the only number greater than or equal to zero that is equal to a number that is less than or equal to zero, is zero. Hence $a = c$ and $b = d$. □

Corollary 0.2. *If $z_1 = z_2$ where $z_1, z_2 \in \mathbb{C}$, then $\mathrm{Re}\,(z_1) = \mathrm{Re}\,(z_2)$ and $\mathrm{Im}\,(z_1) = \mathrm{Im}\,(z_2)$.*

0.4.4 Algebraic manipulation of complex numbers

In order to manipulate complex numbers simply follow the rules for reals, but adding the rule $i^2 = -1$. Hence for $z_1 = a + ib$ and $z_2 = c + id$, where $a, b, c, d \in \mathbb{R}$, we have that

$$\text{addition/subtraction :} \qquad z_1 + z_2 = (a + ib) \pm (c + id) = (a \pm c) + i\,(b \pm d)\ ; \tag{0.17a}$$

$$\begin{aligned}\text{multiplication :} \qquad z_1\, z_2 &= (a + ib)\,(c + id) = ac + ibc + ida + (ib)\,(id) \\ &= (ac - bd) + i\,(bc + ad)\ ;\end{aligned} \tag{0.17b}$$

$$\text{inverse :} \qquad z_1^{-1} = \frac{1}{z} = \frac{1}{a + ib}\,\frac{a - ib}{a - ib} = \frac{a}{a^2 + b^2} - \frac{ib}{a^2 + b^2}. \tag{0.17c}$$

Remark. All the above operations on elements of $\mathbb{C}$ result in new elements of $\mathbb{C}$. This is described as *closure*: $\mathbb{C}$ is closed under addition and multiplication.

Exercises.

(i) For z_1^{-1} as defined in (0.17c), check that $z_1\, z_1^{-1} = 1 + i.0$.

(ii) Show that addition is *commutative* and *associative*, i.e.

$$z_1 + z_2 = z_2 + z_1 \quad \text{and} \quad z_1 + (z_2 + z_3) = (z_1 + z_2) + z_3\,. \tag{0.18a}$$

(iii) Show that multiplication is *commutative* and *associative*, i.e.

$$z_1 z_2 = z_2 z_1 \quad \text{and} \quad z_1(z_2 z_3) = (z_1 z_2) z_3\,. \tag{0.18b}$$

(iv) Show that multiplication is *distributive* over addition, i.e.

$$z_1(z_2 + z_3) = z_1 z_2 + z_1 z_3\,. \tag{0.18c}$$

Chapter 2

Discrete Mathematics

2.1 Combinatorics

Ways to Arrange, Select, or Distribute *r* Objects from *n* Items or into *n* Boxes

	Arrangement (Ordered Outcome) or Distribution of Distinct Objects	*Combination (Unordered Outcome) or Distribution of Identical Objects*
No repetition	$P(n, r)$	$C(n, r)$
Unlimited repetition	n^r	$C(n+r-1, r)$
Restricted repetition	$P(n; r_1, r_2, \ldots r_m)$	—

Theorem (Subtuples)**.**
The number of k-tuples that can be formed from a set of size n without replacement is

$$(n)_k := n \cdot (n-1) \cdots (n-k+1) = \frac{n!}{(n-k)!}.$$

Remark. As a special case, the number of n-tuples (i.e. permutations/arrangements) is $n!$. (This is also the number of $n-1$ tuples.)

Theorem (Subsets)**.**
The number of subsets of size k that can be formed from a set of size n is

$$C(n,k) = \binom{n}{k} := \frac{(n)_k}{k!} = \frac{n!}{(n-k)!\ k!}.$$

Proof. Each distinct k-subset gives rise to $k!$ k-tuples by assigning position labels. Therefore $(n)_k = \binom{n}{k} k!$. □

Theorem (Multiset arrangements)**.**
Consider a multiset comprising n distinct elements, with $r_i \geq 1$ repeats of the i-th element. The number of n-tuples that can be formed from such a multiset is

$$\begin{aligned} P(n; r_1, \ldots, r_k) :=& \binom{n}{r_1}\binom{n-r_1}{r_2}\cdots\binom{n-r_1\cdots-r_{n-1}}{r_n} \\ =& \frac{n!}{r_1! r_2! \cdots r_n!}. \end{aligned}$$

Proof. The r_1 copies of the first element must all go somewhere. $\binom{n}{r_1}$ counts the number of distinct positions they can occupy. Then there are $n - r_1$ empty positions left. Etc. □

Remark. The number $n!$ of permutations of a set is a special case of this with $r_i = 1$ for all i.

Example.

1. **How many ways are there to assign 100 different diplomats to five different continents?**
   5^{100}

2. **How many ways if 20 diplomats must be assigned to each continent?**
   $P(100; 20, 20, 20, 20, 20)$. Arrange the 100 diplomats in an arbitrary order. Now we have a multiset of country labels with 20 repeats of each label. Given the fixed ordering of the diplomats, there's a one-to-one correspondence between distinct permutations of the multiset and assignments of diplomats to countries.

3. **How many ways are there to distribute 20 (identical) sticks of red licorice and 15 (identical) sticks of black licorice among five children?**
   $\binom{20+5-1}{5-1}\binom{15+5-1}{5-1}$.

Theorem. *How many k-tuples for $k \leq n$ can be formed from such a multiset?*

TODO

Theorem (Stars and bars)**.**
Consider the number of ways that n identical objects can be put into k buckets, recording only the counts in each bucket (not the identities of the objects).

With no empty buckets, the answer is

$$\binom{n-1}{k-1} \qquad (k-1 \text{ bars to be placed in } n-1 \text{ gaps between } n \text{ stars}).$$

With empty buckets allowed, the answer is

$$\binom{n+k-1}{k-1} = P(n+k-1; n, k-1) \qquad (\text{number of arrangements of } n \text{ stars and } k-1 \text{ bars}).$$

Proof. Represent this as n unlabeled stars, and $k-1$ bars representing the partition of the stars into different buckets.

With no empty buckets allowed, there are $n-1$ gaps where the bars can be placed, hence $\binom{n-1}{k-1}$ ways of dividing up the items.

With empty buckets allowed, there could be multiple bars in the same position. The number of $(n+k-1)$-tuples that can be formed from the star and bar symbols is

$$\begin{aligned} P(n+k-1; n, k-1) &= C(n+k-1, k-1)C(n, n) \\ &= C(n+k-1, k-1) \\ &= C(n+k-1, n)C(k-1, k-1) \\ &= C(n+k-1, n). \end{aligned}$$

Note that $\binom{n-1}{k-1}$ for the no-empty-buckets version can also be derived as follows:

1. Place one item into each bucket.
2. Now there are $n-k$ items into k buckets and empty buckets are allowed for the subsequent allocations. So the answer is $\binom{(n-k)+k-1}{k-1} = \binom{n-1}{k-1}$ by the empty-buckets-allowed theorem.

□

Theorem (Stars and bars)**.**
The number of ways that n items can be put into k buckets, with empty buckets allowed, recording only the counts in each bucket (not the identities of the items), is

Theorem (Partitions)**.**
The number of ways that n items can be put into k buckets, with no empty buckets, recording the identities of the items in each bucket, is the number of partitions of size k of a set of size n. It is equal to the Stirling number of the second kind:

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^n. \quad \textit{(check this)}$$

Proof. TODO □

Theorem (Identities)**.**

$$\binom{m+n}{r} = \sum_{i=0}^{r}\binom{m}{i}\binom{n}{r-i}$$

2.1.1 Tucker - Applied Combinatorics - Exercises

(5.1) **General Counting Method for Arrangements and Selections**

(37) **If three distinct dice are rolled, what is the probability that the highest value is twice the smallest value?**

$\frac{(3\times2\times3)+(3\times3!)}{6^3}$

An outcome is a 3-tuple such as $(1,1,1)$. Outcomes that match the criterion belong to two disjoint subsets:

i. Outcomes with two distinct values, such as $(1,1,2)$. There are $3\times2\times3$ such outcomes (3 choices of unordered pairs of numbers, each with two alternative labelings and 3 distinct permutations).

ii. Outcomes with three distinct values, such as $(2,3,4)$. There are $3\times3!$ such outcomes ($1+2$ unordered triples of numbers, each with 3! distinct permutations)

(5.2) **Simple arrangements and selections**

(Example 2) **How many ways are there to arrange the 7 letters of the word SYSTEMS...**

i. **...?**

$$7_{(7-3)} = 7 \cdot 6 \cdot 5 \cdot 4 \qquad \text{(Choose positions of the other 4 letters, then Ss determined.)}$$

ii. **...with the 3 Ss consecutive?**

$$5_{(5)} = 5! \qquad \text{(Consider as 5-letter word S}^3\text{YTEM.)}$$

iii. **...with E before M?**

$$\binom{7}{2} 5_{(5-3)} = \binom{7}{2} 5 \cdot 4 \qquad \text{(Choose position of E,M, then choose position of non-Ss.)}$$

iv. **...with E before M and 3 Ss consecutive?**

$$\binom{5}{2} 3! \qquad \text{(Consider as 5-letter word S}^3\text{YTEM, choose position of E,M, then choose positions for remaining}$$

(Example 6) How

2.1.2 Generating functions

Definition (Generating function)**.** *Let a_r be the number of ways to select r objects in some counting procedure. Then $g(x)$ is a generating function for a_r if $g(x)$ has the polynomial expansion*

$$a_0 + a_1 x + \ldots + a_n x^n.$$

Example. Find the generating function for a_r, the number of ways to select r balls from 3 green, 3 white, 3 blue, and 3 gold balls.

2.2 Relations and partitions

A relation on a set A is a subset of A^2. Thus for a pair $(a_1, a_2) \in A^2$ the relation says whether a_1 is related to a_2.

An equivalence relation is a relation that is reflexive, symmetric, and transitive, and thus makes sense as defining a partitioning of the set into groups of equivalent elements.

The equivalence relation doesn't tell you explicitly which group a pair belongs to (it just tells you that they are in the same group). But the information is there: the groups are the connected components in the graph in which two vertices are connected if they are related. There are fewer equivalence relations than assignments to labeled buckets, since the equivalence relation does not identify the buckets. How many equivalence relations are there, compared to Stirling II number and stars-and-bars count configurations?

2.3 Pythagorean triples

Project Euler question 9

> A Pythagorean triplet is a set of three natural numbers, $a < b < c$, for which $a^2 + b^2 = c^2$. For example, $3^2 + 4^2 = 9 + 16 = 25 = 5^2$.
>
> There exists exactly one Pythagorean triplet for which $a + b + c = 1000$. Find the product abc.

Proof.
Let $m, n \in \mathbb{N}$.

Recall that $|m + ni| := \sqrt{m^2 + n^2}$ and that $|wz| = |w||z|$ for $w, z \in \mathbb{C}$.

Note that $|(m + ni)^2| = |(m^2 - n^2) + 2mni| = m^2 + n^2 \in \mathbb{Z}$.

Therefore $(m^2 - n^2, 2mn, m^2 + n^2)$ is a pythagorean triple for all $m, n \in \mathbb{N}$. (Claim: all pythagorean triples are of this form.)

Therefore we seek $m, n \in \mathbb{Z}$ such that $m > n$ and

$$\begin{aligned} m^2 - n^2 + 2mn + m^2 + n^2 &= 1000 \\ m^2 + mn &= 500 \\ \left(m + \frac{n}{2}\right)^2 - \frac{n}{4} - 500 &= 0 \\ m &= \sqrt{\frac{n}{4} + 500} - \frac{n}{2} \end{aligned}$$

Therefore (?) $\sqrt{\frac{n}{4} + 500} \in \mathbb{Z}$. So $\frac{n}{4} + 500 = a^2$ for some $a \in \mathbb{Z}$.

□

Chapter 3

Abstract algebra

3.1 Definitions

Definition (Group)**.** *A group is a set, together with a binary operation that maps any two elements to another element in the set. I.e. it is a triple $(S, \cdot, I)$ specifying the set, the operation and the identity element respectively. It satisfies the group axioms:*

1. *existence of identity*
2. *existence of an inverse for each element*
3. *associativity*

If the operation is commutative it is said to be "abelian".

Definition (Homomorphism)**.** *A structure-preserving map between two groups*

Definition (Endomorphism)**.** *A homomorphism from a group to itself.*

Definition (Field)**.** *A field is a set $\mathbb{F}$ for which both $(\mathbb{F}, +, 0)$ and $(\mathbb{F}, \times, 1)$ are abelian groups.*

Definition (Vector space)**.** *A vector space V over a field $\mathbb{F}$ is an abelian group $(V, +, 0)$ for which multiplication by "scalars" from $\mathbb{F}$ is defined, and additionally satisfies*

1. *Linear combinations using scalars from $\mathbb{F}$ remain within the vector space:*
   $au + bv \in V$ for all $a, b \in \mathbb{F}$ and $u, v \in V$.

Definition (Ring)**.** [1] *A ring is an abelian group $(R, +, 0)$ which additionally has a multiplication operation. The multiplication may or may not be commutative, and does not necessarily have inverses. Both distributive laws must hold unless we're assuming commutativity: $a(b + c))$ and $(b + c)a$.*

Examples

- zero ring $\{0\}$ (multiplicative identity is $1 = 0$ in this ring; in any other ring $1 \neq 0$.)
- Any field is a ring
- $\mathbb{Z}$
- $\mathbb{Z}/n\mathbb{Z}$
- Set of $n \times n$ matrices over a field

Subrings must contain $0, 1$.

Subring of $\mathbb{C}$: Gaussian integers $\mathbb{Z} + i\mathbb{Z}$.

What about sets of lines through the origin in complex plane such that angles are closed under addition? That's a subring of $\mathbb{C}$ too right?

"Best way to get rings": **endomorphism ring**. Start with an abelian group $(A, +, 0)$. The endomorphism ring is the set of all group homomorphisms $A \to A$ [2]

$$\text{End } (A) = \{f : A \to A\}$$

[1] Gross, Abstract Algebra, lecture 24
[2] aren't these called automorphisms?

Addition and multiplication are defined by

$$(f+g)(a) = f(a) + g(a)$$
$$(fg)(a) = f(g(a)).$$

It must have an additive identity. This must be the constant zero function $0(a) = 0$. Is that a homomorphism? Yes: $0(a+b) = 0 = 0(a) + 0(b)$.

And additive inverse: $(-f)(a) = -(f(a))$. Is that a homomorphism? Yes:

$$(-f)(a+b) = -(f(a+b)) = -(f(a) + f(b)) = -f(a) + -f(b)$$

The multiplicative identity is just the identity homomorphism.

Multiplication (i.e. composition) is not necessarily commutative.

It would be a field if there were multiplicative inverses: do these exist? Only for those homomorphisms that are isomorphisms.

To construct the ring $\mathbb{Z}$: it's (isomorphic to) the endomorphism ring of the group $(\mathbb{Z}, +, 0)$. What's the correspondence between group homomorphisms and integers? Well, consider group homomorphism f. The entire homomorphism is determined by $f(1)$! Since

$$\begin{aligned} f(1) &= k \\ f(2) &= f(1+1) = f(1) + f(1) = 2k \\ &\dots \\ f(n) &= kn. \end{aligned}$$

So we map the homomorphism f to the integer $f(1)$.

And if $f(1) = k_1$ and $g(1) = k_2$, then multiplication of integers is

$$(f \times g)(n) = f(g(n)) = k_1 k_2 n.$$

This is the reason why the product of two negative integers is positive: a negative number corresponds to a homomorphism that maps positive integers to negative.

Similarly,

$$\mathbb{Z}/n\mathbb{Z} = \text{ End } (\mathbb{Z}/n\mathbb{Z}, +, 0)$$

since we identify 1 with...

This is a phenomenon of cyclic groups. I.e. $\mathbb{Z}/n\mathbb{Z}$ (finite) and $\mathbb{Z}$ (infinite). There are no other cyclic groups.

3.1.1 Polynomial

A polynomial is $P(x) = a_0 + a_1 x^1 + \ldots + a_n x^n$.

The coefficients a_i must come from some ring R, and the set of all such polynomials is written $R[x]$.

If R is a commutative ring, so is $R[x]$.

Therefore we can write a polynomial in two variables as $R[x][y]$, i.e. the coefficients of the polynomial in y are themselves polynomials in x.

If $R = \mathbb{C}$ this leads towards algebraic geometry. Rings like integers, Gaussian integers, etc are the subject of number theory.

The variable/"indeterminate" x must also come from some ring, since it is involved in both addition and multiplication (?).

Multiplication of polynomials e.g. coefficients from $R = \mathbb{Z}/2\mathbb{Z}$:

$$(x+1)(x+1) = x^2 + 2x + 1 = x^2 + 1$$

since $2 = 0 \mod 2$.

3.2 Vector Spaces

3.2.1 Definitions

Linear independence

A set of vectors $\{v_1, v_2, \ldots\}$ are linearly independent if the only solution to

$$a_1 v_1 + a_2 v_2 + \ldots = 0$$

is $a_1 = a_2 = \ldots = 0$.

Span

The span of a set of vectors is the set of all vectors that can be formed from them by linear combination.

Basis

$E \subset V$ is a basis for V if

1. E spans V
2. if the addition of any further $v \in V \setminus E$ to E would cause E to lose its linear dependence.

Coordinates

The coordinate of a vector v in basis $E = \{e_1, e_2, \ldots, e_n\}$ is the unique list of scalars $a_1, a_2, \ldots, a_n$ such that $v = a_1 e_1 + a_2 e_2 + \ldots + a_n e_n$.

So although a vector v may live in an n-dimensional space, v does not consist of a list of n components until it is represented by its coordinates in a particular basis.

Linear transformation

A linear transformation $f : V \to W$ is a homomorphism between the vector spaces, preserving linear transformation: for $u, v \in V$

$$f(au + bv) = af(u) + bf(v).$$

Having fixed a basis E for V, to specify f it's sufficient to specify $f(e)$ for all $e \in E$. That's what a matrix is: it contains $f(e_j)$ in column j.

3.2.2 The space of functions $f : \mathbb{N} \to \mathbb{R}$

The elements of the vector space are functions (i.e. subsets of $\mathbb{N} \times \mathbb{R}$).

(They are sequences.)

This is a group under addition of functions: $f + g : \mathbb{N} \to \mathbb{R}$ where $(f + g)(n) = f(n) + g(n)$. The identity is the zero function $f(n) = 0$.

So we have addition of functions, but multiplication of functions plays no role. Is the set of all functions a field?

The vector space is over a field, say $\mathbb{R}$. For $a \in \mathbb{R}$, $(af)(n) = a(f(n))$.

Once we have established a basis E, then each function can be assigned numerical coordinates in the (infinite-dimensional) vector space.

So what would be a basis for this vector space?

3.3 Groups

Consider a set A with n elements.

Definition. *A **permutation** is a bijection: $A \to A$. There are $n!$ permutations of A.*

Definition. *The **symmetry group** S_n is the set of permutations of a set of size n, denoted $S_n = Sym(\{1, 2, \ldots, n\})$. It is a group, under composition.*

Definition. *Consider an n-sided regular polygon. A **symmetry** of the polygon is an isometry of the plane which preserves the polygon.*

Definition. *The **dihedral group** D_{2n} is the group of symmetries under composition. It is of order $2n$. Let r be a clockwise rotation through $360/n$ degrees, and let s be a reflection about a chosen axis of symmetry. The elements of the group are*

$$e, r, r^2, \ldots, r^{n-1}, s, sr, sr^2, \ldots, sr^{n-1}.$$

Remark. The symmetry group contains permutations; some of these are not symmetries.

The set of symmetries is a proper subset of the set of permutations of the vertices: for example the permutation which interchanges two adjacent vertices but leaves all other vertices unchanged is not a symmetry.

Intuition. A k-**cycle** is a permutation that cycles k elements and leaves the rest unchanged. Its order is k.

Definition. *A permutation $\sigma \in S_n$ is a k-**cycle** if there exist k distinct elements $a_1, \ldots, a_k \in S_n$ such that*

$$\begin{cases} a_i\sigma = a_{i+1} \text{ for } 1 \leq i < k \\ a_k\sigma = a_1 \\ x\sigma = x \text{ for } x \notin \{a_1, \ldots, a_k\}. \end{cases}$$

The cycle σ is written as e.g. $(a_1, a_2, \ldots, a_k)$, but note that e.g. $(a_2, a_3, \ldots, a_k, a_1)$ is the same thing.

Example. Let

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 3 & 1 & 5 \end{pmatrix} \quad \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \end{pmatrix} \quad \gamma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 3 & 1. \end{pmatrix}$$

Determine the product $\alpha\beta\gamma$, the inverse of β and the order of γ.

In cycle notation,

$$\alpha = (124) \quad \beta = (13524) \quad \gamma = (125)(34).$$

By following each element in turn through the 3 permutations (using either representation),

$$\alpha\beta\gamma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 4 & 5 \end{pmatrix} = (13).$$

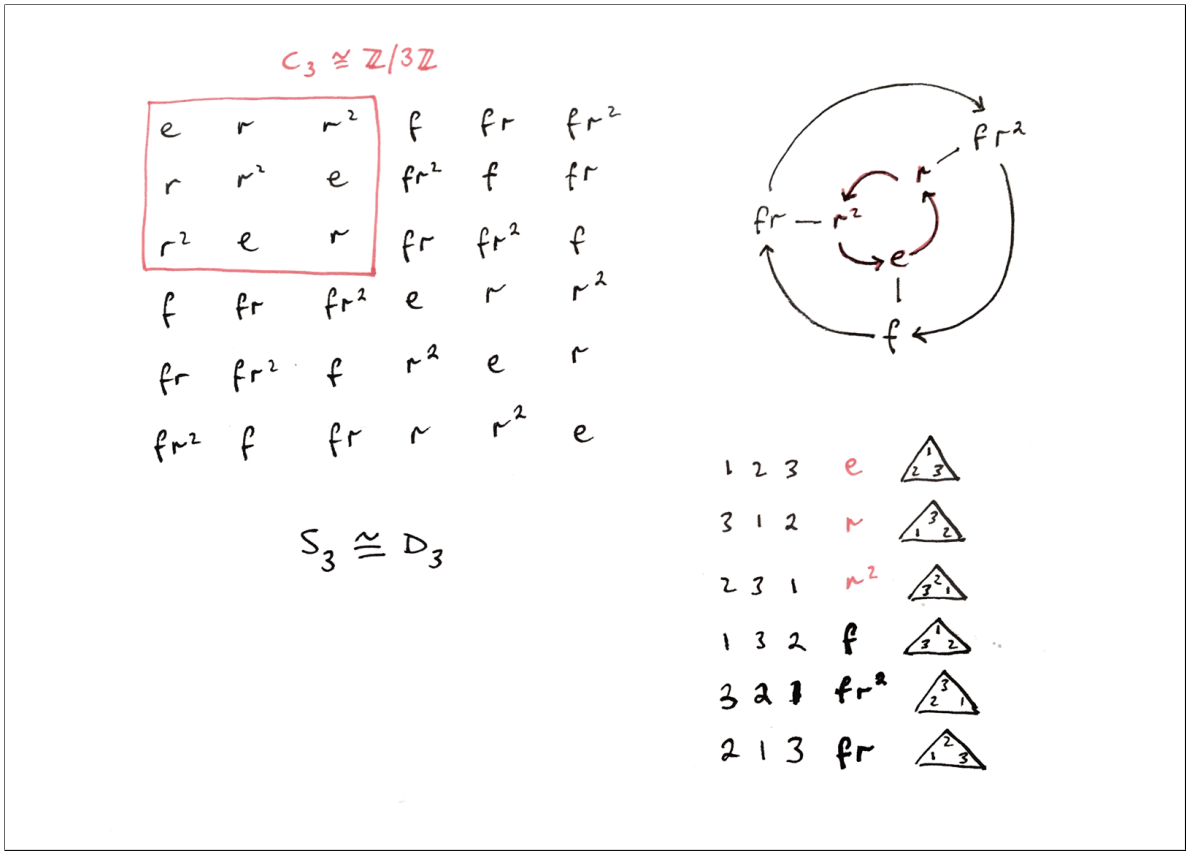
$$\beta^{-1} = \begin{pmatrix} 3 & 4 & 5 & 1 & 2 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 1 & 2 & 3 \end{pmatrix} = (14253).$$

The order of γ is 6.

3.4 Examples of groups, homomorphisms and quotients

3.4.1 Finite order

- S_2: the set of permutations of two objects, where the operation is composition of functions. There are just two elements in the group: the do-nothing permutation and the switch-the-elements permutation: $\{e, \tau\}$.
- $S_3 \cong D_3$: the symmetry group S_3 is the set of permutations of three objects. There are 6 elements: the identity, 3 transitions[3] and two cyclic permutations. It's isomorphic to the dihedral group D_3, i.e. the group of symmetries of an equilateral triangle: the two cyclic permutations correspond to rotation by $\pi/3$ and $2\pi/3$ radians, and the three transitions correspond to the 3 possible reflections (each of which leaves one vertex unchanged).



- $\{1, i, -1, -i\}$ where the operation is multiplication of complex numbers.

3.4.2 Infinite order

- The group $\mathbb{Z}$, i.e. integers under addition
  - ◇ A homomorphism is $f : \mathbb{Z} \to \{0, 1\}$, given by $f(n) = \begin{cases} 0, & n \text{ is even} \\ 1, & n \text{ is odd} \end{cases}$, with the operation on

[3] A transition is a permutation that switches two elements and leaves all other alone

$\{0, 1\}$ being addition mod 2. The kernel is $evens = 2\mathbb{Z}$, and the resulting quotient group is $\mathbb{Z}/2\mathbb{Z} = \{evens, odds\} = \{2\mathbb{Z}, 2\mathbb{Z} + 1\}$.

- The ring $\mathbb{Z}$ - example of homomorphism.
- $\mathbb{Z}^+_{>0}$ positive integers under addition
- Similarly, $\mathbb{Q}^+$, $\mathbb{Q}^\times$, $\mathbb{C}^+$, $\mathbb{C}^x$, $\mathbb{R}^+$, $\mathbb{R}^x$ etc
- $GL_n(\mathbb{F})$: the set of $n \times n$ matrices with entries from the field $\mathbb{F}$, under matrix multiplication.
  - ◇ A homomorphism is $f : GL_n(\mathbb{F}) \to \mathbb{F}^\times$ given by $A \mapsto \det(A)$. This is a homomorphism since $\det(AB) = \det(A)\det(B)$. The kernel is the set $SL_n(\mathbb{F})$ of matrices with determinant 1. If the field is $\mathbb{R}$ then these rotate or flip space without stretching it. The resulting quotient group is the set $\{\mathcal{A}(x) \mid x \in \mathbb{F}\}$, where $\mathcal{A}(x)$ is the set of matrices with determinant x.
- The set $\mathbb{F}[x]$ of polynomials with coefficients in a field $\mathbb{F}$ can be a vector space, and a ring. (Not a field, since multiplicative inverses don't exist for degree ≥ 1.)

  As a vector space, differentiation is a linear transformation (homomorphism). This is non-injective (polynomials differing only by an additive constant additive are sent to the same polynomial). The kernel is the set of degree 0 polynomials ($\mathbb{F}$). The quotient space $\mathbb{F}[x]/\mathbb{F}$ contains cosets of the form $p(x) + \mathbb{F}$, i.e. a set of polynomials differing only by an additive constant.

  But differentiation does not preserve multiplication of polynomials, so it is not a ring homomorphism.
- $SL_n(\mathbb{R})$: set of $n \times n$ matrices with determinant 1 (kernel of the determinant homomorphism $GL_n(\mathbb{R}) \to \mathbb{R}^\times$ and therefore a normal subgroup of $GL_n(\mathbb{R})$)

3.5 Homomorphism

A **homomorphism** is a map from one group to another. If it is bijective, it is an **isomorphism**. If it is bijective and from a group to itself (i.e. a permutation of the group elements) then it is an **automorphism**. The critical feature of these concepts is that they "preserve group structure", i.e. they preserve the relationships among group elements defined by the group operation. Suppose that they map from group G to group G'. Then the preservation-of-structure criterion is that the map sends a product $g_1 \circ g_2$ to the product of whatever the separate elements are sent to:

$$f(g_1 \circ g_2) = f(g_1) \circ f(g_2)$$

There the composition on the left is happening in G and the composition on the right is happening in G'. (For an automorphism, $G = G'$.)

Another way of saying this is that "the following diagram commutes":

$$\begin{pmatrix} g_1, g_2 & \xrightarrow{f} & f(g_1), f(g_2) \\ \downarrow & & \downarrow \\ g_1 g_2 & \xrightarrow{f} & f(g_1 g_2) = f(g_1) f(g_2) \end{pmatrix},$$

i.e. it does not matter whether you first perform the internal structure operation on the left-hand side and then apply f, or alternatively apply f first and perform the internal structure operation on the right-hand side.

Note that an element such as g_1 that is being sent somewhere by a morphism may itself already be a map of sorts, e.g. if it is a permutation in S_3. This is potentially confusing, since an automorphism can be thought of as a permutation of group elements. So an automorphism on S_3 is a permutation of group elements that are themselves permutations of some generic labeled objects.

The definition of homomorphism implies that $f(g^{-1}) = f(g)^{-1}$ since $f(gg^{-1}) = f(g)f(g^{-1}) = f(e)$.

3.6 Kernel, Nullspace, Bijection and Congruency

Consider a homomorphism f with kernel N.

Theorem: a and b are sent to the same place by f if and only if $b = an$ for some $n \in N$.

Corollary: f is a bijection (isomorphism) if and only if the kernel contains only the identity element.

Example: Consider the absolute value homomorphism mapping complex numbers under multiplication to positive reals under multiplication. The equivalence classes are concentric circles around the origin. Two complex numbers have the same absolute value iff one can be obtained from the other by rotation only (no scaling). This is multiplication by a complex number with absolute value 1, and such a complex number is in the kernel.

Proof: Clearly, if $b = an$ then b is sent to the same place as a, since

$$f(b) = f(an) = f(a)f(n) = f(a).$$

However we need to demonstrate the converse, i.e. that the *only* way that b can be sent to the same place as a is if $b = an$ for some $n \in N$.

Two almost identical ways of showing that:

(1) Show that if $f(a) = f(b)$ then $b = an$ for some $n \in N$

In linear algebra, you can always get from u to v by adding $v - u = -u + v$, so the claim is that $L(u) = L(v)$ implies $-u + v$ is in the nullspace, which is true:

$$L(-u + v) = L(-u) + L(v) = L(-u) + L(u) = 0.$$

For a group homomorphism, b can be written as $aa^{-1}b$, so the claim is that $f(a) = f(b)$ implies $a^{-1}b \in N$, which is true:

$$f(a^{-1}b) = f(a^{-1})f(b) = f(a)^{-1}f(a) = e.$$

(2) Show that if it is not the case that $b = an$ for some $n \in N$, then $f(a) \neq f(b)$

In linear algebra, you can always get from u to v by adding $v - u = -u + v$, so if $-u + v$ is not in the nullspace then

$$L(v) = L(u + (-u + v)) = L(u) + L(-u + v) \neq L(u).$$

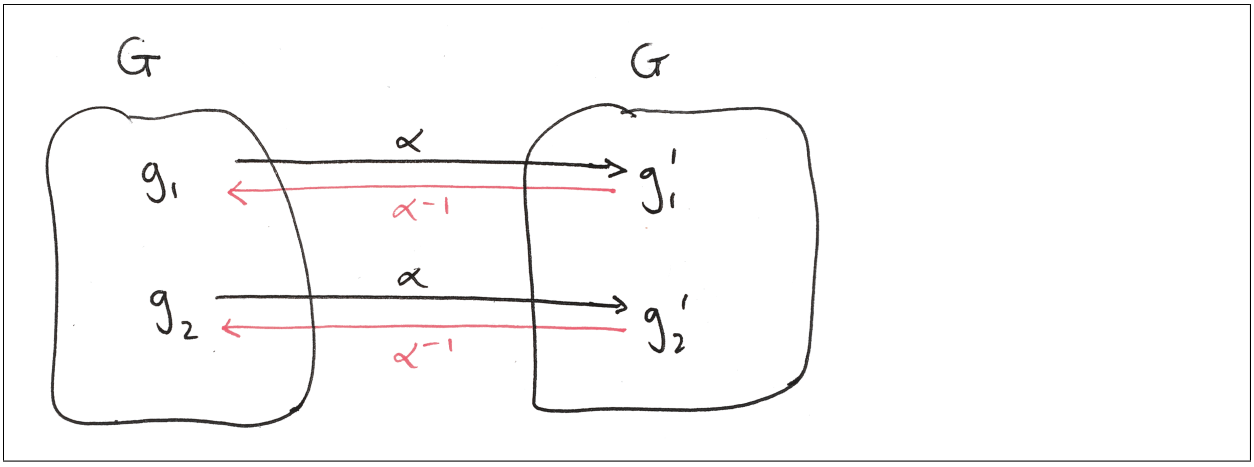
For a group homomorphism, b can be written as $aa^{-1}b$, so if $a^{-1}b$ is not in the kernel then

$$f(b) = f(aa^{-1}b) = f(a)f(a^{-1}b) \neq f(a)$$

3.7 Inverse of an automorphism is an automorphism

[Artin 2.3.11: show that Aut(G) is a group]

Suppose α is an automorphism that sends g_1 and g_2 to g_1' and g_2', respectively.



We need to show that α^{-1} preserves structure, i.e. that when α^{-1} acts on an element which is a product, say $g_1'g_2'$, it sends it to the product of whatever it send the individual factors to:

$$\alpha^{-1}(g_1'g_2') = \alpha^{-1}(g_1')\alpha^{-1}(g_2').$$

Firstly, we know that $\alpha^{-1}(g_1')$ and $\alpha^{-1}(g_2')$ exist, i.e. some elements are taken to them by a, because a is an automorphism and therefore surjective. So we'll call those g_1 and g_2, and the equality we need to demonstrate has become

$$\alpha^{-1}(\alpha(g_1)\alpha(g_2)) = g_1 g_2.$$

Since α is an automorphism, it preserves structure, therefore $\alpha(g_1)\alpha(g_2) = \alpha(g_1 g_2)$. So,

$$\alpha^{-1}(\alpha(g_1)\alpha(g_2)) = \alpha^{-1}(\alpha(g_1 g_2)) = g_1 g_2,$$

as required.

3.8 Quotient groups

3.8.1 Quotient groups and the first isomorphism theorem in plain English

1. You have groups G and H and a homomorphism $f : G \to H$. That is special; it is not just any map.
2. You use the values of f to define an equivalence relation $\sim$ on G. That's not special, you could do that with any map.
3. Note that this will only be interesting if f is non-injective, i.e. if the equivalence relation does actually group some elements together.
4. You define a group operation on the equivalence classes of G. This is "inherited from the underlying group"[4].

   So far everything has been straightforward; here is the only subtle point:

   It is essential that the operation defined on the equivalence classes is well-defined. Fortunately, it will be. Ultimately, the reason is that the equivalence classes were defined by the values of a homomorphism, not just any arbitrary labeling.
5. So now you have a new group $G/\sim$ containing equivalence classes. There are four interesting things about it:
   (a) Obvious: It is smaller (simpler) than the original group G: you have "modded out" by the equivalence relation.
   (b) Somewhat obvious: All information about the original group structure on G is preserved in the group structure on $G/\sim$. This is because we decided that the group operation on the equivalence classes would be inherited from G.
   (c) Obvious: There is a one-to-one correspondence between the equivalence classes and the image of the homomorphism (the elements of the image "label" the equivalence classes).
   (d) Somewhat obvious: this one-to-one correspondence is actually an isomorphism.

[4] What this means is that the rule for combining equivalence classes is as follows:
(a) Pick an arbitrary element of each equivalence class.
(b) Combine those two elements according to the group operation.
(c) Declare the result of the operation on equivalence classes to be the equivalence class of the group-element level result.

Why? We started off with a non-injective homomorphism $G \to H$. Then we did two things: (1) we coalesced into a single new element all the elements in G that mapped to the same element in H; (2) we declared that the new coalesced element on the left

Theorem 1 (First Isomorphism theorem: statement I)**.**
Let $f : G \to H$ be a group homomorphism.

Let $\sim$ be the equivalence relation on G defined by $g_1 \sim g_2 \iff f(g_1) = f(g_2)$.

Then the set $G/\sim$ of equivalence classes "inherits the structure" of G in the following sense:

Let $C_1, C_2 \in G/\sim$ be equivalence classes with $g_1 \in C_1$ and $g_2 \in C_2$. Define $C_1C_2 :=$ (equivalence class of g_1g_2). Then this is well-defined and $G/\sim$ is a group under this operation.

TODO: state something about isomorphism of $G/\sim$ and $\operatorname{Im} f$. The group operation in H could be anything; all we know is that $f : G \to H$ is a homomorphism. And the group operation in $G/\sim$ is inherited from G. The point here is that the map $G/\sim \to \operatorname{Im} f$ is a homomorphism, just as the original map $f : G \to H$ was. In fact, it's an isomorphism, because it is a bijection.

Remark.
Note that the equivalence class of g_1, i.e. the preimage of $f(g_1)$, is

$$f^{-1}(f(g_1)) = g_1 \cdot \operatorname{Ker} f.$$

This is true since, (reverse direction) if $k \in \operatorname{Ker} f$, then $f(g_1k) = f(g_1)e = f(g_1)$; and (forwards direction) if $f(g_2) = f(g_1)$ then (TODO: prove subset in this direction).

Therefore an equivalent definition of the equivalence relation is:

$G/\operatorname{Ker} f := G/\sim$, where $g_1 \sim g_2 \iff (g_1 \cdot \operatorname{Ker} f) = (g_2 \cdot \operatorname{Ker} f)$.

$G/\operatorname{Ker} f$ is a set of cosets of $\operatorname{Ker} f$, and may also be thought of as a set of equivalence classes of $\sim$.

This gives rise to the conventional statement of the theorem:

Theorem 2 (First Isomorphism theorem: statement II)**.**
Let $f : G \to H$ be a group homomorphism.

Then the set $G/\operatorname{Ker} f$ "inherits the structure" of G in the following sense:

Let $C_1, C_2 \in G/\operatorname{Ker} f$ be cosets of $\operatorname{Ker} f$, with $g_1 \in C_1$ and $g_2 \in C_2$. Define $C_1C_2 := (g_1g_2 \cdot \operatorname{Ker} f)$. Then this is well-defined and $G/\operatorname{Ker} f$ is a group under this operation.

TODO: state something about isomorphism of $G/\operatorname{Ker} f$ and $\operatorname{Im} f$.

3.8.2 Summary

A quotient group can be formed by:

1. Identify a subgroup
2. Form cosets
3. Inherit operation on cosets from operation on original group elements

But only if the subgroup is normal: that's what's required for inheriting the group operation to result in a well-defined operation on the cosets (i.e. when performing an operation involving members of two different cosets, all choices of members to act as exemplars of the cosets give the same result.)

3.8.3 Modular arithmetic

The canonical example of a quotient group comes from "modular arithmetic" on the integers. For example, consider the integers, mod 4. This means that every integer is mapped to whatever its remainder is after dividing by 4. The integers mod 4 is a group, which contains 4 elements: $\{\overline{0}, \overline{1}, \overline{2}, \overline{3}\}$. So $5 \to \overline{1}$, $14 \to \overline{2}$, $-1 \to \overline{3}$, etc.

We said that the integers mod 4 are a group, so what is the group operation? The answer is that we define an addition law on the elements: for example, $\overline{1} + \overline{3} = \overline{1+3} = \overline{4} = \overline{0}$. In words, to find the result of combining $\overline{1}$ with $\overline{3}$, you first add 1 and 3 as usual to get 4, then see where 4 is mapped to, and that's the answer. This corresponds to the fairly familiar notion that e.g. $5 + 23 = 28 = 0 \pmod 4$, but it is a bit subtle/slippery, and it helps to pause and consider exactly what's going on.

Let's be explicit about what $\overline{1}$ actually is: it's the set of all integers that have 1 as their remainder when divided by 4. So, $5 \in \overline{1}$ for example. What we've just done is define an addition operation on these *sets* (as opposed to addition of integers). The operation works as follows. To add $\overline{2}$ and $\overline{3}$, you do the following:

1. Take any number that has remainder 2 (mod 4) and any number that has remainder 3 (mod 4). 2. Add them together using normal integer addition and find the remainder (mod 4) of the result.

There are two important points here.

Firstly, we didn't need anything beyond what we had to come up with this operation: it uses the addition operation that's *already defined* on the main group to define an operation on the subsets.

Secondly, it is only well-defined if you always get the same answer regardless of which integers you pick in step (1). In this case that is true.

So we have an example of a "quotient group": $\{\overline{0}, \overline{1}, \overline{2}, \overline{3}\}$ under this addition operation. Let's recap and start putting this in group theoretic terminology.

3.8.4 A quotient group is a group of cosets

$\overline{0} \pmod 4$ is the following subset of the integers $\mathbb{Z}$ under addition: $\{\ldots, -12, -8, -4, 0, 4, 8, 12, \ldots\}$. It's not only a subset, but a *subgroup* (it contains the identity element 0, every element has an additive inverse, and addition stays within the subset). It is written as $4\mathbb{Z}$ (or in general, $n\mathbb{Z}$ for mod n). However, we will often use H for a subgroup, so let's call it H.

$\overline{1} = \{\ldots, -11, -7, -3, 1, 5, 9, 13, \ldots\}$ is not a subgroup of $\mathbb{Z}$ because it does not contain the identity element. What it is is a *coset* of the subgroup H: the set comprising all the results you get by adding 1 to elements of H. We can write this as $1 + H$. In fact, it's usually written $1H$; we just have to remember that the operation here is additive rather than multiplicative.

Of course, $\overline{2}$ and $\overline{3}$ are cosets defined in the same way. $\{\overline{0}, \overline{1}, \overline{2}, \overline{3}\}$ are the only distinct cosets: for example, $\overline{4} = 4 + H$ is exactly the same set of integers as $\overline{0}$. Similarly, $5 + H = \overline{1}$, etc.

So we arrived at the (integers mod 4) quotient group as follows:

1. We started with the group of integers under addition, $\mathbb{Z}^+$.
2. We identified a subgroup H.
3. We identified the cosets of H: $\{H, 1 + H, 2 + H, 3 + H\}$
4. We defined an operation on the cosets: $(i + H) + (j + H) = (i + j) + H$.

[4] Multiplication preserves structure also: $\mathbb{Z}/n\mathbb{Z}$ is a field iff n is prime.

5. We noted that it was only well-defined because

$$\begin{pmatrix}\text{any number with}\\ \text{remainder } i\end{pmatrix}+\begin{pmatrix}\text{any number with}\\ \text{remainder } j\end{pmatrix}=\begin{pmatrix}\text{a number with the same}\\ \text{remainder as } i+j\end{pmatrix}.$$

Note that (3) and (4) can equally be written like this, which is how it's likely to be written when considering subgroups and cosets more abstractly:

1. We identified the cosets of H: $\{H, 1H, 2H, 3H\}$
2. We defined an operation on the cosets: $(iH)+(jH)=(ij)H$.

3.8.5 Notational digression

The integers mod 4 is written $\mathbb{Z}/4\mathbb{Z}$. It's an example of a quotient group. You read that as (some group)/(some subgroup). In this case the group is the integers under addition, and the subgroup is $4Z=\{\ldots,-12,-8,-4,0,4,8,12,\ldots\}$[5].
In general, one writes G/H to refer to the quotient group of "G mod H".

3.8.6 A second example of a quotient group

Here's an example of a (simple) problem from an undergraduate textbook on group theory:

> Identify the quotient group $\mathbb{R}^\times/P$, where P denotes the subgroup of positive real numbers.

What does this mean and how does one do it? Well, let's try to follow the same steps as for the integers mod 4 example above.

Our starting group is the non-zero real numbers under multiplication $\mathbb{R}^\times$: this plays the role of $\mathbb{Z}^+$ in the modular arithmetic example. And the subgroup is the positive real numbers P.

What are the cosets of P? To get one example of a coset, you pick a number x from the main group, and you form a set by combining x with each element of the subgroup P in turn. So that's the set $\{xp|p\in P\}$. We can see that we're either going to get all the positive reals (if x is positive), or all the negative reals (if x is negative). So the set of cosets has those two sets as its elements: $\{P,-1P\}$.

OK, so we've done steps (1)-(3). Now, what's the group operation that's going to combine two cosets and produce another coset? Well, the whole point is that this group operation is inherited from the original group: that's what we did in the integers mod 4 example; we used the standard addition of integers to define the result of adding cosets $i+H$ and $j+H$ to be the coset $(i+j)+H$. The analogous definition here would be to use the standard multiplication of real numbers to say that $(xP)(yP)=(xy)P$. That's going to lead us to the following intuitively reasonable multiplication table:

	P	$-1P$
P	P	$-1P$
$-1P$	$-1P$	P

And we conclude that the quotient group is isomorphic to the group of size 2 (there's only one – the one with this multiplication table).

[5] In this context, $4\mathbb{Z}$ always means multiplication, even if the group operation is addition! So it's the set $\{4z|z\in\mathbb{Z}\}$. It is *not* the same as the coset $4+\mathbb{Z}=\{4+z|z\in\mathbb{Z}\}$. This is a well-established notational inconsistency.

The only question is (5): is the operation on cosets well-defined? In this case, the answer is yes: for example, any positive number $x \in P$, multiplied by any negative number $y \in -1P$, is going to give a negative number $xy \in -1P$.[6].

3.8.7 Quotient groups of arbitrary groups

What about in general? If we have a subgroup, can we just identify the cosets of the subgroup, and define a composition law on them using the composition law from the main group? Will it always be well-defined in the sense answered above? The answer is: yes if and only if the subgroup is "normal".

A normal subgroup H is defined to be a subgroup that is closed under conjugation. This means that you can take any element g of the main group, form the product ghg^{-1} using any element h of the subgroup, and the result will always be in the subgroup. One can prove that if and only if this is true, then the composition of cosets is well-defined, in which case the prescription above for forming a quotient group can be followed (find the cosets of H, define the operation on the cosets).

So, if you need to find a quotient group of some subgroup, you need to show that the subgroup is normal. There are two ways of doing that:

1. Show that it is closed under conjugation.
2. Show that it is the kernel of a homomorphism.

[6] We can prove it easily here because the group is commutative: $(xP)(yP) = (Px)(yP) = P(xy)P = (xy)PP = (xy)P$. In addition to commutativity those steps make use of associativity and closure.

3.8.8 Quotient groups

A mapping f preserves structure if, for example:

$$\begin{aligned} a &\mapsto f(a) \\ b &\mapsto f(b) \\ ab &\mapsto f(a)f(b) \end{aligned}$$

An isomorphism is a bijection that preserves structure.

A homomorphism is a mapping that preserves structure but isn't necessarily a bijection.[7]

Theorem 3. *The kernel of a homomorphism is a subgroup* [8].

Example: the group of *rotations* of $\mathbb{R}^3$ is a subgroup of the group of rigid motions that fix the origin (the latter includes reflections). Now the $\det : \mathrm{GL}_3(\mathbb{R}) \to \mathbb{R}$ mapping is a homomorphism, since $\det(T_1T_2) \equiv \det(T1)\det(T_2)$. The rotations are those mappings with determinant 1, hence they are the kernel of a homomorphism.

Theorem 4. *Some subgroups are not the kernel of any homomorphism*

The counter example given in the proof (below) is an element of the form $g^{-1}hg$ for h in the subgroup and g outside the subgroup. Basically, we observe that

$$\varphi(g^{-1}hg) = \varphi(g^{-1}) \cdot \varphi(h) \cdot \varphi(g) = \varphi(g^{-1}) \cdot e \cdot \varphi(g) = \varphi(g^{-1}) \cdot \varphi(g) = \varphi(e) = e,$$

and therefore that the subgroup, if it is to be a kernel, must *contain* all products of the form $g^{-1}hg$ (conjugation by an element outside the subgroup).

So,

$$\textbf{kernel of homomorphism} \implies \textbf{closed under conjugation}.$$

But does

$$\textbf{closed under conjugation} \implies \textbf{kernel of homomorphism}\ ?$$

Yes. Closure under conjugation implies that the **subgroup is the kernel of the homomorphism which maps g to its coset, with the operation on cosets inherited from the group**: $(g_1H)\cdot(g_2H) = g_1g_2H$. Justification of this claim follows.

Suppose that $\varphi : G \to K$ is a homomorphism from G to some group K, and that the kernel of φ is H and that H is closed under conjugation. What can we deduce about φ?

Theorem 5.
The left and right cosets of H coincide, and φ is constant on the cosets, taking different values on each coset.

This means that there is a bijection between the cosets of H and the image of φ. So, we can say that the image of φ *is* the cosets of H.

---

[6] Notes based on Tim Gowers' blog https://gowers.wordpress.com/2011/11/20/normal-subgroups-and-quotient-groups/

[7] I.e. it might not be injective (might send different inputs to the same output), or might not be surjective (might fail to hit certain elements).

[8] **Proof.**
Kernel is $\{a : f(a) = e\}$.
Contains identity? Yes, homomorphisms always send the identity to the identity ($f(ea) = f(e)f(a)$ but this must equal $f(a)$, hence $f(e)$ is the identity.) Contains inverses? Yes, $f(aa^{-1}) = f(a)f(a^{-1}) = ef(a^{-1}) = e$, so $f(a^{-1})$ must also be e.

Theorem 6.
If $\varphi(g_1H) = a_1$ and $\varphi(g_2H) = a_2$ then $\varphi(g_1g_2H) = a_1a_2$.

This allows us to define the group operation on the elements of the image of φ: it implies that

$$(g_1H) \cdot (g_2H) = g_1g_2H.$$

Theorem 7.

$$\textbf{\textit{kernel of homomorphism}} \iff \textbf{\textit{closed under conjugation}} \iff gH = Hg \quad \forall g \in G$$

Theorem. *Some subgroups are not the kernel of any homomorphism.*

Proof. Counter-example: consider the permutation group $S_3 = \{e, (12), (13), (23), (231), (312)\}$, and its subgroup $\{e, (12)\}$.

Suppose this subgroup is the kernel of a homomorphism. I.e. $e \mapsto e, (12) \mapsto e$, but nothing else is sent to the identity.

Now consider $(13)(12)(13)$:

$$123 \to 321 \to 231 \to 132,$$

i.e. $(13)(12)(13) = (23)$.

But

$$\varphi\Big((13)(12)(13)\Big) = \varphi\Big((13)\Big)\varphi\Big((12)\Big)\varphi\Big((13)\Big) = \varphi\Big((13)\Big)\varphi\Big((13)\Big) = \varphi\Big((13)(13)\Big) = e,$$

so $(23) \mapsto e$, which is a contradiction. Therefore $\{e, (12)\}$ isn't the kernel of any homomorphism. □

Theorem. *The left and right cosets of H coincide, and φ is constant on the cosets, taking different values on each coset.*

Proof. TODO □

Theorem.
If $\varphi(g_1H) = a_1$ and $\varphi(g_2H) = a_2$ then $\varphi(g_1g_2H) = a_1a_2$.

Proof. TODO □

3.8.9 First isomorphism theorem

Every normal subgroup is the kernel of the homomorphism that sends a group element to its coset.

Can two distinct homomorphisms share the same kernel?

Let $f : G \to G'$ be a homomorphism with kernel N. $e \in N$, therefore every $g \in G$ is in some coset gN, so the set of cosets partitions the domain. What about the image? Consider two elements gn_1 and gn_2 of the same coset. These both get sent to the same value, since $f(gn_i) = f(g)f(n_i) = f(g)$.

So is it possible to have homomorphisms f and φ with the same kernel N but with $f(g) \neq \varphi(g)$ for some $g \in G$? If that were true [...]

3.9 Exercises - Harvard E122

Exercises from Artin *Algebra* 1st edition.

Harvard E122 (http://wayback.archive-it.org/3671/20150528171650/https://www.extension.harvard.edu/open-learning-initiative/abstract-algebra)

Harvard 122 (http://www.math.harvard.edu/ ctm/home/text/class/harvard/122/02/html/hw.html)

3.9.1 E122 Homework 1

** 1. Read 1.1, pp. 38-42 **

** 1.1.7 Find a formula for $\begin{pmatrix} 1 & 1 & 1 \\ & 1 & 1 \\ & & 1 \end{pmatrix}^n$ and prove it by induction. **

** 1.1.16 **

** 1.1.17 **

3.9.2 E122 Homework 2

Read 2.1, 2.2

** 2.1.5 **

** 2.1.7 **

** 2.2.1 **

** 2.2.14 (122) Let G be a cyclic group of order n, and let r be an integer dividing n. Prove that G contains exactly one subgroup of order r. **

If n is prime then the only subgroups of G are $\{e\}$ and G. The integers which divide n are 1 and n and so it is true that for every such integer r there is one subgroup of order r.

If n is not prime then G has non-trivial subgroups for every integer that divides it.

For example, if G has order 8, the claim is that G has only one subgroup of order 2. Well, all subgroups of order 2 are isomorphic to $\{e, \tau\}$. So the claim implies that groups of permutations of order 4, 6, 8 etc (i.e. S_4, S_6, S_8) contain only one transposition.

A solution is given in the [Harvard 122 materials](http://www.math.harvard.edu/ ctm/home/text/class/harvard/122/02/html

[8] https://theoremoftheweek.wordpress.com/2010/05/20/theorem-26-the-first-isomorphism-theorem/

The starting point is to note that $< g^{n/r} >$ is a subgroup of order r (n/r fits r times into each chunk of n items). So in the example with G of order 8, $< g^4 >$ has order 2. The question is whether there is any other subgroup of order 2. Intuitively it seems obvious that the answer is no, since only 4 divides up chunks of 8 in that way.

Formally, the proof proceeds by supposing that $H' =< g^m >$ also has order r (using a lemma that every subgroup of a cyclic group is cyclic). It then shows that m must divide n (otherwise leads to a contradiction). Therefore the order of H' is n/m. Therefore $m = n/r$ which shows that H' is H.

** 2.2.15 **

** 2.2.20(a) **

3.9.3 E122 Homework 3

** 2.3.1 **

** 2.3.11 **

** 2.3.12 **

** 2.4.3 **

** 2.4.5 (122) **

** 2.4.6 **

** 2.4.8 (122) **

** 2.4.11 **

** 2.4.16 (122) **

** 2.4.23 (122) **

3.9.4 E122 Homework 4

** 1. Read Artin 1.4.**

** 2. Let V denote the Klein 4-group. Show that $\mathrm{Aut}(V)$ is isomorphic to S_3.**

Every automorphism sends the identity to itself. This is because an automorphism must preserve structure, therefore we require $\rho(e) = \rho(ee^{-1}) = \rho(e)\rho(e^{-1}) = \rho(e)\rho(e)^{-1}$ which implies $\rho(e) = e$.

Therefore, elements of $\mathrm{Aut}(V)$ are distinguished by their effect on the 3 non-identity elements and there is a 1-1 correspondence f between elements of $\mathrm{Aut}(V)$ and S_3.

We need to show that f is a homomorphism, i.e. that $f(\rho_1\rho_2) = f(\rho_1)f(\rho_2)$. The operation is composition of permutations in both groups...the identity seems obvious.

3.

Define $f : \mathrm{GL}_n(\mathbb{R}) \to \mathrm{GL}_n(\mathbb{R})$ by $f(A) = {}^tA^{-1}$ (where ${}^tA$ is the transpose of A). Show that f is an automorphism, but not an inner automorphism for $n \geq 1$.

To show that f is an automorphism we need to show that it preserves structure and is a bijection.

Preservation of structure: We require that ${}^t(AB)^{-1} = {}^tA^{-1}\,{}^tB^{-1}$. The RHS is ${}^tA^{-1}\,{}^tB^{-1} = ({}^tB\,{}^tA)^{-1} = {}^t(AB)^{-1}$ as required.

Bijection: $f^{-1}(A) = {}^tA$. Since an inverse mapping exists, the original mapping must be bijective.

Finally, we need to show that f is not an inner automorphism for $n \geq 1$. An inner automorphism is an automorphism defined by $\rho(A) = BAB^{-1}$ for some fixed B. So we need to show that there is no B for which ${}^tA^{-1} = BAB^{-1}$ for all A. BAB^{-1} corresponds to the action of A performed in a basis defined by B. The determinant of A is invariant under change of basis. However, $\det {}^tA^{-1} = (\det A)^{-1}$. Therefore ${}^tA^{-1} = BAB^{-1}$ is not true in general since it is not true when $\det A \neq \pm 1$.

http://math.stackexchange.com/questions/98378/fx-tx-1-is-an-automorphism-of-gl-n-mathbbr

I'm also watching these lectures and trying to do the homework, so by no means an expert. In the final part of the proof, a variant would be to use determinants:

Assume that f is an inner automorphism. Therefore for some fixed B, ${}^tA^{-1} = BAB^{-1}$ for all A. But this implies $(\det A)^{-1} = \det A$ which is true only for $\det A = \pm 1$. Therefore f is not an inner automorphism.

4. 1.4.5 Prove that the transpose of a permutation matrix P is its inverse.

3.9.5 E122 Homework 5

2.5.1 Prove that the nonempty fibres of a map form a partition of the domain.

We need to show

1. That every element of the domain is in some fibre, and 2. That no element is in more than one fibre

Let f be a map from set S to set T. A fibre ϕ_t of f is the set $\{s : f(s) = t\}$.

For any element s in the domain, $f(s)$ exists and s is in fibre $\phi_{f(s)}$. This proves (1).

To prove (2), suppose that s belongs to fibres ϕ_{t_1} and ϕ_{t_2}. Then $f(s) = t_1$ and $f(s) = t_2$ which is a contradiction showing that no element belongs to two fibres.

2.5.6

(a) Prove that the relation x conjugate to y in a group G is an equivalence relation on G.

x conjugate to y means $y = gxg^{-1}$ for some $g \in G$. We need to show that the relation is reflexive, transitive and symmetric.

Reflexivity: x is equal to x conjugated with the identity element: $x = exe^{-1}$.

Symmetry: Let $y = gxg^{-1}$. Then multiplying on the right by g and on the left by g^{-1} gives $g^{-1}yg = x$.

Transitivity: Let $y = g_1xg_1^{-1}$ and $z = g_2yg_2^{-1}$. Then $z = g_2(g_1xg_1^{-1})g_2^{-1} = (g_2g_1)x(g_1^{-1}g_2^{-1})$

(b) Describe the elements a whose conjugacy class (= equivalence class) consists of the element a alone.

2.6.2 Prove directly that distinct cosets do not overlap.

Let H be a subgroup of G and let $g_1H = \{g_1h : h \in H\}$ be a coset of H in G. Consider an element $g_3 \in g_1H$. This means that $g_3 = g_1h$ for a unique $h \in H$.

Suppose that g_3 is also in another coset g_2H. Then $g_3 = g_1h = g_2h'$ for $h, h' \in H$ and therefore $g_1hh'^{-1} = g_2$. But $hh'^{-1} \in H$ so g_2 is in coset g_1H, which shows that cosets g_1H and g_2H are the same.

2.6.4 Give an example showing that left cosets and right cosets of $\mathrm{GL}_2(\mathbb{R})$ in $\mathrm{GL}_2(\mathbb{C})$ are not always equal.

We don't even need to consider matrices outside $\mathrm{GL}_2(\mathbb{R})$. Let A = $\left(\begin{smallmatrix}1 & 0\\ 0 & 0\end{smallmatrix}\right)$ and let $H = \mathrm{GL}_2(\mathbb{R})$.

Geometrically, this matrix projects points in 2D onto the x-axis.

The left coset of $\mathrm{GL}_2(\mathbb{R})$ containing A is the set of matrices of the form

$$\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}a & b\\ c & d\end{pmatrix} = \begin{pmatrix}a & b\\ 0 & 0\end{pmatrix}$$

where $a, b, c, d \in \mathbb{R}$. Geometrically, this composition first sends the basis vectors to new locations with x-coordinates a and b, and then projects onto the x axis.

The right coset of $\mathrm{GL}_2(\mathbb{R})$ containing A is the set of matrices of the form

$$\begin{pmatrix}a & b\\ c & d\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix} = \begin{pmatrix}a & 0\\ c & 0\end{pmatrix}$$

Geometrically, this composition first sends the $\begin{bmatrix}0\\1\end{bmatrix}$ basis vector to the origin, and then performs an arbitrary linear transformation. But once sent to the origin, the $\begin{bmatrix}0\\1\end{bmatrix}$ basis vector stays there, regardless of the nature of the second transformation.

2.6.5 Let H, K be subgroups of a group G of orders 3, 5 respectively. Prove that $H \cap K = \{e\}$.

Since H and K are of prime order, they are isomorphic to the cyclic groups of orders 3 and 5 respectively. Therefore the non-identity elements of H are of order 3, and the non-identity elements of K are of order 5 (because the order of an element in a group must divide the order of the group). Therefore, while e is in $H \cap K$, no other element is.

** 2.6.10 (122) **

** 2.6.11 (122) **

** 2.8.2 (122) **

** 2.8.10 (122) **

3.9.6 E122 Homework 6

** 2.9.2 2. **

** (a) Prove that the square a^2 of an integer a is congruent to 0 or 1 modulo 4. **

Suppose a is even, so $a = 2n$ for some $n \in \mathbb{Z}$. Then $a^2 = 4n^2 \equiv 0 \pmod 4$. Alternatively, suppose a is odd, i.e. $a = 2n + 1$ for some $n \in \mathbb{Z}$. Then $a^2 = 4n^2 + 4n + 1 \equiv 1 \pmod 4$.

** (b) What are the possible values of a^2 modulo 8? **

If a is even, then $a^2 = 4n^2$. If n^2 is odd then $a^2 \equiv 4 \pmod 8$ and if even then $a^2 \equiv 0 \pmod 8$. If a is odd, then $a^2 = 4n(n+1) + 1 \equiv 1 \pmod 8$. So the possible values are 0, 1, 4.

** 2.9.4 Prove that every integer a is congruent to the sum of its decimal digits modulo 9. **

Let $a = d_0 + d_1 10 + d_2 10^2 + \ldots$.

[*Good solution*](https://github.com/AMouri/artin-algebra/blob/3083860baf553b472495fd01ef62489db9a261ee/ChapterNot
that $10 \equiv 1 \pmod 9$. So $a \equiv d_0 + d_1 + d_2 + \ldots \pmod 9$.

My solution: we require that the difference is a multiple of 9. The difference is $\sum_{i=0}^{\infty} d_i 10^i - \sum_{i=0}^{\infty} d_i = \sum_{i=0}^{\infty} d_i(10^i - 1)$...which is a multiple of 9.

** 2.9.5 Solve the congruence $2x \equiv 5$ **

** (a) modulo 9 **

[*Good solution*](https://github.com/AMouri/artin-algebra/blob/3083860baf553b472495fd01ef62489db9a261ee/Chapter$x =$
$2^{-1}5 \equiv 5 \cdot 5 \equiv 7 \pmod 9$

The possible values of $2x \pmod 9$ are 0, 2, 4, 6, 8, 1, 3, 5, 7, 9. So 2 (mod 9) does have a multiplicative inverse (5).

My solution: $14 = 2 \cdot 7 \equiv 5 \pmod 9$. So the solution is $x \equiv 7 \pmod 9$.

** (b) modulo 6 ** [*Good solution*](https://github.com/AMouri/artin-algebra/blob/3083860baf553b472495fd01ef62489db9a
possible values of $2x \pmod 6$ are 0, 2, 4. Therefore $2x \equiv 5$ has no solution.

I.e. 2 (mod 6) has no multiplicative inverse?

My solution: $\overline{5} \pmod 6 = 5 + 6\mathbb{Z} = 1 + 2 \cdot 2 + 2 \cdot 3\mathbb{Z} = 1 + 2(2 + 3\mathbb{Z})$ is an equivalence class of odd numbers. Therefore $2x \equiv 5 \pmod 6$ has no solutions.

2.9.8 Use Proposition (2.6) to prove the Chinese Remainder Theorem: Let m, n, a, b be integers, and assume that the greatest common divisor of m and n is 1. Then there is an integer x such that $x \equiv a \pmod m$ and $x \equiv b \pmod n$.

Proposition 2.6 is the theorem describing the fact that, for two integers i and j, $\gcd(i, j)$ generates the subgroup $i\mathbb{Z} + j\mathbb{Z}$.

Here's the proof from [A Mouri's solutions](https://github.com/AMouri/artin-algebra/blob/3083860baf553b472495fd01ef6248

From Proposition 2.6, $1 = an + bm$. Then

$1 \equiv an + bm \pmod m \to 1 \equiv an \pmod m$.

Since $\gcd(n, m) = 1$, then $n^{-1} \equiv a \mod m$. Similarly, $m^{-1} \equiv b \mod n$.

My attempt:

We need to show that for any values a, b, we can always find an integer x which exceeds the previous multiple of m by a and exceeds the previous multiple of n by b.

So we want to show that the intersection of $a + m\mathbb{Z}$ and $b + n\mathbb{Z}$ is non-empty. I.e. we need to show that we can always find integers r, s satisfying $a + rm = b + sn$. That's equivalent to $rm - sn = b - a$. We know that $m\mathbb{Z} + n\mathbb{Z} = \gcd(m, n)\mathbb{Z}$, and thus since $\gcd(m, n) = 1$ we know that $m\mathbb{Z} + n\mathbb{Z}$ is all of $\mathbb{Z}$. Therefore the integer $b - a$ must be reachable by taking some number b of steps of length m and some number $-a$ of steps of length n.

3.9.7 E122 Homework 7

** 2.10.1 (E122 & 122) Let G be the group of invertible real upper triangular 2 x 2 matrices. Determine whether or not the following conditions describe normal subgroups H of G. If they do, use the First Isomorphism Theorem to identify the quotient group G/H. **

$$G = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \right\}$$

** (a) $a_{11} = 1$ **

$$H = \left\{ \begin{pmatrix} 1 & b \\ 0 & d \end{pmatrix} \right\}$$

This is closed, contains the identity and contains inverses, so it is a subgroup. And closed under conjugation, so normal.

Geometrically, the elements of the subgroup H are matrices representing a stretch in the direction of the second basis vector ("Y axis") and shearing parallel to the first basis vector ("X axis). Conjugation yields matrices ghg^{-1} which correspond to doing the same transformation in a different basis. However, this class of change-of-basis operations preserves the direction of the first basis vector. Therefore the transformation,

in the new basis, also corresponds to stretching in the direction of the second basis vector and shearing parallel to the first basis vector, and thus the conjugated matrix is still in the subgroup.

$$\begin{aligned}
&\begin{pmatrix} e & f \\ 0 & h \end{pmatrix} \begin{pmatrix} 1 & b \\ 0 & d \end{pmatrix} \begin{pmatrix} h & -f \\ 0 & e \end{pmatrix} \frac{1}{eh} \\
=&\begin{pmatrix} e & f \\ 0 & h \end{pmatrix} \begin{pmatrix} h & -f+be \\ 0 & de \end{pmatrix} \frac{1}{eh} \\
=&\begin{pmatrix} eh & e(-f+be)+fde \\ 0 & hde \end{pmatrix} \frac{1}{eh}
\end{aligned}$$

The theory of quotient groups involves identifying establishing a group structure on a set of cosets of a normal subgroup: this set of cosets is then a quotient group G/H. There's a homomorphism $f : G \to G/H$ defined by $f(g) =$ (coset containing g), and the multiplication law between cosets is

(coset containing g) x (coset containing g') = (coset containing gg')

The normal subgroup is the kernel of this homomorphism, since $f(1) =$ (coset containing 1) $= 1H$.

So we have a subgroup H of the group G of real upper-triangular 2x2 matrices. We've determined that the normal subgroup is H, a particular subset of real upper-triangular matrices. Each coset is therefore, for some fixed $g \in G$, the set of matrices that can be obtained by multiplying g by some $h \in H$. So there are infinitely many cosets and each has infinitely many elements. Each coset contains matrices of the form

$$\begin{pmatrix} 1 & b \\ 0 & d \end{pmatrix} \begin{pmatrix} e & f \\ 0 & h \end{pmatrix} = \begin{pmatrix} e & b+df \\ 0 & dh \end{pmatrix}.$$

Since b and d can be any real numbers, we could write this as

$$\begin{pmatrix} 1 & * \\ 0 & * \end{pmatrix} \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} = \begin{pmatrix} a & * \\ 0 & * \end{pmatrix},$$

thus the cosets are distinguished by the real number in their top-left entry. So the quotient group G/H is $\left\{ \left(\begin{smallmatrix} a & * \\ 0 & * \end{smallmatrix}\right) : a \in \mathbb{R} \right\}$; it is isomorphic to the positive real numbers (elements of G would not be invertible if a were 0). Viewed as transformations of the plane, the elements of G/H are distinguished by the magnitude of their stretch in the x-direction.

** (b) $a_{12} = 0$ **

$$H = \left\{ \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix} \right\}$$

This is closed, contains identity and inverses, hence is a subgroup. However again, it is not closed under conjugation, so not normal.

Geometrically, the matrix represents a stretch in orthogonal x- and y-directions. The non-normality is because the matrix that performs the transformation in the changed basis does not stretch in the directions of the basis vectors in the original basis.

$$\begin{aligned}
&\begin{pmatrix} e & f \\ 0 & h \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix} \begin{pmatrix} h & -f \\ 0 & e \end{pmatrix} \frac{1}{eh} \\
=&\begin{pmatrix} e & f \\ 0 & h \end{pmatrix} \begin{pmatrix} ah & -af \\ 0 & de \end{pmatrix} \frac{1}{eh} \\
=&\begin{pmatrix} eah & -eaf + fde \\ 0 & hde \end{pmatrix} \frac{1}{eh}
\end{aligned}$$

** (c) $a_{11} = a_{22}$ **

$$H = \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \right\}$$

Closed and contains identity and inverses, so a subgroup, but again not closed under conjugation.

Geometrically, the matrix represents a stretch by the same amount in orthogonal x- and y-directions, plus a shear. The non-normality is because the matrix that performs the transformation in the changed basis does not stretch equally in the directions of the basis vectors in the original basis.

$$\begin{aligned}
&\begin{pmatrix} e & f \\ 0 & h \end{pmatrix} \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \begin{pmatrix} h & -f \\ 0 & e \end{pmatrix} \frac{1}{eh} \\
=&\begin{pmatrix} e & f \\ 0 & h \end{pmatrix} \begin{pmatrix} ah & -af + be \\ 0 & de \end{pmatrix} \frac{1}{eh} \\
=&\begin{pmatrix} eah & e(-af + be) + fde \\ 0 & hde \end{pmatrix} \frac{1}{eh}
\end{aligned}$$

** (d) $a_{11} = a_{22} = 1$ **

$$H = \left\{ \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \right\}$$

Closed, contains identity and inverses. It performs a shear, with no stretching. Closed under conjugation and therefore normal.

Geometrically, the matrix represents a shear, with no stretching. The normality is because the matrix that performs that transformation in the changed basis also performs a shear without stretching in the original basis.

$$\begin{aligned}
&\begin{pmatrix} e & f \\ 0 & h \end{pmatrix} \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} h & -f \\ 0 & e \end{pmatrix} \frac{1}{eh} \\
=&\begin{pmatrix} e & f \\ 0 & h \end{pmatrix} \begin{pmatrix} h & -f + be \\ 0 & e \end{pmatrix} \frac{1}{eh} \\
=&\begin{pmatrix} eh & e(-f + be) + fe \\ 0 & eh \end{pmatrix} \frac{1}{eh}
\end{aligned}$$

Each coset contains matrices of the form

$$\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e & f \\ 0 & h \end{pmatrix} = \begin{pmatrix} e & f + bh \\ 0 & h \end{pmatrix}.$$

which could be written as

$$\begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e & f \\ 0 & h \end{pmatrix} = \begin{pmatrix} e & * \\ 0 & h \end{pmatrix}.$$

The elements of the quotient group G/H are thus distinguished by their diagonal entries. I.e., G/H is isomorphic to $\mathbb{R}^2$.

** 2.10.3 Let P be a partition of a group G with the property that for any pair of elements A, B of the partition, the product set AB is contained entirely within another element C of the partition. Let N be the element of P which contains 1. Prove that N is a normal subgroup of G and that P is the set of its cosets. **

** Prove that N is a normal subgroup of G: **

We know that $1 = 1^{-1}$ is in N. Therefore we know that the product set NN must lie entirely within N, since one example of a product (i.e. $1 \cdot 1^{-1}$) is in N. I.e. N is closed. Does it contain inverses? Consider an arbitrary element n of N. Let n^{-1} be in some element $A \in P$ of the partition. Then the product set NA must lie entirely within N, since nn^{-1} does. So either $A = N$, or N and A are distinct elements of the partition but with the property that $NN = NA$. I'm going to say that $NN = NA$ implies $N = A$ by the definition of the problem, and thus that N contains inverses, but I may be missing something here.

So, N contains the identity, is closed and contains inverses. It inherits associativity from G. So N is a subgroup. Is it normal? Consider gNg^{-1} for some fixed g. Since $1 \in N$, we conclude that gN is a subset of the element A of the partition that contains g. And since $g \in gN$, $(gN)g^{-1}$ must be a subset of N. Therefore N is normal.

** Prove that P is the set of its cosets: **

The set of cosets of N are $\{gN : g \in G\}$ by definition.

We need to show that exactly one partition satisfying the definition of P exists, and that it is the set of cosets of N.

First, we show that the set of cosets satisfy the definition of P. Consider two cosets gN and hN. Their product set is $(gN)(hN)$. We know that $g \in gN$ and $h \in hN$ so we conclude that $(gN)(hN)$ is a subset of the element of the partition that contains gh.

So the set of cosets satisfy the definition of P, showing that at least one such partition exists. Now we show that any such partition must be the set of cosets. Consider subsets A, B of G that satisfy the definition of P. We need to show that A and B are cosets of N. Fix an arbitrary element a of A. Since $1 \in N$ and $a1 \in A$ we conclude that the product set AN is equal to A and therefore that A is equal to the coset aN. Similarly, B is a coset of N.

** 2.10.5 Identify the quotient group $\mathbb{R}^\times/P$, where P denotes the subgroup of positive real numbers. **

The cosets of P are the positive and negative reals, $\{P, -1P\}$. We define composition of cosets to be $(xP)(yP) = (xy)P$. The group operation is commutative, therefore P is normal and the operation on cosets is well-defined. Therefore the quotient group is isomorphic to the group of size 2: P is the identity and $(-1P)(-1P) = P$.

** 2.10.6 Let $H = \{\pm 1, \pm i\}$ be the subgroup of $G = \mathbb{C}^\times$ of fourth roots of unity. Describe the cosets of H in G explicitly, and prove that G/H is isomorphic to G. **

The cosets of H in G are sets of the form $zH = \{\pm z, \pm zi\}$, where $z \in \mathbb{C}^\times$. Geometrically, they are the four points of a cross in the complex plane; a rotated and scaled version of the unit cross corresponding to $\{\pm 1, \pm i\}$.

If $z \neq z'$ then $zH \neq z'H$. So there is a bijection between G/H and G. I.e. G/H is isomorphic to G.

** 2.10.10 (122) Describe the quotient groups $\mathbb{C}^\times/P$ and $\mathbb{C}^\times/U$, where U is the subgroup of complex numbers of absolute value 1 and P denotes the positive reals. **

** $\mathbb{C}^\times/P$ **

The cosets of P are the set of radial lines emanating from the origin in the complex plane. So a single coset is $e^{i\theta}P = \{pe^{i\theta}|p \in P\}$ and there is one coset for every value of $\theta \in [0, 2\pi)$. We define the following operation on the set of cosets: $(e^{i\theta_1}P)(e^{i\theta_2}P) = e^{i(\theta_1+\theta_2)}P$. The group operation is commutative, therefore P is normal and the operation is well-defined.

So the quotient group is isomorphic to the group of angles $[0, 2\pi)$ under addition, and to U. This is the "circle group" T.

** $\mathbb{C}^\times/U$ **

The cosets of U are concentric circles in the complex plane. So a single coset is $re^{i\phi}U = rU = \{re^{i\theta}|0 \leq \theta < 2\pi\}$, and there is one coset for every $r \in P$. We define the following operation on the set of cosets: $(r_1U)(r_2U) = r_1r_2U$. The group operation is commutative, therefore U is normal and the operation is well-defined.

So the quotient group is isomorphic to P, the positive reals under multiplication.

3.9.8 E122 Homework 8

** 3.1.1 Which of the following subsets of the vector space of real $n \times n$ matrices is a subspace?**

** (a) symmetric matrices ($A = A^t$) **

This is a subspace: it's closed under addition, contains the identity (zero matrix), and contains additive inverses, and is closed under scalar multiplication.

** (b) invertible matrices **

This is not a subspace, since the additive identity (zero matrix) is not an invertible matrix. (It's closed under scalar multiplication but I'm not sure if it contains additive inverses.)

** (c) upper-triangular matrices **

This is a subspace, since the additive identity (zero matrix) is upper triangular, and it is closed under scalar multiplication and contains additive inverses.

** 3.1.5 Prove that the classification of subspaces of $\mathbb{R}^3$ stated after (1.2) is complete **

> The subspaces of $\mathbb{R}^3$ are of four types > > (i) The zero vector > > (ii) Vectors lying in a line through the origin > > (iii) Vectors lying in a plane through the origin > > (iv) The whole space $\mathbb{R}^3$

Subspaces must contain the identity. So $\{0\}$ is one subspace. Now consider a subspace that contains $x \neq 0$ in addition to 0. In order to be closed under scalar multiplication the subspace must also contain all vectors $x' = cx$ for every $c \in \mathbb{R}$. So that's a line through the origin and x. One such subspace exists for every distinct direction that a line through the origin can take. Next planes...

** 3.2.1 Prove that the set of numbers of the form $a + b\sqrt{2}$, where a, b are rational numbers, is a field. **

A field is by definition a set with an addition and a multiplication operation. For each operation there must be associativity, closure, commutativity, identity, inverses. A distributive law must hold: $(a + b)c = ac + bc$.

More briefly, for a set $\mathbb{F}$ to be a field, there must be an addition law that turns $\mathbb{F}$ into an Abelian group (identity 0), and there must be a multiplication law that turns $\mathbb{F} - \{0\}$ into an Abelian group (identity 1). Distributivity must hold: $(a + b)c = ac + bc$.

Take the **addition** operation first. Closure, commutativity and associativity obviously hold. The identity is $0 = 0 + 0\sqrt{2}$. The inverse is $(a + b\sqrt{2})^{-1} = (-a + -b\sqrt{2})$ (this is unique since the product of a rational and irrational is irrational).

Now **multiplication**.

Closure and commutativity: $(a + b\sqrt{2})(c + d\sqrt{2}) = (c + d\sqrt{2})(a + b\sqrt{2}) = (ac + 2bd) + (cb + ad)\sqrt{2}$.

Associativity:

$$\begin{aligned}
\Big((a + b\sqrt{2})(c + d\sqrt{2})\Big)\,(e + f\sqrt{2}) = \\
\Big((ac + 2bd) + (cb + ad)\sqrt{2}\Big)\,(e + f\sqrt{2}) = \\
\big((ac + 2bd)e + 2(cb + ad)f\big) + \big((cb + ad)e + (ac + 2bd)f\big)\,\sqrt{2} = \\
(ace + 2adf + 2bcf + 2bde) + (acf + ade + bce + 2bdf)\,\sqrt{2} =
\end{aligned}$$

$$\begin{aligned}
(a + b\sqrt{2})\,\Big((c + d\sqrt{2})(e + f\sqrt{2})\Big) = \\
(a + b\sqrt{2})\,\Big((ce + 2df) + (cf + de)\sqrt{2}\Big) = \\
\big(a(ce + 2df) + 2b(cf + de)\big) + \big(a(cf + de) + b(ce + 2df)\big)\,\sqrt{2} = \\
(ace + 2adf + 2bcf + 2bde) + (acf + ade + bce + 2bdf)\,\sqrt{2}
\end{aligned}$$

Identity is $1 + 0\sqrt{2} = 1$.

Inverses: $(a + b\sqrt{2})(a - b\sqrt{2}) = (a^2 - 2b^2) \implies (a + b\sqrt{2})^{-1} = \frac{1}{(a^2 - 2b^2)}(a - b\sqrt{2})$

Distributivity: ...bored now.

** 3.2.7 Define homomorphism of fields, and prove that every homomorphism of fields is injective. **

A homomorphism between fields F and G is a function $f : F \to G$ such that $f(a + b) = f(a) + f(b)$ and $f(ab) = f(a)f(b)$.

Suppose that f is not injective. Then $f(a) = f(b)$ for some $a \neq b$. So $f(a - b) = f(a) + -f(b) = 0$. Then $f((a - b)(a - b)^{-1}) = f(a - b)f((a - b)^{-1}) = 0$ since the product of 0 with anything is 0; and yet $f((a-b)(a-b)^{-1}) = f(1) = 1$. This is a contradiction which proves that homomorphisms must be injective.

(So what about linear maps with non-empty nullspace / non full rank? Presumably they are not homomorphisms?)

** 3.2.15 **

** (a) By pairing elements with their inverses, prove that the product of all nonzero elements of $\mathbb{F}_p$ is -1. **

$$\prod aa^{-1}bb^{-1}...$$

** (b) Let p be a prime integer. Prove Wilson's Theorem: $(p - 1)! \equiv -1 \pmod{p}$. **

Chapter 4

Linear Algebra

4.1 Examples of vector spaces

1. The set $\mathbb{R}^n$ of $n-$tuples of real numbers, under componentwise addition and componentwise multiplication by real scalars.
2. Complex numbers
   (a) $\mathbb{C}$ under addition with multiplication by scalars from $\mathbb{C}$ is a field, and therefore a vector space. It is one-dimensional (1 and i are not linearly independent).
   (b) The set $\mathbb{C}^n$ of $n-$tuples of complex numbers, under componentwise addition and componentwise multiplication by complex numbers.
   (c) $\mathbb{C}$ under addition with multiplication by real scalars is equivalent to $\mathbb{R}^2$.
3. Matrices & linear transformations:
   (a) The set $M_{m\times n}(\mathbb{R})$ of $m \times n$ matrices is a vector space, under componentwise addition and multiplication by real scalars.
   (b) The set $\mathrm{Hom}(V, W)$ of linear transformations from vector space V to vector space W is a vector space: for scalar a, define $(aT)(v) := a(T(v))$, and $(S + T)v := S(v) + T(v)$.
4. The set $\mathbb{R}_n[x]$ of polynomials of degree $\leq n$ is a real vector space.
5. The set $\mathbb{R}^X$ of real-valued functions on any set X is a real vector space. Examples:
   (a) Let $[n] = \{1, 2, \ldots, n\}$. Note that the function space $\mathbb{R}^{[n]}$ is the same as $\mathbb{R}^n$ (both are sets of $n-$tuples of reals).
   (b) Similarly, $\mathbb{R}^{[m]\times[n]}$ is the same as $M_{m\times n}(\mathbb{R})$.
   (c) $\mathbb{R}^{\mathbb{R}}$, the set of all functions $\mathbb{R} \to \mathbb{R}$.
   (d) The set of continuous functions $\mathbb{R} \to \mathbb{R}$, and differentiable functions $\mathbb{R} \to \mathbb{R}$, under pointwise addition and pointwise scalar multiplication (from any field?).
   (e) Set of solutions of a homogeneous linear ODE
6. Sequences (a_n) of real numbers, under term-wise addition and term-wise scalar multiplication, form a vector space, identifiable with the function space $\mathbb{R}^{\mathbb{N}}$. Examples:
   (a) Set of convergent sequences
7. The set of solutions of a system of *homogeneous* linear equations in n variables is a subspace V of $\mathbb{R}^n$. (Let A be the matrix representing the system and let u and v be solutions. Then $Au = Av = 0$ and V is a subspace since $A(u + v) = Au + Av = 0$, and $A(\lambda u) = \lambda Au = 0$.)

4.2 Linear systems

Consider the linear systems

$$\begin{cases} x = 0 \\ y = 0 \end{cases} \qquad\qquad \begin{cases} x - y = 0 \\ x + y = 1 \\ x - z = 0 \end{cases}$$

A "solution" is an assignment of values to the n variables which makes all m equations true.

In other words, we notice that the equations involve n variables, and consider the set of n-tuples $S = \{(x, y, \ldots) \mid x, y, \ldots \in \mathbb{R}\}$.

The set of solutions is the subset of S for which all the equations are true.

Geometrically, we think of the 2-tuple (a, b) as a point in the $\mathbb{R}^2$ plane. Specifically, if our basis is $\mathbf{e}_1, \mathbf{e}_2$, then (a, b) is the point $a\mathbf{e}_1 + b\mathbf{e}_2$. We might imagine that the basis is the standard orthogonal basis, but that's not necessary.

The linear equations define hyperplanes (lines, planes etc) in S.

The set of solutions is the intersection of these hyperplanes: another hyperplane or the empty set.

So at this point, we do not treat the ambient space as a vector space (we're not adding or scaling points), and neither the equation hyperplanes nor the solution hyperplane, need be a subspace (since it need not contain the origin).

Next, we rewrite the linear system as a matrix applied to a vector, $Ax = b$:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

The equation coefficients are now represented by a linear transformation $A : \mathbb{R}^n \to \mathbb{R}^m$.

This matrix equation is saying:

1. Let the x coefficients be a vector $\mathbf{a}_1 \in \mathbb{R}^m$. And let the y coefficients be another vector $\mathbf{a}_2 \in \mathbb{R}^m$, and so on.
2. So now you have n vectors spanning some subspace of $\mathbb{R}^m$.
3. Is b in their span? If so, for what values of $x, y, \ldots$ does $x\mathbf{a}_1 + y\mathbf{a}_2 + \cdots = b$?

From Frenkel's Multivariable Calculus lectures:

> *The dimensionality of an object is equal to the dimensionality of the ambient space, minus the number of independent equations.*

So, basically, suppose there are n variables. Then the solution set is a subset (hyperplane) of $\mathbb{R}^n$, and

Independent equations	Solution set
1	$(n-1)$-dimensional hyperplane
2	$(n-2)$-dimensional hyperplane
$\vdots$	$\vdots$
$n-1$	line
n	point
$n+1$	impossible
$\vdots$	$\vdots$

So when do we get no solutions? That's when

$$\begin{aligned} &\text{the } n \text{ columns of } A \text{ do not span } \mathbb{R}^m \\ \iff &\operatorname{Rank} A < (\text{number of equations}) \\ \iff &\text{not all equations independent,} \end{aligned}$$

and b is not in their span.

In other words, suppose we have a linear system involving n variables.

Suppose that all the m equations are independent: full row rank.

Then $m \leq n$.

Now we introduce a dependent equation into the system.

One error above is that its only the coefficients of the equation that we're considering when we say the rows are dependent/independent. So it's not correct to talk about "independent equations".

4.3 Subspaces

A subspace U of V is a subset of V for which

1. $0 \in U$
2. For any finite subset $U^* \subset U$, the set of all linear combinations of U^* is also a subset of U.

4.4 Span, basis, dimension

Theorem 8. *Every basis has the same size.*

Proof. Let $v_1, \ldots, v_n$ be a basis for a vector space V.

□

Theorem 9. *A spanning set that is the same size as a basis is also a basis.*

Proof. Let $v_1, \ldots, v_n$ be a basis for a vector space V, and let $u_1, \ldots, u_n$ span V.

We know that $v_1, \ldots, v_n$ are linearly independent and that if we remove any one of them they will cease to span.

We want to show that $u_1, \ldots, u_n$ are linearly independent.

Suppose, that the u_i are not linearly independent and that $u_2, \ldots, u_n$ span V. Thus there are $n-1$ vectors in this spanning set. But the Steinitz Exchange Lemma states that if $v_1, \ldots, v_n$ are linearly independent and $u_1, \ldots, u_m$ span, then $n \leq m$. This contradiction proves that the u_i are linearly independent. □

Theorem 10. *Let U, V be vector spaces, let $f : U \to V$ be an invertible linear map, and let $e_1, \ldots, e_n$ be a basis for U. Then $f(e_1), \ldots, f(e_n)$ is a basis for V.*

Proof. We need to show that the $f(e_i)$ are linearly independent and spanning.

1. **Linear independence**
   Suppose $\sum_{i=1}^{n} \lambda_i f(e_i) = 0$.

   Therefore $f\left(\sum_{i=1}^{n} \lambda_i e_i\right) = 0$ since f is linear.

   Therefore $\sum_{i=1}^{n} \lambda_i e_i = f^{-1}(0) = 0$, since the preimage of 0 is $\{0\}$ for an invertible linear map.

   But the e_i are linearly independent, therefore $\lambda_i = 0$ for all $i = 1, \ldots n$, as required.

2. **Spanning**
   Let $v \in V$.

   Then $v = f(u)$ for some $u \in U$, since f is surjective.

   Therefore $v = f\left(\sum_{i=1}^{n} \lambda_i e_i\right) = \sum_{i=1}^{n} \lambda_i f(e_i)$ for some $\lambda_1, \ldots, \lambda_n$, as required.

□

4.5 Linear transformations and matrices

A linear transformation is completely specified by

1. Some basis vectors i and j
2. Where those basis vectors are taken to by the transformation.

How the transformation affects any other point follows from those two pieces of information.

So i might be taken to $ai + bj$, and j might be taken to $ci + dj$. In this case we would use the following matrix to describe the transformation:

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

Some examples are

stretch by a in the i-direction	$\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}$
stretch by a in the i-direction and shear right	$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$
rotate anticlockwise 90°	$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$

Note that we haven't said what i and j are yet; they *define* the 2-dimensional space that we're considering. But, we can think of them for now as the usual orthogonal unit vectors in 2D space.

So the matrix tells us where the basis vectors have been taken to. Any other vector $fi + gj$ is taken to wherever that is using the transformed basis vectors:

$$fi + gj \longrightarrow f\begin{bmatrix} a \\ b \end{bmatrix} + g\begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} fa + gc \\ fb + gd \end{bmatrix}$$

And that's how matrix multiplication is defined:

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{bmatrix} f \\ g \end{bmatrix} = \begin{bmatrix} fa + gc \\ fb + gd \end{bmatrix}$$

A matrix represents a linear transformation by showing where the basis vector are taken to.

Theorem 11. *The inverse of a 2x2 matrix is...*

4.6 Geometric interpretation of matrix operations

https://math.stackexchange.com/questions/37398/what-is-the-geometric-interpretation-of-the-transpose
https://math.stackexchange.com/questions/598258/determinant-of-transpose/636198#636198

4.7 Commutativity

4.7.1 Examples of transformations that don't commute

Let A be reflection around the first coordinate axis $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and let B be 90° anticlockwise rotation $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

Then $BA = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \neq AB = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$.

Note that $A^{-1} = A = A^T$ and $B^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = B^T$.

Therefore these are both orthogonal (unitary) matrices.

4.8 Eigenvalues, eigenvectors, characteristic polynomial

Let V be a vector space and let $T : V \to V$ be a linear transformation.

Definition (eigenvalue)**.** *λ is an eigenvalue of T iff there exists $v \in V$ such that $Tv = \lambda v$.*

Definition (eigenspace)**.** *$E_\lambda = \{v \mid Tv = \lambda v\}$ is an eigenspace of T.*

Definition (eigenvector)**.** *An eigenvector is a non-zero element of an eigenspace.*

Definition (characteristic polynomial)**.** *The characteristic polynomial of T is $\chi_T(x) = \det(T - xI)$. Note that λ is an eigenvalue of T iff $x = \lambda$ is a root of $\chi_T(x)$.*

Intuition.
Decompose T as the sum of two transformations: $T = \lambda I + T^*$. This means that the effect of applying T to a vector is the same as applying λI to the vector, and separately applying T^* to the same vector, and adding the two results.

Note that applying λI to a vector just scales the vector by λ.

Note that $T^* = T - \lambda I$.

Therefore what $T - \lambda I$ does to a vector is: whatever remains to be done after scaling by λ, in order to have the same effect as T.

Suppose λ is an eigenvalue. Then there exists an eigenspace E_λ (a line, at least) containing vectors which are simply stretched by a factor λ. So for $v \in E_\lambda$, nothing remains to be done after scaling by λ, and so we have $(T - \lambda I)(v) = 0$.

Therefore

- If $T - \lambda I$ has a nullspace containing a non-zero element, then λ is an eigenvalue and the nullspace is the eigenspace for λ.
- The roots of $\det(T - xI)$ are the eigenvalues of T.

Remark (Repeated eigenvalues).
If two eigenvectors share the same eigenvalue then they are in the same eigenspace.

Proof. Suppose that $Tv_1 = \lambda v_1$ and $Tv_2 = \lambda v_2$ and $v_1 \neq v_2$, and let a be a scalar.

Then $T(v_1 + av_2) = T(v_1) + aT(v_2) = \lambda v_1 + a\lambda v_2 = \lambda(v_1 + av_2)$. □

4.9 Change of basis

Suppose person B uses some other basis vectors to describe locations in space. Specifically, in our coordinates, their basis vectors are $\begin{bmatrix}2\\1\end{bmatrix}$ and $\begin{bmatrix}-1\\1\end{bmatrix}$.

When they state a vector, what is it in our coordinates?

If they say $\begin{bmatrix}-1\\2\end{bmatrix}$, what is that in our coordinates?

Well, if they say $\begin{bmatrix}1\\0\end{bmatrix}$, that's $\begin{bmatrix}2\\1\end{bmatrix}$ in our coordinates. And if they say $\begin{bmatrix}0\\1\end{bmatrix}$, that's $\begin{bmatrix}-1\\1\end{bmatrix}$ in our coordinates. So the matrix containing *their basis vectors expressed using our coordinate system* transforms a point expressed in their coordinate system into one expressed in ours. That last sentence is critical, so hopefully it makes sense! So, the answer is

$$\begin{pmatrix}2 & -1\\ 1 & 1\end{pmatrix}\begin{bmatrix}-1\\2\end{bmatrix} = \begin{bmatrix}-4\\1\end{bmatrix}.$$

When we state a vector, what is it in their coordinates?

We give the vector $\begin{bmatrix}3\\2\end{bmatrix}$. What is that in their coordinate system? By definition, the answer is the weights that scales their basis vectors to hit $\begin{bmatrix}3\\2\end{bmatrix}$. So, the solution to

$$\begin{pmatrix}2 & -1\\ 1 & 1\end{pmatrix}\begin{bmatrix}a\\b\end{bmatrix} = \begin{bmatrix}3\\2\end{bmatrix}.$$

Computationally, we can see that we can get the solution by multiplying both sides by the inverse:

$$\begin{bmatrix}a\\b\end{bmatrix} = \begin{pmatrix}2 & -1\\ 1 & 1\end{pmatrix}^{-1}\begin{bmatrix}3\\2\end{bmatrix}.$$

Conceptually, we have

$$\begin{pmatrix}2 & -1\\ 1 & 1\end{pmatrix} = \begin{pmatrix}\text{matrix converting their}\\ \text{representation to ours}\end{pmatrix}$$

where "their representation" means the vector expressed using their coordinate system. So the role played by the inverse is

$$\begin{bmatrix}a\\b\end{bmatrix} = \begin{pmatrix}\text{matrix converting our}\\ \text{representation to theirs}\end{pmatrix}\begin{bmatrix}3\\2\end{bmatrix}.$$

When we state a transformation, what is it in their coordinates?

We state a 90° anticlockwise rotation of 2D space:

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

what is that transformation in their coordinates? The answer is

$$\begin{pmatrix} \text{matrix converting our} \\ \text{representation to theirs} \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \text{matrix converting their} \\ \text{representation to ours} \end{pmatrix}$$

since the composition of those three transformations defines a single transformation that takes in a vector expressed in their coordinate system, converts it to our coordinate system, transforms it as requested, and then converts back to theirs.

Let

$$P = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}$$

be the change-of-basis matrix . Then the matrix, in their coordinates, of the rotation transformation is

$$P^{-1} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} P.$$

What about the uniform stretch transformation? In our coordinates this has matrix $\lambda I = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$. In their coordinates, it has matrix

$$P^{-1}\lambda I P = \lambda P^{-1} P = \lambda I.$$

I.e. a uniform stretch transformation represented by a diagonal matrix has the same matrix in any basis. That's because – forget about introducing any basis – there is only one "uniform stretch transformation": it's the transformation that acts on space like it's a balloon being inflated uniformly. Whatever basis vectors you choose, each one $\mathbf{e}_i$ is going to be taken to $\lambda\mathbf{e}_i$. That means the matrix of the transformation, in whatever basis, is $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$, because the vector

"one unit in the $\mathbf{e}_1$ direction, zero units in the $\mathbf{e}_2$ direction"

is going to be taken to the vector

"λ units in the $\mathbf{e}_1$ direction, zero units in the $\mathbf{e}_2$ direction".

What about a non-uniform stretch transformation?

Consider $\mathbb{R}^2$. Fix a first basis vector e_1.

Consider the map $T : \mathbb{R}^2 \to \mathbb{R}^2$ which stretches space by a factor of 2 in the direction of e_1, and by a factor of 3 in the orthogonal direction.

Suppose that the second basis vector e_2 is orthogonal to e_1 and has the same magnitude.

Then the matrix of T is $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ with respect to this basis.

Now consider an alternative basis $\{f_1, f_2\}$ where f_2 intersects with f_1 at $45°$.

Specifically, with respect to basis $\{e_1, e_2\}$, we have $f_1 = (1, 0)$ and $f_2 = (1, 1)$.

Then the matrix of T with respect to basis $\{f_1, f_2\}$ is

$$\begin{aligned} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} &= \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 0 & 3 \end{pmatrix} \\ &= \begin{pmatrix} 2 & -1 \\ 0 & 3 \end{pmatrix}. \end{aligned}$$

(It's obvious that $f_1 \mapsto (2, 0)$; that $f_2 \mapsto (-1, 3)$ is clear in a diagram.)

The eigenvalues of T are clearly 2 and 3, independent of basis.

The eigenspaces are the line through e_1, and the line through e_2.

So with respect to basis $\{e_1, e_2\}$, the eigenspaces are $\{(a, 0) \mid a \in \mathbb{R}\}$ and $\{(0, a) \mid a \in \mathbb{R}\}$.

And with respect to basis $\{f_1, f_2\}$, the eigenspaces are $\{(a, 0) \mid a \in \mathbb{R}\}$ and $\{(-a, a) \mid a \in \mathbb{R}\}$.

Consider the map which stretches space by a factor of 2 in one direction, and a factor of 3 in another direction.

Then there exists a basis for which the map has matrix $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$.

What are the eigenspaces of this map?

The characteristic polynomial (basis independent) is $\det(A - \lambda I) = 0$ where A is the matrix of the map wrt some basis.

4.10 Symmetric matrices

Spectral theorem for symmetric matrices

Symmetric $n \times n$ matrix A (real).

$A^{-1} = A^{\mathbf{T}}$

n orthogonal eigenvectors with real eigenvalues.

Orthonormal matrix U containing normalized eigenvectors.

$A = U\Lambda U^{-1} = U\Lambda U^{\mathbf{T}}$

(Eigenvalues are uniquely determined by matrix. Eigenvalues can be repeated, in which case any linear combination of their eigenvalues is also an eigenvalue.)

4.11 Inner Product Spaces

Note that if $f(\cdot)$ is linear:

1. $f(ax + by) = f(ax) + f(by)$.

Definition (Bilinear form)**.**
A bilinear form is a binary function $f(\cdot, \cdot)$ such that:

1. $f(ax + by, z) = f(ax, z) + f(by, z)$
2. $f(z, ax + by) = f(z, ax) + f(z, by)$.

Claim. *The dot product in $\mathbb{F}^n$ is bilinear.*

Proof.

$$\begin{aligned}\langle ax + by, z\rangle :=& \sum_i (ax + by)_i z_i \\ =& \sum_i (ax_i + by_i) z_i \\ =& \sum_i ax_i z_i + \sum_i by_i z_i \\ =& \langle ax, z\rangle + \langle by, z\rangle \\ \langle z, ax + by\rangle :=& \dots\end{aligned}$$

□

Note that $\langle x, y\rangle = x \cdot y = x^T y = x^T I y$.

And note that the "quadratic form" $ax^2 + 2bxy + cy^2$ can be written as

$$\mathbf{x^T} A\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}^{\mathbf{T}} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

This is a scalar. In general, a quadratic form for symmetric matric A is

$$\mathbf{x^T} A\mathbf{y} = \sum_{jk} A_{jk} x_j y_k.$$

These quadratic forms are also bilinear forms: the dot product is a quadratic form using the identity matrix.

Definition (Gram matrix)**.** *Take a collection of vectors $v_1, \ldots, v_n$. A Gram matrix is the $n \times n$ matrix $(\langle v_i, v_j \rangle)$.*

Theorem. *Every bilinear form is of the form $\langle u, v \rangle = u^T A v$ for some Gram matrix.*

Definition. *A bilinear form is symmetric if $\langle u, v \rangle = \langle v, u \rangle$.*

Theorem. *The bilinear form $\langle u, v \rangle := u^T A v$ is symmetric if and only if A is symmetric.*

Definition. *A bilinear form is positive definite if $\langle u, v \rangle > 0$ for all $v \in V \setminus \{0\}$.*

Definition (Inner product)**.**
An inner product is a bilinear form that is symmetric and positive definite.

An inner product space is a vector space equipped with an inner product.

In an abstract inner product space we define the angle between u and v to be $\cos^{-1}\left(\frac{\langle u,v\rangle}{\|u\|\|v\|}\right)$.

In a real inner product space we define the norm to be $\|u\| := \sqrt{\langle u, u \rangle}$.

Theorem (Cauchy-Schwartz inequality)**.**
Let V be an inner product space and let $u, v \in V$. Then $\langle u, v \rangle \leq \|u\|\|v\|$.

Proof. Define $f(t) := \langle tu + v, tu + v \rangle = \|tu + v\|^2$.

Use bilinearity and symmetry to show that $f(t) = t^2\langle u, u \rangle + 2t\langle u, v \rangle + \langle v, v \rangle$. (How?)

The Cauchy-Schwartz inequality follows by noting that the determinant of this quadratic must be negative. □

4.12 Complex vector spaces

When viewed as a real vector space (i.e. with real scalars), $\mathbb{C}$ is two-dimensional, e.g. $\{1, i\}$ is a basis.

When viewed as a complex vector space (i.e. with complex scalars), $\mathbb{C}$ is one-dimensional: $\{1\}$ is a basis; $\{1, i\}$ are no longer linearly independent.

Definition. *Let V be a complex vector space.*

$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$ is a sesquilinear form if

1. $\langle au + bv, z \rangle = a\langle u, z \rangle + b\langle v, z \rangle$
2. $\overline{\langle u, u \rangle} = \langle u, u \rangle$ *(therefore $\langle u, u \rangle \in \mathbb{R}$).*

Definition (Hermitian space)**.**
Let V be a complex vector space (i.e. complex scalars).

A Hermitian form is a sesquilinear form that is symmetric and positive definite.

A complex inner product space, or Hermitian space, is a complex space equipped with a Hermitian form as an inner product.

[0] Essence of Linear Algebra video series by Grant Sanderson / 3blue1brown

4.13 Finding the nth Fibonacci number via an eigenvector change of basis

This is the problem given at the end of the eigenvectors video in the Essence of Linear Algebra[1] series by 3blue1brown[2].

Introduction

Consider the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

The first few powers are

$$\begin{aligned} A^1 &= \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} &&= \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \\ A^2 &= \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} &&= \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \\ A^3 &= \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} &&= \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} \\ A^4 &= \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} &&= \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix} \end{aligned}$$

The Fibonacci sequence is the sequence you get by starting with $0, 1$ and after that always forming the next number by adding the two previous ones: $F_0, F_1, F_2, F_3, F_4, F_5, F_6, F_7, ... = 0, 1, 1, 2, 3, 5, 8, 13, ...$

The matrix powers are generating the Fibonacci sequence:

$$A^n = \begin{pmatrix} F_{n-1} & F_n \\ F_n & F_{n+1} \end{pmatrix}$$

So if there were a way to compute the n^{th} power of that matrix "directly", that would also be a way to compute the n^{th} Fibonacci number "directly", i.e. without computing all the preceding Fibonacci numbers *en route*.

How can we do this? To state the problem in a different way, we need to construct a new matrix that performs exactly the same transformation as A^n, but which somehow does the exponentiation step "in one go" rather than by multiplying A with itself n times.

Solution outline

Matrices represent transformations, so we can talk about them as taking in some vector and producing some other vector. The approach we're going to take is to re-express the A^n transformation as follows:

1. Convert the input vector to its representation in an alternative basis which uses the eigenvectors as the basis vectors (it's called an "eigenbasis").

[1] https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab
[2] http://www.3blue1brown.com/

2. In this alternative basis, compute the new position of the vector after carrying out the A^n transformation.
3. Convert the resulting vector back to its representation in our original basis.

I.e., we're going to compute the overall transformation as this product of matrices (remember that one reads these things right-to-left):

$$\begin{pmatrix}\text{matrix converting their}\\ \text{representation to ours}\end{pmatrix}\begin{pmatrix}\text{matrix that does the A transformation}\\ \text{in the alternative basis}\end{pmatrix}^n\begin{pmatrix}\text{matrix converting our}\\ \text{representation to theirs}\end{pmatrix}$$

The crux of all this is that the exponentiation is efficient in the eigenbasis. That's because, in the eigenbasis, the transformation is just stretching space in the directions of the two basis vectors. So to do the transformation n times in the eigenbasis, you just stretch by the stretch-factor raised to the n^{th} power, rather than doing n matrix multiplications.

Solution details

Let's suppose we've already found the eigenvectors, and that there are two of them, and that we've arranged them as the two columns of a matrix V. V holds the basis vectors of the alternative basis, and therefore we know from the [change of basis](./linear-algebra.html#change-of-basis) notes that V is the matrix that takes as input a vector expressed in the alternative basis and outputs its representation in our basis.

So, step (3) is done by V, and step (1) is done by V^{-1}, and the matrix performing all three steps is going to look like

$$V\begin{pmatrix}\text{matrix that does the A transformation}\\ \text{in the alternative basis}\end{pmatrix}^n V^{-1}$$

OK, so what is the matrix in the middle? The [change of basis](./linear-algebra.html#change-of-basis) notes tell us that we can compute it as

$$\begin{pmatrix}\text{matrix converting our}\\ \text{representation to theirs}\end{pmatrix}A\begin{pmatrix}\text{matrix converting their}\\ \text{representation to ours}\end{pmatrix}$$

In other words the matrix in the middle is

$$V^{-1}AV$$

and the entire transformation is

$$V\left(V^{-1}AV\right)^nV^{-1}$$

Put back into words, that's

$$\begin{pmatrix}\text{matrix converting their}\\ \text{representation to ours}\end{pmatrix}\left(\begin{pmatrix}\text{matrix converting our}\\ \text{representation to theirs}\end{pmatrix}A\begin{pmatrix}\text{matrix converting their}\\ \text{representation to ours}\end{pmatrix}\right)^n\begin{pmatrix}\text{matrix converting our}\\ \text{representation to theirs}\end{pmatrix}$$

Recall that above we observed that the n^{th} power of A is a matrix with the nth Fibonacci number in its bottom left and top right entries. So the following tasks remain:

1. Find the eigenvectors and put them in a matrix V.
2. Find the inverse of V.
3. Compute the matrix product $V^{-1}AV$.

4. Compute the result of raising that to the n^{th} power.
5. Plug the result of that into the overall expression.
6. Take the entry in the bottom left or top right (they should be the same!).

The result should be an expression giving the n^{th} Fibonacci number as a function of n. It should be possible to give as input to that function the number one million, and have it output the one millionth Fibonacci number directly, without it having to go through the preceding 999,999 Fibonacci numbers.

The answer without showing the calculations

The eigenvectors are

$$V = \begin{pmatrix} 2 & 2 \\ 1+\sqrt{5} & 1-\sqrt{5} \end{pmatrix}$$

which has inverse

$$V^{-1} = \frac{-1}{4\sqrt{5}} \begin{pmatrix} 1-\sqrt{5} & -2 \\ -1-\sqrt{5} & 2 \end{pmatrix}$$

Therefore

$$V^{-1}AV = \frac{1}{2} \begin{pmatrix} 1+\sqrt{5} & 0 \\ 0 & 1-\sqrt{5} \end{pmatrix}$$

and

$$(V^{-1}AV)^n = \frac{1}{2^n} \begin{pmatrix} (1+\sqrt{5})^n & 0 \\ 0 & (1-\sqrt{5})^n \end{pmatrix}$$

and

$$V\left(V^{-1}AV\right)^n V^{-1} = \begin{pmatrix} \frac{\left((1+\sqrt{5})^{n-1}-(1-\sqrt{5})^{n-1}\right)}{2^{n-1}\sqrt{5}} & \frac{\left((1+\sqrt{5})^{n}-(1-\sqrt{5})^{n}\right)}{2^{n}\sqrt{5}} \\ \frac{\left((1+\sqrt{5})^{n}-(1-\sqrt{5})^{n}\right)}{2^{n}\sqrt{5}} & \frac{\left((1+\sqrt{5})^{n+1}-(1-\sqrt{5})^{n+1}\right)}{2^{n+1}\sqrt{5}} \end{pmatrix}$$

Therefore the nth Fibonacci number is

$$F_n = \frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n\sqrt{5}}$$

Does this actually work?

Yes.

```
from math import sqrt

def fib(n):
    return (
        ( (1 + sqrt(5))**n - (1 - sqrt(5))**n )
        /
        float(2**n * sqrt(5)))

for i in range(10):
    print(i, fib(i))
```

```
0 0.0
1 1.0
2 1.0
3 2.0
4 3.0
5 5.0
6 8.0
7 13.0
8 21.0
9 34.0
```

History

The formula is known as Binet's formula (https://en.wikipedia.org/wiki/Fibonacci_number#Closed-form_expression) (1843) but was apparently known to Euler, Daniel Bernoulli and de Moivre more than a century earlier. It can be derived without using linear algebra techniques; I don't know when the style of proof attempted here would first have been done. The result can be written as

$$F_n = \frac{\phi^n - (1-\phi)^n}{\sqrt{5}}$$

where $\phi = \frac{1+\sqrt{5}}{2}$ is the golden ratio (https://en.wikipedia.org/wiki/Golden_ratio).

Calculations

1. Find the eigenvectors

We follow the textbook approach: We have

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

An eigenvector v satisfies $Av = \lambda v$ for some scalar λ. That equation can be rearranged as follows

$$\begin{aligned} A\mathbf{v} &= \lambda I\mathbf{v} \\ A\mathbf{v} - \lambda I\mathbf{v} &= \mathbf{0} \\ (A - \lambda I)\mathbf{v} &= \mathbf{0} \end{aligned}$$

which means that the matrix $A - \lambda I$ is a transformation that takes some non-zero vector $\mathbf{v}$ to the zero vector (i.e. it has a non-empty "null space"). This means that the transformation cannot be reversed, i.e. the matrix has no inverse, i.e. its determinant is zero. So, use that last fact to find the eigenvectors λ:

$$\det(A - \lambda I) = 0$$

$$\det \begin{pmatrix} -\lambda & 1 \\ 1 & 1-\lambda \end{pmatrix} = 0$$

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

Using the quadratic formula we have $a = 1, b = -1, c = -1$ and

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{1 \pm \sqrt{5}}{2}$$

which are the two eigenvalues.

To find eigenvectors associated with the eigenvalues, go back to the equations

$$(A - \lambda I)\mathbf{v} = \mathbf{0}$$

$$\begin{pmatrix} -\lambda & 1 \\ 1 & 1-\lambda \end{pmatrix} \mathbf{v} = \mathbf{0}$$

Let an eigenvector v be $\left[\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right]$. The matrix equation corresponds to this system of equations:

$$\begin{cases} -\lambda v_1 + v_2 & = 0 \\ v_1 + (1-\lambda)v_2 & = 0 \end{cases}$$

From the first equation we have $v_2 = \lambda v_1$. There are infinitely many eigenvectors (a line of them) associated with any given eigenvalue, so we can pick an arbitrary value for v_1. If we choose $v_1 = 2$ then we have eigenvectors $\left[\begin{smallmatrix} 2 \\ 1+\sqrt{5} \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 2 \\ 1-\sqrt{5} \end{smallmatrix}\right]$. The matrix containing the eigenvectors is

$$V = \begin{pmatrix} 2 & 2 \\ 1+\sqrt{5} & 1-\sqrt{5} \end{pmatrix}$$

2. Find inverse of V

The inverse of a 2x2 matrix is given by

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}^{-1} = \frac{1}{\det} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$$

where $\det = ad - cb$. Therefore

$$V^{-1} = \frac{1}{2(1-\sqrt{5}) - 2(1+\sqrt{5})} \begin{pmatrix} 1-\sqrt{5} & -2 \\ -(1+\sqrt{5}) & 2 \end{pmatrix}$$

$$= \frac{-1}{4\sqrt{5}} \begin{pmatrix} 1-\sqrt{5} & -2 \\ -(1+\sqrt{5}) & 2 \end{pmatrix}$$

3. Find the matrix product $V^{-1}AV$

Before we get lost in the calculation, let's remember what this is. It's a matrix that does the A transformation, but *in the coordinate system defined by A's eigenvectors*. So, the resulting matrix *must* do nothing other than stretch space in the direction of one or both basis vectors in that coordinate system. That's because (1) we represent a transformation with a matrix saying where each of the basis vectors are taken to, (2) the definition of an eigenvector of a transformation is that it is a vector which is simply stretched by the transformation with no change in direction, therefore (3) if the eigenvectors are the basis vectors, then the matrix representing the transformation must just stretch space in the two directions. A matrix which stretches space in the direction of the basis vectors looks like $\left(\begin{smallmatrix} a & 0 \\ 0 & b \end{smallmatrix}\right)$, i.e. it is diagonal. Therefore, $V^{-1}AV$ *must* be diagonal.

$$V^{-1}AV = \frac{-1}{4\sqrt{5}} \begin{pmatrix} 1-\sqrt{5} & -2 \\ -(1+\sqrt{5}) & 2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 1+\sqrt{5} & 1-\sqrt{5} \end{pmatrix}$$

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

$$= \frac{-1}{4\sqrt{5}} \begin{pmatrix} -4-2(3+\sqrt{5}) & 6-2\sqrt{5}-2(3-\sqrt{5}) \\ -(6+2\sqrt{5})+2(3+\sqrt{5}) & 4+2(3-\sqrt{5}) \end{pmatrix}$$

$$= \frac{-1}{2\sqrt{5}} \begin{pmatrix} -2-3-\sqrt{5} & 3-\sqrt{5}-3+\sqrt{5} \\ -3-\sqrt{5}+3+\sqrt{5} & 2+3-\sqrt{5} \end{pmatrix}$$

$$= \frac{-1}{2\sqrt{5}} \begin{pmatrix} -5-\sqrt{5} & 0 \\ 0 & 5-\sqrt{5} \end{pmatrix}$$

$$= \frac{1}{2} \begin{pmatrix} 1+\sqrt{5} & 0 \\ 0 & 1-\sqrt{5} \end{pmatrix}$$

4. Compute $(V^{-1}AV)^n$

The matrix is diagonal so this is straightforward. Note that this is the whole point of converting to the eigenbasis: the exponentiation at this step just involves the usual operations of raising scalar numbers to a power; no need to multiply matrices together. A computer will be able to compute the n^{th} power of a diagonal matrix much faster than that of a non-diagonal matrix.

$$(V^{-1}AV)^n = \frac{1}{2^n}\begin{pmatrix}(1+\sqrt{5})^n & 0\\ 0 & (1-\sqrt{5})^n\end{pmatrix}$$

5. Plug the n^{th} power into the overall expression

$$V\Big(V^{-1}AV\Big)^nV^{-1} = \frac{-1}{4\sqrt{5}}\frac{1}{2^n}\begin{pmatrix}2 & 2\\ 1+\sqrt{5} & 1-\sqrt{5}\end{pmatrix}\begin{pmatrix}(1+\sqrt{5})^n & 0\\ 0 & (1-\sqrt{5})^n\end{pmatrix}\begin{pmatrix}1-\sqrt{5} & -2\\ -(1+\sqrt{5}) & 2\end{pmatrix}$$

$$= \frac{-1}{4\sqrt{5}}\frac{1}{2^n}\begin{pmatrix}2 & 2\\ 1+\sqrt{5} & 1-\sqrt{5}\end{pmatrix}\begin{pmatrix}(1-\sqrt{5})(1+\sqrt{5})^n & -2(1+\sqrt{5})^n\\ -(1+\sqrt{5})(1-\sqrt{5})^n & 2(1-\sqrt{5})^n\end{pmatrix}$$

$$= \frac{-1}{4\sqrt{5}}\frac{1}{2^n}\begin{pmatrix}2(-4)\big((1+\sqrt{5})^{n-1}-(1-\sqrt{5})^{n-1}\big) & -4\big((1+\sqrt{5})^n-(1-\sqrt{5})^n\big)\\ -4\big((1+\sqrt{5})^n-(1-\sqrt{5})^n\big) & -2\big((1+\sqrt{5})^{n+1}-(1-\sqrt{5})^{n+1}\big)\end{pmatrix}$$

$$= \frac{1}{4\sqrt{5}}\begin{pmatrix}4\frac{\left((1+\sqrt{5})^{n-1}-(1-\sqrt{5})^{n-1}\right)}{2^{n-1}} & 4\frac{\left((1+\sqrt{5})^{n}-(1-\sqrt{5})^{n}\right)}{2^{n}}\\ 4\frac{\left((1+\sqrt{5})^{n}-(1-\sqrt{5})^{n}\right)}{2^{n}} & \frac{\left((1+\sqrt{5})^{n+1}-(1-\sqrt{5})^{n+1}\right)}{2^{n-1}}\end{pmatrix}$$

$$= \begin{pmatrix}\frac{\left((1+\sqrt{5})^{n-1}-(1-\sqrt{5})^{n-1}\right)}{2^{n-1}\sqrt{5}} & \frac{\left((1+\sqrt{5})^{n}-(1-\sqrt{5})^{n}\right)}{2^{n}\sqrt{5}}\\ \frac{\left((1+\sqrt{5})^{n}-(1-\sqrt{5})^{n}\right)}{2^{n}\sqrt{5}} & \frac{\left((1+\sqrt{5})^{n+1}-(1-\sqrt{5})^{n+1}\right)}{2^{n+1}\sqrt{5}}\end{pmatrix}$$

4.14 Polynomials, rings, minimal and characteristic polynomials

Let $f(x) \in \mathbb{F}[x]$ be a polynomial: $f(x) = a_k x^k + \ldots + a_0$.

Let $A \in M_n(\mathbb{F})$ be an $n \times n$ matrix over a field $\mathbb{F}$.

We can evaluate the polynomial on the matrix: $f(A) = a_k A^k + \ldots + a_0 I$.

Theorem. *For all $A \in M_n(\mathbb{F})$, there exists $f(x) \in \mathbb{F}[x]$ such that $f(A) = 0$.*

Proof. Note that $\dim M_n(\mathbb{F}) = n^2$.[3]

Let $k > n^2$. Then $A^k, A^{k-1}, \ldots, I$ is linearly dependent. Therefore there exists a k-th degree polynomial $f(x) \in \mathbb{F}[x]$ such that $f(A) = 0$. □

Theorem. *The assignment $E_A : f(x) \to f(A)$ is a ring homomorphism.*

It's not an isomorphism because some $f(A) = g(A)$ for $f \neq g$? I.e. it's non-injective. So the kernel is the set of polynomials $p(x)$ such that $p(A) = 0$. It contains the minimal and characteristic polynomials.

Proof. Let $f, g \in \mathbb{F}[x]$ with $f(x) = a_J x^J + \ldots + a_0$ and $g(x) = b_J x^J + \ldots + b_0$. (If f and g are not of the same degree then pad the lower degree one with zero coefficients to make it the same degree as the higher one.)

Addition:

$$\begin{aligned} E_A\Big((f+g)(x)\Big) &= (f+g)(A) \\ &= f(A) + g(A) \quad \text{(by definition of addition of polynomials)} \\ &= E_A\Big(f(x)\Big) + E_A\Big(g(x)\Big) \end{aligned}$$

Multiplication:

$$\begin{aligned} E_A\Big((fg)(x)\Big) &= (fg)(A) \\ &= f(A)g(A) \quad \text{(by definition of multiplication of polynomials)} \\ &= E_A\Big(f(x)\Big) E_A\Big(g(x)\Big) \end{aligned}$$

□

Definition (Minimal polynomial)**.** *Let V be a finite-dimensional vector space over $\mathbb{F}$, and let A be a matrix of a linear transformation $T : V \to V$.*

The minimal polynomial *$m_A(x)$ is the monic polynomial $p(x)$ of minimal degree such that $p(A) = 0$.*

Theorem.

1. *The minimal polynomial is unique.*
2. *Let $f(x)$ be a polynomial. If $f(A) = 0$ then $m_A | f$.*

[3] Let $\Delta_{ij} \in M_n(\mathbb{F})$ be the matrix with (i,j)-th entry 1, and 0 elsewhere. Then $\{\Delta_{ij} \mid i, j \leq n\}$ is a basis.

4.15 Quotient spaces, induced maps

Theorem. *Let $T : V \to W$ be an isomorphism[4] between vector spaces V and W, and let $A \subseteq V, B \subseteq W$ be subspaces. Then the formula $\overline{T}(v + A) = T(v) + B$ gives a well-defined linear map $\overline{T} : V/A \to W/B$ if and only if $T(A) \subseteq B$.*

Therefore

Theorem. *Let $T : V \to V$ with U a subspace of V. If U is T-invariant, then T induces a linear map of quotients $\overline{T} : V/U \to V/U$ given by $v + U \mapsto T(v) + U$.*

4.16 Cross product

Definition. *The cross product is defined for 3-dimensional vectors only. It can be written as a formal determinant*

$$u \times v = \left|\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}\right|,$$

which can be computed using the cofactor expansion:

$$u \times v = (u_2v_3 - u_3v_2)\mathbf{i} - (u_1v_3 - u_3v_1)\mathbf{j} + (u_1v_2 - u_2v_1)\mathbf{k}.$$

[4] note: not linear; but why not homomorphism?

4.17 Matousek – 33 Miniatures

4.17.1 Fibonacci - matrix multiplication

Definition. *The Fibonacci sequence* $0, 1, 1, 2, 3, 5, 8, \ldots$ *is defined by*

$$\begin{aligned} x_0 &= 0 \\ x_1 &= 1 \\ x_n &= x_{n-1} + x_{n-2}, \qquad n \geq 2. \end{aligned}$$

Remark. The sequence can be generated by taking $\begin{pmatrix} x_1 \\ x_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ as the initial state and multiplying repeatedly by $A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, yielding the sequence

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \begin{pmatrix} 8 \\ 5 \end{pmatrix}, \ldots.$$

Thus $\begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix} = A^n \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.[5]

[5] The book describes a trick for efficiently raising a matrix A to an integer power n involving using the binary expansion of n to determine the computations to perform. So $\log_2(n)$ matrix multiplications rather than n.

4.17.2 Fibonacci - sequence space

Let V be the vector space containing all sequences of real numbers $u_0, u_1, \ldots$.

Let W be the subspace of V containing sequences such that $u_{n+2} = u_{n+1} + u_n$ for all $n \geq 0$.

Claim. *A basis for W is*

$$e_1 = 0, 1, 1, 2, 3, 5, 8, \ldots$$
$$e_2 = 1, 0, 1, 1, 2, 3, 5, \ldots.$$

Proof. We need to show that e_1 and e_2 are linearly independent and spanning. Note that every sequence $u \in W$ is determined by the first two values (u_0, u_1). Define the projection $P : W \to \mathbb{R}^2$ by $p(u) := (u_0, u_1)$. Note that P is linear [6], injective and invertible. Let $i = (0, 1) \in \mathbb{R}^2$ and $j = (1, 0) \in \mathbb{R}^2$. Therefore $P^{-1}(i), P^{-1}(j) = e_1, e_2$ is a basis for W, by theorem (10). □

Now we look for a different basis of W. Specifically, we have an inspiration: we seek sequences $u \in W$ of the form $u_n = \tau^n$ for some τ. Thus τ must satisfy $\tau^{n+2} = \tau^{n+1} + \tau^n$ for all $n \geq 0$. We solve this for $n = 0$. We have $\tau^2 = \tau + 1$, therefore $\tau_1 = \frac{1+\sqrt{5}}{2}$ and $\tau_2 = \frac{1-\sqrt{5}}{2}$.

Define two new sequences $e'_1 = \tau_1^0, \tau_1^1, \ldots$ and $e'_2 = \tau_2^0, \tau_2^1, \ldots$.

Claim. *e'_1, e'_2 is another basis for W.*

Proof. We need only they are linearly independent. If they are, then they span W since W is 2-dimensional. So suppose $\lambda_1 e'_1 + \lambda_2 e'_2 = 0$. Then, from considering the first two elements, we have $\begin{cases} \lambda_1 + \lambda_2 = 0 \\ \lambda_1 \tau_1 + \lambda_2 \tau_2 = 0. \end{cases}$

Therefore $\lambda_1(\tau_1 - \tau_2) = 0$, so $\lambda_1 = \lambda_2 = 0$, as required. □

Therefore for all $u \in W$ there exist λ_1, λ_2 such that $u = \lambda_1 e'_1 + \lambda_2 e'_2$.

In particular, there exist λ_1, λ_2 such that $e_1 = 0, 1, 1, 2, 3, 5, 8, \ldots = \lambda_1 e'_1 + \lambda_2 e'_2$.

We can use the first two elements of the sequence to solve for λ_1 and λ_2. We have $\begin{cases} 0 = \lambda_1 + \lambda_2 \\ 1 = \lambda_1 \tau_1 + \lambda_2 \tau_2 \end{cases}$,

therefore $\lambda_1 = \frac{1}{\tau_1 - \tau_2} = \frac{1}{\sqrt{5}}$ and $\lambda_2 = \frac{-1}{\sqrt{5}}$.

The n-th element of the Fibonacci sequence is therefore

$$\lambda_1 \tau_1^n + \lambda_2 \tau_2^n = \frac{1}{\sqrt{5}} \left(\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right).$$

6

$$P(\lambda u) = (\lambda u_0, \lambda u_1) = \lambda P(u)$$
$$P(u + v) = (u_0 + v_0, u_1 + v_1) = (u_0, u_1) + (v_0, v_1) = P(u) + P(v).$$

Fibonacci - eigenbasis

(Not in book.)

Recall from (4.17.1) that the Fibonacci sequence can be generated by taking $\begin{pmatrix}1\\0\end{pmatrix}$ as the initial state and multiplying repeatedly by $A = \begin{pmatrix}1 & 1\\1 & 0\end{pmatrix}$. Thus $\begin{pmatrix}x_{n+1}\\x_n\end{pmatrix} = A^n \begin{pmatrix}1\\0\end{pmatrix}$.

We can compute the matrix power via "diagonalization": i.e., in a basis defined by two eigenvectors the matrix of the linear transformation is diagonal and thus the matrix power can be computed trivially.

$|A| \neq 0$ therefore A is full rank and has two linearly independent eigenvectors. Let these be $\begin{pmatrix}v_{11}\\v_{21}\end{pmatrix}$ and $\begin{pmatrix}v_{12}\\v_{22}\end{pmatrix}$ and let τ_1, τ_2 be associated eigenvectors. Define $V := \begin{pmatrix}v_{11} & v_{12}\\v_{21} & v_{22}\end{pmatrix}$ and $T := \begin{pmatrix}\tau_1 & 0\\0 & \tau_2\end{pmatrix}$. Then

$$A^n = V^{-1} T^n V.$$

The characteristic polynomial of A is $\tau^2 - \tau - 1 = 0$, therefore the eigenvalues are $\tau_1 = \frac{1+\sqrt{5}}{2}$ and $\tau_2 = \frac{1-\sqrt{5}}{2}$.

Let $v(\tau) = \begin{pmatrix}v_1\\1\end{pmatrix}$ be an eigenvector associated with eigenvalue τ. (Since there is a line of eigenvectors for each eigenvalue, we make an arbitrary choice of $v_2 = 1$ to find particular eigenvectors.)

We have

$$\begin{aligned}
(A - \tau I)\, v &= 0\\
\begin{pmatrix}1-\tau & 1\\1 & -\tau\end{pmatrix}\begin{pmatrix}v_1\\v_2\end{pmatrix} &= 0\\
&\begin{cases} v_1(1-\tau) + v_2 = 0 & (1)\\ v_1 - \tau v_2 = 0 & (2)\end{cases}
\end{aligned}$$

From (2) therefore $v(\tau) = \begin{pmatrix}\tau\\1\end{pmatrix}$ and the eigenvectors are the columns of $V := \begin{pmatrix}\tau_1 & \tau_2\\1 & 1\end{pmatrix}$.

Note that $\tau_1\tau_2 = -4$ therefore $\tau_2 = \frac{-4}{\tau_1}$.

Therefore

$$\begin{aligned}
A^n &= V^{-1}T^nV \\
&= \frac{1}{\tau_1-\tau_2}\begin{pmatrix}1 & -\tau_2\\ -1 & \tau_1\end{pmatrix}\begin{pmatrix}\tau_1^n & 0\\ 0 & \tau_2^n\end{pmatrix}\begin{pmatrix}\tau_1 & \tau_2\\ 1 & 1\end{pmatrix} \\
&= \frac{1}{\tau_1+\frac{2^2}{\tau_1}}\begin{pmatrix}1 & \frac{2^2}{\tau_1}\\ -1 & \tau_1\end{pmatrix}\begin{pmatrix}\tau_1^n & 0\\ 0 & \frac{-2^{2n}}{\tau_1^n}\end{pmatrix}\begin{pmatrix}\tau_1 & \frac{-2^2}{\tau_1}\\ 1 & 1\end{pmatrix} \\
&= \frac{1}{\tau_1+\frac{2^2}{\tau_1}}\begin{pmatrix}1 & \frac{2^2}{\tau_1}\\ -1 & \tau_1\end{pmatrix}\begin{pmatrix}\tau_1^{n+1} & -2^2\tau_1^{n-1}\\ \frac{-2^{2n}}{\tau_1^n} & \frac{-2^{2n}}{\tau_1^n}\end{pmatrix} \\
&= \frac{1}{\tau_1+\frac{2^2}{\tau_1}}\begin{pmatrix}\tau_1^{n+1}-\frac{2^{2(n+1)}}{\tau_1^{n+1}} & -2^2\tau_1^{n-1}-\frac{2^{2(n+1)}}{\tau_1^{n+1}}\\ -\tau_1^{n+1}-\frac{2^{2n}}{\tau_1^{n-1}} & 2^2\tau_1^{n-1}-\frac{2^{2n}}{\tau_1^{n-1}}\end{pmatrix} \\
&= \frac{1}{\tau_1^{n+2}+2^2\tau_1^n}\begin{pmatrix}\tau_1^{2(n+1)}-2^{2(n+1)} & -2^2\tau_1^{2n}-2^{2(n+1)}\\ -\tau_1^{2(n+1)}-2^{2n}\tau_1^2 & 2^2\tau_1^{2n}-2^{2n}\tau_1^2\end{pmatrix}
\end{aligned}$$

Recall that $\begin{pmatrix}x_{n+1}\\ x_n\end{pmatrix} = A^n\begin{pmatrix}1\\ 0\end{pmatrix}$. Therefore

$$x_n = \frac{-\tau_1^{2(n+1)}-2^{2n}\tau_1^2}{\tau_1^{n+2}+2^2\tau_1^n}$$

4.17.3 Fibonacci - generating function

Definition (Fibonacci sequence)**.** *The Fibonacci sequence is defined by* $a_0 = 1, a_1 = 1, a_n = a_{n-1} + a_{n-2}$.

Definition (Generating function)**.** *The generating function of the sequence* $(a_n)_{n=0}^{\infty}$ *is* $\sum_{n=0}^{\infty} a_n x^n$.

Theorem. *The n-th Fibonacci number is...*

Proof. The generating function of the Fibonacci sequence is

$$\begin{aligned} f(x) &= a_0 + a_1 x + \sum_{n=2}^{\infty} a_n x^n \\ &= 1 + x + \sum_{n=2}^{\infty} (a_{n-2} + a_{n-1}) x^n \\ &= 1 + x + x^2 \sum_{n=2}^{\infty} a_{n-2} x^{n-2} + x \sum_{n=2}^{\infty} a_{n-1} x^{n-1} \\ &= 1 + x + x^2 f(x) + x \left(f(x) - 1 \right) \\ &= 1 + f(x)(x^2 + x), \end{aligned}$$

so

$$\begin{aligned} f(x) &= \frac{1}{1 - x - x^2} \\ &= \frac{1}{\left(x - \frac{1+\sqrt{5}}{2} \right) \left(x - \frac{1-\sqrt{5}}{2} \right)} \\ &= \frac{4}{(2x - 1 - \sqrt{5})(2x - 1 + \sqrt{5})}. \end{aligned}$$

Performing a partial fraction decomposition,

$$\frac{4}{(2x-1-\sqrt{5})(2x-1+\sqrt{5})} = \frac{A}{2x-1-\sqrt{5}} + \frac{B}{2x-1+\sqrt{5}}$$
$$4 = x(2A+2B) - (A+B) + \sqrt{5}(A-B)$$
$$\begin{cases} A+B=0 \\ A-B=\frac{4}{\sqrt{5}} \end{cases}$$

so $A = \frac{1}{\sqrt{5}}$, $B = -\frac{1}{\sqrt{5}}$ and the generating function is

$$f(x) = \frac{1}{\sqrt{5}}\left(\frac{1}{2x-1-\sqrt{5}} + \frac{1}{2x-1+\sqrt{5}}\right).$$

Taking derivatives,

$$\sqrt{5}f'(x) = -\frac{2}{(2x-1-\sqrt{5})^2} - \frac{2}{(2x-1+\sqrt{5})^2}$$
$$\sqrt{5}f''(x) = \frac{4}{(2x-1-\sqrt{5})^3} + \frac{4}{(2x-1+\sqrt{5})^3}$$
$$\sqrt{5}f^{(n)}(x) = (-1)^n\left(\frac{2^n}{(2x-1-\sqrt{5})^{n+1}} + \frac{2^n}{(2x-1+\sqrt{5})^{n+1}}\right)$$
$$\sqrt{5}f^{(n)}(x) = (-2)^n\left(\frac{1}{(2x-1-\sqrt{5})^{n+1}} + \frac{1}{(2x-1+\sqrt{5})^{n+1}}\right),$$

and so the Maclaurin expansion of $f(x)$ is

$$\sqrt{5}f(x) = \sum_{n=0}^{\infty} \frac{(-2)^n}{n!(2x-1-\sqrt{5})^{n+1}}x^n$$

□

4.17.4 The Clubs of Oddtown

Oddtown has two rules about clubs:

- Every club must have an *odd* number of members.
- The number of members in the intersection of any two distinct clubs must be *even*.

Theorem 12. *Under these rules, the number of clubs is less than the number of members.*

Proof. Let m be the number of clubs and n be the number of members.

Let $A = (a_{ij})$ be a matrix over the finite field F_2, where $a_{ij} = \begin{cases} 1 & \text{if person } j \text{ is a member of club } i, \\ 0 & \text{otherwise.} \end{cases}$

Thus rows correspond to clubs, and the (i, j)-th entry of AA^T is 1 if the intersection of club i and club j is odd, and 0 if even.

Therefore the rules imply that $AA^T = I_m$. I.e. the rank of AA^T is m.

But the rank of a product of matrices cannot be larger than the minimum rank of any one factor, so $m \leq n$. □

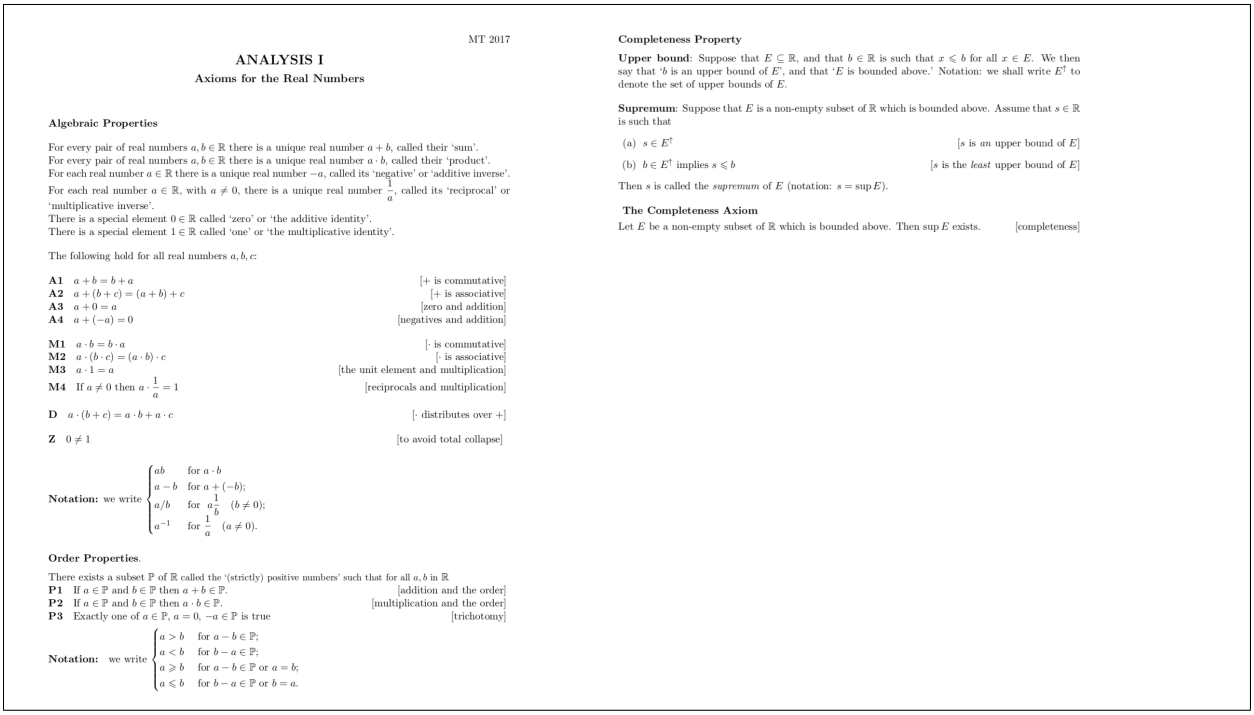
Chapter 5

Real Analysis

5.1 Sequences and Series

Notes from Oxford - M1 - Sequences and Series.

5.1.1 Axioms for the real numbers



MT 2017

ANALYSIS I

Axioms for the Real Numbers

Algebraic Properties

For every pair of real numbers $a, b \in \mathbb{R}$ there is a unique real number $a+b$, called their 'sum'.
For every pair of real numbers $a, b \in \mathbb{R}$ there is a unique real number $a \cdot b$, called their 'product'.
For each real number $a \in \mathbb{R}$ there is a unique real number $-a$, called its 'negative' or 'additive inverse'.
For each real number $a \in \mathbb{R}$, with $a \neq 0$, there is a unique real number $\frac{1}{a}$, called its 'reciprocal' or 'multiplicative inverse'.
There is a special element $0 \in \mathbb{R}$ called 'zero' or 'the additive identity'.
There is a special element $1 \in \mathbb{R}$ called 'one' or 'the multiplicative identity'.

The following hold for all real numbers a, b, c:

A1 $a+b=b+a$ [+ is commutative]
A2 $a+(b+c)=(a+b)+c$ [+ is associative]
A3 $a+0=a$ [zero and addition]
A4 $a+(-a)=0$ [negatives and addition]

M1 $a \cdot b = b \cdot a$ [$\cdot$ is commutative]
M2 $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ [$\cdot$ is associative]
M3 $a \cdot 1 = a$ [the unit element and multiplication]
M4 If $a \neq 0$ then $a \cdot \frac{1}{a} = 1$ [reciprocals and multiplication]

D $a \cdot (b+c) = a \cdot b + a \cdot c$ [$\cdot$ distributes over +]

Z $0 \neq 1$ [to avoid total collapse]

Notation: we write $\begin{cases} ab & \text{for } a \cdot b \\ a-b & \text{for } a+(-b); \\ a/b & \text{for } a\frac{1}{b} \ (b \neq 0); \\ a^{-1} & \text{for } \frac{1}{a} \ (a \neq 0). \end{cases}$

Order Properties.

There exists a subset $\mathbb{P}$ of $\mathbb{R}$ called the '(strictly) positive numbers' such that for all a, b in $\mathbb{R}$
P1 If $a \in \mathbb{P}$ and $b \in \mathbb{P}$ then $a+b \in \mathbb{P}$. [addition and the order]
P2 If $a \in \mathbb{P}$ and $b \in \mathbb{P}$ then $a \cdot b \in \mathbb{P}$. [multiplication and the order]
P3 Exactly one of $a \in \mathbb{P}$, $a=0$, $-a \in \mathbb{P}$ is true [trichotomy]

Notation: we write $\begin{cases} a > b & \text{for } a-b \in \mathbb{P}; \\ a < b & \text{for } b-a \in \mathbb{P}; \\ a \geqslant b & \text{for } a-b \in \mathbb{P} \text{ or } a=b; \\ a \leqslant b & \text{for } b-a \in \mathbb{P} \text{ or } b=a. \end{cases}$

Completeness Property

Upper bound: Suppose that $E \subseteq \mathbb{R}$, and that $b \in \mathbb{R}$ is such that $x \leqslant b$ for all $x \in E$. We then say that 'b is an upper bound of E', and that 'E is bounded above.' Notation: we shall write $E^{\uparrow}$ to denote the set of upper bounds of E.

Supremum: Suppose that E is a non-empty subset of $\mathbb{R}$ which is bounded above. Assume that $s \in \mathbb{R}$ is such that

(a) $s \in E^{\uparrow}$ [s is *an* upper bound of E]

(b) $b \in E^{\uparrow}$ implies $s \leqslant b$ [s is the *least* upper bound of E]

Then s is called the *supremum* of E (notation: $s = \sup E$).

The Completeness Axiom

Let E be a non-empty subset of $\mathbb{R}$ which is bounded above. Then $\sup E$ exists. [completeness]

5.1.2 Approximation property of supremum

Theorem. *Let $S \subset \mathbb{R}$ be non-empty and bounded above (so $\sup S$ exists). For all $\delta > 0$, there exists $s_\delta \in S$ such that*

$$\sup S - \delta < s_\delta \leq \sup S.$$

Intuition. The supremum is either a member of S or it is "touching" an element of S with "no gap".

Proof. If $\sup S \in S$ then we can take $s_\delta = \sup S$ for all δ and we are done.

So assume $\sup S \notin S$. For a contradiction, suppose the negation of the claim, i.e. that there exists $\delta > 0$ such that for all $s \in S$ either $s \leq \sup S - \delta$ or $s > \sup S$. Since $s > \sup S$ is impossible by definition of sup, we have that $s \leq \sup S - \delta$ for all $s \in S$. But then $\sup S - \delta$ is an upper bound for S and $\sup S - \delta < \sup S$, a contradiction. □

5.1.3 Archimedean Property of $\mathbb{N}$

Theorem.

1. $\mathbb{N}$ *has no upper bound.*

2. *For all* $\epsilon > 0$ *there exists* $n \in \mathbb{N}$ *such that* $\frac{1}{n} < \epsilon$.

Proof.

1. Suppose $\mathbb{N}$ has an upper bound. Then $\sup \mathbb{N}$ exists. By the Approximation Property there exists $n \in \mathbb{N}$ such that $\sup \mathbb{N} - \frac{1}{2} < n \leq \sup \mathbb{N}$. But then $n + 1 \in \mathbb{N}$ and $n + 1 > \sup N$, a contradiction. Therefore $\sup \mathbb{N}$ does not exist, therefore $\mathbb{N}$ has no upper bound.

2. Since $\mathbb{N}$ has no upper bound, there exists $n \in \mathbb{N}$ such that $n > 1/\epsilon$, i.e. $1/n < \epsilon$.

□

5.1.4 Well-ordered property of $\mathbb{N}$

Theorem. *Every nonempty subset of* $\mathbb{N}$ *has a minimum.*

Proof. Let $\varnothing \neq S \subseteq \mathbb{N} \subset \mathbb{R}$. Note that S is bounded below by 0, therefore $\inf S$ exists. Suppose $\inf S \notin S$. By the Approximation Property, there exists $n_1 \in S$ such that $\inf S \leq n_1 < \inf S + 1$.

We claim that $\inf S = n_1$. Suppose for a contradiction that $\inf S \neq n_1$. Then $n_1 = \inf S + \delta$ for some $0 < \delta < 1$. By the Approximation property again, there exists $n_2 \in S$ such that $\inf S \leq n_2 < n_1 < \inf S + 1$.

But since $n_1 > n_2$ we have $n_1 \geq n_2 + 1$, therefore $n_1 \geq \inf S + 1$ which contradicts $n_1 < \inf S + 1$. Therefore $\inf S = n_1 \in S$ and $\min S$ exists. □

Remark. Similarly:

1. Every nonempty subset of $\mathbb{Z}$ that is bounded below has a minimum.

2. Every nonempty subset of $\mathbb{Z}$ that is bounded above has a maximum.

Intuition. Because of the "gappiness" of $\mathbb{N}$ and $\mathbb{Z}$, bounded subsets must contain their suprema/infima.

5.1.5 Existence of ceil and floor

Definition (floor and ceil)**.** *Let* $x \in \mathbb{R}$. *Then floor of* x *is* $\lfloor x \rfloor = \max\{n \in \mathbb{Z} \mid n \leq x\}$ *and ceil of* x *is* $\lceil x \rceil = \min\{n \in \mathbb{Z} \mid n \geq x\}$.

Theorem ($\lfloor x \rfloor$ and $\lceil x \rceil$ exist)**.**
Let $x \in \mathbb{R}$. *Define* $S = \{n \in \mathbb{Z} \mid n \geq x\} \subset \mathbb{R}$. *Note that* S *is bounded below by* x. *Also* S *is non-empty by the Archimedean Property of* $\mathbb{N}$, *since otherwise* x *would be an upper bound for* $\mathbb{N}$. *Therefore* $\lceil x \rceil = \min S$ *exists by Well-Ordering.*

Similarly, $\lfloor x \rfloor$ *exists.*

5.1.6 Existence of $\sqrt{2}$

Theorem. *There exists a unique $a \in \mathbb{R}$ such that $a^2 = 2$.*

Remark. The only thing that ties the proof to the reals is that it relies on completeness (sup exists). We know that $\sqrt{2} \notin \mathbb{Q}$, therefore $\mathbb{Q}$ is not complete.

Proof. Let $S = \{s \in \mathbb{R} \mid s^2 < 2\}$. Since S is bounded above, $a := \sup S$ exists. We show that $a^2 = 2$ by showing that $a^2 < 2$ and $a^2 > 2$ lead to contradictions.

Note that $1 \in S$, therefore $a \geq 1$.

1. **Suppose** $a^2 < 2$. We seek an $h > 0$ such that $(a+h)^2 < 2$ since this would contradict the definition $a := \sup S$. Note that

$$\begin{aligned}(a+h)^2 - 2 &= a^2 + 2ah + h^2 - 2 \\ &< a^2 - 2 + 3ah \qquad \text{if } h < a \\ &< 0 \qquad\qquad\quad \text{if } h < (2-a^2)/3a.\end{aligned}$$

   Therefore if we take $h < \min\left(a, \frac{2-a^2}{3a}\right)$ then $a + h \in S$ which contradicts the definition $a := \sup S$.

2. **Suppose** $a^2 > 2$. By the Approximation Property for all $0 < h < 1$ we can find $s \in S$ such that $a - h < s$. Therefore $(a-h)^2 < s^2 < 2$. We seek a value of h such that $(a-h)^2 \geq 2$, which would be a contradiction. Note that $a^2 - 2ah < (a-h)^2$. If we take $h = (a^2-2)/2a$ then we have $a^2 - 2ah = 2 < (a-h)^2 < 2$, the desired contradiction.

Finally to show that a is unique, suppose that there exists $b \in \mathbb{R}$ with $b^2 = 2$. Then $0 = a^2 - b^2 = (a+b)(a-b)$ therefore $a = b$. □

5.1.7 Connection between sequences and functions

Theorem. *The following two statements are equivalent:*

1. *$f(x) \to L$ as $x \to a$.*
2. *For every sequence (x_n) such that $x_n \neq a$*

$$\left(\lim_{n\to\infty} x_n = a\right) \implies \left(\lim_{n\to\infty} f(x_n) = f(a)\right)$$

5.1.8 Limit of product is product of limits

Theorem.
Let $\lim_{x\to a} f(x) = L_f$ and $\lim_{x\to a} g(x) = L_g$. Then $\lim_{x\to a} f(x)g(x) = L_f L_g$.

Proof. Note that

$$\begin{aligned}\lim_{x\to a} f(x)g(x) &= \lim_{x\to a}\Big((f(x)-L_f)(g(x)-L_g) + L_f g(x) + L_g f(x) - L_f L_g\Big) \\ &= L_f L_g + \lim_{x\to a}(f(x)-L_f)(g(x)-L_g),\end{aligned}$$

so we need to show that $\lim_{x\to a}(f(x)-L_f)(g(x)-L_g)=0$. Fix $\epsilon>0$. Since $\lim_{x\to a}(f(x)-L_f)=\lim_{x\to a}(g(x)-L_g)=0$, there exists δ (pick the minimum of the two δs) such that whenever $|x-a|<\delta$

$$|(f(x)-L_f)|<\sqrt{\epsilon} \quad \text{and} \quad |(g(x)-L_g)|<\sqrt{\epsilon},$$

therefore $|(f(x)-L_f)(g(x)-L_g)-0|<\epsilon$ as required. □

5.1.9 Limit of quotient is quotient of limits

Theorem.
Let $\lim_{x\to a} f(x)=L_f$ *and* $\lim_{x\to a} g(x)=L_g\neq 0$. *Then*

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\frac{L_f}{L_g}.$$

Proof. TODO

$$\lim_{x\to a}\frac{f(x)}{g(x)}-\frac{L}{M}=\lim_{x\to a}\frac{f(x)}{g(x)}-\frac{1}{g(x)}+\frac{1}{g(x)}-\frac{L}{M}$$

Let $L_f=\lim_{x\to a} f(x)$ and $L_g=\lim_{x\to a} g(x)\neq 0$.

Fix $\epsilon>0$ and let δ_f and δ_g be such that

$$|x-a|<\delta_f \implies |f(x)-L_f|<\epsilon$$
$$|x-a|<\delta_g \implies |g(x)-L_g|<\epsilon.$$

Let $\delta=\min(\delta_f,\delta_g)$. Then

$$\frac{|f(x)-L_f|}{|g(x)-L_f|}$$

□

5.2 Continuity and Differentiability

Note from Oxford - M2 - Continuity and Differentiability.

5.2.1 Limit point

Definition. *Let* $E\subset\mathbb{R}$. *A point* $p\in\mathbb{R}$ *is a limit point of* E *iff for all* $\delta>0$ *there exists* $x\in E$ *such that* $0<|x-p|<\delta$.

Intuition. A deleted ball, of arbitrarily small radius, placed over p, will capture at least one point of E.

5.2.2 Limit, Convergence

Definition (Limit of a sequence (x_n))**.**
$\lim_{n\to\infty} x_n = L$ *iff for all* $\epsilon > 0$ *there exists* $N \in \mathbb{N}$ *such that* $n > N \implies |x_n - L| < \epsilon$. *The sequence is then said to converge to* L.

Definition (Limit of a function $f : \mathbb{R} \to \mathbb{R}$)**.**
$\lim_{x\to a} f(x) = L$ *means: for all* $\epsilon > 0$ *there exists* $\delta > 0$ *such that* $0 < |x - a| < \delta \implies |f(x) - L| < \epsilon$.

Equivalent notation: $f(x) \to L$ as $x \to a$

Remark.

1. The value of f at a is irrelevant (f need not be defined at a).
2. f must tend to L from both sides.

5.2.3 Limits of functions - Examples

Example 13. Let $E = \mathbb{R} \setminus \{0\}$ and define $f : E \to \mathbb{R}$ by $f(x) = L$. Then 0 is a limit point of E and $f(x) \to L$ as $x \to 0$.

Proof. Fix $\delta > 0$. Then $\exists x\ 0 < |x - 0| < \delta$ is true since we can choose $x = \frac{\delta}{2}$. Therefore 0 is a limit point of E.

Fix $\epsilon > 0$. Let $\delta = 1$. Then $0 < |x - 0| < \delta \implies |f(x) - L| = 0 < \epsilon$. □

5.2.4 Continuity of a function f

Definition. f *is continuous at* a *if* $\lim_{x\to a} f(x) = f(a)$.

Therefore, using the definition of limit, f is continuous at a iff for all $\epsilon > 0$ there exists $\delta > 0$ such that $|x - a| < \delta \implies |f(x) - f(a)| < \epsilon$.

5.2.5 Uniform convergence and uniform continuity

Definition (Uniform convergence)**.** *A sequence of functions* $\{f_n\}_{n\geq 0}$ *has a limit* f *iff for every point* x *in the input set the sequence* $\{f_n(x)\}_{n\geq 0}$ *has limit* $f(x)$.

They converge uniformly to f *iff the same* m *works for all input values.*

Definition (Uniform continuity)**.** *A function* f *is uniformly continuous iff the same* δ *works for all* x_0.

A function f *is uniformly continuous iff for all* ϵ*, no matter how small, a* δ *exists such that for all* $x_0 \in U$*, if* x *is within* δ *of* x_0 *then* $f(x)$ *is within* ϵ *of* $f(x_0)$.

5.2.6 Intermediate value theorem

Theorem. *Let $a, b \in \mathbb{R}$ with $b > a$, and $f : [a, b] \to \mathbb{R}$ be continuous. Let u lie strictly between $f(a)$ and $f(b)$. Then there exists $c \in (a, b)$ such that $f(c) = u$.*

Proof. Define $S := \{x \in [a, b] \mid f(x) < u\}$. Since $a \in S$, S is non-empty. By completeness of reals $c := \sup S$ exists. The theorem now follows from continuity of f at c. (Fix $\epsilon > 0$ and consider points $a^* \in (c - \delta, c)$ and $a^{**} \in (c, c + \delta)$, noting whether they are in S and the $\epsilon - \delta$ continuity criterion.) □

5.2.7 Mean-value theorem

Theorem. *Let $a, b \in \mathbb{R}$ with $b > a$, and $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $x \in (a, b)$ such that $f'(x) = \frac{f(b)-f(a)}{(b-a)}$.*

5.2.8 Differentiability implies continuity

Theorem. *Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable. Then f is continuous.*

Proof.
Let $a \in \mathbb{R}$. The claim is that $\lim_{x \to a} f(x) - f(a) = 0$. Since f is differentiable,

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

exists. Therefore by (5.1.8)

$$\lim_{x \to a} f(x) - f(a) = \lim_{x \to a} (x - a) \frac{f(x) - f(a)}{x - a} = 0 \cdot f'(a) = 0.$$

□

Remark. Intuitively it seems that differentiability implies continuity because, for the derivative to exist, the numerator $f(x) - f(a)$ must get small as $x \to a$, as the denominator $x - a$ does.

5.3 Metric Spaces

5.3.1 Metric space

Definition 14. *Let X be a set. Suppose $d : X \times X \to \mathbb{R}$ satisfies positivity, symmetry and the triangle equality. Then d is a metric and (X, d) is a metric space.*

5.3.2 Open ball

Definition 15. *Let (X, d) be a metric space, $x \in X$ and $\delta > 0$. Then $B(x, \delta) := \{x \in X \mid d(x, x) < \delta\}$ iss an open ball of radius δ centred at x.*

Remark. Also closed ball, $\leq$. E.g. singleton set.

5.3.3 Ball-based continuity criterion

Lemma 16. *f is continuous at x if for all $\epsilon > 0$ there exists $\delta > 0$ such that $f\left(B(x, \delta)\right) \subseteq B(f(x), \epsilon))$.*

Equivalently, $B(x, \delta) \subseteq f^{-1}\left(B(f(x), \epsilon)\right)$.

5.3.4 Neighbourhood

Definition 17. *Let (X, d) be a metric space. $N \subseteq X$ is a neighbourhood of $x \in X$ iff there exists $\delta > 0$ such that $B(x, \delta) \subseteq N$.*

Remark. N is a neighbourhood of x if a ball can be placed at x without poking outside N.

5.3.5 Open and closed subsets of a metric space

Definition 18. *Let (X, d) be a metric space. Then $U \subseteq X$ is open iff it is a neighbourhood of all of its elements.*

$V \subseteq X$ is closed iff its complement in X is open.

5.3.6 Topology on a metric space

Definition 19. *Let (X, d) be a metric space. The collection $\mathcal{T}$ of all open sets in the metric space is called the topology of X.*

Remark. Note that the definitions so far have the following dependency:

(open set) $\leftarrow$ (neighbourhood) $\leftarrow$ (ball) $\leftarrow$ (metric),

so they apply to metric spaces only.

5.3.7 Open set-based continuity criterion

Theorem 20. *Let X and Y be metric spaces and let $f : X \to Y$. Then*

f is continuous at x iff for every neighbourhood $N \subseteq Y$ of $f(x)$, the preimage $f^{-1}(N)$ is a neighbourhood of $x \in X$.

f is continuous iff for every open set U of Y, $f^{-1}(U)$ is an open set of X.

Remark. So we have defined continuity in terms of open sets (the topology). This means that the metric is only relevant insofar as it induces the topology; two metric spaces with the same topology have the same notion of continuity.

Proof.
Let f be continuous at $x \in X$, and let $N \subseteq Y$ be a neighbourhood of $f(x)$.

Then by definition of neighbourhood there exists a ball at $f(x)$ that stays within N.

By continuity of f the preimage of that ball is a superset of a ball at x.

So the preimage of the ball is a neighbourhood of x. Therefore the preimage of N is also.

Conversely, ... similar.

Let f be continuous on X. Now every open set U of Y contains a ball around some point y... □

5.3.8 Topology on a set, topological space

Definition 21. *A topology on a set X is a collection $\mathcal{T}$ of subsets of X, which are called the open sets. They must satisfy*

1. *closed under arbitrary unions. In particular, $\varnothing$ is an open set of X.*
2. *closed under finite intersections. In particular, X is an open set of X.*

A topological space is a pair $(X, \mathcal{T})$.

Remark. Criteria for closed sets follow by applying de Morgan's laws (closure under finite unions and arbitrary intersections).

$f : X \to Y$ closed iff $f^{-1}(V)$ is closed for all closed sets $V \subseteq Y$.

5.3.9 Limit point

Definition 22. *Let (X, d) be a metric space and $Z \subseteq X$ be any subset.*

$x \in X$ is a limit point of Z if for all $\delta > 0$ the deleted open ball $B(x, \delta) \setminus \{x\}$ has non-empty intersection with Z.

If z is not a limit point of Z, then it is an isolated point.

The set of limit points of Z is denoted Z', and it is clear that $Z_1 \subseteq Z_2 \implies Z_1' \subseteq Z_2'$.

Intuition. $x \notin Z$ is a limit point of Z iff it "touches" Z.

$z \in Z$ is a limit point of Z if it "lies in a contiguous region of Z"

An isolated point of Z is what it sounds like.

Example.
Let $Z = (0, 1] \cup \{2\}$.

Intuitively, 0 is a limit point of Z because it "touches" Z.

Formally, 0 is a limit point of Z because for all $\delta > 0$ the deleted open ball $B(0, \delta)$ contains a point $z > 0 \in Z$.

Intuitively, 2 is an isolated point.

Formally, 2 is not a limit point because $\big(B(2, 0.5) \setminus \{2\}\big) \cap Z = \varnothing$. And yet $2 \in Z$, therefore 2 is an isolated point.

5.3.10 Open sets theorems

1. An open ball is open

5.3.11 Closed sets theorems

1. A closed ball is closed

5.3.12 Continuity theorems

1. $f : X \to Y$ is continuous if for every open ball in Y there is an open ball in X that maps inside it.
2. $f : X \to Y$ is continuous if the preimage of $B(f(x), \epsilon)$ in Y is a ball $B(x, \delta)$ in X.
3. $f : X \to Y$ is continuous if the preimage of the neighbourhood of $f(x)$ is a neighbourhood of x.
4. $f : X \to Y$ is continuous if the preimage of every open set in Y is an open set in X.

5.3.13 Continuity of a linear map

Theorem 23. *Let $f : V \to W$ be a linear map between normed vector spaces. Then f is continuous if and only if $\{\big\|f(x)\big\| : \|x\| \leq 1\}$ is bounded.*

Proof.
Let $v \in V$.

Note that $f(v) = f(v) - f(0)$ since f is linear.

Suppose f is continuous. Then it is continuous at 0.

Therefore for every $\epsilon > 0$ there exists $\delta > 0$ such that $\|v\| < \delta \implies \big\|f(v)\big\| < \epsilon$.

$\vdots$

For the converse, suppose that $\|v\| \leq 1 \implies \big\|f(v)\big\| < M$.

Let $\epsilon > 0$ be given.

Pick $\delta > 0$ such that $\delta M < \epsilon$.

Now consider two points $u, v \in V$ where $\|u - v\| < \delta$. We have

$$\big\|f(u) - f(v)\big\| = \big\|f(u - v)\big\| = \delta \left\| f\left(\frac{u-v}{\delta}\right)\right\|.$$

Note that $\big\|\frac{u-v}{\delta}\big\| < 1$, therefore $\left\|f\left(\frac{u-v}{\delta}\right)\right\| < M$. Therefore we have

$$\big\|f(u) - f(v)\big\| < \delta M < \epsilon$$

as required. □

5.3.14 Norm of linear map is bounded

Theorem 24. $\{\big\|f(x)\big\| : \|x\| \leq 1\}$ *is bounded for linear map* f*, under the Euclidean norm* $\|\|_2$.

Proof. See Oxford A2 Sheet 1 exercises. □

Chapter 6

Calculus

6.1 Overview

Differential calculus is a way to compute quantities related to functions by treating the smooth curve or surface of function output values as being comprised of many local linear functions. Each linear approximation applies over a tiny (arbitrarily small) local interval; the linear approximation in the next interval will in general have a slightly different gradient.

A central concept in differential calculus is the *differential*: the change in output value caused by a small change in the input value, at some starting input value. This describes the way in which the function output changes in response to changes in input. Differentials are often used to compute a *derivative*: the ratio of change in output value to the change in some input value. Derivatives define a local *linear approximation* to the function: over a small local region we consider the real function to be approximated by a line with gradient equal to the derivative at that point.

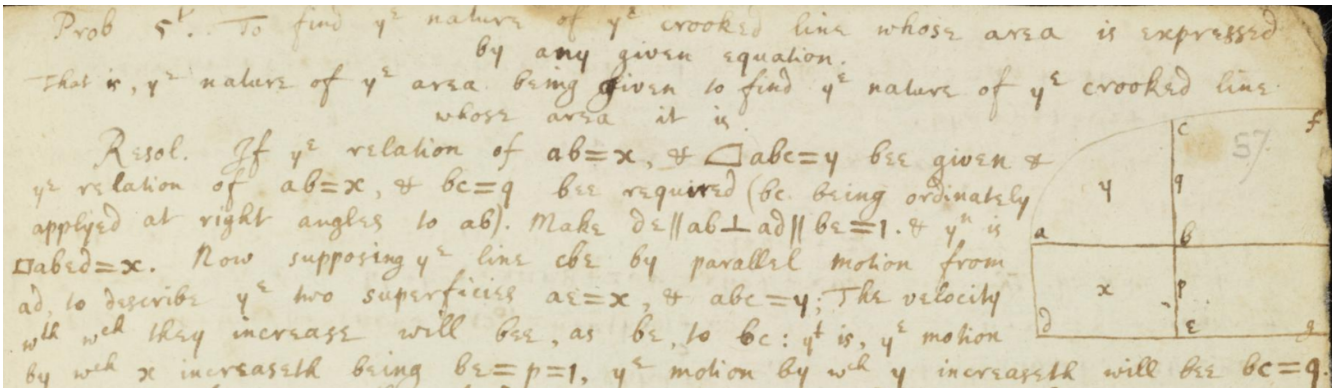
The above is differential calculus. Integral calculus is concerned with "summing" the output values of a function associated with some region in the input space. In the familiar case, the input space is a section of the real number line, and the output values are also real numbers. So "summing" the output values corresponds to calculating the area under a curve (i.e. under the graph of the function).

Now allow the input space to be a higher dimensional Euclidean space, e.g. some region of the plane $\mathbb{R}^2$, but keep the output values as being simply real numbers. One question is what is the value of the integral along some 1-dimensional *path* through the input space. We imagine dividing the input space up into many small sections (vectors) Δx_i, as usual. However, when computing the contribution from one such infinitisimal section, it is not sufficient to say simply that this is $f(x_i)|\Delta x_i|$. The reason is that the appropriate contribution might depend not only on the position x_i but on the direction of the infinitisimal displacement vector Δx_i. Therefore, we define ω_{x_i} to be the linear mapping that takes as input Δx_i and outputs the "height" $f(x_i)$.

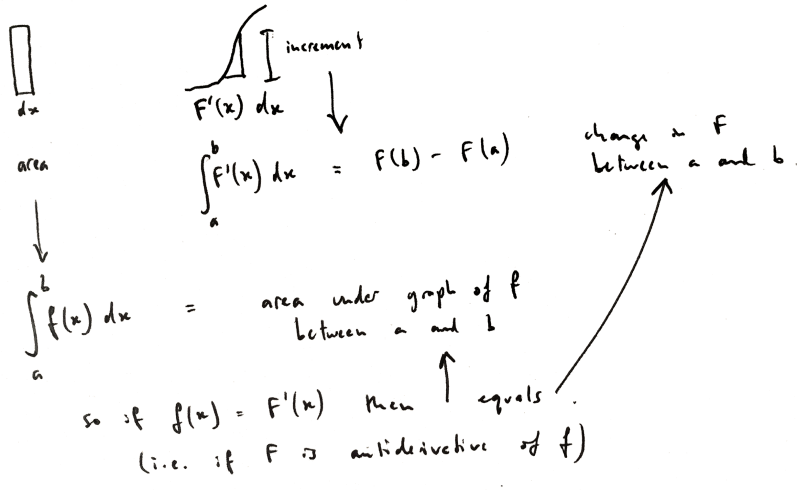
What does this look like in the simple case where the answer is insensitive to the direction of the infinitisimal displacement vector Δx_i? I think ω would depend on $|\Delta x_i|$ only and not otherwise on Δx_i?

Another question is what is the value of the integral over some higher dimensional region of input space (e.g. a subset of the plane).

6.2 The Fundamental Theorem of (Integral) Calculus



Newton's October 1666 Tract on Fluxions.
"...the motion by which y increaseth will bee bc = q."



Recall that the definition of $\int_a^b f(x)\,\mathrm{d}x$ is the area under the graph, computed as the limit of approximating rectangles (Riemann sums).

Consider an "accumulation function", or "area-so-far function" F defined as

$$F(x) = \int_0^x f(u)\,\mathrm{d}u.$$

$F(x)$ is the amount that has accumulated when we are at point x in the input space.

The FTC comes in two parts. Part I states that the derivative of the area-so-far function is the original function of interest:

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x).$$

Note that this is the first time we have connected integration with differentiation: F was defined as a definite integral (area-so-far); nothing in its definition involved differentiation.

[0] https://cudl.lib.cam.ac.uk/view/MS-ADD-03958/109

Part II states that the definite integral $\int_a^b f(x)\,\mathrm{d}x$ can be computed as

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

I think that this is obvious from the definition of F as area-so-far, but the point is that part I has shown us that F might be obtainable as an antiderivative of f rather than via some explicit area calculation (e.g. Riemann sums).

So how do we prove this? What exactly is it we need to prove anyway? We have a definition for derivative, and we have a definition for area-so-far (limit of Riemann sums). So, first, using the definition of derivative,

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) := \lim_{h\to 0}\frac{F(x+h)-F(x)}{h}.$$

In the numerator is the area above a horizontal section of width h. Intuitively, this is approximately $hf(x)$, giving

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = \lim_{h\to 0}\frac{hf(x)}{h} = f(x),$$

as desired. How to make this rigorous? Using the Riemann sums definition of area,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}F(x) &= \lim_{h\to 0}\frac{\lim_{N\to\infty}\sum_i^N \frac{h}{N} f\left(x+\frac{ih}{N}\right)}{h} \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_i^N \lim_{h\to 0} f\left(x+\frac{ih}{N}\right) \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_i^N f(x) \\
&= f(x).
\end{aligned}$$

But in fact real proofs use the Extreme Value Theorem. I am told that one error in the above proof is that it is not valid to exchange the order of the two limits.

TODO FTC – moving away from thinking that an integral "just has to end with d-something". Why does one seek the antiderivative of the part without the d-something?

Examples

In all the following examples, some quantity is "accumulating"[1].

1. $F(x)$ is the area under a graph to the left of x.
   $f(x)$ is the height of the graph at x.

2. $F(x)$ is the volume of a vase between the base and height x.
   $f(x)$ is the cross-sectional area at height x.

3. $F(r)$ is the area of a circle with radius r.
   $f(r)$ is the diameter of a circle with radius r.

[1] "Accumulating" can involve decreasing as well as increasing. For example if the particle starts moving back towards the origin, or if the vase is being filled with a tube and someone starts sucking on it rather than dispensing water.

4. $F(t)$ is the volume of water in a vase that is being filled, at time t.
   $f(t)$ is the rate of filling at time t.

5. $F(t)$ is the position of a moving particle at time t, relative to the origin.
   $f(t)$ is the velocity of the particle at time t.

6. $F(t)$ is the number of bacteria at time t.
   $f(t)$ is the rate at which new bacteria are produced at time t.

Constant rate

1. The height of the graph is constant at h (a rectangle).
   The area to the left of x is hx.

2. $F(x)$ is the volume of a vase between the base and height x.
   The cross-sectional area is constant at a (a cylinder).
   $F(x) = ax$

3. $F(t)$ is the volume of water in a vase that is being filled, at time t.
   Water enters at a constant rate v liters/sec.
   $F(t) = vt$

4. $F(t)$ is the displacement of a moving particle at time t, relative to the origin.
   The velocity of the particle is constant at v m/sec.
   $F(t) = vt$.

5. $F(t)$ is the number of bacteria at time t.
   Bacteria are produced at a constant rate v bacteria/sec.
   $F(t) = vt$.

The amount-so-far can be computed manually:

1. If the rate of increase is constant at v, then the amount to the left of x is simply vx.

2. If the rate of increase at time t is ct (proportional to t), then the amount-so-far graph is a triangle, so the amount to the left of t is $\frac{1}{2} \cdot ct \cdot t = \frac{1}{2}ct^2$.

3. If the rate of increase at point r is $2\pi r$ (the outer edge of a growing disc), then the amount-so-far graph is a triangle again, and the area of the disc is $\frac{1}{2} \cdot r \cdot 2\pi r = \pi r^2$.

What about if the rate of increase is a more complex function? We can still compute the area so far manually, as a limit of Riemann sums:

Compare

$$\begin{aligned}
\int_0^2 (2-x^2)\,\mathrm{d}x &= \lim_{N\to\infty} \sum_{i=1}^{N} \frac{2}{N}\left(2-\left(\frac{2i}{N}\right)^2\right) \\
&= \lim_{N\to\infty} \sum_{i=1}^{N} \frac{4}{N} - \frac{8i^2}{N^3} \\
&= \lim_{N\to\infty} \left(4 - \frac{8}{N^3}\sum_{i=1}^{N} i^2\right) \\
&= \lim_{N\to\infty} \left(4 - \frac{8}{N^3}\frac{N(N+1)(2N+1)}{6}\right) \\
&= \lim_{N\to\infty} \left(4 - 8\frac{(N+1)(2N+1)}{6N^2}\right) \\
&= \lim_{N\to\infty} \left(4 - 8\frac{2+3N^{-1}+N^{-2}}{6}\right) \\
&= 4 - \frac{8}{3} = \frac{4}{3}
\end{aligned}$$

with the solution using antiderivatives:

$$\begin{aligned}
\int_0^2 (2-x^2)\,\mathrm{d}x &= \left[2x - \frac{x^3}{3}\right]_0^2 \\
&= 4 - \frac{8}{3} = \frac{4}{3}.
\end{aligned}$$

Let's fix a physical example for discussing FTC: a moving object. The key quantity here is the distance from the starting point.

Next, before writing the equations that state the FTC, let's be clear about the objects that are going to be involved in those equations. The most important object is a function that gives the distance from the starting point as a function of time.

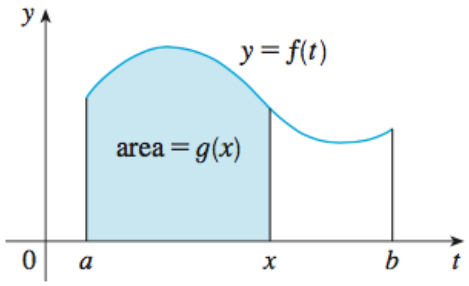
More generally, this is an "accumulation function", or "area-so-far function".

Now, let's introduce some notation. The notation $\int_3^4 f(t)\,\mathrm{d}t$ is *defined* to mean the area under the curve f, between 3 and 4. It's really important to be clear here: the definition of $\int_3^4 t^2\,\mathrm{d}t$ is simply that it is the area under the t^2 curve between those two points. (In particular, note that the definition does *not* involve $\frac{1}{3}t^3$).

Similarly, $\int_0^4 f(t)\,\mathrm{d}t$ is the area under the curve between 0 and 4. The answer is a number. The answer doesn't involve t: t is just a variable used internally in that expression.

Now comes a slightly less obvious point: if the upper limit is not a fixed number, but a variable, as in $\int_0^x f(t)\,\mathrm{d}t$, then that entire expression represents a function of x: it takes in an x value and outputs the area under the curve, between 0 and x. We can give the new function a name, g, and write the definition of g as

$$g(x) = \int_0^x f(t)\,\mathrm{d}t.$$



Functions like g are "accumulation functions", or "area-so-far functions", because they tell you the area up to x, i.e. the area to the left of x.

The FTC is usually split into two parts. The first part states

> At any point x, the rate of change of the area-so-far function at that point is the same as the height of the curve at that point.

This is what Newton was saying when he wrote "...the motion by which y increaseth will bee q.": in his diagram, y is the area, and q is the height of the curve[2].

6.3 Differentiation basics

Theorem (Quotient rule). $\left(\frac{f}{g}\right)^{-1} = \frac{gf' - fg'}{g^2}$

[2] He actually wrote "$bc = q$"; bc is a line in his diagram with length q.

6.3.1 Derivatives of trigonometric functions

Claim. $\tan' = \frac{1}{\cos^2} =: \sec^2$

Proof. $\tan = \frac{\sin}{\cos}$, so by the quotient rule

$$\tan' = \frac{\cos^2 + \sin^2}{\cos^2} = \frac{1}{\cos^2} = \sec^2 .$$

□

Claim. *What is the derivative of* $\sin^{-1}$*?*

Proof.

$$\frac{\mathrm{d}\, sin^{-1} a}{\mathrm{d}\, a} = \frac{\mathrm{d}\,\theta}{\mathrm{d}\sin\theta} = \frac{1}{\cos\theta} = \frac{1}{\sqrt{1-\sin^2\theta}} = \frac{1}{\sqrt{1-a^2}}$$

□

Claim. *What is the derivative of* $\tan^{-1}$*?*

Proof.

$$\frac{\mathrm{d}\tan^{-1}(a)}{\mathrm{d}\, a} = \frac{\mathrm{d}\,\theta}{\mathrm{d}\tan(\theta)} = \cos^2(\theta) = \cos^2(\tan^{-1} a)$$

Note that a right-angle triangle with angle $\tan^{-1} a$ has opposite length a relative to adjacent length 1. Therefore $\cos(\tan^{-1} a) = \frac{1}{\sqrt{1+a^2}}$.

Therefore the derivative of $\tan^{-1}(a)$ is $\frac{1}{1+a^2}$. □

6.4 Berkeley Math 53 (Frenkel)

6.4.1 Curves and surfaces

A function is a rule associating input values from one set with output values from another; a function is a set of (input, output) pairs in which each input value occurs at most once.

A curve in d dimensions is a set of d-dimensional points that form a "connected" 1-dimensional object.

A surface is a similar concept to a curve, but is 2-dimensional.

The dimensionality of an object is equal to the dimensionality of the ambient space, minus the number of independent equations.

6.4.2 Specifying a curve or surface

Cartesian equation: A curve can be specified as the set of points satisfying some condition (e.g. $x^2+y^2 = R^2$) or by specifying that one dimension records the value of a function whose inputs are the other dimensions ($z = 3 + 1.5(x-1) - 2.7(y-2)$).

Graph: Let f be $\mathbb{R} \to \mathbb{R}$. The graph of f is the set of points (x, y) satisfying $y = f(x)$. This defines a curve in 2D (which never "turns back on itself"; the tangent line to the curve is never vertical.)

A curve in 3D would require two equations (to reduce the dimensionality of the ambient space to that of the object being specified; i.e. the intersection of two surfaces). In practice, curves in 3D are usually specified in parametric form.

Parametric form: For a curve in 2D, suppose the x-coordinate is given by $f(t)$ and the y-coordinate by $g(t)$. Then the curve is the set of points $\big(f(t), g(t)\big)$ for some range of the parameter t. E.g. a line represented in parametric form using vector notation: $\mathbf{r} = \mathbf{r}_0 + \mathbf{v}t$. (A surface would require 2 parameters, so they are often specified using Cartesian equations.)

6.4.3 Area under a curve

What is the area A under the curve from $t = a$ to $t = b$? It's just $\int_\alpha^\beta y\,\mathrm{d}x$ as usual[3], but how do we express this as an integral with respect to t?

Well, $y = g(t)$; what about $\mathrm{d}x$? $x = f(t)$ (displacement), therefore $\mathrm{d}x = \mathrm{d}t f'(t)$ (velocity $\times$ time; local linear approximation). So, the area under the curve bounded by start and end t-values is $A = \int_a^b g(t)f'(t)\,\mathrm{d}t$.

Thus, if the x-coordinate is increasing rapidly with t, then the area is larger.

6.4.4 Length of a curve

The length of a curve is $L = \int \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$, over some interval.

This can be expressed as an integral with respect to x (non-parametric form): $L = \int_\alpha^\beta \sqrt{1 + (\frac{\mathrm{d}y}{\mathrm{d}x})^2}\,\mathrm{d}x$.

Or it can be expressed as an integral over an interval of t values (parametric form): $L = \int_a^b \sqrt{(\frac{\mathrm{d}x}{\mathrm{d}t})^2 + (\frac{\mathrm{d}y}{\mathrm{d}t})^2}\,\mathrm{d}t$

[3] $(\alpha, \beta) = (f(a), f(b))$

6.4.5 Area and volume of revolution of a curve

Suppose a curve is revolved around the x-axis.

Volume
This is computed as a sum of discs with width $\mathrm{d}x$:

$$V = \int_{x=\alpha}^{x=\beta} \pi y^2 \,\mathrm{d}x.$$

Area
This is computed as a sum of strips (using the hypotenuse rather than the rectangular strips used for the volume[4]):

$$A = \int_{x=\alpha}^{x=\beta} 2\pi y \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$$

6.4.6 Polar coordinates

E.g. the curve $r = \cos(\theta)$ is a circle of radius 1 centered at $(x, y) = (\frac{1}{2}, 0)$. (?)

Area of a sector bounded by a curve

What's the area of the sector bounded by the two rays and a curve, between $\theta = a$ and $\theta = b$?

Note that the area of a sector of ϕ radians of a circle is $\pi r^2 \times \frac{\phi}{2\pi} = \frac{1}{2}\phi r^2$.

We're considering a curve defined by $r = f(\theta)$. We divide it up into many sectors each with angle $\mathrm{d}\theta$. The area is $\int_a^b \frac{1}{2} f(\theta)^2 \,\mathrm{d}\theta$.

6.4.7 Surfaces

Planes

Given a normal vector $\mathbf{n} = \begin{bmatrix} d \\ e \\ f \end{bmatrix}$, and a point in the plane $P = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}$, an equation specifying the plane is

$$\begin{aligned} d(x - x_0) + e(y - y_0) + f(z - z_0) &= 0 \\ dx + ey + fz &= C. \end{aligned}$$

So the normal vector can be read off from the equation.

Similarly the general equation of a line in 2D is

[4] Why exactly do we construct these strips using the hypotenuse, whereas when approximating the area under a graph we construct rectangles $y\,\mathrm{d}x$? See
https://math.stackexchange.com/questions/1691147/why-is-surface-area-not-simply-2-pi-int-ab-y-dx-instead-of-2-pi-in
https://math.stackexchange.com/questions/1074986/surface-area-of-a-solid-of-revolution-why-does-not-int-ba-2-pi-fx
https://math.stackexchange.com/questions/12906/is-value-of-pi-4

$$d(x - x_0) + e(y - y_0) = 0,$$

(TODO: explain this and other content towards end of L11)

so $\begin{bmatrix} d \\ e \end{bmatrix}$ is a normal vector to the line.

Quadric surfaces

Ellipsoids, hyperboloids, paraboloids. Also cylinders (one variable not specified, e.g. $x^2 + y^2 = 1$), and cones (e.g. $z^2 = x^2 + y^2$).

6.4.8 Tangent spaces

Tangent lines

E.g. a tangent vector is given by differentiating the parametric equation for a curve, giving an equation for the tangent line:

$$\mathbf{r} = \mathbf{r}_0 + \begin{bmatrix} x'(t_0) \\ y'(t_0) \\ z'(t_0) \end{bmatrix} s = \mathbf{r}_0 + \mathbf{v}'s.$$

Tangent planes

$$(z - z_o) = (x - x_0)f_x(x_0, y_0) + (y - y_0)f_y(x_0, y_0)$$

And what's the normal vector to that tangent plane? It's $\begin{bmatrix} f_x(x_0, y_0) \\ f_y(x_0, y_0) \\ -1 \end{bmatrix}$.

6.4.9 Limits (L8)

$\frac{x^2}{x^2+y^2}$ has no limit at $(0, 0)$. Easy to prove by exhibiting paths with different limits: e.g. along x-axis vs. y-axis. Lack of limit related to degree of numerator and denominator being same.

But $\frac{2x^3}{x^2+y^2}$ does have a limit at $(0, 0)$.

Proof: consider a disk of radius r. For points in this disk, $x^2 + y^2 \leq r^2$ and so $x \leq r$. Now

$$\left| \frac{2x^3}{x^2 + y^2} \right| = 2|x| \left| \frac{x^2}{x^2 + y^2} \right| \leq 2r,$$

so for any desired closeness to the limiting value 0, we can find an r that will do it.

6.4.10 Partial derivatives (L8)

Clairot's theorem: equality of mixed partials under certain continuity conditions.

> "Same commutative structure as multiplication"; all that matters is how many times you have differentiated w.r.t. x, and to y; "differentiation is in a sense opposite to multiplication".

6.4.11 Differentials (L8)

> "The differential is the function whose graph the tangent line (plane) is, but with the coordinate axes shifted to the point at which it is being evaluated."

A differential, defined at a particular point in the input space, is the function describing the linear approximation at that point: it maps a displacement in the input space to a displacement in the output space.

It's the function whose graph is the tangent space at that point, in a coordinate space shifted to have its origin at that point. So in 1D, if $z = f(x)$, then the differential at x_0 is

$$\mathrm{d}z(x) = (x - x_0)f'(x_0).$$

Not to be confused with Δf — the increment in the *actual function* value — whereas the differential refers to the increment in the linear approximation.

6.4.12 Directional derivatives (L11)

TODO Note: Defining directional derivative as being a function of a *unit* vector is controversial; see e.g. https://math.stackexchange.com/questions/2291302/why-isnt-the-directional-derivative-generally-scaled-d
The majority view is, contra Stewart, that the directional derivative should be defined as a function of a vector of any magnitude. The interpretation of that is that it gives the rate of change of the function as you move past the point with velocity given by the vector u. One motivation is that this makes it linear in u: $dd(u + v) = dd(u) + dd(v)$ etc.

Theorem 25. *The directional derivative of $f(x, y)$ in the direction of a unit vector $u = \begin{bmatrix} a \\ b \end{bmatrix}$ is*

$$D_u f = a\frac{\partial f}{\partial x} + b\frac{\partial f}{\partial y} = \nabla f \cdot \mathbf{u}.$$

Proof. Since u is unit length, $\begin{bmatrix} ha \\ hb \end{bmatrix}$ is a displacement of length h in the direction of u. Then[5]

$$\begin{aligned} D_u f(x_0, y_0) &:= \lim_{h\to 0} \frac{f(x_0 + ha, y_0 + hb) - f(x_0, y_0)}{h} \\ &= \lim_{h\to 0} \frac{f(x_0, y_0) + ha\frac{\partial f}{\partial x}(x_0, y_0) + hb\frac{\partial f}{\partial y}(x_0, y_0) - f(x_0, y_0)}{h} \\ &= a\frac{\partial f}{\partial x}(x_0, y_0) + b\frac{\partial f}{\partial y}(x_0, y_0) \quad \square \end{aligned}$$

[5] The proof in the lecture and in Stewart is slightly different, involving defining these quantities as functions of h and considering the derivative w.r.t. h.

6.4.13 Gradient

$\nabla f(x_0, y_0)$ is normal to the level curve that cuts f at $z = z_0$.

Recall that $\begin{bmatrix} f_x(x_0, y_0) \\ f_y(x_0, y_0) \\ -1 \end{bmatrix}$ is a normal vector to the tangent plane at (x_0, y_0).

6.5 Linear and quadratic approximations to a function

[6]

We construct first- and second-order approximations to a differentiable function $f : \mathbb{R}^2 \to \mathbb{R}$. The approximation is made at some point $(x_0, y_0) = \mathbf{x}_0 \in \mathbb{R}^2$; we demand that the value of the approximation, and the first and second derivatives, match those of f exactly at that point.

6.5.1 Linear approximation to a function $f(x, y)$ near (x_0, y_0):

$$L(x, y) = f(x_0, y_0) + (x - x_0) f_x(x_0, y_0) + (y - y_0) f_y(x_0, y_0)$$

$$= f(\mathbf{x}) + (\mathbf{x} - \mathbf{x}_0) \cdot \nabla_f(\mathbf{x}_0)$$

Note that, at (x_0, y_0), the first partial derivatives of L are equal to those of f, as they must be. (In fact, we could say that the coefficients are determined by this requirement; see the quadratic case below. But the linear case is obvious without "deriving" the coefficients.)

6.5.2 Quadratic approximation to a function $f(x, y)$ near (x_0, y_0):

The j-th component of the gradient of $q(\mathbf{x}) = \mathbf{x}^\mathbf{T} A \mathbf{x}$ is $\frac{\partial q}{\partial x_j} = 2 \sum_k A_{jk} x_k$, so

$$\nabla \, \mathbf{x}^\mathbf{T} \mathbf{A} \mathbf{x} = 2 \mathbf{A} \mathbf{x}.$$

$$Q(x, y) = f(\mathbf{x}_0) + (x - x_0) f_x(\mathbf{x}_0) + (y - y_0) f_y(\mathbf{x}_0) + \frac{1}{2} f_{xx}(\mathbf{x}_0)(x - x_0)^2 + f_{xy}(\mathbf{x}_0)(x - x_0)(y - y_0) + \frac{1}{2} f_{yy}(\mathbf{x}_0)(y - y_0)^2$$

$$= f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) \cdot \nabla f(\mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\mathbf{T} \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0),$$

where $\nabla^2 f(\mathbf{x}_0)$ is the Hessian matrix $\begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix}$ evaluated at $\mathbf{x}_0$.

[6] khanacademy - Grant Sanderson - second partial derivative test

6.5.3 Second partial derivative test and positive definiteness of Hessian

The second partial derivative test for a function of two variables states that we examine the determinant of the Hessian evaluated at the critical point:

$$D = \det \nabla^2 f(\mathbf{x}_0) = f_{xx}(\mathbf{x}_0) f_{yy}(\mathbf{x}_0) - f_{xy}(\mathbf{x}_0)^2.$$

Notice that $D \geq 0$ implies that the sign of f_{xx} and f_{yy} agree (because we're subtracting the square of the mixed partial f_{xy}, i.e. a positive number).

D	roots	f_{xx}		Hessian
$+$	no real roots	$+$	minimum	positive definite
$+$	no real roots	$-$	maximum	negative definite
0	one real root	$+$	minimum	positive semidefinite
0	one real root	$-$	maximum	negative semidefinite
$-$	two real roots	n/a	saddle point	-

Explanation

At a critical point $\mathbf{x}_0$, the gradient is zero and the quadratic approximation is therefore

$$Q(x, y) = f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^{\mathbf{T}} \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

So if this is a minimum (concave-up paraboloid) then this quadratic form is positive for all $\mathbf{x} \neq \mathbf{x}_0$ (and if it's a maximum then it's negative for all $\mathbf{x} \neq \mathbf{x}_0$).

Basically the argument is that, instead of analyzing the function f itself, we analyze its quadratic approximation at the critical point. So the question comes down to: how do we determine whether a quadratic form is always positive, always negative, or takes positive and negative values?

To answer that, consider a generic quadratic form $ax^2 + 2bxy + cy^2$. Let y be constant at y_0; then we have a quadratic in x, the roots of which are

$$x = \frac{-2by_0 \pm \sqrt{4b^2y_0^2 - 4acy_0^2}}{2a} = y_0 \frac{-b \pm \sqrt{b^2 - ac}}{a}.$$

So, whether this is a saddle point or a minimum/maximum depends on whether the quadratic form has real roots. If there are no real roots, then whether it's a minimum or a maximum depends on the sign of f_{xx} (this sign will be the same as that of f_{yy} in the no real roots case).

6.5.4 Derivation of quadratic approximation coefficients

$$\begin{aligned} Q(x, y) =& f(\mathbf{x}_0) + (x - x_0) f_x(\mathbf{x}_0) + (y - y_0) f_y(\mathbf{x}_0) + \\ & a(x - x_0)^2 + b(x - x_0)(y - y_0) + c(y - y_0)^2 \end{aligned}$$

What are the coefficients a, b, c? They are determined by the requirement that the second partial derivatives are identical at the point of approximation $\mathbf{x}_0$.

First look at the first partial derivatives:

$$Q_x = f_x(\mathbf{x}_0) + 2a(x - x_0) + b(y - y_0)$$
$$Q_y = f_y(\mathbf{x}_0) + b(x - x_0) + 2c(y - y_0)$$

so the quadratic approximation is an exact first-order approximation at $\mathbf{x}_0$, as required:

$$Q_x(\mathbf{x}_0) = f_x(\mathbf{x}_0)$$
$$Q_y(\mathbf{x}_0) = f_y(\mathbf{x}_0),$$

Now look at the second derivatives:

$$Q_{xx} = 0 + 2a + 0$$
$$Q_{xy} = 0 + 0 + b$$
$$Q_{yx} = 0 + b + 0$$
$$Q_{yy} = 0 + 0 + 2c$$

Since we require that these match those of f exactly at $\mathbf{x}_0$, we have

$$a = \frac{1}{2} f_{xx}(\mathbf{x}_0)$$
$$b = f_{xy}(\mathbf{x}_0) = f_{yx}(\mathbf{x}_0)$$
$$c = \frac{1}{2} f_{yy}(\mathbf{x}_0),$$

so the quadratic approximation is

$$\begin{aligned} Q(x, y) = & f(\mathbf{x}_0) + (x - x_0) f_x(\mathbf{x}_0) + (y - y_0) f_y(\mathbf{x}_0) + \\ & \frac{1}{2} f_{xx}(\mathbf{x}_0)(x - x_0)^2 + f_{xy}(\mathbf{x}_0)(x - x_0)(y - y_0) + \frac{1}{2} f_{yy}(\mathbf{x}_0)(y - y_0)^2 \end{aligned}$$

6.6 Oxford M5 Multivariable calculus

6.6.1 Integrals in two dimensions

Example (5).

Proof.

$$\begin{aligned}\int\int_R (x+y^2)\,\mathrm{d}x\,\mathrm{d}y &= \int_1^3 \int_0^2 (x+y^2)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_1^3 \left(\frac{x^2}{2}+xy^2\right)\Big|_{x=0}^{x=2} \mathrm{d}y \\ &= \int_1^3 2+2y^2\,\mathrm{d}y \\ &= 2y+\frac{2y^3}{3}\Big|_1^3 \\ &= 6+18-2-\frac{2}{3} \\ &= \frac{64}{3}\end{aligned}$$

□

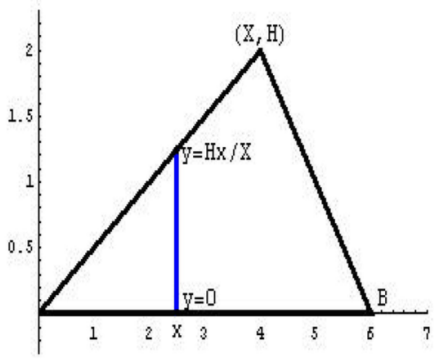
Example (6). Let R be the unit square. Determine $\int\int_R y\cos^2(\pi xy)dA$.

TODO

Proof.

$$\int\int_R y\cos^2(\pi xy)dA = \int_0^1 \int_0^1 y\cos^2(\pi xy)\,\mathrm{d}x\,\mathrm{d}y$$
$$= \qquad\qquad\qquad\qquad =$$

□



Example 7 *Calculate the area of the triangle with vertices* $(0,0)$, $(B,0)$ *and* (X,H).

Proof. (I. sum of two one-dimensional integrals) □

Proof. (II. sum of two integrals over area)
The triangle is composed of a piecewise linear function:

$$\text{height}(x) = \begin{cases} x\frac{H}{X}, & 0 \le x \le X \\ (x-B)\frac{H}{(X-B)}, & X < x \le B. \end{cases}$$

$$\begin{aligned}
\text{area} &= \int\int_R dA \\
&= \int_0^X \int_0^{xH/X} \mathrm{d}y\,\mathrm{d}x + \int_X^B \int_0^{(x-B)\frac{H}{(X-B)}} \mathrm{d}y\,\mathrm{d}x \\
&= \int_0^X xH/X\,\mathrm{d}x + \int_X^B (x-B)\frac{H}{(X-B)}\,\mathrm{d}x \\
&= \frac{H}{X}\int_0^X x\,\mathrm{d}x + \frac{H}{(X-B)}\int_X^B (x-B)\,\mathrm{d}x \\
&= \frac{HX^2}{2X} + \frac{H}{(X-B)}\Big[\frac{(x-B)^2}{2}\Big]_X^B \\
&= \frac{HX}{2} - \frac{H}{(X-B)}\frac{(X-B)^2}{2} \\
&= \frac{HX}{2} - \frac{H(X-B)}{2} \\
&= \frac{BH}{2}
\end{aligned}$$

□

6.6.2 Change of variables and Jacobians

Let $R, S \subset \mathbb{R}$ and $u : R \to S$.

Define $\psi : R \to \mathbb{R}$ and $\Psi : S \to \mathbb{R}$, such that $\Psi(f(x)) = \psi(x)$ for all $x \in R$.

One definition of the integral is to divide R into segments of length δx, let ψ_i be the value of ψ at the start of the i-th segment, and define

$$\int_{x \in R} \psi(x) \, \mathrm{d}x = \lim_{\delta x \to 0} \sum_i \psi_i \delta x.$$

Now let u_i' be the value of the derivative at the start of the i-th line segment.

Then the length of the i-th segment of S is $u_i' \ \delta x$.

Therefore the integral over S is

$$\begin{aligned} \int_{u \in S} \Psi(u) \, \mathrm{d}u &= \lim_{\delta x \to 0} \sum_i \psi_i u_i' \ \delta x \\ &= \int_{x \in R} \psi(x) \frac{\mathrm{d}u}{\mathrm{d}x} \, \mathrm{d}x. \end{aligned}$$

Theorem (Integration by substitution)**.**
Let $u = h(x)$. Then

$$\int g(h(x))h'(x)\,\mathrm{d}x = \int g(u)\,\mathrm{d}u.$$

Proof. Let $G' = g$, i.e. G is an antiderivative of g.

Recall the chain rule:

$$(G \circ h)' = G'h'$$

Integrating both sides with respect to x gives

$$G \circ h + C = \int G'h'\,\mathrm{d}x = \int gh'\,\mathrm{d}x.$$

Let $u = h(x)$. Then

$$G(u) + C = \int g(u)\,\mathrm{d}u = \int \frac{\mathrm{d}G}{\mathrm{dh}}\frac{\mathrm{d}u}{\mathrm{d}x}\,\mathrm{d}x.$$

□

Proof. Let $G' = g$, i.e. G is an antiderivative of g.

Recall the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}x}G(h(x)) = \frac{\mathrm{d}G}{\mathrm{dh}}\frac{\mathrm{dh}}{\mathrm{d}x}.$$

Integrating both sides with respect to x gives

$$G(h(x)) + C = \int \frac{\mathrm{d}G}{\mathrm{dh}}\frac{\mathrm{dh}}{\mathrm{d}x}\,\mathrm{d}x.$$

Let $u = h(x)$. Then

$$G(u) + C = \int g(u)\,\mathrm{d}u = \int \frac{\mathrm{d}G}{\mathrm{dh}}\frac{\mathrm{d}u}{\mathrm{d}x}\,\mathrm{d}x.$$

□

Definition (Jacobian)**.** *Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $f(x, y) := (u(x, y), v(x, y))$.*

The Jacobian of f is $\frac{\partial(u,v)}{\partial(x,y)} = \det\begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} = \det\begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix}$.

It is defined analogously in 3D.

Theorem 26. *The Jacobian of a map is the factor by which the map stretches space locally.*

Proof. (Sketch)
Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a differentiable function given by $(x, y) \mapsto \Big(u(x, y), v(x, y)\Big)$.

Consider a small rectangular area with bottom-left corner (x, y) and top-right corner $(x + \delta x, y + \delta y)$.

Let u_x, u_y, v_x, v_y be the partial derivatives evaluated at (x, y).

The linear approximation to f at (x, y) is

$$f(x, y) \approx f(x, y) + \begin{bmatrix} u_x \delta x + u_y \delta y \\ v_x \delta x + v_y \delta y \end{bmatrix}$$

So the bottom-right and top-left corners are mapped as follows:

$$\text{bottom right:} \quad (x, y) \mapsto f(x, y) + \begin{bmatrix} u_x \delta x \\ v_x \delta x \end{bmatrix}$$
$$\text{top left:} \quad (x, y) \mapsto f(x, y) + \begin{bmatrix} u_y \delta y \\ v_y \delta y \end{bmatrix}$$

Thus the image of the original rectangular area is a parallelogram spanned by the vectors $\delta x \begin{bmatrix} u_x \\ v_x \end{bmatrix}$ and $\delta y \begin{bmatrix} u_y \\ v_y \end{bmatrix}$. The area of this parallelogram is given by the cross product:

$$\text{area} = \left| \delta x \begin{bmatrix} u_x \\ v_x \end{bmatrix} \times \delta y \begin{bmatrix} u_y \\ v_y \end{bmatrix} \right| = |(u_x v_y - u_y v_x)\mathbf{k}| \delta x \delta y = \det \begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix} \delta x \delta y.$$

□

Example. Let $x = r\cos\theta$ and $y = r\sin\theta$, where r and θ are polar co-ordinates. Then

$$\begin{aligned} \frac{\partial(x, y)}{\partial(r, \theta)} &= \det \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \\ &= r(\cos^2\theta + \sin^2\theta) \\ &= r. \end{aligned}$$

Example. In reverse, $r(x, y) = \sqrt{x^2 + y^2}$ and $\theta(x, y) = \tan^{-1}(y/x)$.

Note that $\frac{\partial\theta}{\partial x} = \frac{1}{1+\frac{y^2}{x^2}} \frac{-y}{x^2} = \frac{-y}{x^2+y^2}$, and $\frac{\partial\theta}{\partial y} = \frac{1}{1+\frac{y^2}{x^2}} \frac{1}{x} = \frac{x}{x^2+y^2}$.

So

$$\begin{aligned} \frac{\partial(r, \theta)}{\partial(x, y)} &= \det \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{pmatrix} \\ &= \frac{x^2 + y^2}{(x^2 + y^2)^{2/3}} \\ &= \frac{1}{r}. \end{aligned}$$

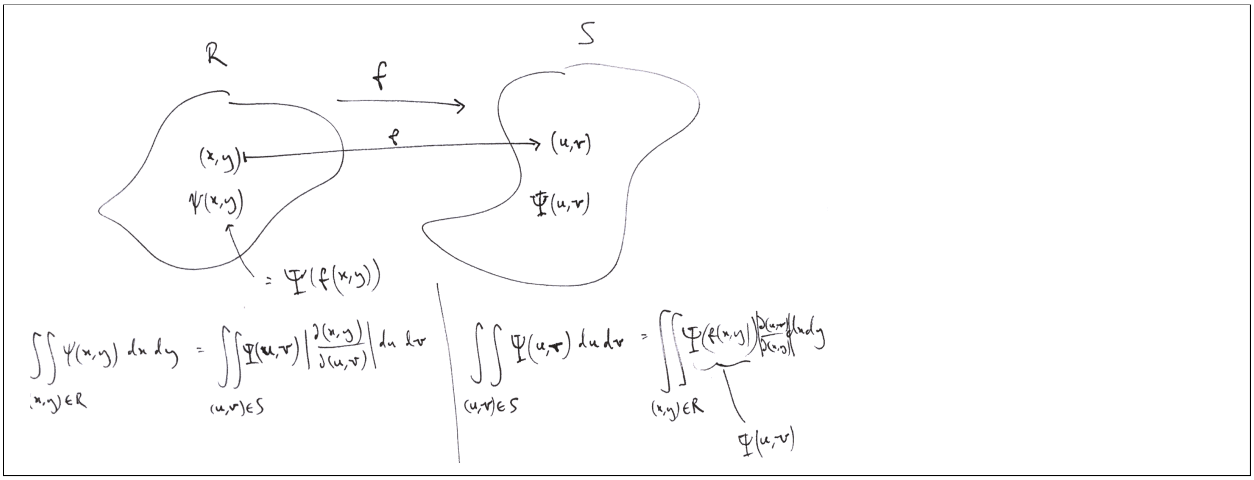
Theorem.
Let:

1. $R, S \subseteq \mathbb{R}^2$
2. $f : R \to S$ *given by* $f(x, y) = (u(x, y), v(x, y))$

3. $\psi(x,y) = \Psi(u,v) = \Psi(u(x,y), v(x,y))$.

Then

$$\int\int_{(x,y)\in R} \psi(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int\int_{(u,v)\in S} \Psi(u,v)\left|\frac{\partial(x,y)}{\partial(u,v)}\right|\mathrm{d}u\,\mathrm{d}v$$
$$\int\int_{(u,v)\in S} \Psi(u,v)\,\mathrm{d}u\,\mathrm{d}v = \int\int_{(x,y)\in R} \psi(x,y)\left|\frac{\partial(u,v)}{\partial(x,y)}\right|\mathrm{d}x\,\mathrm{d}y.$$



Intuition. If f stretches space locally, then a local value $\psi(x,y)$ over R contributes more when accumulating Ψ values over S.

Proof. (Sketch)
Divide R into N small squares.

Let u_x, u_u, v_x, v_y be the partial derivatives evaluated at the center of the i-th square.

Note from theorem (26) above that the image of the i-th square is a parallelogram with area $\begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} \delta x \delta y$.

An approximation for the integral over S is

$$\begin{aligned}\int\int_{(u,v)\in S} \Psi(u,v)\,\mathrm{d}u\,\mathrm{d}v &\approx \sum_i \Psi_i \mathrm{Area}(\mathrm{Parallelogram}_i) \\ &= \sum_i \psi_i \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} \delta x \delta y,\end{aligned}$$

which on taking the limit $N \to \infty$ gives

$$\int\int_{(u,v)\in S} \Psi(u,v)\,\mathrm{d}u\,\mathrm{d}v = \int\int_{(x,y)\in R} \psi(x,y)\left|\frac{\partial(u,v)}{\partial(x,y)}\right|\mathrm{d}x\,\mathrm{d}y.$$

□

Exercise 13 *Evaluate*

$$\iint_{\mathbb{R}^2} \exp[-(x^2+y^2)]\mathrm{d}A.$$

Hence, determine $\int_{-\infty}^{\infty} \exp[-p^2]\mathrm{d}p$.

First of all, note that $\int_{-\infty}^{\infty} e^{-x}\,\mathrm{d}x$ does not converge, and that it is not obvious how to calculate $\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x$.

Proof. Let $f(x,y) = (r,\theta) = \left(\sqrt{x^2+y^2}, \tan^{-1}(y/x)\right)$.

$$\int\int_{\mathbb{R}^2} \exp\{-(x^2+y^2)\}\,\mathrm{d}A = \int\int_{(r,\theta)} \exp\{-\}\,\mathrm{d}A$$

□

Proof. □

6.7 3blue1brown - Essence of Calculus

6.7.1 The paradox of the derivative

6.7.2 Derivatives formulas through geometry

6.7.3 Visualizing the chain rule and product rule

More complex functions can be formed by addition, multiplication and composition of simpler functions. How do we compute derivatives of such more complex functions?

6.7.4 Sum rule

Suppose $f(x) = g(x) + h(x)$. Visualize an input parameter x represented by the x-axis, and the graphs of g and h, and a third graph of f whose height at every point is the sum of the other two.

A horizontal nudge $\mathrm{d}x$ to the input causes vertical changes $\mathrm{d}\,g(x)$ and $\mathrm{d}\,h(x)$. The resulting vertical change to $f(x)$ is

$$\mathrm{d}\,f(x) = \mathrm{d}\,g(x) + \mathrm{d}\,h(x),$$

or equivalently

$$\frac{\mathrm{d}\,f(x)}{\mathrm{d}x} = \frac{\mathrm{d}\,g(x)}{\mathrm{d}x} + \frac{\mathrm{d}\,h(x)}{dx}.$$

6.7.5 Product rule

Suppose $f(x) = g(x) \cdot h(x)$. Consider an input parameter x and visualize a rectangle with one side length $g(x)$ and the other side length $h(x)$. $f(x)$ is the area of the rectangle.

A nudge $\mathrm{d}x$ to the input causes the sides to grow by $\mathrm{d}\,g(x)$ and $\mathrm{d}\,h(x)$ respectively. Therefore the change to the area is approximately

$$\mathrm{d}\,f(x) = h(x)\,\mathrm{d}\,g(x) + g(x)\,\mathrm{d}\,h(x),$$

or equivalently

$$\frac{\mathrm{d}\,f(x)}{\mathrm{d}x} = h(x)\frac{\mathrm{d}\,g(x)}{dx} + g(x)\frac{\mathrm{d}\,h(x)}{dx}.$$

6.7.6 Integration by Parts

TODO: graphical intuition (see wikipedia page)

6.7.7 Chain rule: function composition

Suppose $f(x) = g(h(x))$. Visualize 3 real number lines: at the top the input parameter x; in the middle $h(x)$ and at the bottom $g(h)$.

A nudge $\mathrm{d}x$ to the input causes a change $\mathrm{d}\,h = \frac{\mathrm{d}\,h}{\mathrm{d}x}\,\mathrm{d}x$, which in turn causes a change $\mathrm{dg} = \frac{\mathrm{dg}}{\mathrm{dh}}\,\mathrm{dh}$. So we have

$$\mathrm{df} = \mathrm{d}\,g(h(x)) = \frac{\mathrm{dg}}{\mathrm{dh}}\frac{\mathrm{dh}}{\mathrm{d}x}\,\mathrm{d}x,$$

or equivalently

$$\frac{\mathrm{df}}{\mathrm{d}x} = \frac{\mathrm{d}\,g(h(x))}{\mathrm{d}x} = \frac{\mathrm{dg}}{\mathrm{dh}}\frac{\mathrm{dh}}{\mathrm{d}x}.$$

Example

$f(x) = \sin(x^2) = g(h(x))$. So the middle number line shows $h(x) = x^2$ and the output number line at the bottom shows $g(h) = \sin(h)$.

We know that for the outer function, $\mathrm{dg} = \cos h\,\mathrm{dh}$, and for the inner function $\mathrm{dh} = 2x\,\mathrm{d}x$, so

$$\mathrm{d}\,g(h(x)) = \cos(h) \cdot 2x\,\mathrm{d}x = \cos(x^2) \cdot 2x\,\mathrm{d}x.$$

6.7.8 Implicit differentiation

Consider the circle defined by $x^2 + y^2 = 5$. Here, on the face of it, we don't have a function with input and an output; we just have a set of points in 2D defined by some condition which they satisfy (an implicit curve).

How do we find the tangent to the circle at the point $(3, 4)$? We want $\frac{\mathrm{d}y}{\mathrm{d}x}$.

Consider a related problem. A ladder of length 5 is leaned against a wall, with initial height 4, and is slipping down at 1 m/s. Define $y(t)$ to be its height at time t, so $\frac{\mathrm{d}y}{\mathrm{d}t} = -1$. What is $\frac{\mathrm{d}x}{\mathrm{d}t}$?

Clearly the starting point is that $x(t)^2 + y(t)^2 = 5^2$. One solution is

$$x(t) = \left(5^2 - y(t)^2\right)^{1/2}$$

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{-2y\frac{\mathrm{d}y}{\mathrm{d}t}}{2\left(5^2 - y(t)^2\right)^{1/2}} = \frac{y}{x}.$$

Another solution is to note that the sum of the squares is constant:

$$\frac{\mathrm{d}\left(x(t)^2 + y(t)^2\right)}{\mathrm{d}t} = 0$$

$$2x\,\mathrm{d}x + 2y\,\mathrm{d}y = 0$$

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{y\,\mathrm{d}y}{x\,\mathrm{d}t} = \frac{y}{x}.$$

In the case of the ladder problem, it was clear what was going on since we could differentiate $x(t)^2 + y(t)^2$ with respect to t.

Going back to the implicit curve $x^2 + y^2 = 5$, there is in fact a function there: a function of two variables:

$$z(x, y) = x^2 + y^2$$

We want $\frac{\mathrm{d}y}{\mathrm{d}x}$. What is $\frac{\mathrm{d}y}{\mathrm{d}x}$? It's a ratio of two nudges to the two input variables. OK, but those nudges could be anything; the ratio is not determined. But we have a condition: the two nudges must stay on a tangent line to the circle. So,

$$\begin{aligned}(x + \mathrm{d}x)^2 + (y + \mathrm{d}y)^2 &= 5\\ x^2 + 2x\,\mathrm{d}x + y^2 + 2y\,\mathrm{d}y &= 5\\ 2x\,\mathrm{d}x + 2y\,\mathrm{d}y &= 0\\ \frac{\mathrm{d}y}{\mathrm{d}x} &= -\frac{x}{y}\end{aligned}$$

The derivative of z in the direction of the vector $\mathbf{u} = \begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix}$ is

$$\begin{aligned} \mathrm{D}_u z &= \frac{\partial z}{\partial x}\,\mathrm{d}x + \frac{\partial z}{\partial y}\,\mathrm{d}y \\ &= 2x\,\mathrm{d}x + 2y\,\mathrm{d}y. \end{aligned}$$

And the condition for staying on the tangent to the circle is that z stays constant:

$$\begin{aligned} \mathrm{D}_u z &= 2x\,\mathrm{d}x + 2y\,\mathrm{d}y = 0 \\ \frac{\mathrm{d}y}{\mathrm{d}x} &= \frac{-x}{y}. \end{aligned}$$

$$y^2 \sin x = x$$

Chapter 7

Differential Equations

Let y be the position of a particle in one dimension, and let t be time. So there is just a single input variable: t.

An Ordinary Differential Equation is an equation relating the input variable t to y and its derivatives. So, in general,

$$f\left(t, y, y', y'', \ldots\right) = 0.$$

A first-order ODE involves first derivatives only.

Consider the subset[1] of first-order ODEs that specify a velocity $v(t, y)$ at each point in (t, y) space. Thus this ODE contains all the information needed to animate the motion of the particle, starting from any point (t_0, y_0). So the statement of the initial condition problem is

$$\frac{\mathrm{d}y}{\mathrm{d}t} = v(t, y) \qquad\qquad y(t_0) = y_0.$$

The solution to an ODE is a function $y = \varphi(t)$ that describes a motion of the particle having the specified velocities at each point it passes through. I.e., if $y = \varphi(t)$ is a solution, then

$$\frac{\mathrm{d}\,\varphi}{\mathrm{d}t} = v\Big(t, \varphi(t)\Big) \qquad\qquad \text{for all } t.$$

We can think of v as a surface over the (t, y) plane. A solution is a curve in the plane whose derivative is equal to the height of the surface v, at every point on the curve.

The phase space of this problem is the set of all possible (y, v) values.?

7.1 Taxonomy

7.1.1 Linear DEs

A **linear DE** can be written as $Ly = f$, where L is a linear operator. The domain of f is the same as the domain of y.

Basically this means that derivatives of y of any degree may appear, but they may not be multiplied together. I.e.

$$\sum_{n=0}^{d} P_i(t) \frac{\mathrm{d}^n y}{\mathrm{d}t^n}(t) = Q(t)$$

is a linear DE of degree d. (The 0-th derivative is the function y itself.) The $P_i(t)$ and $Q(t)$ may be any (?) functions of the independent variable.

Theorem: Linear combinations of solutions of linear DEs are themselves solutions.

If $Q(t) = 0$ then it is a **homogeneous linear DE**.

[1] I.e. $f(t, y, y') = 0$ can be rearranged to give y' as a function of t, y.

7.1.2 First-order linear DEs: integrating factors

A **first-order linear DE** can be written in the form

$$y'(t) + P(t)y(t) = Q(t).$$

First-order linear DEs can be solved by use of an **integrating factor**: we seek $I(t)$ such that

$$\Big(I(t)y(t)\Big)' = I(t)\Big(y'(t) + P(t)y(t)\Big),$$

since then

$$y(t) = \frac{1}{I(t)} \int I(t)Q(t)\,\mathrm{d}t + C.$$

To find I, we want:

$$\Big(I(t)y(t)\Big)' = I(t)\Big(y'(t) + P(t)y(t)\Big),$$

i.e.

$$\begin{aligned} I(t)y'(t) + I'(t)y(t) &= I(t)y'(t) + I(t)P(t)y(t) \\ I'(t) &= I(t)P(t) \\ \int \frac{1}{I(t)}\,\mathrm{d}\,I &= \int P(t)\,\mathrm{d}t \\ I &= Ae^{\int P(t)\,\mathrm{d}t}, \end{aligned}$$

so we use $I(t) = e^{\int P(t)\,\mathrm{d}t}$.

7.2 Special cases

7.2.1 Velocity depends on time only

$$\frac{\mathrm{d}y}{\mathrm{d}t} = v(t)$$

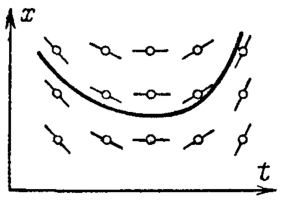


Fig. 4. A field invariant with respect to vertical translations

To find functions that solves this, one possibility is that we can find the antiderivative explicitly:

$$\int \frac{\mathrm{d}y}{\mathrm{d}t}\,\mathrm{d}t := y(t) + C = \int v(t)\,\mathrm{d}t.$$

7.2.2 Velocity depends on location only (autonomous)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = v(y)$$

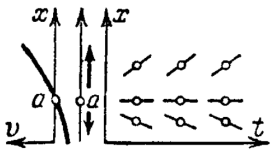


Fig. 6. The vector field and the direction field for the equation $\dot{x} = v(x)$

7.3 Examples

7.3.1 C^{14} dating

In a living organism the amount of C^{14}, as a proportion of all the C^{12} and C^{14}, is expected to be a known constant p_0. After death, C^{14} decays to C^{12}. How old is a specimen with proportion p_1 of C^{14}?

Let λ be the rate at which one atom of C^{14} decays in atoms/sec. So in a sample of N atoms, the expected number to decay in one second is $N\lambda$.

Let $N(t)$ be the number of C^{14} atoms remaining at time t. We can specify the model as a first-order ODE:

$$\frac{\mathrm{d}\,N}{\mathrm{d}t} = -N\lambda.$$

Equivalently, dividing by the constant total number of carbon atoms,

$$\frac{\mathrm{d}\,p}{\mathrm{d}t} = -p\lambda,$$

where $p(t)$ is the proportion of C^{14} at time t.

It's easy to find a family of functions $p(t)$ that satisfies this differential equation. Since

$$\frac{1}{p(t)}\frac{\mathrm{d}\,p}{\mathrm{d}t} = -\lambda,$$

it must be the case that their antiderivatives are the same, up to a constant:

$$\log(p(t)) = -\lambda t + C$$
$$p(t) = Ae^{-\lambda t}.$$

Further, the expected proportion in a living organism determines a particular function as the solution:

$$p(0) = p_0 = Ae^{-\lambda.0}$$

so $A = p_0$ and the solution is

$$p(t) = p_0 e^{-\lambda t}.$$

So the estimated age of a sample with proportion p_1 is

$$t = \frac{1}{\lambda} \log\left(\frac{p_0}{p_1}\right).$$

Lemma 27. *[Replacement Lemma] Suppose $f : [a, b] \to \mathbb{R}$ is continuous. Then for $x \in [a, b]$*

$$\int_a^t \left(\int_a^{\tau'} f(\tau)\, d\tau \right) d\tau' = \int_a^t (t - \tau) f(\tau)\, d\tau.$$

7.4 Integral equations

Consider the differential equation

$$y''(t) + \lambda y(t) = g(t),$$

where $y, g : [0, L] \to \mathbb{R}$, g is a known continuous function, and $\lambda > 0$.

Integration from 0 to t once gives

$$y'(t) - y'(0) + \lambda \int_0^t y(\tau)\, \mathrm{d}\tau = \int_0^t g(\tau)\, \mathrm{d}\tau,$$

and a second time gives

$$y(t) - y(0) - y'(0)t + \lambda \int_0^t \int_0^{\tau'} y(\tau)\, \mathrm{d}\tau\, \mathrm{d}\tau' = \int_0^t \int_0^{\tau'} g(\tau)\, \mathrm{d}\tau\, \mathrm{d}\tau'.$$

By the Replacement Lemma (27),

$$y(t) - y(0) - y'(0)t + \lambda \int_0^t (t - \tau) y(\tau)\, \mathrm{d}\tau = \int_0^t (t - \tau) g(\tau)\, \mathrm{d}\tau.$$

Now impose the initial conditions $y(0) = 0$, $y'(0) = v_0$. Then

$$y(t) = f(t) - \lambda \int_0^t (t - \tau) y(\tau)\, \mathrm{d}\tau,$$

where $f(t)$ is the known function

$$f(t) = v_0 t + \int_0^t (t - \tau) g(\tau)\, \mathrm{d}\tau.$$

[1] Collins, Differential Equations, ch. 1

7.5 Picard's Existence Theorem

Consider again the initial value problem

$$\frac{\mathrm{d}y}{\mathrm{d}t} = v(t, y) \qquad y(t_0) = y_0.$$

The ODE could also be written as

$$y(t) = \int v\Big(t, y(t)\Big)\,\mathrm{d}t + C,$$

but this is merely an equivalent restatement, since the definition of indefinite integral is antiderivative. If we can find an antiderivative, then fine. If not, note that by FTC, the following definite integral describes a solution:

$$y(t) = y(t_0) + \int_{t_0}^{t} v\Big(\tau, y(\tau)\Big)\,\mathrm{d}\,\tau.$$

But this specifies $y(t)$ in terms of itself, since the velocity v depends not only on t but also on the current position[2].

7.5.1 Definition: Lipschitz condition

$v(t, y)$ is Lipschitz in the y direction if there exists an upper bound L on the absolute value of the straight line slope between any two points lying on a vertical line. I.e. $\exists L > 0$ such that

$$\Big|v(t, y_1) - v(t, y_0)\Big| \leq L\Big|y_1 - y_0\Big|$$

for all pairs of points (t, y_0), $v(t, y_1)$.

7.5.2 Theorem: Picard's existence theorem

Let R be a rectangle of width $2h$ and height $2k$ and let (t_0, y_0) be the center of the rectangle. Suppose

1. Within R, $v(t, y)$ is continuous, with $|v(t, y)| \leq M$
2. $Mh \leq k$
3. Within R, $v(t, y)$ is Lipschitz in the y direction, with bound L on the absolute value of the straight line slope between any two points.

Then the initial value problem

$$\frac{\mathrm{d}y}{\mathrm{d}t} = v(t, y) \qquad y(t_0) = y_0$$

has a unique solution in R.

7.5.3 Examples

In these cases, $|y'|$ and $\frac{\partial v}{\partial y}$ are bounded in any rectangle.

[2] for example, the rate of change of the proportion on carbon-14 depends on the current proportion of carbon-14.

A

$y' = v(x, y) = x^2 + y^2 \qquad y(0) = 0$

So it can be approximated by Picard iterates. Is an explicit solution possible here?

B

$y' = (1 - 2x)y \qquad y(0) = 1$

This can be solved explicitly by separation-of-variables:

$$\log(y) = x - x^2 + C$$
$$y = Ae^{x(1-x)}.$$

7.5.4 Non-examples

$|y'|$ is bounded in any rectangle for all these examples. However, $\frac{\partial v}{\partial y}$ is not. Picard's theorem guarantees unique solutions only in rectangles excluding such problematic points.

A

$y' = v(x, y) = 3y^{2/3} \qquad y(0) = 0$

$\frac{\partial v}{\partial y} = 2y^{-1/3} \to \pm\infty$ at $y = 0$.

B

$y' = v(x, y) = x^2 y^{1/5} \qquad y(0) = b$

$\frac{\partial v}{\partial y} = \frac{1}{5}x^2 y^{-4/5}$ which is not defined at $y = 0$.

C

$y' = v(x, y) = y^2 \qquad y(0) = 1$

$\frac{\partial v}{\partial y} = 2y$, so seems like it should be fine. Solve by separation-of-variables:

$$\int y^{-2} y' \, \mathrm{d}x = x + C$$
$$-y^{-1} = x + C$$
$$y = \frac{1}{C - x}.$$

The solution passing through the initial value $y(0) = 1$ is

$$y = \frac{1}{1 - x},$$

which does not exist for all x in the rectangle.

7.5.5 Gronwall's inequality

Theorem (Gronwall's inequality)**.** *Let t_0, Y_0 and λ be known constants, and let Y be a non-negative continuous function. Suppose that*

$$Y(t) \leq Y_0 + \lambda \left| \int_{t_0}^{t} Y(\tau)\, d\tau \right|.$$

Then

$$Y(t) \leq Y_0 e^{\lambda |t - t_0|}.$$

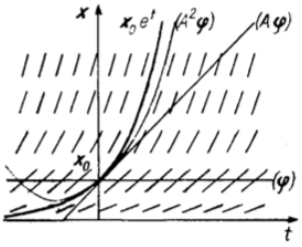


Fig. 217. The Picard approximation for the equation $\dot{x} = x$

Remark. I think something close to the following is true[3]: The above diagram from Arnold is strongly suggestive of Gronwall's inequality. If the Picard successive approximation procedure maps a function φ onto itself, then that's a solution. The only other possibility is that it maps φ onto a function $A\varphi$ which is strictly greater. In that case, φ is bounded above by the true solution.

Remark. If we had equality instead of the inequality, then differentiation would give

$$Y'(t) = \lambda Y(t).$$

This is the just the differential equation version of the integral equation with which we started. Recall that the general form of an ODE is

$$Y'(t) = v(t, Y(t)),$$

so here we have $v(t, Y(t)) = AY(t)$. In other words, the direction field does not depend on t, as in Arnold's diagram.

The solution to this ODE, with initial state $Y(t_0) = Y_0$, is

$$Y(t) = Y_0 e^{\lambda(t - t_0)}.$$

So Gronwell's inequality is saying that Y is bounded above by the solution to the differential equation that results from replacing the inequality with equality.

Proof. TODO. Uses an integrating factor e^{-At}. □

[3] I think this is being careless about sign; note that Gronwall's inequality concerns a non-negative function, like the absolute value of a solution to a DE.

7.5.6 Continuous dependence of solution on initial state

Consider y and z, respectively solutions to two different IVPs[4]. The IVPs specify the same DE but different initial state:

$$y(t_0) = y_0$$
$$z(t_0) = z_0.$$

We are interested in how the difference between the solutions depends on the difference $|y_0 - z_0|$ in initial state. We have

$$y(t) - z(t) = y_0 - z_0 + \int_{t_0}^{t} v(\tau, y(\tau)) - v(\tau, z(\tau)) \,\mathrm{d}\,\tau,$$

therefore, for $t > t_0$,

$$\begin{aligned}\Big|y(t) - z(t)\Big| &\leq \Big|y_0 - z_0\Big| + \int_{t_0}^{t} \Big|v(\tau, y(\tau)) - v(\tau, z(\tau))\Big| \,\mathrm{d}\,\tau \\ &\leq \Big|y_0 - z_0\Big| + \int_{t_0}^{t} L\Big|y(\tau) - z(\tau)\Big| \,\mathrm{d}\,\tau,\end{aligned}$$

and by Gronwall's inequality

$$\begin{aligned}\Big|y(t) - z(t)\Big| &\leq \Big|y_0 - z_0\Big| e^{L(t-t_0)} \\ &\leq \Big|y_0 - z_0\Big| e^{Lh}\end{aligned}$$

Therefore the solutions depend continuously on the initial state since, for arbitrary $\epsilon > 0$,

$$\Big|y_0 - z_0\Big| < e^{-Lh}\epsilon \implies \Big|y(t) - z(t)\Big| < \epsilon.$$

7.5.7 Contraction mapping theorem

Definition (Contraction)**.** *Let M be a metric space with some norm $\|\cdot\|$. A mapping $T : M \to M$ is a contraction iff there exists $0 < K < 1$ such that for all $u \in M$*

$$\|T(u)\| \leq K\|u\|.$$

The following images are from Arnold, *Ordinary Differential Equations*.

[4] initial-value problems

In this section we construct a contraction mapping of a complete metric space whose fixed point defines the solution of a given differential equation.

1. The Successive Approximations of Picard

Consider the differential equation $\dot{\boldsymbol{x}} = \boldsymbol{v}(t, \boldsymbol{x})$, defined by the vector field $\boldsymbol{v}$ in some domain of the extended phase space $\boldsymbol{R}^{n+1}$ (Fig. 214).

We define the *Picard mapping* to be the mapping A that takes the function $\varphi : t \mapsto \boldsymbol{x}$ to the function $A\varphi : t \mapsto \boldsymbol{x}$, where

$$(A\varphi)(t) = \boldsymbol{x}_0 + \int_{t_0}^{t} \boldsymbol{v}(\tau, \varphi(\tau))\, d\tau.$$

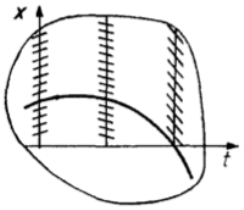


Fig. 214. An integral curve of the equation $\dot{\boldsymbol{x}} = \boldsymbol{v}(t, \boldsymbol{x})$

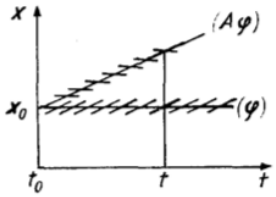


Fig. 215. The Picard mapping A

Geometrically, passing from φ to $A\varphi$ (Fig. 215) means constructing with respect to a curve (φ) a new curve $(A\varphi)$ whose tangent for each t is parallel to a given direction field, only not on the curve $(A\varphi)$ itself – for then $A\varphi$ would be a solution – but at the corresponding point of the curve (φ). We have

$$\begin{array}{c} \varphi \text{ is a solution} \\ \text{with the initial condition} \end{array} \Leftrightarrow (\varphi = A\varphi). \quad \varphi(t_0) = \boldsymbol{x}_0$$

Motivated by the contraction mapping theorem, we consider the sequence of *Picard approximations* φ, $A\varphi$, $A^2\varphi, \ldots$ (starting, say, with $\varphi = x_0$).

Example 1. $\dot{x} = f(t)$ (Fig. 216). $(A\varphi)(t) = x_0 + \int_{t_0}^{t} f(\tau)\,d\tau$. In this case the first step already leads to the exact solution.

Example 2. $\dot{x} = x$, $t_0 = 0$ (Fig. 217). The convergence of the approximations in this case can be observed directly. At the point t

$$\begin{aligned} \varphi &= 1, \\ A\varphi &= 1 + \int_0^t d\tau = 1 + t, \\ A^2\varphi &= a + \int_0^t (1+\tau)\,d\tau = 1 + t + t^2/2, \\ &\cdots\cdots\ \cdots \\ A^n\varphi &= 1 + t + t^2/2 + \cdots + t^n/n!, \\ \lim_{n\to\infty} A^n\varphi &= e^t. \end{aligned}$$

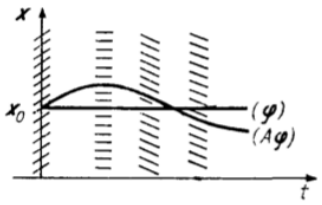


Fig. 216. The Picard approximation for the equation $\dot{x} = f(t)$

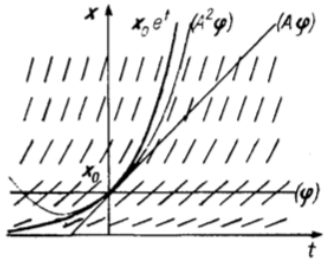


Fig. 217. The Picard approximation for the equation $\dot{x} = x$

Remark 1. Thus the two definitions of the exponential

$$1)\ e^t = \lim_{n\to\infty}\left(1 + \frac{t}{n}\right)^n, \quad 2)\ e^t = 1 + t + \frac{t^2}{2!} + \cdots$$

correspond to two methods of approximating the solutions of the the very simple differential equation $\dot{x} = x$: the broken line method of Euler, and the method of successive approximations of Picard. Historically the original definition of the exponential was simple:

3) e^t is the solution of the equation $\dot{x} = x$ with initial condition $x(0) = 1$.

7.5.8 Proof of Picard's existence theorem

Consider the sequence of functions

$$\begin{aligned} y_0(t) &= y_0 \\ y_n(t) &= y_0 + \int_{t_0}^{t} v\Big(\tau, y_{n-1}(\tau)\Big)\,\mathrm{d}\,\tau. \end{aligned}$$

We will show that

1. the $y_n(t)$ converge to a function $y_\infty(t)$;
2. $y_\infty(t)$ is a solution;
3. $y_\infty(t)$ is the only solution.

Proof that the $y_n(t)$ converge uniformly to a function $y_\infty(t)$

The basic idea is to write the limiting function $y_\infty(t)$ as a telescoping sum, and then to show that the series thus defined converges.

Define

$$e_n(t) = y_{n+1}(t) - y_n(t), \qquad n = 0, 1, 2, \ldots.$$

Then the limiting function that is our objective is

$$y_\infty(t) = y_0 + \sum_{n=0}^{\infty} e_n(t),$$

if the series converges.

We are going to use the Weierstrass M-test to show that the series of functions $\sum_{n=0}^{\infty} e_n(t)$ converge uniformly. So, we need to show that each e_n is bounded in absolute value by some constant W_n, and that the series $\sum_{n=0}^{\infty} W_n$ converges.

For $n \geq 1$ each term is

$$e_n(t) = \int_{t_0}^{t} v\Big(\tau, y_n(\tau)\Big) - v\Big(\tau, y_{n-1}(\tau)\Big)\,\mathrm{d}t.$$

Now, by assumption, v is Lipschitz in the y direction with bound L. (Informally, this means that the absolute value of the straight line slope between any two points lying on a vertical line is bounded by L). Therefore

$$\Big|v\Big(t, y_n(t)\Big) - v\Big(t, y_{n-1}(t)\Big)\Big| \leq L\Big|y_n(t) - y_{n-1}(t)\Big|.$$

And since $\left|\int_a^b f(t)\,\mathrm{d}t\right| \leq \int_a^b |f(t)|\,\mathrm{d}t$,

$$\begin{aligned} |e_n(t)| &\leq L\left|\int_{t_0}^{t} \Big|y_n(\tau) - y_{n-1}(\tau)\Big|\,\mathrm{d}\,\tau\right| \\ &= L\left|\int_{t_0}^{t} \Big|e_{n-1}\Big|\,\mathrm{d}\,\tau\right|. \end{aligned}$$

[5]

For the Weierstrass M-test we need to express the RHS as a constant W_n, depending only on L, M, n, t_0, y_0. We will do this by induction.

For the first few terms we have

$$\begin{aligned}
|e_0(t)| &= \left| \int_{t_0}^{t} v(\tau, y_0)\,\mathrm{d}\,\tau \right| \\
&\leq M|t - t_0| && \text{(by assumption that } v \text{ is bounded by } M\text{)} \\
&\leq Mh \\
|e_1(t)| &= L\left| \int_{t_0}^{t} |e_0(\tau)|\,\mathrm{d}\,\tau \right| && \text{(by assumption that } v \text{ is Lipschitz in } y\text{)} \\
&\leq L\left| \int_{t_0}^{t} M|\tau - t_0|\,\mathrm{d}\,\tau \right| \\
&= LM\frac{|\tau - t_0|^2}{2}\Big|_{t_0}^{t} \\
&= LM\frac{|t - t_0|^2}{2} \\
|e_2(t)| &= L\left| \int_{t_0}^{t} |e_1(\tau)|\,\mathrm{d}\,\tau \right| && \text{(by assumption that } v \text{ is Lipschitz in } y\text{)} \\
&\leq L\left| \int_{t_0}^{t} LM\frac{|t - t_0|^2}{2}\,\mathrm{d}\,\tau \right| \\
&= L^2M\frac{|t - t_0|^3}{3!}.
\end{aligned}$$

So it seems that

Lemma 28. *Suppose*

1. $|e_0(t)| \leq Mh$,
2. $|e_n(t)| \leq L\left| \int_{t_0}^{t} \left|e_{n-1}\right| d\tau \right|$ *for* $n \geq 1$.

Then

$$|e_n(t)| \leq L^n M\frac{h^{n+1}}{(n+1)!} =: W_n.$$

Furthermore, $\lim_{n\to\infty} W_n = 0$.

[5] The outer modulus is required to handle the case $t < t_0$.

Proof. To prove this, note that we know it is true of e_0. So suppose it is true of e_n. Then the next term is

$$\begin{aligned}
|e_{n+1}(t)| &:= \Big|y_{n+2}(t) - y_{n+1}(t)\Big| \\
&\le L\left|\int_{t_0}^{t} \Big|y_{n+1}(\tau) - y_n(\tau)\Big| \,\mathrm{d}\,\tau\right| \\
&= L\left|\int_{t_0}^{t} |e_n(\tau)| \,\mathrm{d}\,\tau\right| \\
&= L\left|\int_{t_0}^{t} L^n M \frac{|\tau - t_0|^{n+1}}{(n+1)!} \,\mathrm{d}\,\tau\right| \\
&= L^{n+1} M \frac{|t - t_0|^{n+2}}{(n+2)!} \\
&\le L^{n+1} M \frac{h^{n+2}}{(n+2)!},
\end{aligned}$$

so

$$|e_n(t)| \le L^n M \frac{h^{n+1}}{(n+1)!}$$

for all $n \ge 0$ by induction.

According to the Ratio Test for convergence of a series, we examine

$$\lim_{n\to\infty} \frac{W_{n+1}}{W_n} = \lim_{n\to\infty} \frac{L^{n+1} M \frac{h^{n+2}}{(n+2)!}}{L^n M \frac{h^{n+1}}{(n+1)!}} = \lim_{n\to\infty} \frac{Lh}{n+2} = 0,$$

proving that the series $\sum_{n=0}^{\infty} W_n$ converges. □

To summarize:

1. Each $e_n(t)$ is bounded in absolute value by $W_n = L^n M \frac{h^{n+1}}{(n+1)!}$
2. The series $\sum_{n=0}^{\infty} W_n$ converges, by the Ratio Test.
3. Therefore the series $\sum_{n=0}^{\infty} e_n(t)$ converges uniformly, by the Weierstrass M-test.
4. Therefore the sequence $(y_n)_{n\ge 0}$ converges uniformly to a limiting function $y_\infty(t)$, since $\sum_{n=0}^{\infty} e_n(t) = y_\infty(t) - y_0$.

Proof that $y_\infty(t)$ is a solution

To prove that the limiting function y_∞ is a solution, we need to show that

$$y'_\infty(t) = v(t, y_\infty(t)) \qquad \text{and} \qquad y_\infty(t_0) = y_0.$$

Recall the definition of the Picard successive approximations:

$$y_n(t) = y_0 + \int_{t_0}^{t} v\Big(\tau, y_{n-1}(\tau)\Big) \,\mathrm{d}\,\tau.$$

Certainly, $y_\infty(t_0) = y_0$. And

$$y_\infty(t) = \lim_{n\to\infty} y_n = y_0 + \int_{t_0}^{t} v(\tau, y_\infty(\tau))\,\mathrm{d}\,\tau$$

as long as it is justified to take the limit inside the integral.

This would be justified if $v(\tau, y_n(\tau))$ converges uniformly to $v(\tau, y_\infty(\tau))$. These are two different functions, both mapping t to the first derivative. Let's write them as $v_{y_n}(t)$ and $v_{y_\infty}(t)$. We're looking for uniform convergence of the former to the latter, i.e. uniform over all values of t. The definition of uniform convergence is that there exists real $\epsilon > 0$ and integer $N \geq 0$ such that for all t, if $n > N$ then $|v_{y_\infty}(t) - v_{y_n}(t)| < \epsilon$.

By assumption v is Lipschitz in the y direction, so

$$|v(t, y_1) - v(t, y_2)| \leq L|y_1 - y_2| \leq 2Lk \qquad \forall y_1, y_2 \in [-k, k],$$

giving the bound needed to prove uniform convergence.

Therefore

$$y_\infty(t) = y_0 + \int_{t_0}^{t} v(\tau, y_\infty(\tau))\,\mathrm{d}\,\tau,$$

and therefore, by differentiating both sides,

$$y'_\infty(t) = v(t, y_\infty(t)).$$

Proof that $y_\infty(t)$ is the unique solution

We've shown that $(y_n)_{n\geq 0}$ converges to a solution y_∞. Now we need to show that if Y is a solution then $Y = y_\infty$.

Recall that the proof of convergence relied on the following:

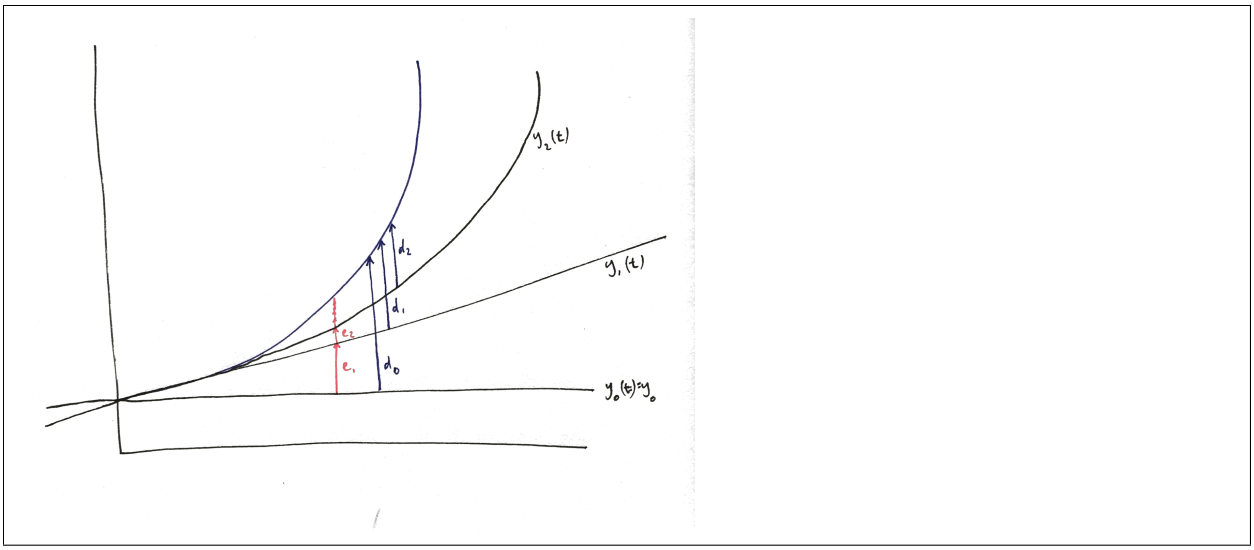
1. The definition of e_n meant that e_n could be expressed in terms of e_{n-1}.
2. A bound for the first term e_1 was provided by the assumption that v was bounded. Informally, this placed a bound on the amount of v height difference that could be accumulated by y_1 between t_0 and t.
3. A bound for subsequent terms could be expressed in terms of the bound for the previous term. Informally, this was because the subsequent terms involved differences in v height, which are bounded due to the Lipschitz assumption.

We now need to do something similar to demonstrate that if Y is a solution, then $y_n \to Y$ as $n \to \infty$.

So suppose Y is a solution. Then

$$Y(t) = y_0 + \int_{t_0}^{t} v(\tau, Y(\tau))\,\mathrm{d}\,\tau.$$

Define $d_n(t) = Y(t) - y_n(t)$. We need to show that $|d_n(t)| \to 0$ for all t as $n \to \infty$.



The first term is

$$d_0(t) = \int_{t_0}^{t} v(\tau, Y(\tau))\,\mathrm{d}\,\tau.$$

As in the convergence proof, the fact that v is assumed to be bounded provides a bound:

$$|d_0(t)| \leq M|t - t_0|.$$

Subsequent terms are

$$\begin{aligned}
|d_n(t)| &= \Big|Y(t) - y_n(t)\Big| \\
&= \left| \int_{t_0}^{t} v(\tau, Y(\tau)) - v(\tau, y_{n-1}(\tau))\,\mathrm{d}\,\tau \right| \\
&\leq \left| \int_{t_0}^{t} \Big|v(\tau, Y(\tau)) - v(\tau, y_{n-1}(\tau))\Big|\,\mathrm{d}\,\tau \right|.
\end{aligned}$$

As in the convergence proof, this involves a difference in v height, so we can use the Lipschitz assumption to express d_n in terms of d_{n-1}:

$$\begin{aligned}
|d_n(t)| &\leq \left| \int_{t_0}^{t} L\Big|Y(\tau) - y_{n-1}(\tau)\Big|\,\mathrm{d}\,\tau \right| \\
&= \left| L \int_{t_0}^{t} \Big|d_{n-1}(\tau)\Big|\,\mathrm{d}\,\tau \right|.
\end{aligned}$$

Therefore we can apply Lemma 28 with $d_n(t)$ substituted for $e_n(t)$, to conclude that $d_n(t) \to 0$ as $n \to \infty$ for all t, proving that if Y is a solution, then $y_\infty = Y$. This completes the proof of Picard's existence theorem. □

7.6 Simmons

7.6.1 Picard's theorem

For every point (t, y) in a rectangle, the ODE

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y)$$

has a solution passing through that point if $\frac{\partial f}{\partial y}$ is Lipschitz continuous in that rectangle.

7.6.2 Families of curves

For a family of curves, say the family of circles

$$x^2 + y^2 = c^2 \tag{7.1}$$

we can obtain a differential equation by implicit differentiation:

$$2x + 2y\frac{\mathrm{d}y}{\mathrm{d}x} = 0. \tag{7.2}$$

Alternatively (eoc),

$$\begin{aligned}(x + \mathrm{d}x)^2 + (y + \mathrm{d}y)^2 &= c^2\\ x^2 + 2x\,\mathrm{d}x + y^2 + 2y\,\mathrm{d}y &= c^2\\ 2x\,\mathrm{d}x + 2y\,\mathrm{d}y &= 0.\end{aligned}$$

7.6.3 Orthogonal trajectories

What's the family of curves each of which is equal to every circle in (7.1)?

Well, we know that their gradients are negative the inverse of the circle gradients. So if we let $\frac{\mathrm{d}y}{\mathrm{d}x}$ now be the gradient of the orthogonal trajectories, then from (7.2),

$$2x - 2y\frac{\mathrm{d}x}{\mathrm{d}y} = 0$$

is an ODE specifying the family of orthogonal trajectories. Thus

$$\begin{aligned}\frac{\mathrm{d}y}{\mathrm{d}x} &= \frac{y}{x}\\ \log(y) &= \log(x) + C\\ y &= Ax,\end{aligned}$$

so the orthogonal trajectories are lines through the origin, as expected.

7.6.4 Use of polar coordinates to make a problem tractable (separable)

TODO

7.7 Arnold - Problems

7.7.1

At what altitude is the density of the air one half of that at the surface of the Earth? Regard temperature as constant. One cubic meter of air at the Earth's surface weighs 1250g.

$$\rho(0) = 1250$$
$$=$$

Chapter 8

Complex Analysis

Useful results

Geometric series $a + aw + \ldots + aw^n = \frac{a(1-w^{n+1})}{1-w}$

8.1 Complex Numbers

I.2.1 Prove that $\mathbb{C}$ obeys the associative law for multiplication and the distributive law.

Let $u, v, w \in \mathbb{C}$ with $u = a + bi$, $v = c + di$, and $w = f + gi$.

Multiplication is associative since

$$\begin{aligned} uv &= (ac - bd) - (ad + bc)i \\ &= (ca - db) - (cb + da)i = vu. \end{aligned}$$

Multiplication is left-distributive over addition since

$$\begin{aligned} u(v + w) &= (a + bi)\big((c + f) + (d + g)i\big) \\ &= (ac + af - bd - bg) + (ad + ag + bc + bf)i \\ &= (ac - bd) + (ad + bc)i + (af - bg) + (ag + bf)i \\ &= (a + bi)(c + di) + (a + bi)(f + gi) \\ &= uv + uw. \end{aligned}$$

Since multiplication is commutative, multiplication is also right-distributive over addition.

I.2.2 Find the multiplicative inverses of the complex numbers (0, 1) and (1, 1)

I.2.3 Think of $\mathbb{C}$ as a vector space over $\mathbb{R}$. Let $c = (a, b)$ be in $\mathbb{C}$, and regard multiplication by c as a real linear transformation T_c. Find the matrix M_c for T_c with respect to the basis $(1, 0), (0, 1)$. Observe that the map $c \mapsto M_c$ preserves addition and multiplication. Conclude that the algebra of two-by-two matrices over $\mathbb{R}$ contains a replica of $\mathbb{C}$.

Background:

What does "$\mathbb{C}$ as a vector space over $\mathbb{R}$" mean? A vector space is a set of tuples. The elements of the tuples are elements of the field ($\mathbb{R}$ in this case). Vector spaces support addition and scalar multiplication, where the scalars come from the field. So this means that $\mathbb{C}$ is a set of ordered pairs of reals, supporting addition of pairs and multiplication of a pair by a real scalar. What it does *not* imply is that pairs can be multiplied, although, in the case of $\mathbb{C}$, they can, since $\mathbb{C}$ is a field.

A linear transformation is a function from one vector space U to another, W, such that $f(u + w) = f(u) + f(w)$, and $f(au) = af(u)$ for $u \in U$, $w \in W$ and a in the field. In other words, the linear transformation preserves the two vector space operations, addition and scalar multiplication; it is a homomorphism on the vector space.

OK, so the operation that was ignored by conceiving of $\mathbb{C}$ as a vector space over $\mathbb{R}$, multiplication of the vectors, we're going to regard as a "real linear transformation", i.e. a function of $\mathbb{R}^2$. Find the matrix for it with respect to the basis $((1, 0), (0, 1))$. The first basis vector (1) is transformed as $(1, 0) \mapsto (a, b)(1, 0) = (a, b)$. The second basis vector (i) is transformed as $(0, 1) \mapsto (a, b)(0, 1) = (-b, a)$. Therefore the matrix of T_c is

$$M_c = \begin{pmatrix} a & -b \\ b & a \end{pmatrix},$$

which is a rotation + scaling transformation of $\mathbb{R}^2$.

Observe that the map $c \mapsto M_c$ preserves addition and multiplication.

Let $f : \mathbb{R}^2 \to$ (2x2 matrices) denote the map $c \mapsto M_c$. Then

$$f(c_1 + c_2) = \begin{pmatrix} a_1 + a_2 & -b_1 - b_2 \\ b_1 + b_2 & a_1 + a_2 \end{pmatrix} = f(c_1) + f(c_2),$$

and

$$f(c_1 c_2) = \begin{pmatrix} a_1 b_1 - a_2 b_2 & -a_1 b_2 - a_2 b_1 \\ a_1 b_2 + a_2 b_1 & a_1 b_1 - a_2 b_2 \end{pmatrix},$$

while

$$f(c_1) f(c_2) = \begin{pmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{pmatrix} \begin{pmatrix} a_2 & -b_2 \\ b_2 & a_2 \end{pmatrix} = \begin{pmatrix} a_1 a_2 - b_1 b_2 & -a_1 b_2 - a_2 b_1 \\ a_2 b_1 + a_1 b_2 & a_1 a_2 - b_1 b_2 \end{pmatrix}$$

? That's not preserving multiplication of complex vectors. Does it mean preserving scalar multiplication?

Conclude that the algebra of two-by-two matrices over $\mathbb{R}$ contains a replica of $\mathbb{C}$

The "algebra of two-by-two matrices over $\mathbb{R}$" refers to the fact that $2x2$ matrices can be added, and multiplied by a scalar from the field $\mathbb{R}$ (they form a vector space), and can also be multiplied.

I.4.3 Prove that if a polynomial with real coefficients has the complex root z, then it also has $\overline{z}$ as a root.

Let $P : \mathbb{C} \to \mathbb{C}$ defined by $P(c) = r_0 + r_1 c^1 + \ldots + r_k c^k$ be a k-th degree polynomial of a complex variable c, with real coefficients r_k, and let $z = a + bi$ be a root, i.e. $P(z) = 0$. The claim is that $P(\overline{z}) = 0$.

To show this, take the complex conjugate of both sides of the equation $P(z) = 0$:

$$\overline{P(z)} = \overline{r_0 + r_1 z^1 + \ldots + r_k z^k} = \overline{0}.$$

Then, since $\overline{z_1 + z_2} = \overline{z}_1 + \overline{z}_2$ and $\overline{rz} = r\overline{z}$,

$$r_0 + r_1 \overline{z^1} + \ldots + r_k \overline{z^k} = P(\overline{z}) = 0,$$

proving that if z is a root then $\overline{z}$ is a root also.

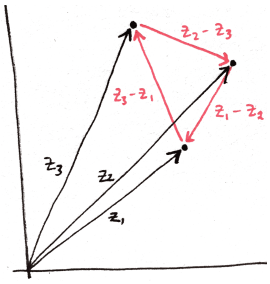
I.7.4 Prove that the distinct complex numbers z_1, z_2, z_3 are the vertices of an equilateral triangle if and only if

$$z_1^2 + z_2^2 + z_3^2 = z_1 z_2 + z_2 z_3 + z_3 z_1$$

The condition can be rewritten as

$$(z_1 - z_2)^2 + (z_2 - z_3)^2 + (z_3 - z_1)^2 = 0,$$

Each of the three terms on the left side is the square of a complex number which, when viewed as a vector in $\mathbb{R}^2$, forms one side of a triangle.



So the original claim is equivalent to the claim that the three vectors

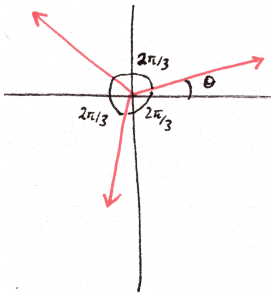
$$(z_1 - z_2)^2,\ (z_2 - z_3)^2,\ (z_3 - z_1)^2$$

sum to 0 if and only if the triangle is equilateral.

First let's prove that if the triangle is equilateral, then the three vectors sum to 0. Translate each of the unsquared vectors

$$(z_1 - z_2),\ (z_2 - z_3),\ (z_3 - z_1)$$

so that they originate at the origin; they are of equal magnitude and they divide the circle into 3 sectors of equal angle $\frac{2\pi}{3}$. Let $\theta < \frac{2\pi}{3}$ be the arbitrary angle between one of the vectors and the first coordinate axis. Interpreted as complex numbers, we see that their arguments are θ, $\frac{2\pi}{3} + \theta$, and $\frac{4\pi}{3} + \theta$.



Now we form their squares

$$(z_1 - z_2)^2,\ (z_2 - z_3)^2,\ (z_3 - z_1)^2.$$

Since $(z_1 - z_2)$, $(z_2 - z_3)$, and $(z_3 - z_1)$ are of equal magnitude, so are their squares. And the arguments of their squares are 2θ, $\frac{4\pi}{3} + 2\theta$, and $\frac{8\pi}{3} + 2\theta \equiv \frac{2\pi}{3} + 2\theta \mod 2\pi$. Therefore the three squared side vectors, when translated so that they originate at the origin, also divide up the circle into sectors of equal angle $\frac{2\pi}{3}$: the geometrical picture differs from the previous one only by a uniform scaling and relabeling of the vectors, and we conclude that these squared vectors also sum to zero (return to the origin when placed head-to-tail). I.e. the equilaterality assumption implies

$$(z_1 - z_2)^2 + (z_2 - z_3)^2 + (z_3 - z_1)^2 = 0,$$

proving one direction of the equivalence.

To prove the other direction, we need to show that if

$$z_1^2 + z_2^2 + z_3^2 = z_1 z_2 + z_2 z_3 + z_3 z_1,$$

or equivalently,

$$(z_1 - z_2)^2 + (z_2 - z_3)^2 + (z_3 - z_1)^2 = 0,$$

then the triangle is equilateral. For example, it would suffice to show that

$$|z_1 - z_2| = |z_2 - z_3| = |z_3 - z_1|,$$

but I haven't found a way to do so.

I.10.1 Use de Moivre's formula to find expressions for $\cos 5\theta$ and $\sin 5\theta$ as polynomials in $\cos\theta$ and $\sin\theta$.

From de Moivre's formula we have $(\cos\theta + i\sin\theta)^5 = \cos 5\theta + i\sin 5\theta$. The left hand side expands as

$$\begin{aligned}(\cos\theta + i\sin\theta)^5 = &\cos^5\theta \\ &+ 5i\cos^4\theta\sin\theta \\ &- 10\cos^3\theta\sin^2\theta \\ &- 10i\cos^2\theta\sin^3\theta \\ &+ 5\cos\theta\sin^4\theta \\ &+ i\sin^5\theta.\end{aligned}$$

Equating real and imaginary components from the right side and the expansion of the left side we have

$$\begin{aligned}\cos 5\theta &= \cos^5\theta - 10\cos^3\theta\sin^2\theta + 5\cos\theta\sin^4\theta \\ \sin 5\theta &= \sin^5\theta - 10\cos^2\theta\sin^3\theta + 5\cos^4\theta\sin\theta\end{aligned}$$

We can write these as polynomials in $\cos\theta$ and $\sin\theta$ respectively by using the identity $\cos^2\theta = 1 - \sin^2\theta$:

$$\begin{aligned}\cos 5\theta &= \cos^5\theta - 10\cos^3\theta(1 - \cos^2\theta) + 5\cos\theta(1 - 2\cos^2\theta + \cos^4\theta) \\ &= 16\cos^5\theta - 20\cos^3\theta + 5\cos\theta, \\ \sin 5\theta &= \sin^5\theta - 10(1 - \sin^2\theta)\sin^3\theta + 5(1 - 2\sin^2\theta + \sin^4\theta)\sin\theta \\ &= 16\sin^5\theta - 20\sin^3\theta + 5\sin\theta.\end{aligned}$$

I.11.4 Prove that the sum of the n-th roots of 1 equals 0, ($n > 1$).

Let $w = \cos\frac{2\pi}{n} + i\sin\frac{2\pi}{n}$ be the n-th root of 1 with smallest argument, other than 1 itself. Then the sum of the roots is $1 + w + w^2 + \ldots + w^{n-1}$. This is the first n terms of a geometric series with constant ratio w, and is therefore equal to $\frac{1-w^n}{1-w} = \frac{1-1}{1-w} = 0$.

I.11.5 Let w be an n-th root of 1 different from 1 itself. Establish the formulas

$$1 + 2w + 3w^2 + \ldots + nw^{n-1} = \frac{n}{w-1},$$

$$1 + 4w + 9w^2 + \ldots + n^2w^{n-1} = \frac{n^2}{w-1} - \frac{2n}{(w-1)^2}.$$

[Note: my answers to this question appear to be wrong.]

The sum of the first $n+1$ terms of a geometric series with first term 1 and constant ratio w is

$$1 + w + w^2 + \ldots + w^n = \frac{1 - w^{n+1}}{1 - w} = \frac{1 - w}{1 - w} = 1,$$

since $w^{n+1} = w$.

This equation is true for w in n-th root, but not for any w.

Taking derivatives of both sides gives

$$1 + 2w + 3w^2 + \ldots + nw^{n-1} = 0,$$

which does not agree with the given formula, so something's wrong.

To take derivatives of both sides need to write a function identity, not an identity between two numbers.

Nevertheless, if it were the case that

$$1 + 2w + 3w^2 + \ldots + nw^{n-1} = \frac{n}{w-1}$$

then we could multiply by w, giving

$$w + 2w^2 + 3w^3 + \ldots + nw^n = \frac{nw}{w-1},$$

and differentiate with respect to w again, giving

$$1 + 4w + 9w^2 + \ldots + n^2w^{n-1} = \frac{(w-1)n - nw}{(w-1)^2} = \frac{n}{w-1} - \frac{nw}{(w-1)^2},$$

which also doesn't agree with the given formula.

Again, to take a derivative, need a function on RHS that agrees with LHS $\forall w$.

I.13.1 Stereographic projection

The stereographic projection maps z onto the surface of a sphere according to

$$z \mapsto \frac{(2\ \mathrm{Re}\ z, 2\,\mathrm{Im}\, z, |z|^2 - 1)}{|z|^2 + 1}.$$

I.13.1 Establish the following formula for the spherical metric

$$\rho(z_1, z_2) = \frac{2|z_1 - z_2|}{\sqrt{|z_1|^2 + 1}\sqrt{|z_2|^2 + 1}}$$

$\rho(z_1, z_2)$ is the Euclidean distance between the image points of z_1 and z_2 on the Riemann sphere, therefore

$$\begin{aligned}\rho(z_1, z_2) &= \left| \frac{(2\ \mathrm{Re}\ z_1, 2\,\mathrm{Im}\, z_1, |z_1|^2 - 1)}{|z_1|^2 + 1} - \frac{(2\ \mathrm{Re}\ z_2, 2\,\mathrm{Im}\, z_2, |z_2|^2 - 1)}{|z_2|^2 + 1} \right| \\ &= \left| \frac{(2\ \mathrm{Re}\ z_1, 2\,\mathrm{Im}\, z_1, |z_1|^2 - 1)(|z_2|^2 + 1) - (2\ \mathrm{Re}\ z_2, 2\,\mathrm{Im}\, z_2, |z_2|^2 - 1)(|z_1|^2 + 1)}{(|z_1|^2 + 1)(|z_2|^2 + 1)} \right|\end{aligned}$$

Meanwhile,

$$|z_1 - z_2| = \sqrt{(\ \mathrm{Re}\ z_1 - \ \mathrm{Re}\ z_2)^2 + (\mathrm{Im}\, z_1 - \mathrm{Im}\, z_2)^2}$$

I.14.1 Establish the formula

$$\rho(z, \infty) = \frac{2}{\sqrt{|z|^2 + 1}}$$

$\rho(z,\infty)$ is the Euclidean distance between the image point of z and the north pole $(0,0,1)$:

$$\begin{aligned}\rho(z,\infty) &= \sqrt{\left(\frac{2\ \mathrm{Re}\ z}{|z|^2+1}-0\right)^2+\left(\frac{2\,\mathrm{Im}\,z}{|z|^2+1}-0\right)^2+\left(\frac{|z|^2-1}{|z|^2+1}-1\right)^2}\\ &= \frac{\sqrt{4\,(\ \mathrm{Re}\ z)^2+4\,(\mathrm{Im}\,z)^2+4}}{|z|^2+1}\\ &= \frac{2}{\sqrt{|z|^2+1}}.\end{aligned}$$

8.2 Complex Differentiation

Consider z approaching z_0. $z-z_0$ is a vector pointing from z_0 to z, and $f(z)-f(z_0)$ is a vector pointing between the image points for some complex-valued function f. The derivative of f at z_0 is the rotation + scaling linear transformation (i.e. the complex number c) that takes $z-z_0$ as close as possible to $f(z)-f(z_0)$. Note that the transformation must be the *same* regardless of the path taken by z as it approaches z_0. In other words, the action of f on *all* vectors in an infinitisimal disc around z_0 is the same as multiplying by a complex number c.

The transformation f can be described by two surfaces over the complex plane: $u(x,y)$ and $v(x,y)$, so that $f:\begin{bmatrix}x\\y\end{bmatrix}\mapsto\begin{bmatrix}u(x,y)\\v(x,y)\end{bmatrix}$. If f is differentiable at (x_0,y_0) then it has a local linear approximation with sublinear error. That linear approximation is

$$\begin{bmatrix}u(x,y)\\v(x,y)\end{bmatrix}\approx\begin{bmatrix}u(x_0,y_0)\\v(x_0,y_0)\end{bmatrix}+\begin{bmatrix}(x-x_0)\frac{\partial u}{\partial x}+(y-y_0)\frac{\partial u}{\partial y}\\(x-x_0)\frac{\partial v}{\partial x}+(y-y_0)\frac{\partial v}{\partial y}\end{bmatrix}$$

This is more succinctly expressed using the Jacobian:

$$\begin{bmatrix}u(x,y)\\v(x,y)\end{bmatrix}\approx\begin{bmatrix}u(x_0,y_0)\\v(x_0,y_0)\end{bmatrix}+\begin{pmatrix}u_x & u_y\\v_x & v_y\end{pmatrix}\begin{bmatrix}x-x_0\\y-y_0\end{bmatrix}.$$

Note that this "linear approximation" form

$$y\approx y_0+y'(x-x_0)$$

could just as well be written

$$y-y_0\approx y'(x-x_0)$$

showing that one way of describing the derivative is "whatever you have to multiply a small displacement in the input space by to get the displacement in the output space".

Recall that the derivative of a complex function f is defined to be a complex number,

$$f'\left(\begin{bmatrix}x_0\\y_0\end{bmatrix}\right)=\lim_{(x,y)\to(x_0,y_0)}\frac{\begin{bmatrix}u(x,y)\\v(x,y)\end{bmatrix}-\begin{bmatrix}u(x_0)\\v(y_0)\end{bmatrix}}{\begin{bmatrix}x\\y\end{bmatrix}-\begin{bmatrix}x_0\\y_o\end{bmatrix}},$$

i.e.

$$f'(z_0)=\lim_{z\to z_0}\frac{f(z)-f(z_0)}{z-z_0},$$

i.e. the derivative is whatever complex number you multiply the vector $z-z_0$ by to get its image vector $f(z)-f(z_0)$, in the limit as $z\to z_0$.

The partial derivatives of the complex-valued f in the real and imaginary directions are the complex numbers

$$f_x = u_x + iv_x$$
$$f_y = u_y + iv_y$$

or

$$f_x = \begin{bmatrix} u_x \\ v_x \end{bmatrix}$$
$$f_y = \begin{bmatrix} u_y \\ v_y \end{bmatrix}$$

The geometric interpretation of these is that they define how the image vector $f(z)$ changes in response to a small change to z.

u can be approximated by a local tangent plane. That's what u_x and u_y do. And so can v; that's what v_x and v_y do. But when we consider the effect of a small displacement in the 2D input space on the 2D output space, we describe the two tangent plane approximations jointly as a linear transformation of the input plane, defined by the Jacobian. The thing is, the linear transformation must have the same effect as multiplication by a complex number.

The derivative is "what you have to multiply the input displacement by to get the output displacement". That's true for a single-variable function $\mathbb{R} \to \mathbb{R}$

$$u(x) - u(x_0) = f'(x_0) \cdot (x - x_0)$$

and it's true for a surface over the plane ($\mathbb{R}^2 \to \mathbb{R}$)

$$u(x, y) - u(x_0, y_0) = \frac{\partial u}{\partial x} \cdot (x - x_0) + \frac{\partial u}{\partial y} \cdot (y - y_0)$$

so presumably something analogous holds for a linear transformation of the plane ($\mathbb{R}^2 \to \mathbb{R}^2$), i.e.

$$\mathbf{z} - \mathbf{z}_0 = \frac{\partial \mathbf{z}}{\partial x} \cdot (x - x_0) + \frac{\partial \mathbf{z}}{\partial y} \cdot (y - y_0).$$

or

$$\begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix} - \begin{bmatrix} u(x_0, y_o) \\ v(x_0, y_0) \end{bmatrix} = \begin{bmatrix} u_x \\ v_x \end{bmatrix} \cdot (x - x_0) + \begin{bmatrix} u_y \\ v_y \end{bmatrix} \cdot (y - y_0).$$

That's exactly the same as the equation involving the Jacobian above

$$\begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix} - \begin{bmatrix} u(x_0, y_0) \\ v(x_0, y_0) \end{bmatrix} = \begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}.$$

So how are we to make sense of the equation relating f' and the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$? Clearly in some sense the Jacobian *is* f', or at least, the complex number that does what the Jacobian does is f'. And

$$\frac{\partial f}{\partial x} = \begin{bmatrix} u_x \\ v_x \end{bmatrix} = u_x + iv_x$$
$$\frac{\partial f}{\partial y} = \begin{bmatrix} u_y \\ v_y \end{bmatrix} = u_y + iv_y,$$

and so from the Cauchy-Riemann constraint

$$\frac{\partial f}{\partial y} = -v_x + iu_x = i\frac{\partial f}{\partial x},$$

i.e. the partial derivative w.r.t. y points at 90° to the x partial derivative.

So if the local linear approximation to the transformation f behaves exactly as multiplication by a complex number, then the Jacobian must have the form of a rotation+scale matrix, $\left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right)$. Therefore the Jacobian must satisfy the Cauchy-Riemann equations

$$\begin{cases} u_x = v_y \\ v_x = -u_y. \end{cases}$$

The Jacobian that effects the local linear rotation+scale transformation, together with the equivalent complex number, is

$$\begin{pmatrix} u_x & -v_x \\ v_x & u_x \end{pmatrix} \qquad u_x + iv_x$$

or

$$\begin{pmatrix} v_y & u_y \\ -u_y & v_y \end{pmatrix} \qquad v_y - iu_y.$$

So we can write

$$\begin{aligned} f' &= u_x + iv_x = f_x \\ &= v_y - iu_y = -if_y, \end{aligned}$$

therefore as above, another expression of the Cauchy-Riemann criterion is

$$f_x = -if_y.$$

Question: what is the intuition for the fact that the complex number representing the partial derivative with respect to x is the *same* as the complex number that effects the full linear transformation? (and at 90° to the partial with respect to y) And what's the intuition for the fact that $\frac{\partial f}{\partial z} = \frac{\partial f}{\partial x}$, while $\frac{\partial f}{\partial \overline{z}} = 0$?

f is differentiable iff the error in the linear transformation goes to 0 as $(x, y) \to (x_0, y_0)$ (i.e. real partial derivatives of u and v exist) and the partial derivatives satisfy the Cauchy-Riemann equations.

Partial derivatives in the z and $\overline{z}$ directions

The (fixed) x, y and (varying) $z, \overline{z}$ directions are related by

$$\begin{aligned} x &= (z + \overline{z})/2 \\ y &= (z - \overline{z})/2i. \end{aligned}$$

So by the chain rule,

$$\begin{aligned} \frac{\partial f}{\partial z} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial z} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial z} \\ &= (u_x + iv_x)\frac{1}{2} + i(u_x + iv_x)\frac{1}{2i} \\ &= u_x + iv_x \\ &= \frac{\partial f}{\partial x} \end{aligned}$$

and

$$\begin{aligned}\frac{\partial f}{\partial \overline{z}} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial \overline{z}} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \overline{z}} \\ &= (u_x + iv_x)\frac{1}{2} + i(u_x + iv_x)\frac{-1}{2i} \\ &= \frac{1}{2}\Big((u_x - u_x) + i(v_x - v_x)\Big) \\ &= 0.\end{aligned}$$

II.8.1(b,d) Let the function f be holomorphic in the open disc D. Prove that each of the following conditions forces f to be constant:

Let $f(z) = u(z) + iv(z)$.

(a) $f' = 0$ throughout D

Informally: $f' = 0$ throughout D means that the best linear approximation of $f(z) - f(z_0)$ is $0(z - z_0)$ which implies that $f(z) = f(z_0)$ everywhere, so f is constant.

Formally: Since f is holomorphic, $f' = u_x + iv_x = 0$. Equating real and imaginary parts shows that $u_x = v_x = 0$ and therefore that $v_y = u_x = 0$ and $u_y = -v_x = 0$. Since the Jacobian of f is the zero matrix, f is constant.

(b) f is real-valued in D

f is real-valued, so $f(z) - f(z_0)$ is real-valued. Therefore the local linear approximation $c(z - z_0)$ collapses the plane onto the real axis, i.e. the Jacobian matrix has the form $\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}$. But f is holomorphic, so the Jacobian must also have the form $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$. Therefore the Jacobian is the zero matrix, i.e. all partial derivatives are zero, $u_x = u_y = v_x = v_y = 0$, so f is constant.

(c) $|f|$ is constant in D

Informally: $|f|$ is constant means that it collapses all points in the open disc D onto a circle. Therefore the Jacobian of f has determinant $0 = u_x^2 + v_x^2 = v_y^2 + u_y^2$. Therefore the Jacobian is the zero matrix and the function f is constant.

Formally: $|f|$ is constant, therefore $|f|^2 = f\overline{f} = u^2 + v^2$ is constant. Therefore the following two partial derivatives are constant:

$$\begin{cases} \frac{\partial}{\partial x}|f|^2 = 2uu_x + 2vv_x = 0 \\ \frac{\partial}{\partial y}|f|^2 = 2uu_y + 2vv_y = 0. \end{cases}$$

Since f is holomorphic, $u_x = v_y$ and $u_y = -v_x$, so

$$\begin{cases} uu_x - vu_y = 0 \\ uu_y + vu_x = 0, \end{cases}$$

Multiplying the first equation by u and the second by v we have

$$\begin{cases} u^2u_x - uvu_y = 0 \\ uvu_y + v^2u_x = 0, \end{cases}$$

and summing these gives

$$u_x(u^2+v^2)=0,$$

which proves that either $u_x=0$ or that f is constant (in which case $u_x=0$ also). Similarly, multiplying the first equation by v and the second by u gives

$$\begin{cases} uvu_x - v^2u_y = 0 \\ u^2u_y + uvu_x = 0, \end{cases}$$

and subtracting the first from the second gives

$$u_y(u^2+v^2)=0.$$

We conclude that $u_x=u_y=0$ and that f is therefore constant.

(d) $\arg f$ **is constant in** D

Let $\arg f=\theta$, constant throughout D. Then $\arg(f(z)-f(z_0))=\theta$, whenever $z\neq z_0$. Therefore the best local linear approximation to f is a linear transformation that collapses the plane onto a line with angle θ. The Jacobian determinant is therefore zero. Since f is holomorphic the Jacobian is of the form $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ and therefore we have $a^2+b^2=0$, so $a=b=0$. Therefore the Jacobian is the zero matrix, i.e. $f'=0$ throughout D, so f is constant.

II.8.2 Let the function f be holomorphic in the open set G. Prove that the function $g(z)=\overline{f(\overline{z})}$ is holomorphic in the set $G^*=\{\overline{z}:z\in G\}$.

Let

$$f: x+iy \mapsto s(x,y)+it(x,y).$$
$$g: x+iy \mapsto u(x,y)+iv(x,y)$$

We want to show that the Jacobian of g exists and and satisfies the Cauchy-Riemann equations. We have

$$\begin{aligned} g(x+iy) &= \overline{s(x,-y)+it(x,-y)} \\ &= s(x,-y)-it(x,-y), \end{aligned}$$

and therefore

$$\begin{aligned} u(x,y) &= s(x,-y) \\ v(x,y) &= -t(x,-y). \end{aligned}$$

Now $f=s+it$ is holomorphic, so $s_x=t_y$ and $s_y=-t_x$. Therefore the partial derivatives of g are

$$\begin{aligned} u_x &= \frac{\partial}{\partial x}s(x,-y) = s_x \\ u_y &= \frac{\partial}{\partial y}s(x,-y) = -s_y = t_x \\ v_x &= -\frac{\partial}{\partial x}t(x,-y) = -t_x \\ v_y &= -\frac{\partial}{\partial y}t(x,-y) = t_y = s_x. \end{aligned}$$

Therefore $u_x = v_y$ and $v_x = -u_y$, showing that the Jacobian of g satisfies the Cauchy-Riemann equations, and therefore that g is holomorphic in its domain.

II.16.4 Prove that, if u is a real-valued harmonic function in an open disk D, then any two harmonic conjugates of u in D differ by a constant.

Let v and w be harmonic conjugates of u, so that

$$\begin{cases} u_x = v_y = w_y \\ u_y = -v_x = -w_x. \end{cases}$$

We want to show that $q = v - w$ is constant, i.e. that $q_x = q_y = 0$, throughout D. From the Cauchy-Riemann equalities above, we have $q_x = v_x - w_x = 0$ and $q_y = v_y - w_y = 0$ as required.

II.16.7 Prove (assuming equality of second-order mixed partial derivatives) that

$$\frac{\partial^2}{\partial \overline{z} \partial z} = \frac{1}{4}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)$$

Thus, Laplace's equation can be written as $\frac{\partial^2 f}{\partial \overline{z} \partial z} = 0$.

First note that x and y are related to $\overline{z}$ via

$$x = \frac{z + \overline{z}}{2}$$
$$y = \frac{z - \overline{z}}{2i},$$

therefore by the chain rule

$$\begin{aligned} \frac{\partial}{\partial \overline{z}} &= \frac{\partial}{\partial x}\frac{\partial x}{\partial \overline{z}} = \frac{1}{2}\frac{\partial}{\partial x} \\ &= \frac{\partial}{\partial y}\frac{\partial y}{\partial \overline{z}} = -\frac{1}{2i}\frac{\partial}{\partial y}. \end{aligned}$$

Now $\frac{\partial}{\partial z}$ is defined by

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right),$$

and taking the partial derivative with respect to $\overline{z}$ gives

$$\begin{aligned} \frac{\partial^2}{\partial \overline{z} \partial z} &= \frac{1}{2}\left(\frac{\partial}{\partial \overline{z}}\frac{\partial}{\partial x} - i\frac{\partial}{\partial \overline{z}}\frac{\partial}{\partial y}\right) \\ &= \frac{1}{2}\left(\frac{1}{2}\frac{\partial}{\partial x}\frac{\partial}{\partial x} - i\left(\frac{-1}{2i}\right)\frac{\partial}{\partial y}\frac{\partial}{\partial y}\right) \\ &= \frac{1}{4}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right). \end{aligned}$$

Laplace's equation is $\frac{\partial^2 f}{\partial x} + \frac{\partial^2 f}{\partial y} = 0$, which can also be written as

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) f = 0,$$

and therefore

$$4\frac{\partial^2}{\partial\overline{z}\partial z}f = 0,$$

i.e.

$$\frac{\partial^2 f}{\partial\overline{z}\partial z} = 0.$$

II.16.8 Prove that if u is a real-valued harmonic function then the function $\frac{\partial u}{\partial z}$ is holomorphic.
As above, first note that x and y are related to $\overline{z}$ via

$$x = \frac{z+\overline{z}}{2}$$
$$y = \frac{z-\overline{z}}{2i} = -i\frac{z-\overline{z}}{2}.$$

By the chain rule

$$\begin{aligned}\frac{\partial u}{\partial z} &= \frac{\partial u}{\partial x}\frac{\partial x}{\partial z} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial z}\\ &= \frac{1}{2}\frac{\partial u}{\partial x} - \frac{i}{2}\frac{\partial u}{\partial y}.\end{aligned}$$

Switching notation, we write this as $\frac{\partial u}{\partial z} = \frac{1}{2}u_x - \frac{i}{2}u_y$.

Define a complex-valued function

$$\begin{aligned}w(x+iy) = \frac{\partial u}{\partial z} &= s(x,y) + it(x,y)\\ &= \frac{1}{2}u_x - \frac{i}{2}u_y.\end{aligned}$$

Then the Jacobian of w is

$$\begin{pmatrix} s_x & s_y \\ t_x & t_y \end{pmatrix} = \frac{1}{2}\begin{pmatrix} u_{xx} & u_{xy} \\ -u_{yx} & -u_{yy} \end{pmatrix}.$$

But since u is harmonic, we know that $u_{xx} + u_{yy} = 0$, therefore the Jacobian of w satisfies the Cauchy-Riemann equations and $w = \frac{\partial u}{\partial z}$ is holomorphic.

8.3 Image of a curve under a transformation

What is the effect of the inversion mapping $z \mapsto w = \frac{1}{z}$ on circles and lines?

Let $z = x + iy$ with image $w = \frac{1}{z} = u + iv$ and note that $\frac{1}{z} = \frac{\overline{z}}{|z|^2}$. Therefore the mapping is

$$x + iy \mapsto \frac{x}{x^2+y^2} - i\frac{y}{x^2+y^2} = u + iv.$$

The general equation of a circle or line in the plane is

$$Ax^2 + Ay^2 + Bx + Cy + D = 0.$$

We use the inverse mapping to establish an equation that holds in the transformed complex plane. Since the inverse mapping is the same as the forward mapping, we have

$$w = u + iv \mapsto \frac{u}{u^2+v^2} - i\frac{v}{u^2+v^2} = x + iy.$$

So points $w = u + iv$ in the transformed complex plane satisfy

$$A\frac{u^2}{(u^2+v^2)^2} + A\frac{v^2}{(u^2+v^2)^2} + B\frac{u}{u^2+v^2} - C\frac{v}{u^2+v^2} + D = 0,$$

i.e.

$$\frac{A}{u^2+v^2} + B\frac{u}{u^2+v^2} - C\frac{v}{u^2+v^2} + D = 0,$$

or

$$A + Bu - Cv + Du^2 + Dv^2 = 0.$$

So we see that, if a circle/line exists in the pre-transformed plane, then...

8.4 Linear-Fractional Transformations

Complex projective space $\mathbf{CP}^1$ is a space of equivalence classes of vectors in $\mathbb{C}^2$. Basically the elements of $\mathbf{CP}^1$ are analogs of lines through the origin in $\mathbb{R}^2$ (one-dimensionsal subspaces): two vectors are equivalent if the ratios between their vector components are equal. And that ratio provides a bijection between $\mathbf{CP}^1$ and $\overline{\mathbb{C}}$.

Since linear transformations of $\mathbb{C}^2$ map lines (in $\mathbb{C}^2$) to lines (in $\mathbb{C}^2$), they induce a bijection on $\mathbf{CP}^1$ and therefore on $\overline{\mathbb{C}}$.

In fact linear-fractional transformations are induced by a two-by-two complex matrix (an element of $\mathrm{GL}_2(\mathbb{C})$) [Do I understand why?]. This makes linear-fractional transformations closed under composition and gives them an identity (the LFT corresponding to the identity matrix) and inverses (given by the matrix inverse). So there is a group of LFTs which is the homomorphic image of $\mathrm{GL}_2(\mathbb{C})$, under the map which sends a two-by-two matrix to its induced LFT. The kernel of the homomorphism contains scalar multiples of the identity matrix I_2. I think that's basically because such uniform scaling matrices leave lines unchanged and therefore leave the one-dimensional subspaces unchanged. Therefore the group of LFTs is isomorphic to the quotient group $\mathrm{GL}_2(\mathbb{C})/(\mathbb{C}\backslash\{0\})I_2$ (each coset is formed by taking a matrix and scaling it by multipying it with a scaled identity matrix from the kernel).

III.5.2

Given four distinct points z_1, z_2, z_3, z_4 in $\overline{\mathbb{C}}$, their cross ratio, which is denoted by $(z_1, z_2; z_3, z_4)$ is defined to be the image of z_4 under the linear-fractional transformation that sends z_1, z_2, z_3 to $\infty, 0, 1$, respectively. Prove that if ϕ is a linear-fractional transformation then

$$\big(\phi(z_1), \phi(z_2); \phi(z_3), \phi(z_4)\big) = (z_1, z_2; z_3, z_4).$$

Let f be the linear-fractional transformation that maps z_1, z_2, z_3 to $\infty, 0, 1$ respectively, so that the cross-ratio is defined to be $(z_1, z_2; z_3, z_4) = f(z_4)$. We want to show that the cross ratio, defined in this way, is invariant under an arbitrary linear-fractional transformation ϕ.

First, let's find an explicit expression for $f(z)$ in terms of z_1, z_2, z_3. We know that $f(z_1) = \infty$ and $f(z_2) = 0$, so perhaps f has the form $f(z) = c\frac{z_2 - z}{z_1 - z}$ for some constant c. We also require $f(z_3) = 1$. One way to achieve that is to choose $c = \frac{z_1 - z_3}{z_2 - z_3}$, so the definition of f becomes

$$f(z) = c\frac{(z_2 - z)(z_1 - z_3)}{(z_2 - z_3)(z_1 - z)}.$$

Defined like this, f is a linear-fractional transformation, and it does send z_1, z_2, z_3 to $\infty, 0, 1$, respectively. Furthermore, by theorem III.5, this is the only linear-fractional transformation that does so.

So we have

$$(z_1, z_2; z_3, z_4) = f(z_4) = \frac{(z_1 - z_3)(z_2 - z_4)}{(z_1 - z_4)(z_2 - z_3)},$$

and we want to show that this quantity is invariant under an arbitrary linear-fractional transformation ϕ. Let $\phi(z) = \frac{az+b}{cz+d}$, with $ad - bc = 1$ (since we are free to scale the coefficients a, b, c, d uniformly as we wish, if $ad - bc \neq 1$ then we scale them all by $\frac{1}{\sqrt{ad-bc}}$). Now consider

$$\begin{aligned}
\phi(z_i) - \phi(z_j) &= \frac{(az_i + b)(cz_j + d) - (az_j + b)(cz_i + d)}{(cz_i + d)(cz_j + d)} \\
&= \frac{z_i z_j(ac - ac) + z_i(ad - bc) + z_j(bc - ad) + (bd - bd)}{(cz_i + d)(cz_j + d)} \\
&= \frac{z_i - z_j}{(cz_i + d)(cz_j + d)}.
\end{aligned}$$

Letting $A_i = cz_i + d$, we see that the cross-ratio of the transformed points is

$$\big(\phi(z_1), \phi(z_2); \phi(z_3), \phi(z_4)\big) = \frac{(z_1 - z_3)(z_2 - z_4)/A_1A_3A_2A_4}{(z_1 - z_4)(z_2 - z_3)/A_1A_4A_2A_3} = (z_1, z_2; z_3, z_4).$$

III.6.3

Prove that a linear-fractional transformation with only one fixed point is conjugate to a translation.

Let $\phi(z) = \frac{az+b}{cz+d}$, with $ad - bc = 1$ (justified in III.5.2 above). The fixed points of this mapping are the solutions of

$$\frac{az + b}{cz + d} = z,$$

which is a quadratic equation

$$cz^2 + (d - a)z - b = 0,$$

with solutions

$$\begin{aligned}
z &= \frac{(a - d) \pm \sqrt{(a - d)^2 + 4bc}}{2c} \\
&= \frac{(a - d) \pm \sqrt{(a + d)^2 - 4(ad - bc)}}{2c} \\
&= \frac{(a - d) \pm \sqrt{(a + d)^2 - 4}}{2c}
\end{aligned}$$

ϕ_1 has only one fixed point, so $a + d = \pm 2$, i.e. $d = 2 - a$ or $d = -2 - a$.

We want to show that there exists a linear-fractional transformation $\phi_2(z) = z + k$, and another linear-fractional transformation ψ, such that $\phi_2 = \psi \circ \phi_1 \circ \psi^{-1}$. In other words, performing the ϕ_1 transformation under the change of basis specified by ψ, yields a translation.

Let's just try to show that by calculation. Let $\psi(z) = \frac{ez+f}{gz+h}$ with $eh - fg = 1$ so that $\psi^{-1}(z) = \frac{hz-f}{-gz+e}$. Then

$$\begin{aligned}
\phi_2(z) &= (\psi \circ \phi_1 \circ \psi^{-1})(z) \\
&= \begin{pmatrix} e & f \\ g & h \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} h & -f \\ -g & e \end{pmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} \\
&= \begin{pmatrix} e & f \\ g & h \end{pmatrix} \begin{pmatrix} ah - bg & -af + be \\ -ch - dg & -cf + de \end{pmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} \\
&= \begin{pmatrix} e(ah - bg) + f(-ch - dg) & e(-af + be) + f(-cf + de) \\ g(ah - bg) + h(-ch - dg) & g(-af + be) + h(-cf + de) \end{pmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix}
\end{aligned}$$

Things we know:

- Trace is invariant under change of basis, so the trace of the product of the 3 matrices is the same as that of ϕ_1: ± 2.
- The determinants of all the matrices and products thereof are 1.

So

$$\begin{aligned}
e(ah - bg) + f(-ch - dg) + g(-af + be) + h(-cf + de) &= \pm 2 \\
(ah - bg)(-cf + de) - (-af + be)(-ch - dg) &= 1 \\
eh - fg &= 1 \\
ad - bc &= 1 \\
a + d &= \pm 2
\end{aligned}$$

A translation has matrix of the form $\begin{pmatrix} x & y \\ 0 & x \end{pmatrix}$. So the question is, can we find e, f, g, h such that

$$\begin{aligned}
g(ah - bg) + h(-ch - dg) &= 0 \\
e(ah - bg) + f(-ch - dg) &= g(-af + be) + h(-cf + de)
\end{aligned}$$

$$\begin{aligned}
ah - bg &= h(ch + dg)/g \\
eh(ch + dg)/g + f(-ch - dg) &= g(-af + be) + h(-cf + de)
\end{aligned}$$

III.9.2

Find the images of the disc $|z| < 1$ and the half-plane $\operatorname{Re} z > 0$ under the linear-fractional transformation that maps ∞ to 1 and has i and $-i$ as fixed points.

The boundary of the disc is the unit circle and therefore must be mapped to either a circle or a line. If it were mapped to a circle then not only i and $-i$ would be fixed points but also all points on the unit circle in the domain would be fixed and, in particular, 1 would be a fixed point. But 1 is not fixed since $\infty \mapsto 1$. Therefore the unit circle is mapped to a line and this line must be the imaginary axis $\operatorname{Re} z = 0$, since it must contain i and $-i$.

The image of the disc $|z| < 1$ must therefore be one of the half-planes either side of the imaginary axis, since connected sets are mapped to connected sets. To determine which, we note that ∞ is mapped to 1. But ∞ was outside the unit disc in the domain, and so in the transformed complex plane, the image of ∞ must remain connected to points outside the image of the unit disc. Therefore the image of the unit disc is the half plane $\text{Re } z < 0$.

The imaginary axis $\text{Re } z = 0$ must be mapped to a circle, since we know that its image contains $i, -i$ and 1, and no line passes through those 3 points. In fact, its image must be the unit circle, i.e. the equator on the Riemann sphere. So the remaining question is whether the image of the half-plane $\text{Re } z > 0$ is the northern or southern hemisphere. To determine which we note that, before the transformation, it was possible to walk North from $-i$, through ∞, to i, with the half-plane in question on our right-hand side. Therefore the image of the half-plane must[1] be on the same side as we perform the corresponding walk between the images of those points. That walk takes us from $-i$, through 1, to i, showing that the image of the half-plane $\text{Re } z > 0$ is the southern hemisphere $|z|$, i.e. the unit disc $|z| < 1$.

[1] This argument is based on a theorem stating that linear-fractional transformations are conformal, and the definition of conformality specifying that orientations of the sort described are preserved.

III.9.5

Prove that the linear-fractional transformations mapping the disc $|z| < 1$ onto itself are those induced by matrices of the form

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix}$$

with $|a|^2 - |b|^2 = 1$.

My initial thought here was the following:

The transformation maps the unit disc (i.e. the southern hemisphere of the Riemann sphere) onto itself. In order for that to be so, I suspect it would have to map the unit circle onto itself (perhaps an argument based on continuity of the transformation here?). And I think that a linear-fractional transformation maps the unit circle onto itself if and only if that mapping is multiplication by a unit-length complex number i.e. rotation of the Riemann sphere around the polar axis. Such mappings have the form

$$f(z) = \frac{az + b}{cz + d}$$

where $b = c = 0$ and $|a| = |d|$. So an answer along those lines would hope to show that a linear-fractional transformation is induced by a matrix of the form

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix}$$

with $|a|^2 - |b|^2 = 1$, if and only if the matrix is of the form

$$\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix},$$

with $|a| = |d|$. But that doesn't look to be a true statement.

III.9.7

For the function

$$f(z) = \left(\frac{z+1}{z-1}\right)^2$$

(defined to equal 1 **at** $z = \infty$ **and** ∞ **at** $z = 1$**), find the images of the following sets:**

(a) The extended real axis.

(b) The extended imaginary axis.

(c) The half-plane Re $z > 0$**.**

$$\begin{aligned} f(z) &= \left(\frac{z+1}{z-1}\right)^2 \\ &= \frac{z^2+2z+1}{z^2-2z+1} \\ &= w \end{aligned}$$

We have

$$\begin{aligned} 0 &\mapsto 1 \\ \infty &\mapsto 1 \\ 1 &\mapsto \infty \\ -1 &\mapsto 0 \\ i &\mapsto -1 \\ -i &\mapsto \frac{i}{i+2} \end{aligned}$$

so the mapping is non-injective and therefore non-invertible.

Also as far as I can see the mapping can not be viewed as a linear-fractional transformation as these are ratios of first-degree, not second-degree, polynomials in z. Therefore I can't use any theorems about linear-fractional transformations such as triple transitivity and preservation of clircles.

One idea would be to find the inverse mapping and use this inverse mapping to find how equations are transformed. E.g. if we let $z = x + iy$ and $f(z) = w = u + iv = u(x, y) + iv(x, y)$ then for part (a) the extended real axis in the domain is defined by $y = 0$. If there were an inverse mapping, then we could establish the following equation in the transformed plane: $\operatorname{Im} f^{-1} = 0$ and rearrange this equation to get an equation that describes the image of the extended real axis.

However, the forward mapping is non-injective, so I don't think we can do that.

Incidentally, I think the non-injectiveness and non-invertibility of the forward mapping can also be seen from this attempt to find the inverse, which leads to a quadratic expression with two solutions.

$$\begin{aligned} f(z) &= \frac{z^2+2z+1}{z^2-2z+1} \\ &= w \end{aligned}$$

$$z^2(1-w) + 2z(1+w) + (1-w) = 0$$

$$
\begin{aligned}
z &= \frac{-2(1+w) \pm \sqrt{4(1+w)^2 - 4(1-w)^2}}{2(1-w)} \\
&= \frac{-(1+w) \pm \sqrt{(1+w)^2 - (1-w)^2}}{1-w} \\
&= \frac{-(1+w) \pm \sqrt{1+2w+w^2-1+2w-w^2}}{1-w} \\
&= \frac{-(1+w) \pm 2\sqrt{w}}{1-w}
\end{aligned}
$$

8.5 Elementary functions

Exponential function

Based on the premise that $(e^z)' = e^z$, we start with the Maclaurin series definition

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots,$$

which converges for all z.

Letting $z = x + iy$, we demand that $e^{x+iy} = e^x e^{iy}$, but what is e^{iy}?

$$
\begin{aligned}
e^{iy} &= \sum_{n=0}^{\infty} \frac{i^n y^n}{n!} \\
&= \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k}}{(2k)!} + i \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k+1}}{(2k+1)!}
\end{aligned}
$$

which are the Taylor series for cos and sin. So

$$e^{x+iy} = e^x(\cos y + i \sin y).$$

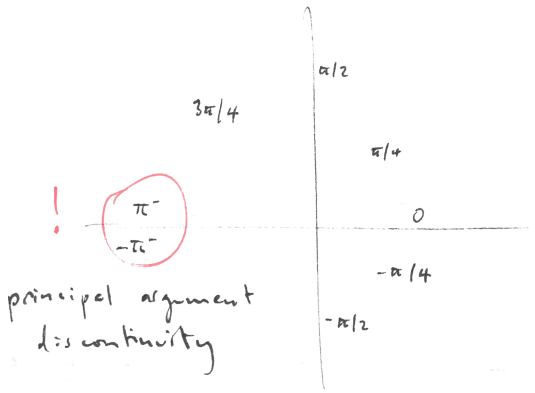
The argument of the image point depends on the imaginary part of the input.

Basically, the exponential map takes a vertical line $\operatorname{Re} z = x$ and wraps it round a circle infinitely many times. The radius of the circle is e^x.

Branches of inverse functions

A given point on the circle is hit by infinitely many points on the line: $\ldots, x+i(y-2\pi), x+iy, x+i(y+2\pi), \ldots$. These are the logarithms of the point on the circle.

The "principal argument" of a complex number z is $\operatorname{Arg} z \in (-\pi, \pi]$. It is continuous at points away from the negative real axis.



The n-th roots of z are

$$\sqrt[n]{|z|}\left(\cos\left(\frac{\text{Arg } z+2k\pi}{n}\right)+i\sin\left(\frac{\text{Arg } z+2k\pi}{n}\right)\right),$$

$k=0,1,\ldots,n-1$. The "principal root" is given by $k=0$:

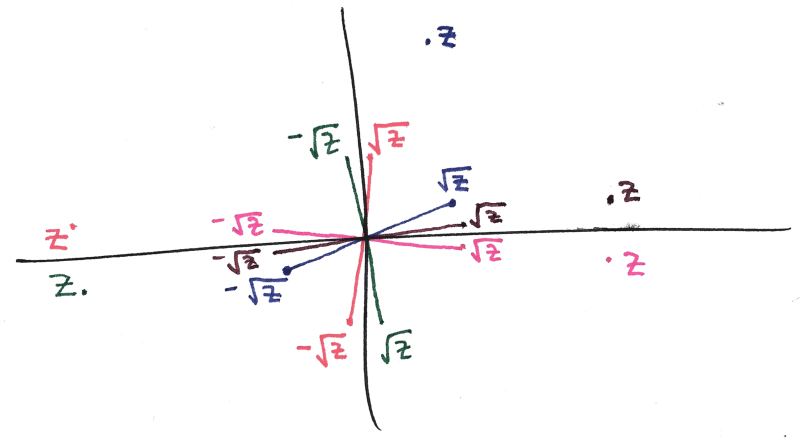
$$\sqrt[n]{|z|}\left(\cos\frac{\text{Arg } z}{n}+i\sin\frac{\text{Arg } z}{n}\right).$$

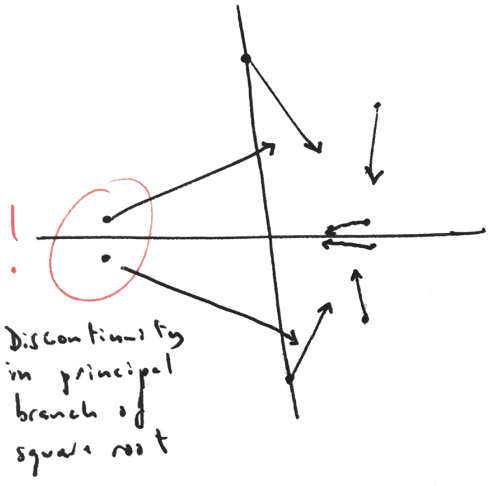
The "principal branch" of the log function is given by

$$\text{Log } z=\ln|z|+i\text{Arg } z.$$

Because they involve Arg , both Log and the principal root function are continuous at points away from the negative real axis.

Here are some diagrams of the principal branch of the square root function. Notice that it is discontinuous at points of the negative real axis. I.e. it is a branch in the domain $\mathbb{C}\backslash[0,-\infty)$





Suppose we want a branch in a domain that includes the negative real axis. Then we can use

$$g(z) = \begin{cases} \sqrt{z}, & \operatorname{Im} z \geq 0 \\ -\sqrt{z}, & \operatorname{Im} z < 0 \end{cases}$$

and we could also use $-g$; there are two branches. However, these are discontinuous at points on the positive real axis, so they are branches in the domain $\mathbb{C}\backslash[0, +\infty)$.

Consider the infinitely wide strip $\{z : -\pi < \operatorname{Im} z < \pi\}$. The image of this under the exponential map is the entire complex plane with 0 removed. For example, consider some complex number w with $|w| = r > 0$. It is hit by a point on the vertical line $x = \ln r$. So this domain-restricted version of the exponential map has an inverse, and that inverse is Log .

For a multivalent function, it's not possible to find an inverse that sends every image point $f(z)$ back to the correct place z on the left-hand side, for all z. That's because many z hit the same $f(z)$. But it is possible to find a "right inverse": a function that sends image points back to one of the possible preimage points. We want this to be continuous, so we often have to restrict the domain on the right-hand side to avoid points of discontinuity, e.g. remove $(-\infty, 0]$ in the case of the Log and principal root inverse functions.

Hyperbolic functions

Just as their real counterparts,

$$\cosh z = (e^z + e^{-z})/2$$
$$\sinh z = (e^z - e^{-z})/2$$

with

$$\tanh = \sinh / \cosh$$
$$\coth = 1/\tanh$$
$$\operatorname{sech} = 1/\cosh$$
$$\operatorname{cosech} = 1/\sinh$$

Trigonometric functions

From $e^{iy} = \cos y + i \sin y$ we have $e^{-iy} = \cos y - i \sin y$ and therefore

$$\cos y = (e^{iy} + e^{-iy})/2$$
$$\sin y = (e^{iy} - e^{-iy})/2i,$$

for real y. cos and sin are defined on $\mathbb{C}$ by substituting a complex variable z in place of y:

$$\cos z = (e^{iz} + e^{-iz})/2$$
$$\sin z = (e^{iz} - e^{-iz})/2i,$$

with $\tan, \cot, \sec$ and cosec also defined as usual.

IV.5.2 Describe the curves $|f| =$ constant and $\arg f =$ constant for the function

$$f(z) = \exp(z^2).$$

Let $z = x + iy$, so

$$\begin{aligned} f(z) &= \exp((x^2 - y^2) + 2ixy) \\ &= e^{x^2-y^2}\left(\cos 2xy + i \sin 2xy\right) \end{aligned}$$

Therefore $|f| = k$ for some constant $k \in \mathbb{R}$ implies that $e^{x^2-y^2} = k > 0$, i.e. $y = \pm\sqrt{x^2 - \log k}$. In other words, the preimage of a circle of radius k centered on the origin is the union of the two curves $y = \pm\sqrt{x^2 - \log k}$.

$\arg f = \theta$ for some constant $0 \leq \theta < 2\pi$ implies that $2xy = \theta$, i.e. $y = \frac{\theta}{2x}$. In other words, the preimage of a ray at angle θ is the graph of $y = \frac{\theta}{2x}$.

IV.9.2 Find all values of $\log(\log i)$

$\log i$ is the following set of image points lying on the imaginary axis:

$$\log i = \left\{ i\left(\frac{\pi}{2} + 2\pi k\right) : k \in \mathbb{Z} \right\}.$$

Fix a particular k. The log of the corresponding image point is the following set of secondary image points, lying on the vertical line through $\frac{\pi}{2} + 2\pi k$:

$$\log\left(i\left(\frac{\pi}{2} + 2\pi k\right)\right) = \left\{ \log\left(\frac{\pi}{2} + 2\pi k\right) + i\left(\frac{\pi}{2} + 2\pi l\right) : l \in \mathbb{Z} \right\},$$

Therefore the set of all values of $\log(\log i)$ is the following rectangular grid of points

$$\log\left(\log i\right) = \left\{ \log\left(\frac{\pi}{2} + 2\pi k\right) + i\left(\frac{\pi}{2} + 2\pi l\right) : k \in \mathbb{Z}, l \in \mathbb{Z} \right\}.$$

IV.13.3 [Not in homework.] Let G be the open set one obtains by removing from $\mathbb{C}$ the interval $[-1, 1]$ on the real axis. Prove that there is a branch of the function $\sqrt{\frac{z+1}{z-1}}$ in G. (Suggestion: What is the image of G under the map $z \mapsto \frac{z+1}{z-1}$?)

The map $z \mapsto \frac{z+1}{z-1}$ maps points as follows:

$$\begin{aligned} -1 &\mapsto 0 \\ 0 &\mapsto -1 \\ 1 &\mapsto \infty \\ i &\mapsto -i \\ -i &\mapsto i \end{aligned}$$

Thus

1. the image of G is $\mathbb{C}$ with the negative real axis removed;
2. the image of the unit circle is the imaginary axis;
3. the image of the unit disc is the left half-plane and the image of the complement of the unit disc is the right half-plane.

IV.13.4 Let G be as in Exercise IV.13.3. Prove that there is a branch of the function $\sqrt{z^2-1}$ in G.

Let $f(z) = \sqrt{z^2-1}$, using the principal square root function defined by

$$\sqrt{w} = \sqrt{|w|}\left(\cos\frac{\operatorname{Arg} w}{2} + i\sin\frac{\operatorname{Arg} w}{2}\right).$$

The principal square root function is discontinuous at points w in $(-\infty, 0]$. Therefore f will be continuous for all $z \in \mathbb{C}$ except where $z^2 - 1 \in (-\infty, 0]$, i.e. $-1 \le z \le 1$. Therefore f will be continuous in $G = \mathbb{C}\backslash[-1, 1]$.

f maps points as follows:

$$\begin{aligned} -1 &\mapsto 0 \\ 0 &\mapsto -1 \\ 1 &\mapsto 0 \\ i &\mapsto -2 \\ -i &\mapsto -2 \\ \infty &\mapsto \infty \end{aligned}$$

Therefore the image of G under f is $\mathbb{C}$ with the interval $[-1, 0]$ removed.

IV.16.1 Find all the values of $(1+i)^i$.

$(1+i)^i$ is the set of values

$$\begin{aligned} \exp\big(i\log(1+i)\big) &= \exp\left(i\left(\log\sqrt{2} + i\left(\frac{\pi}{4} + 2\pi k\right)\right)\right) \\ &= \exp\left(-\pi\left(2k + \frac{1}{4}\right) + i\frac{\log 2}{2}\right) \\ &= e^{-\pi\left(2k+\frac{1}{4}\right)}\left(\cos\frac{\log 2}{2} + i\sin\frac{\log 2}{2}\right) \end{aligned}$$

for $k \in \mathbb{Z}$.

IV.16.3 Prove that if f is a branch of z^c in an open set not containing 0, then f is holomorphic and f' is a branch of cz^{c-1}.

A branch f of z^c, defined on some open set excluding 0, means that $f(z) = e^{c\operatorname{Log} z}$ for some branch Log of the logarithm map.

The branch of the logarithm, the exponential function, and multiplication by a complex number are all holomorphic transformations. Therefore f is holomorphic, because the composition of holomorphic functions with compatible domains and ranges is holomorphic.

The derivative of f is, by the chain rule,

$$f'(z) = ce^{c\operatorname{Log} z}\frac{d\operatorname{Log} z}{dz} = ce^{c\operatorname{Log} z}\frac{1}{z} = c\frac{f(z)}{z},$$

where I have assumed without proof that $\frac{d\operatorname{Log} z}{dz} = \frac{1}{z}$. The final expression above is a branch of cz^{c-1} defined in the same region that f is defined.

8.6 Power Series

V.6.2 Prove that the sequence $(g_n)_{n=0}^{\infty}$ converges locally uniformly in the open set G if and only if it converges uniformly on each compact subset of G.

First, terminology: Sarason states that $(g_n)_{n=0}^{\infty}$ converges locally uniformly in G if each point of G has a neighborhood in which the sequence converges uniformly. I'm going to take that to mean "if and only if".

For the forward direction, we need to show that if

(A): each point of G has a neighborhood in which (g_n) converges uniformly

then

(B): (g_n) converges uniformly on each compact subset of G.

I don't have a proof, but a suggested approach for how to prove this is by contradiction:

1. Suppose (A) is true but that (B) is not, so that there exists some compact subset S of G on which (g_n) does not converge uniformly.
2. Show that there exists a point of S which lacks any neighborhood within which convergence is uniform. □

For the reverse direction, we need to show that if

(B): (g_n) converges uniformly on each compact subset of G.

then

(A): each point of G has a neighborhood in which (g_n) converges uniformly

Again I don't have a proof, but a suggested approach for how to prove this is:

1. Consider a point z of G.
2. Show that there is a compact subset S of G that contains z.
3. Show that z has a neighborhood which is a subset of S. □

V.7.2 Prove that the series $\sum_{n=0}^{\infty}\left(\frac{z-1}{z+1}\right)^n$ converges locally uniformly in the half-plane $\mathbf{Re}\, z > 0$, and find the sum.
(No attempt)

V.14.1(b) Find the radius of convergence of the following series:

$$\sum_{n=0}^{\infty}\frac{(n!)^3}{(3n)!}z^{3n}$$

We use the ratio test:

$$\begin{aligned}\lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right| &= \lim_{n\to\infty}\left|\frac{(n!)^3 z^{3n}}{(3n)!}\frac{(3n+3)!}{((n+1)!)^3 z^{3n+3}}\right| \\ &= |z^{-3}|\lim_{n\to\infty}\left|\frac{(3n+3)(3n+2)(3n+1)}{(n+1)^3}\right| \\ &= |z^{-3}|\lim_{n\to\infty}\left|\frac{27+o(n^{-1})}{1+o(n^{-1})}\right| \\ &= 27|z^{-3}|\end{aligned}$$

So the series converges when $|z^3| > 27$, i.e. outside a disc of radius 3 centered at the origin. The radius of convergence is infinite.

V.16.2 What function is represented by the power series $\sum_{n=1}^{\infty} n^2 z^n$?
(No attempt)

V.18.1 Use the scheme above to determine the power series with center 0 representing the function $f(z) = \frac{1}{1+z+z^2}$ near 0. What is the radius of convergence of this series?

Assume $f(z)$ can be represented as a power series $\sum_{n=0}^{\infty} a_n z^n$. We can write f as the ratio

$$f(z) = \frac{1}{1+z+z^2} = \frac{\sum_{n=0}^{\infty} b_n z^n}{\sum_{n=0}^{\infty} c_n z^n} =: \frac{g(z)}{h(z)},$$

where

$$b_n = \begin{cases} 1, & n = 0 \\ 0, & \text{otherwise} \end{cases}$$

$$c_n = \begin{cases} 1, & 0 \le n \le 2 \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} g(z) = \sum_{n=0}^{\infty} b_n z^n &= f(z)h(z) \\ &= \left(\sum_{n=0}^{\infty} a_n z^n\right)\left(\sum_{n=0}^{\infty} c_n z^n\right) \\ &= \sum_{n=0}^{\infty} z^n \sum_{k=0}^{n} a_k c_{n-k}\end{aligned}$$

8.7 Complex Integration

VI.7.2 Derive the formula

$$\frac{1}{2\pi}\int_0^{2\pi} \cos^{2n} t \; dt = \frac{1\cdot 3\cdot 5\cdots(2n-1)}{2\cdot 4\cdot 6\cdots(2n)}$$

by integrating the function $\frac{1}{z}\left(z+\frac{1}{z}\right)^{2n}$ around the unit circle, parameterized by the curve $\gamma(t) = e^{it} (0 \le t \le 2\pi)$.
Here are two slightly different attempts:

We can write the integral as

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi} \cos^{2n} t \; dt &= \frac{1}{2^{2n+1}\pi}\int_0^{2\pi} (e^{it}+e^{-it})^{2n} \; dt \\ &= \frac{1}{2^{2n+1}\pi}\int_\gamma (z+z^{-1})^{2n} \; dz\end{aligned}$$

Now consider the related integral

$$\begin{aligned}\int_\gamma z^{-1}\left(z+z^{-1}\right)^{2n} \; dz &= \int_\gamma z^{-1}\sum_{k=0}^{2n}\binom{2n}{k} z^{2n-k}z^{-k} \; dz \\ &= \sum_{k=0}^{2n}\binom{2n}{k}\int_\gamma z^{2(n-k)-1} \; dz.\end{aligned}$$

If $k \neq n$, then $z^{2(n-k)-1}$ is the derivative of $\frac{z^{2(n-k)}}{2(n-k)}$, in which case $\int_\gamma z^{2(n-k)-1} \; dz = 0$ since γ is a closed curve. Therefore the only terms remaining in the summation are those for which $k = n$:

$$\begin{aligned}\int_\gamma z^{-1}\left(z+z^{-1}\right)^{2n} \; dz &= \binom{2n}{n}\int_\gamma z^{-1} \; dz \\ &= \binom{2n}{n}\int_0^{2\pi} e^{-it}ie^{it} \; dz \\ &= \binom{2n}{n}2\pi i.\end{aligned}$$

Returning to the original problem, we now know the value of a similar integral:

$$\begin{aligned}\frac{1}{2^{2n+1}\pi}\int_\gamma z^{-1}(z+z^{-1})^{2n} \; dz &= \frac{1}{2^{2n+1}\pi}\binom{2n}{n}2\pi i \\ &= \frac{1}{2^{2n+1}\pi}\;\frac{(2n)!}{2(n!)}\;2\pi i \\ &= \frac{(n+1)\cdot(n+2)\cdots 2n}{2^{2n+1}}i\end{aligned}$$

Alternatively we can write the integral as

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi}\cos^{2n}t\ dt &= \frac{1}{2^{2n+1}\pi}\int_0^{2\pi}(e^{it}+e^{-it})^{2n}\ dt\\ &= \frac{1}{2^{2n+1}\pi}\int_\gamma (z+z^{-1})^{2n}\ dz\\ &= \frac{1}{2^{2n+1}\pi}\int_\gamma \sum_{k=0}^{2n}\binom{2n}{k}z^{2n-k}z^{-k}\ dz\\ &= \frac{1}{2^{2n+1}\pi}\sum_{k=0}^{2n}\binom{2n}{k}\int_\gamma z^{2(n-k)}\ dz.\end{aligned}$$

Now if $2n \neq k$, then $z^{2(n-k)}$ is the derivative of $\frac{z^{2(n-k)+1}}{2(n-k)+1}$, in which case $\int_\gamma z^{2(n-k)}\ dz = 0$.

We can view the integral on the right side as integrating the function $(z+z^{-1})^{2n}$ around the unit circle:

VI.8.1 Let z_1 and z_2 be distinct points of $\mathbb{C}$. Evaluate $\int_{[z1,z2]} z^n dz$ and $\int_{[z1,z2]} \overline{z}^n dz$ for $n = 0, 1, 2, \ldots$.

Let $\gamma(t) = z_1 + t(z_2 - z_1)$ for $t \in [0,1]$ represent the curve $[z_1, z_2]$. We have

$$\begin{aligned}\int_{[z1,z2]} z^n dz &= \int_0^1 \gamma(t)^n \gamma'(t) dt\\ &= (z_2 - z_1)\int_0^1 \big(z_1 + t(z_2 - z_1)\big)^n\, dt.\\ &= (z_2 - z_1)\sum_{k=0}^{n}\binom{n}{k} z_1^{n-k}(z_2 - z_1)^k \int_0^1 t^k dt\\ &= \sum_{k=0}^{n}\frac{\binom{n}{k}}{k+1} z_1^{n-k}(z_2 - z_1)^{k+1}.\end{aligned}$$

And for $\overline{z}$ we have

$$\begin{aligned}\int_{[z1,z2]} \overline{z}^n dz &= \int_0^1 \Big(\overline{\gamma(t)}\Big)^n \gamma'(t) dt\\ &= (z_2 - z_1)\int_0^1 \big(\overline{z_1} + t(\overline{z_2} - \overline{z_1})\big)^n\, dt\\ &= (z_2 - z_1)\sum_{k=0}^{n}\binom{n}{k} \overline{z_1}^{n-k}(\overline{z_2} - \overline{z_1})^k \int_0^1 t^k dt\\ &= \sum_{k=0}^{n}\frac{\binom{n}{k}}{k+1} \overline{z_1}^{n-k}(z_2 - z_1)^{k+1}\end{aligned}$$

VI.8.3 Let the complex-valued function f be defined and continuous in the disc $|z - z_0| < R$.

For $0 < r < R$ let C_r denote the circle $|z - z_0| = r$, with counterclockwise orientation.

VI.8.4 Assume that f is of class C^1. Prove that

$$\lim_{r\to 0} \frac{1}{r^2} \int_{C_r} f(z)dz = 2\pi i \frac{\partial f}{\partial \overline{z}}(z_0).$$

Let $\gamma(\theta) = z_0 + re^{i\theta}$ for $\theta \in [0, 2\pi]$ represent the curve C_r. Then

$$\begin{aligned}\lim_{r\to 0} \frac{1}{r^2} \int_{C_r} f(z)dz &= \lim_{r\to 0} \frac{1}{r^2} \int_0^{2\pi} f(\gamma(\theta))\gamma'(\theta)d\theta \\ &= \lim_{r\to 0} \frac{i}{r} \int_0^{2\pi} f(z_0 + re^{i\theta})e^{i\theta}d\theta\end{aligned}$$

(Not sure where to go from here.)

$$\frac{\partial f}{\partial \overline{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)$$

VI.12.2 Evaluate the integrals $\int_0^\infty \cos t^2 d\theta$ and $\int_0^\infty \sin t^2 d\theta$ (the Fresnel intervals) by integrating e^{-z^2} in the counterclockwise direction around the boundary of the region $\{z : |z| < R, 0 \leq \mathbf{Arg}\ z \leq \frac{\pi}{4}\}$ and letting $R \to \infty$.

We represent the specified curve as $\gamma(\theta) = Re^{i\theta}$ for $\theta \in [0, \frac{\pi}{4}]$, in which case the specified integral is

$$\int_0^{\pi/4} e^{-e^{2i\theta}} Rie^{i\theta}d\theta = Ri \int_0^{\pi/4} e^{i\theta - e^{2i\theta}}d\theta.$$

Give an example of two convergent series whose product diverges.

Chapter 9

Classical Mechanics

9.1 Newton's Laws of Motion

9.1.1 Basics

The basic object of interest is a moving particle. Its position at time t is $\mathbf{r}$. It has that arrow over it because it is a vector. A vector is something that specifies a direction and a magnitude. Think of $\mathbf{r}$ as an arrow from the origin pointing to the current position. Don't think of $\mathbf{r}$ yet as a column vector containing numbers, because we haven't said what coordinate system we're using. Regardless of what coordinate system we use, $\mathbf{r}$ is always a vector pointing from the origin to the current position.

The particle is moving, i.e. the position changes over time. So instead of just writing $\mathbf{r}$, we write $\mathbf{r}(t)$ which says that it's a function of time. Think of that as giving the answer to a question: "At a given time t, what is the position?". The answer (position) is a vector, so we can say that this is a "vector-valued function" (i.e. whatever output it gives, it's always a vector).

Its velocity is a function $\mathbf{v}(t)$ whose value is also a vector (at time t it's going at some speed in some direction). The velocity function $\mathbf{v}(t)$ is the derivative with respect to time of the position function $\mathbf{r}(t)$. That sounds very familiar, but what exactly is the derivative of a vector-valued function?

In normal, non-vector, calculus we imagine some curve like $y = x^2$. So y is a function of x. The value of that function is not a vector; it's just a number (a scalar). The derivative of that function with respect to x is saying: at a particular point along the x-axis, if I start advancing x a tiny bit, how fast is y changing? So, it's the slope of the curve at that point (also just a number, not a vector).

In vector calculus, the derivative of $\mathbf{r}(t)$ with respect to t is saying: at some particular time t, if I start advancing time a tiny bit, where is the position going and how fast is it going there? So the derivative of a vector-valued function is a vector – an arrow with direction and magnitude (speed).

9.1.2 Coordinate systems

Thinking of $\mathbf{r}(t)$ as an arrow with direction and magnitude is correct but a bit abstract. How specifically do we use numbers to represent position? The chapter covers two main coordinate systems. Let's say the particle is moving in 2D space for now.

- **Cartesian coordinates**: we write down how far the particle currently is in the x-direction, $x(t)$, and how far it currently is in the y-direction, $y(t)$.

- **Polar coordinates**: we write down how far the particle currently is, $r(t)$, in the current direction to the particle.

Note that $x(t)$, $y(t)$, and $r(t)$ were not written with arrows. They are just numbers, saying how far the particle is *in some direction*. The "in some direction" part corresponds to the concept of a *unit vector*. A "unit vector" is basically a vector where the direction is of interest, but the magnitude is just set to 1 for convenience.

Cartesian coordinates use two directions to specify the position. We'll write these directions as the unit vectors $\mathbf{e_x}$ and $\hat{y}$. So in Cartesian coordinates, the position is

<table style="width:100<tr> <td>

$$\mathbf{r}(t) = x(t)\mathbf{e_x} + y(t)\hat{y}$$

[1] http://www.amazon.com/Classical-Mechanics-John-R-Taylor/dp/189138922X

</td> <td>

Go $x(t)$ units in the $\mathbf{e_x}$ direction and $y(t)$ units in the $\hat{y}$ direction

</td> </tr> </table>

In contrast, polar coordinates just use one direction to specify the position: the direction of a direct line to the particle's current position. This direction is the unit vector $\mathbf{e_r}(t)$. So in polar coordinates, the position is

<table style="width:100<tr> <td>

$$\mathbf{r}(t) = r(t)\mathbf{e_r}(t)$$

</td> <td>

Go $r(t)$ units in the $\mathbf{e_r}(t)$ direction

</td> </tr> </table>

Notice (and this is pretty important; it's basically the reason the chapter is covering polar coordinates) that in polar coordinates the unit vector $\mathbf{e_r}(t)$ is a function of time (its direction changes as the particle moves); in contrast, in Cartesian coordinates, $\mathbf{e_x}$ and $\hat{y}$ are constant; they always point in the same direction. The polar unit vector is a function of time because it is the direction to wherever-the-particle-currently-is. The Cartesian unit vectors are not functions of time because they are just the x-axis direction and the y-axis direction and these do not change.

9.1.3 Velocity

We can now differentiate these position functions to get the velocity. Recall that the answer is going to be a vector because it is the derivative of a vector-valued function.

Cartesian coordinates

Because $\mathbf{e_x}$ and $\hat{y}$ are not functions of time, differentiating is straightforward:

$$\mathbf{v}(t) = \frac{d}{dt}\Big(x(t)\mathbf{e_x} + y(t)\hat{y}\Big) = \frac{dx(t)}{dt}\mathbf{e_x} + \frac{dy(t)}{dt}\hat{y}$$

Physicists use a dot to represent derivative-with-respect-to-time. So they might write this as

$$\mathbf{v}(t) = \dot{x}(t)\mathbf{e_x} + \dot{y}(t)\hat{y}$$

Either way, what this is saying is that in Cartesian coordinates, the velocity function is a vector comprised of current x-speed in the x-direction and current y-speed in the y-direction. In other words, it's what you expect.

Polar coordinates

$$\mathbf{v}(t) = \frac{d}{dt}\Big(r(t)\mathbf{e_r}(t)\Big)$$

That's a product of two things that are both a function of time, so we use the "product rule"[ref] The product rule is the thing when you studied differentiation that says: when you're differentiating the product of two functions you differentiate one and keep the other as-is, then you differentiate the other while keeping the first as-is, and you add the two things together: $\frac{d(f(t)g(t))}{dt} = \dot{f}(t)g(t) + f(t)\dot{g}(t)$ [/ref] to differentiate it:

$$\frac{d}{dt}\Big(r(t)\mathbf{e_r}(t)\Big) = \dot{r}(t)\mathbf{e_r}(t) + r(t)\frac{d\mathbf{e_r}(t)}{dt}$$

There's quite a few rs there and it's important at this stage not to get lost in the symbols. We know that the answer (velocity) is a vector. That means we can write it as a bunch of things added together, where each thing is a number times some unit vector. And we're using polar coordinates, so the unit vectors are going to be the polar unit vectors. So the thing on the left $\dot{r}(t)\mathbf{e_r}(t)$ is fine: that's saying that the velocity has one component which is the current radial speed (a number $\dot{r}(t)$) in the current radial direction (the unit vector $\mathbf{e_r}(t)$).

What about the thing on the right? It's the current radial distance times the current derivative of the unit vector function. We've said that in polar coordinates the unit vector $\mathbf{e_r}(t)$ changes over time, so it does make sense that we could ask what its derivative with respect to time is. So what is it? The answer is that it's a vector-valued function whose current value always points at right-angles to the current radial direction, but that requires explaining:

Going back to the informal definition of derivatives above, we're at some point t in time, and we imagine starting to advance time a tiny bit, and we look at the change in where the unit vector points, after this infinitesimally small amount of time passes. A unit vector always has length 1, so it can't grow in length. There's only one thing it can do: it can point in a slightly different direction. What direction has it gone in? It's basically like the hand of a clock. It's not too hard to see that if the hand of a clock changes just a tiny bit, then the tip moves in a direction that's almost a tangent to the circle. Change "tiny" to "infinitesimally small" and the "almost" goes away: so the time derivative of the radial unit vector is a vector pointing at right angles to the radial vector. This unit vector in that direction is called $\mathbf{e}_\phi$, because it points in the direction that you go in when you increase the angle ϕ, as opposed to $\mathbf{e_r}$ which points in the direction you go in if you increase the radius r. How fast does the radial unit vector move in the $\mathbf{e}_\phi$ direction? The answer is that it moves at the speed that the angle is increasing, so $\dot{\phi}$[ref]You can prove this by writing the unit vector in Cartesian coordinates, $cos(\phi)\mathbf{e_x} + sin(\phi)\hat{y}$, and then differentiating it to give $\dot{\phi}\big(-sin(\phi)\mathbf{e_x} + cos(\phi)\hat{y}\big)$ which is $\dot{\phi}$ times a vector orthogonal to the original one.[/ref]. In other words, the time derivative of the radial unit vector is $\dot{\phi}(t)\mathbf{e}_\phi(t)$

The conclusion of all that is that in polar coordinates, the velocity vector is

$$\mathbf{v}(t) = \dot{r}(t)\mathbf{e_r}(t) + r(t)\dot{\phi}(t)\mathbf{e}_\phi(t)$$

Compare this with the expression for velocity in Cartesian coordinates

$$\mathbf{v}(t) = \dot{x}(t)\mathbf{e_x} + \dot{y}(t)\hat{y}$$

and we see it's a bit more complicated in polar coordinates.

I understand the polar coordinates version as follows. At time t the particle might be moving radially, and its angle might also be changing. The velocity vector has two components, one in the radial direction, and one in the tangent direction. In the radial direction, it's moving at whatever speed the radius is changing with. In the tangent direction it's moving at the speed that the angle is changing, multiplied by the current radius. That multiplication by radius makes sense informally, because if you are further out from the center of a circle, and the circle rotates by a few degrees, then you move further in space than if you were closer in to the center.

9.1.4 Acceleration

The acceleration function is the derivative of the velocity function with respect to time. Therefore, it is also a vector: at time t the particle is accelerating by some amount, in some direction.

Cartesian coordinates

Again, because the unit vectors do not change with time, it's as you expect: there's an x-acceleration in the x-direction, and a y-acceleration in the y-direction.

$$\mathbf{a}(t) = \ddot{x}(t)\mathbf{e_x} + \ddot{y}(t)\hat{y}$$

Polar coordinates

Above we saw that because, in polar coordinates, the directions of the coordinate system change with time, the function for velocity was more complicated than when using Cartesian coordinates. For acceleration, we differentiate the velocity expression and of course it gets even more complicated. But basically the answer is still a function of the form

$$\mathbf{a}(t) = \Big(\text{Some function of } t\Big)\mathbf{e_r}(t) + \Big(\text{Another function of } t\Big)\mathbf{e}_\phi(t)$$

The functions of t involve the current radius length, the speed and acceleration in the current radius direction, and the speed and acceleration of the angle parameter ϕ. The full expression is in the footnote[ref]In polar coordinates, if you suppose that you know functions $r(t)$ and $\phi(t)$ giving the angle and distance at time t, then the accelerations in the two orthogonal directions at time t are $\mathbf{a}(t) = \Big(\ddot{r}(t) - r(t)\dot{\phi}(t)^2\Big)\mathbf{e_r}(t) + \Big(2\dot{r}(t)\dot{\phi}(t) + r(t)\ddot{\phi}(t)\Big)\mathbf{e}_\phi(t)$ [/ref].

9.1.5 Newton's second law as a differential equation

A key point seems to be: view Newton's second law $\mathbf{F} = m\mathbf{a}$ as a differential equation[ref]The dot means "differentiated with respect to time". So if r is position as a function of time then $\dot{r}$ is velocity and $\ddot{r}$ is acceleration.[/ref]:

$$m\ddot{\mathbf{r}}(t) = \mathbf{F}$$

I'm understanding this as follows: You know what forces are acting on the body in question. You want to know how the position of the body will evolve through time: $\mathbf{r}(t)$. This is a function satisfying the following differential equation: the second derivative with respect to time of $\mathbf{r}(t)$, times m, is equal to the net force acting on the body.

In practice: in a typical problem you have some expression for $\mathbf{F}$ derived from consideration of a diagram showing forces acting on the body. You might be able to discover $\mathbf{r}(t)$ by finding a function whose second derivative is $\mathbf{F}$.

9.1.6 Example problems

Cartesian coordinates

> 1.37 A student kicks a frictionless puck with initial speed v_0, so that it > slides up a plane that is inclined at an angle θ above the > horizontal. **(a)** Write down Newton's second law for the puck and solve to > give its position as a function of time.

This is a simple example of using the Second Law as a differential equation. We write down the forces acting on the particle, set them equal to $m\ddot{r}(t)$ and integrate twice to get position.

The only force acting on the puck is its weight, i.e. its mass times acceleration due to gravity: mg. The puck can only move along the surface of the plane, so we are only interested in the component of the force that acts parallel to the plane. This component is $-mgsin(\theta)$. So taking x as the direction up the plane, Newton's second law is

$$m\ddot{x}(t) = -mgsin(\theta)$$

Integrating once gives velocity

$$\dot{x}(t) = -gsin(\theta)t + v_0$$

Integrating again gives position

$$x(t) = -\frac{1}{2}gsin(\theta)t^2 + v_0t + x_0$$

and $x_0 = 0$ since we start measuring from its starting position.

> **(b)** How long will the puck take to return to its starting point?

The puck is at its starting point whenever $x = 0$:

$$0 = t\bigg(-\frac{1}{2}gsin(\theta)t + v_0 \bigg)$$

The solutions of that are either $t = 0$ (which we already knew) or (the solution we want)

$t = \frac{2v_0}{gsin(\theta)}$

Polar coordinates

> A "halfpipe" at a skateboard park consists of a concrete trough with a > semicircular cross section of radius $R = 5m$. I hold a frictionless > skateboard on the side of the trough pointing down toward the bottom and > release it. Discuss the subsequent motion using Newton's second law. In > particular, if I release the skateboard just a short way from the bottom, how > long will it take to come back to the point of release?

Conceptually, we do the same thing as for the problem using Cartesian coordinates: we write down Newton's second law resolved into two orthogonal directions. It's just that with polar coordinates, these orthogonal directions are constantly changing.

The weight of the skateboard acts downwards. This results in a tangent force causing the skateboard to move along the halfpipe, and also presses the skateboard into the halfpipe a bit, with an associated reaction force. We ignore the force/reaction force between the skateboard and the pipe and focus only on the tangent force: $-mgsin(\phi)$.

The equation for acceleration says that, at time t, acceleration in the current tangent direction is $R\ddot{\phi}(t)$ (halfpipe radius times current angular acceleration[ref]To see this, start with the $\mathbf{e}_\phi(t)$ (tangent direction) part of the full expression for acceleration and note that the radial distance of the skateboard is fixed by the

presence of the half-pipe, so speed $\dot{r}(t)$ (and acceleration) in the radial direction is zero.[/ref]). So Newton's second law in this context is the differential equation

$$mR\ddot{\phi}(t) = -mgsin(\phi(t))$$

We read this as saying:

> We don't know how the angle is changing over time $\phi(t)$ – that is > precisely what we want to know. But what we do know is that whatever that > function is, its second derivative at time t is equal to the sin of the > current angle (times g/R and with a minus sign because the way we've > defined the angle it gets smaller as the weight force takes the skateboard > towards the bottom).

Once we've got to that point, finding the angle function $\phi(t)$ is just math. It turns out that the only function for which it is true that the second derivative has this property[ref]Actually the solution is a function with second derivative having a different property, but one which is very similar to the desired property as long as we're restricting ourselves to the angle being fairly small.[/ref] is

$$\phi(t) = \phi_0 cos\left(\sqrt{\frac{g}{R}}t\right)$$

where ϕ_0 is the angle that the skateboard was released at at time $t = 0$. This is the "solution" of the differential equation: a function matching the criteria that the differential equation specified.

So we have our answer: the forces acting on the skateboard imply (via Newton's second law) that the way the angle of the skateboard changes is a cosine function of time. So the skateboard angle does what cosines do: it starts off at its maximum, decreases to zero, crosses zero and becomes negative for a while, starts turning back towards zero, crosses zero and becomes positive again and gets back to its maximum where it turns around again.

9.1.7 Conservation of momentum

Momentum is mass times velocity, $\mathbf{p}(t) = m\dot{\mathbf{r}}(t)$, so another way of stating the second law is: rate of change of momentum is equal to force. In a multi-particle system the forces-and-reaction-forces of the third law cancel each other out when summing the rate of change of momentum of the whole system. So, total momentum doesn't change due to internal forces (but it does if there are external forces).

pp 21-23 show that conservation of momentum does not hold when considering magnetic and electrostatic forces between charged particles moving close to the speed of light. However I am unfamiliar with those forces and with the "right-hand rule" for fields/forces and I haven't understood this section.

Chapter 10

Machine Learning

- n sample points $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$
- $d = 2$ where not stated.

10.1 Overview

Linear regression lays down a linear surface over $\mathbb{R}^d$. The parameters of that surface ($\mathbf{w}$) are scored according to the sum of squared distances of the y values from the surface.

Logistic regression lays down a logistic surface over $\mathbb{R}^d$. The parameters of that surface ($\mathbf{w}$) are scored according to the probability of drawing the y values from the probability distribution given by the surface.

In both cases, the score (cost) is a measure of distance of the y values from the surface, i.e. a distance between $\mathbf{y}$ and the predictions $\hat{\mathbf{y}}$.

The surface over $\mathbb{R}^d$ maps the n sample points $\mathbf{x}_i \in \mathbb{R}^d$ to their predictions y_i in $\mathbb{R}$ or $[0, 1]$.

10.2 Neural networks

A neural net with one hidden layer of K units first maps $\mathbf{x}_i \in \mathbb{R}^d \to \mathbb{R}^K$ using parameter matrix $\mathbf{V}$, and then maps $\mathbb{R}^K \to \mathbb{R}$ using parameters $\mathbf{w}$. Again, the cost associated with parameters $(\mathbf{V}, \mathbf{w})$ is a distance between $\mathbf{y}$ and the $\hat{\mathbf{y}}$ values in the output layer.

10.2.1 Backpropagation algorithm

No hidden layers: linear regression

$$\mathbf{x} \xrightarrow{\mathbf{w}} (\hat{y} = \mathbf{x}^\mathbf{T}\mathbf{w}) \to L$$

We want to do gradient descent on $\mathbf{w}$.

$$\begin{aligned}\frac{\partial L}{\partial w_j} &= \sum_i \frac{\partial L}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial w_j} \\ &= \sum_i 2(\hat{y}_i - y_i)x_{ij} \\ &= \sum_i 2(\mathbf{x}_i^\mathbf{T}\mathbf{w} - y_i)x_{ij}\end{aligned}$$

Alternatively we can compute the gradient using the chain rule:

$$L(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 = (\mathbf{X}\mathbf{w} - \mathbf{y})^\mathbf{T}(\mathbf{X}\mathbf{w} - \mathbf{y})$$

so

$$\nabla_\mathbf{w} L = 2\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{w} - 2\mathbf{X}^\mathbf{T}\mathbf{y}$$

the j-th component of which is

$$\frac{\partial L}{\partial w_j} = 2w_j 2X_{\cdot j} \cdot \mathbf{y}$$

One hidden layer

Consider classification using a neural net with one hidden layer $\mathbf{h}$ of H units.

We consider one sample point x at a time.

There are K possible categories, and the predictions $\hat{y}$ in the output layer can be interpreted as the probabilities each category given the input x.

Model specification:

K possible output categories; one hidden layer of H units; tanh activation in the hidden layer; logistic activation in the output layer. Notation:

		indices	dimensions
Input layer	$\mathbf{x}$	x_j	$d \times 1$
Weights	$\mathbf{V}$	V_{hj}	$H \times d$
Hidden layer	$\mathbf{z} = \tanh(\mathbf{Vx})$	z_h	$H \times 1$
Weights	$\mathbf{W}$	W_{kh}	$K \times H$
Ouput layer	$\hat{\mathbf{y}} = \sigma(\mathbf{Wz})$	$\hat{y}_k$	$K \times 1$
Loss	$L(\hat{\mathbf{y}}, \mathbf{y})$		scalar

where σ is the logistic function $\sigma(x) = (1 - e^{-x})^{-1}$, and tanh and σ act elementwise.

The loss (cost) function is the cross-entropy (log likelihood of training labels given predictions)

$$-L(\hat{y}, \mathbf{y}) = \sum_k y_k \log(\hat{y}_k) + (1 - y_k) \log(1 - \hat{y}_k).$$

Gradient descent algorithm

We want to do gradient descent on the full set $(\mathbf{V}, \mathbf{W})$ of parameters. This involves computing gradients of the loss function $\nabla_V L$ and $\nabla_W L$. We derive the gradients with respect to one row of these matrices at a time, and give code fragments showing how to compute the matrix of derivatives efficiently.

Gradient with respect to weight matrix W

$\mathbf{W}_k$ is one row of $\mathbf{W}$, of length $H + 1$. We have

$$\nabla_{\mathbf{W}_k} L = \frac{\partial L}{\partial \hat{y}_k} \nabla_{\mathbf{W}_k} \hat{y}_k.$$

Now, $\hat{y}_k = \sigma(\mathbf{W}_k \mathbf{z})$, so

$$\nabla_{\mathbf{W}_k} \hat{y}_k = \mathbf{z} \hat{y}_k (1 - \hat{y}_k).$$

This expression is still correct if the offset is implemented as an additional "dimension", in which case the last element of $\mathbf{W}_k$ is the offset and the last element of $\mathbf{z}$ is 1.

The derivative of the loss with respect to $\hat{y}_k$ is

$$\frac{\partial L}{\partial \hat{y}_k} = -\frac{y_k}{\hat{y}_k} + \frac{1 - y_k}{1 - \hat{y}_k} = \frac{\hat{y}_k - y_k}{\hat{y}_k (1 - \hat{y}_k)}.$$

Multiplying these quantities gives

$$\nabla_{\mathbf{W}_k} L = \mathbf{z}(\hat{y}_k - y_k).$$

In code we can compute the full matrix of derivatives $\nabla_{\mathbf{W}}$ using vector/matrix primitives as

$$\text{diag}(\hat{\mathbf{y}} - \mathbf{y})\ \mathbf{Z},$$

where the rows of $\mathbf{Z}$ are each equal to $\mathbf{z}$:

```
grad__L__z = (W.T * (yhat - y)).sum(axis=1)
zz = z.reshape((1, H + 1)).repeat(K, 0)
grad__L__W = diag(yhat - y) @ zz
```

Gradient with respect to weight matrix V

$\mathbf{V}_h$ is one row of $\mathbf{V}$, of length $d + 1$. We have

$$\nabla_{\mathbf{V}_h} L = \frac{\partial L}{\partial \mathbf{z}_h} \nabla_{\mathbf{V}_h} \mathbf{z}_h.$$

Now, $\frac{\partial L}{\partial z_h} = \sum_k \frac{\partial L}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial z_h}$. We've already found $\frac{\partial L}{\partial \hat{y}_k}$ above, and $\frac{\partial \hat{y}_k}{\partial z_h} = W_{kh}\hat{y}_k(1 - \hat{y}_k)$, giving

$$\frac{\partial L}{\partial z_h} = \sum_k W_{kh}(\hat{y}_k - y_k).$$

$\mathbf{z}_h = \tanh(\mathbf{V}_h\mathbf{x})$, so $\nabla_{\mathbf{V}_h} \mathbf{z}_h = \mathbf{x}(1 - z_h^2)$, and multiplying the two quantities gives

$$\nabla_{\mathbf{V}_h} L = \mathbf{x}(1 - z_h^2) \sum_k W_{kh}(\hat{y}_k - y_k).$$

Again, in code we can compute the full matrix of derivatives $\nabla_{\mathbf{V}} L$ using vector/matrix primitives:

```
grad__L__z = (W.T * (yhat - y)).sum(axis=1)
xx = x.reshape((1, d + 1)).repeat(H + 1, 0)
grad__L__V = diag((1 - z ** 2) * grad__L__z) @ xx
```

10.2.2 Other neural network notes

So the objective function is $L(\hat{\mathbf{y}}(\mathbf{z}(\mathbf{x})))$, or

$$\mathbf{x} \xrightarrow{\mathbf{V}} \mathbf{z} \xrightarrow{\mathbf{W}} \hat{\mathbf{y}} \to L$$

We want to compute the gradient vector, i.e. partials $\frac{\partial L}{\partial V_{hj}}$ and $\frac{\partial L}{\partial w_k}$.

Recall that $\sigma' = \sigma(1-\sigma)$, and note that $\hat{y}_k = \sigma(\mathbf{w}_k^{\mathbf{T}}\mathbf{z})$, so

$$\frac{\partial \hat{y}_k}{\partial W_{kh}} = \hat{y}_k(1-\hat{y}_k)\frac{\partial \mathbf{w}_k^{\mathbf{T}}\mathbf{z}}{\partial W_{kh}} = \hat{y}_k(1-\hat{y}_k)z_h.$$

The gradient with respect to $\mathbf{W}$ is

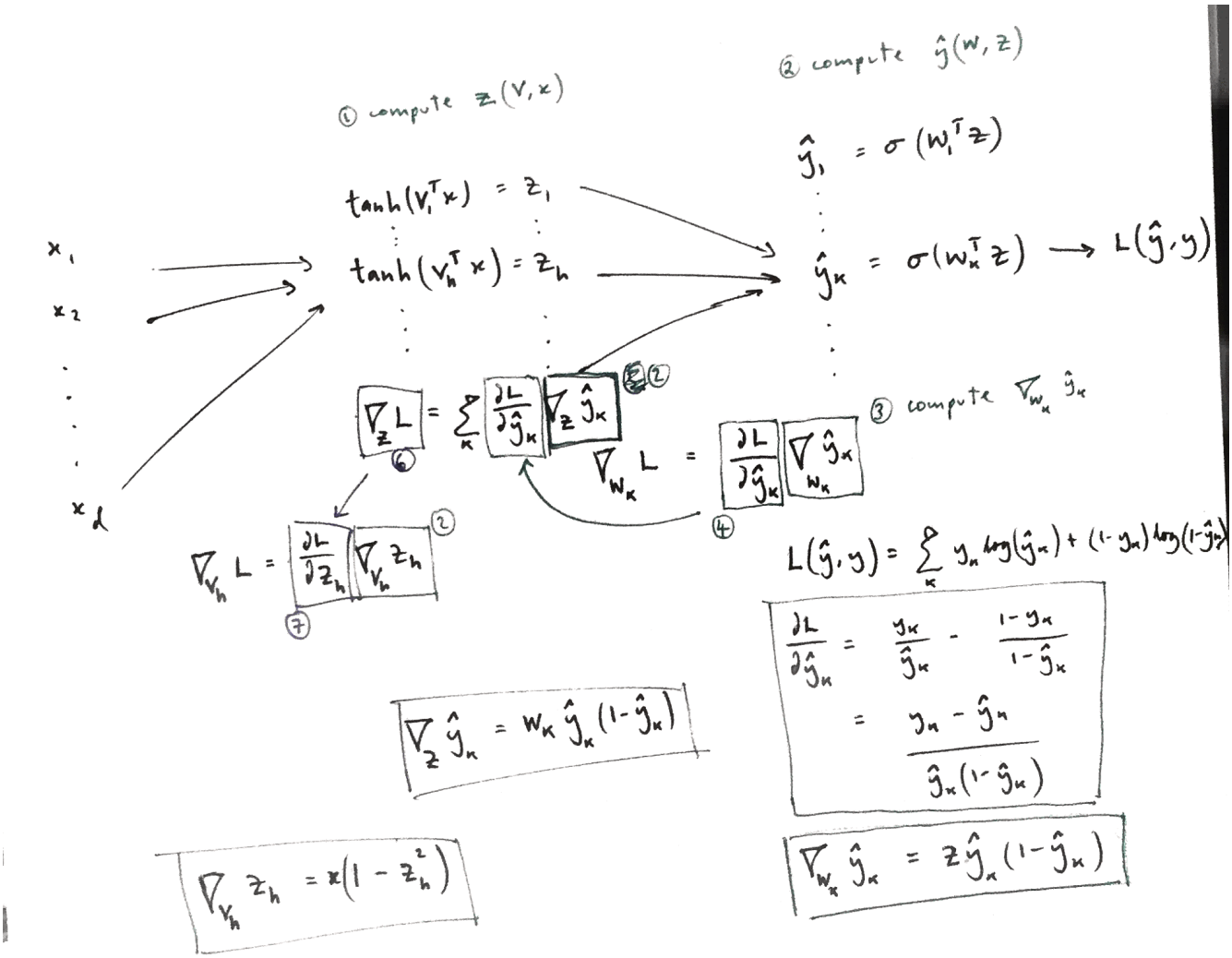
$$\nabla_{\mathbf{w}_k} L = \mathbf{z}(y_k - \hat{y}_k)$$

(proof similar to that in Logistic Regression section), or non-vectorized version:

$$\begin{aligned}
\frac{\partial L}{\partial W_{kh}} &= \frac{\partial}{\partial W_{kh}} \sum_{k'} y_{k'} \log(\hat{y}_{k'}) + (1-y_{k'}) \log(1-\hat{y}_{k'}) \\
&= y_k \frac{\hat{y}_k(1-\hat{y}_k)z_h}{\hat{y}_k} - (1-y_k)\frac{\hat{y}_k(1-\hat{y}_k)z_h}{1-\hat{y}_k} \\
&= z_h\big(y_k(1-\hat{y}_k) - (1-y_k)\hat{y}_k\big) \\
&= z_h\big(y_k - \hat{y}_k\big)
\end{aligned}$$

The gradient with respect to $\mathbf{V}$ is given by

$$\nabla_{\mathbf{V}_h} L = \sum_k \frac{\partial L}{\partial \hat{y}_k} \nabla_{\mathbf{v}_h} \hat{y}_k$$



10.2.3 Trivial case

Forwards

1. 1d input 0 with offset dimension: $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
2. $K = 1$. Label $y = 1$
3. Initial $V = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ (last row ignored)
4. $Vx = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ (last element ignored)
5. 1 hidden unit. $z \leftarrow \begin{bmatrix} \tanh(0) \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
6. Initial $W = \begin{bmatrix} 0 & 0 \end{bmatrix}$

7. $Wz = \begin{bmatrix}0 & 0\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix} = 0$

8. $\hat{y} = s(0) = 0.5$

One iteration of backpropagation

1. $\nabla_{W_k}\,\hat{y}_k = \mathbf{z}\hat{y}_k(1-\hat{y}_k) = \begin{bmatrix}0\\0.25\end{bmatrix}$

$$\nabla_{W_k}\,L = \frac{\partial L}{\partial \hat{y}_k}\,\nabla_{W_k}\,\hat{y}_k$$
$$\frac{\partial L}{\partial \hat{y}_k} = \frac{y_k - \hat{y}_k}{\hat{y}_k(1-\hat{y}_k)}$$
$$\nabla_{W_k}\,\hat{y}_k = z\hat{y}_k(1-\hat{y}_k)$$

$$\nabla_z\,L = \sum_k \frac{\partial L}{\partial \hat{y}_k}\,\nabla_z\,\hat{y}_k$$
$$\nabla_z\,\hat{y}_k = W_k\hat{y}_k(1-\hat{y}_k)$$

$$\nabla_{V_h}\,L = \frac{\partial L}{\partial z_h}\,\nabla_{V_h}\,z_h$$
$$\nabla_{V_h}\,z_h = x(1-z_h^2)$$

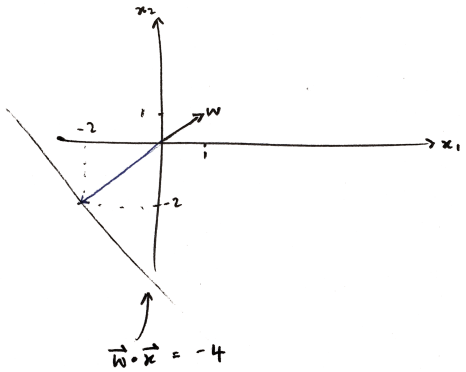
10.3 Classification

A **decision boundary** is a curve separating the plane (sample space) into two regions.

Some classifiers involve a **decision function** f, in which case $f(\mathbf{x}) = 0$ describes the decision boundary.

A **linear classifier** uses a linear decision function $f(x) = \mathbf{w}\cdot\mathbf{x} + \alpha$. This is scalar-valued: it's a plane over the plane (sample space). Its intersection defines a linear decision boundary.

In d-dimensions the decision boundary is a hyperplane ($(d-1)$-dimensional). This still separates the sample space into two regions.



Example: $f(x) = \begin{bmatrix}1\\1\end{bmatrix} \cdot \begin{bmatrix}x_1\\x_2\end{bmatrix} + 4$

- A plane sloping up at 45° in the north-east direction.
- Each input feature has equal influence on the classification.
- Decision boundary is line $x_1 + x_2 = -4$.
- $\mathbf{w}$ is normal to the decision boundary since $\mathbf{w} \cdot (\mathbf{x}_1 - \mathbf{x}_2) = -4 - (-4) = 0$.
- If one feature has a very high weight then $\mathbf{w}$ points close to that axis and the decision boundary is almost perpendicular to that axis (other features almost don't matter).

Distance from the decision boundary to a point: For some point $\mathbf{x}_i$, the height of the decision function plane above $\mathbf{x}_i$ is $\mathbf{w} \cdot \mathbf{x}_i + \alpha$. At the decision boundary, this height is zero. Looking "straight up" the slope of the decision function, its gradient is $\sqrt{w_1^2 + w_2^2} = |\mathbf{w}|$. So the distance of a point $\mathbf{x}_i$ from the hyperplane is $\frac{\mathbf{w} \cdot \mathbf{x}_i + \alpha}{|\mathbf{w}|}$. If $\mathbf{w}$ is not a unit vector, the problem can be rescaled so that it is, in which case the distance is $\mathbf{w} \cdot \mathbf{x}_i + \alpha$.

Examples of linear classifiers:

- **Centroid method**: Decision boundary perpendicular to and bisects line connecting means of labeled training points.
- **Perceptron**:
- **Maximum margin classifier**:
- **LDA**: Fit Gaussians to each class, same covariance across classes.

10.3.1 Perceptron

Labels $y_i \in \{-1, 1\}$. Assume $\alpha = 0$ for now (decision boundary through origin).

Goal: find line separating points (separating hyperplane). I.e. Find $\mathbf{w}$ such that

$$\begin{cases} \mathbf{x}_i \cdot \mathbf{w} \leq 0, & y_i = -1 \\ \mathbf{x}_i \cdot \mathbf{w} \geq 0, & y_i = +1. \end{cases}$$

This is equivalent to the **constraint** $y_i \mathbf{x}_i \cdot \mathbf{w} \geq 0$.

Cost function: total distance $R(\mathbf{w})$ of misclassified points from the decision boundary.

Optimization problem: Find $\mathbf{w}$ that minimizes

$$R(w) = \sum_i L(\mathbf{x}_i \cdot \mathbf{w}, y_i) = \sum_{i \in V} -y_i \mathbf{x}_i \cdot \mathbf{w},$$

where V are the misclassified points.

Per-training point loss function

$$L(\text{prediction}_i, y_i) = L(\mathbf{x}_i \cdot \mathbf{w}, y_i) = \begin{cases} 0, & \text{correct}, y_i \mathbf{x}_i \cdot \mathbf{w} \geq 0 \\ -y_i \mathbf{x}_i \cdot \mathbf{w}, & \text{misclassified} \end{cases}$$

Gradient descent: Find w that minimizes $R(w)$.

$$\nabla_w R = \begin{bmatrix} -\sum_i y_i X_{i1} \\ \vdots \\ -\sum_i y_i X_{id} \end{bmatrix}$$

- On each iteration, compute the gradient; update $\mathbf{w}$ by taking a step downhill of size ρ: $\mathbf{w} \leftarrow \mathbf{w} + \rho \sum_{i \in V} y_i \mathbf{x}_i$.
- A misclassified data point far out in dimension j will cause the gradient to have a large component $-\sum_i y_i X_{ij}$ in that dimension.
- $\mathbf{w}$ thus becomes more closely aligned with that axis and the decision boundary.
- Decision boundary therefore becomes more perpendicular to that axis (axis becomes more "important").

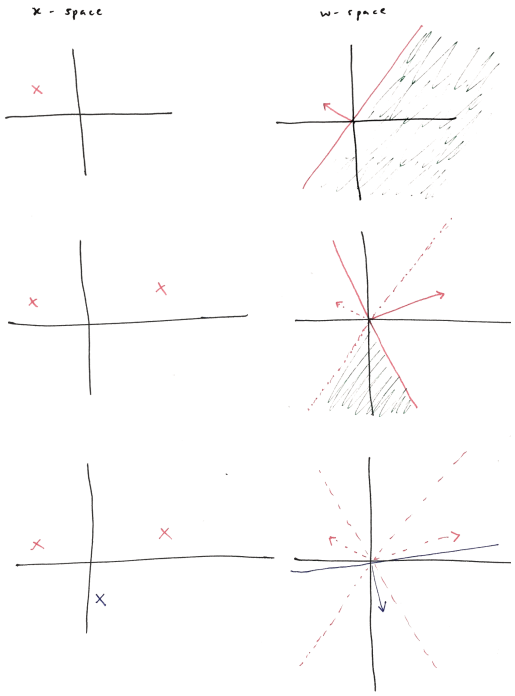
Stochastic gradient descent (Perceptron): on each iteration pick one misclassified point and update $\mathbf{w}$ using gradient for that point: $\mathbf{w} \leftarrow \mathbf{w} + \rho y_{i^*} \mathbf{x}_{i^*}$

Allow decision boundaries that do not pass through origin: add a fictitious dimension so that sample points now lie on the plane $x_{d+1} = 1$ in $(d+1)$ dimensions. Run algorithm as above, just with the new dimensionality.

$$\mathbf{w} \cdot \mathbf{x} + \alpha = 0$$
$$\begin{bmatrix} w_1 \\ w_2 \\ \alpha \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} = 0.$$

10.3.2 Optimization in weight space

$\mathbf{x}$-space	$\mathbf{w}$-space
hyperplane	point $\mathbf{w}$ is normal vector to hyperplane
point	hyperplane whose normal vector is the $\mathbf{x}$ point (? don't understand this yet)



10.3.3 Maximum margin classifiers

Margin is distance from hyperplane to nearest sample point.

Previously, in the perceptron, we used the constraint

$$y_i \mathbf{x}_i \cdot \mathbf{w} \geq 0.$$

Now, we demand that there is a non-zero margin between the decision boundary and the points:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + \alpha) \geq 1,$$

The 1 on the RHS is arbitrary; I think $\mathbf{w}$ and α will adapt to make it true for any positive value, so the point is that we're demanding a strictly non-zero margin.

Optimization problem (quadratic program):
Find $\mathbf{w}, \alpha$ that minimize $|\mathbf{w}|^2$ such that $y_i(\mathbf{x}_i \cdot \mathbf{w} + \alpha) \geq 1$ for all points i.

10.3.4 Soft margin SVMs

[1] [2]

- Still quadratic program but allow points to violate margin via **slack variables** $\xi_i \geq 0$:
- Constraint is $y_i(\mathbf{x}_i \cdot \mathbf{w} + \alpha) \geq 1 - \xi_i$

[1] https://people.eecs.berkeley.edu/~jrs/189/lec/04.pdf
[2] https://www.youtube.com/watch?v=HOZ6ZpPA_Ks

- Find non-linear decision boundaries by introducing new features comprising non-linear functions of base features ("lift points into higher-dimensional space").

Optimization problem:

Find w, α, and ξ_i that minimize $\|w\|^2 + C \sum_{i=1}^{n} \xi_i$		
subject to	$y_i(X_i \cdot w + \alpha) \geq 1 - \xi_i$	for all $i \in [1, n]$
	$\xi_i \geq 0$	for all $i \in [1, n]$

. . . a quadratic program in $d + n + 1$ dimensions and $2n$ constraints.
[It's a quadratic program because its objective function is quadratic and its constraints are linear inequalities.]

$C > 0$ is a scalar regularization hyperparameter that trades off:

	small C	big C
desire	maximize margin $1/\|w\|$	keep most slack variables zero or small
danger	underfitting (misclassifies much training data)	overfitting (awesome training, awful test)
outliers	less sensitive	very sensitive
boundary	more "flat"	more sinuous

10.4 Decision Theory

[3] [4]

Suppose there are two possible **classes**: $\{C, D\}$

Decision rule: $r(\mathbf{x}) : \mathbb{R}^d \to \{C, D\}$

Loss function: E.g. 0-1 loss:

$$L(y_i \to \hat{y}_i) = \begin{cases} 0, & \hat{y}_i = y_i \\ 1, & \text{otherwise} \end{cases} \qquad \text{(correct classification)}$$

Risk: Functional $R(r)$: expected loss for rule r, over $\mathbf{p}(X, Y)$. [5]

So what rule function r minimizes the functional R?

Bayes decision rule: Assign $\mathbf{x}$ to class C if

$$\text{(C posterior at } \mathbf{x}) \times \text{(penalty for misclassifing a true C)}$$

is largest for class C. I.e. if

$$\mathbf{p}(C|\mathbf{x})L(D|C) > \mathbf{p}(D|\mathbf{x})L(C|D).$$

With 0-1 loss, this is: "assign to class with highest posterior".

With 0-1 loss and two classes, it's: "assign to class with posterior > 0.5".

Empirical risk: Discriminative methods (e.g. logistic regression) lack any model for X. How can we estimate expected loss over $p(X, Y)$? Take the observed sample points as defining a discrete, uniform distribution, in which case

$$\hat{R}(r) = \frac{1}{n} \sum L(r(x_i), y_i).$$

This provides a justification for minimizing the sum/mean of per-sample loss.

[3] https://people.eecs.berkeley.edu/~jrs/189/lec/06.pdf
[4] https://www.youtube.com/watch?v=aXkenQ01qYI
[5]

$$\begin{aligned} R(r) =& \pi(Y=-1)\,\mathbb{E}_{\mathbf{X}}\, L(-1 \to r(X)) + \\ & \pi(Y=+1)\,\mathbb{E}_{\mathbf{X}}\, L(+1 \to r(X)) \qquad \text{over } \mathbf{p}(Y)\mathbf{p}(X|Y) \\ =& \sum_X \mathbf{p}(X)\big(\pi(Y=-1)L(-1 \to r(X)) + \\ & \pi(Y=+1)L(+1 \to r(X))\big) \qquad \text{over } \mathbf{p}(X)\mathbf{p}(Y|X) \end{aligned}$$

10.5 Statistical justifications

Regression: want to estimate a function f such that $y_i = f(x_i) + \epsilon$, where ϵ has unknown distribution but mean 0. Ideal would be to estimate f with $h(x_i) = \mathbb{E}(Y|x_i)$ since this is equal to $f(x_i)$.

Likelihood justification for linear regression cost function.

Logistic Regression from Maximum Likelihood

10.6 Bias-Variance Decomposition

$$
\begin{aligned}
&= \mathrm{E}[(h(z) - \gamma)^2] \\
&= \mathrm{E}[h(z)^2] + \mathrm{E}[\gamma^2] - 2\,\mathrm{E}[\gamma\, h(z)] \qquad \text{[Observe that } \gamma \text{ and } h(z) \text{ are independent]} \\
&= \mathrm{Var}(h(z)) + \mathrm{E}[h(z)]^2 + \mathrm{Var}(\gamma) + \mathrm{E}[\gamma]^2 - 2\mathrm{E}[\gamma]\,\mathrm{E}[h(z)] \\
&= (\mathrm{E}[h(z)] - \mathrm{E}[\gamma])^2 + \mathrm{Var}(h(z)) + \mathrm{Var}(\gamma) \\
&= \underbrace{(\mathrm{E}[h(z)] - f(z))^2}_{\text{bias}^2 \text{ of method}} + \underbrace{\mathrm{Var}(h(z))}_{\text{variance of method}} + \underbrace{\mathrm{Var}(\epsilon)}_{\underline{\text{irreducible error}}}
\end{aligned}
$$

10.7 Gaussian discriminant analysis

6 7

Anisotropic:

$$\mathbf{p}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathbf{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

Isotropic:

$$\mathbf{p}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{|\mathbf{x}-\mu|^2}{2\sigma^2}\right)$$

10.7.1 Isotropic Gaussians

Multivariate data $\mathbf{x}$ but features uncorrelated and all features same variance.

QDA

Fit separate Gaussians to the training data in each class. The likelihood is

$$\mathbf{p}(\mathbf{x}|\text{class } C) = \frac{1}{(2\pi)^{d/2}\sigma_C^d} \exp\left(-\frac{|\mathbf{x}-\mu_C|^2}{\sigma_C^2}\right)$$

and we compare the value of $\mathbf{p}(\mathbf{x}|\text{class } C) \cdot \pi_C \cdot L(D|C)$.

The decision boundaries are where the posterior $\times$ loss are equal. It's easier to compare the log of this:

$$Q_C(\mathbf{x}) = -\frac{|\mathbf{x}-\mu_C|^2}{\sigma_C^2} - d\log\sigma_C + \log\pi_C + \log L(D|C)$$

The posterior probability of class C at point $\mathbf{x}$ is[8]

$$\mathbf{p}(C|\mathbf{x}) = \frac{\pi_C\mathbf{p}(\mathbf{x}|C)}{\pi_C\mathbf{p}(\mathbf{x}|C) + \pi_D\mathbf{p}(\mathbf{x}|D)} = \frac{1}{1+e^{-(Q_C(\mathbf{x})-Q_D(\mathbf{x}))}},$$

so logistic in the quadratic expression $Q_C(\mathbf{x}) - Q_D(\mathbf{x})$.

[6] https://people.eecs.berkeley.edu/~jrs/189/lec/07.pdf
[7] https://www.youtube.com/watch?v=4CefboCXxZs
[8] This is assuming 0-1 loss, so the loss doesn't affect $Q_C(\mathbf{x})$

LDA

Estimate separate class means but same variance for all classes. So now

$$\begin{aligned} Q_C(\mathbf{x}) - Q_D(\mathbf{x}) &= \frac{|\mathbf{x}-\mu_D|^2 - |\mathbf{x}-\mu_C|^2}{\sigma^2} + \log\frac{\pi_C}{\pi_D} + \log\frac{L(D|C)}{L(C|D)} \\ &= \frac{(\mathbf{x}-\mu_D)\cdot(\mathbf{x}-\mu_D) - (\mathbf{x}-\mu_C)\cdot(\mathbf{x}-\mu_C)}{\sigma^2} + \log\frac{\pi_C}{\pi_D} + \log\frac{L(D|C)}{L(C|D)} \\ &= \mathbf{x}\cdot\frac{2(\mu_C-\mu_D)}{\sigma^2} + \left(\frac{|\mu_D|^2 - |\mu_C|^2}{\sigma^2} + \log\frac{\pi_C}{\pi_D} + \log\frac{L(D|C)}{L(C|D)}\right) \\ &= \mathbf{x}\cdot\mathbf{w} + \alpha \end{aligned}$$

This means that the decision boundary is linear, and (with 0-1 loss) the posterior is a logistic function which is constant parallel to the decision boundary.

10.8 Symmetric matrices, quadratic forms and eigenvectors

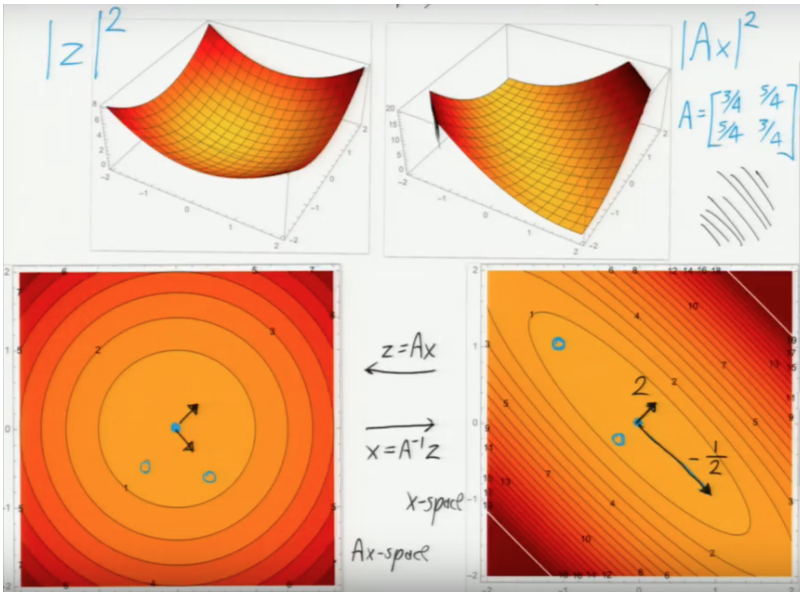
[9]

Spectral theorem: A symmetric matrix has n **orthogonal** eigenvectors[10][11]

To understand a symmetric matrix $\mathbf{A}$, consider its **quadratic form** $|\mathbf{Ax}|^2 = \mathbf{x^T A^2 x}$ (right). Compare this to the graph of $|\mathbf{z}|^2$ (left). The graphs are related by the following changes of coordinates:

$\mathbf{z} \leftarrow \mathbf{Ax}$ changes the elliptical contours into circles; scale by eigenvalues of $\mathbf{A}$.

$\mathbf{A}^{-1}\mathbf{z} \rightarrow \mathbf{x}$ changes circles into ellipses; scale by reciprocal of eigenvalues.



$|\mathbf{Ax}|^2 = 1$ is the equation of an ellipsoid. Its axes are $v_1, \ldots, v_n$ and its radii are $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$,

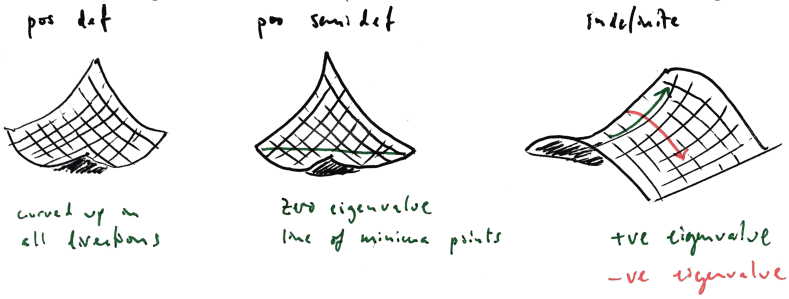
[9] https://people.eecs.berkeley.edu/~jrs/189/lec/08.pdf
[10] There may be more than n (infinite) eigenvectors, but n orthogonal.
[11] Non-symmetric matrices have non-orthogonal eigenvectors in general.

Bigger eigenvalue $\iff$ steeper hill.

Alternate interpretation: the ellipsoids are spheres in a space with a different distance metric. The distance metric (metric tensor) is $\mathbf{M} = \mathbf{A}^2$:

$$d(\mathbf{x}, \mathbf{x}') = |\mathbf{A}\mathbf{x}| - |\mathbf{A}\mathbf{x}'| = \sqrt{(\mathbf{x} - \mathbf{x}')\mathbf{A}^2(\mathbf{x} - \mathbf{x}')}$$

These are diagrams of $\mathbf{x}^\mathbf{T}\mathbf{A}\mathbf{x}$ (not $\mathbf{x}^\mathbf{T}\mathbf{A}^2\mathbf{x}$ since $\mathbf{A}^2$ has no negative eigenvalues):



positive definite	eigenvalues > 0	$\mathbf{x}^\mathbf{T}\mathbf{A}\mathbf{x} > 0 \quad \forall \mathbf{x} \neq \mathbf{0}$
positive semidefinite	eigenvalues ≥ 0	$\mathbf{x}^\mathbf{T}\mathbf{A}\mathbf{x} >= 0 \quad \forall \mathbf{x}$
indefinite	some positive and some negative eigenvalues	
singular	some zero eigenvalue	

Let Λ be a diagonal matrix containing the eigenvalues and $\mathbf{V}$ contain normalized eigenvectors:

$$V = \begin{bmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\ | & | & & | \end{bmatrix}$$

Note that for an **orthonormal** matrix like this:

1. It rotates / reflects the input vectors, without changing their length.
2. $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, therefore $\mathbf{V}^{-1} = \mathbf{V}^\mathbf{T}$.

By the definition of eigenvector we have

$$\mathbf{A}\mathbf{V} = \mathbf{V}\Lambda$$

and therefore the **eigendecomposition** of $\mathbf{A}$

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^\mathbf{T}.$$

So we can perform $\mathbf{A}\mathbf{x}$ as $\mathbf{V}\Lambda\mathbf{V}^\mathbf{T}\mathbf{x}$, and $\mathbf{A}^k\mathbf{x}$ as $\mathbf{V}\Lambda^k\mathbf{V}^\mathbf{T}\mathbf{x}$:

1. $\mathbf{V}^\mathbf{T} = \mathbf{V}^{-1}$ rotates the input vector into axis-aligned coordinates.
2. Λ scales along different axes.
3. $\mathbf{V}$ returns to the original coordinates.

Λ is said to be the diagonalized version of $\mathbf{A}$.

10.9 The Anisotropic Multivariate Normal Distribution, QDA, and LDA

10.10 Regression

10.10.1 Linear Least Squares Regression

Use fictitious dimension trick, so that $\mathbf{w}$ includes the offset term α and $\mathbf{X}$ is $(n \times (d+1))$.

Find $\mathbf{w}$ that minimizes cost function $J(w)$: sum of squared difference between linear predictor and observed training point.

$$J(w) = |\mathbf{Xw} - \mathbf{y}|^2 = \sum_i (\mathbf{x_i^T w} - y_i)^2$$

Solve by differentiating and finding the critical point:

$$\begin{aligned} |\mathbf{Xw} - \mathbf{y}|^2 &= \mathbf{w^T X^T X w} - 2\mathbf{y^T X w} + \mathbf{y^T y} \\ \nabla_{\mathbf{w}} |\mathbf{Xw} - \mathbf{y}|^2 &= 2\mathbf{X^T X w} - 2\mathbf{X^T y} \\ \mathbf{w}^* &= (\mathbf{X^T X})^{-1}\mathbf{X^T y} =: \mathbf{X}^+\mathbf{y} \end{aligned}$$

For a new sample point $\mathbf{x}$, the prediction is $\hat{y} = \mathbf{x} \cdot \mathbf{w}^*$.

Related concepts

- **normal equations**: linear system of d equations in unknown $\mathbf{w}$ resulting from setting the gradient equal to zero: $\mathbf{X^T X w} - \mathbf{X^T y} = \mathbf{0}$
- **pseudoinverse**: The matrix $\mathbf{X}^+ = (\mathbf{X^T X})^{-1}\mathbf{X^T}$ maps $\mathbf{y}$ to $\mathbf{w}^*$. In general there's no $\mathbf{w}$ that solves $\mathbf{Xw} = \mathbf{y}$, but $\mathbf{w}^* = \mathbf{X}^+\mathbf{y}$ makes the LHS as close as possible to $\mathbf{y}$. So it behaves as a "left inverse" of $\mathbf{X}$, since $\mathbf{X}^+\mathbf{X} = \mathbf{I}$ and left-multiplying by $\mathbf{X}^+$ gives the "solution" to $\mathbf{Xw} = \mathbf{y}$.
- **projection matrix** or **hat matrix**: Still focusing on the training phase, the predictions are $\hat{\mathbf{y}} = \mathbf{Xw}^* = \mathbf{X}\mathbf{X}^+\mathbf{y}$. So $\mathbf{X}\mathbf{X}^+$ puts that hat on $\mathbf{y}$, or projects $\mathbf{y}$ onto the hyperplane, in the viewpoint described below.

Projection interpretation

Usually we think of n points in $\mathbb{R}^d$. But instead, consider a separate column of the data for each feature: these are d points in $\mathbb{R}^n$. The observed training data $\mathbf{y}$ is also a point in $\mathbb{R}^n$, and so is the prediction $\hat{\mathbf{y}} = \mathbf{Xw}$.

As we vary $\mathbf{w}$, the prediction $\mathbf{Xw}$ describes a hyperplane spanned by the columns of $\mathbf{X}$.

We want to find the $\mathbf{w}^*$ corresponding to the closest point on the hyperplane to $\mathbf{y}$. So $X\mathbf{w}^* - \mathbf{y}$ must be orthogonal to the hyperplane:

$$\mathbf{X^T} \cdot (\mathbf{Xw}^* - \mathbf{y}) = \mathbf{0}.$$

Which are the normal equations (linear system of d equations), derived differently.

Weighted linear regression

Sample point i has weight b_i. Diagonal $n \times n$ matrix $\mathbf{B}$ contains weights.

$$\begin{aligned} J(\mathbf{w}) &= \sum_i b_i (\mathbf{x}_i^\mathbf{T} \mathbf{w} - y_i)^2 \\ &= (\mathbf{X}\mathbf{w} - \mathbf{y})^\mathbf{T} \mathbf{B} (\mathbf{X}\mathbf{w} - \mathbf{y}) \\ &= \mathbf{w}^\mathbf{T} \mathbf{X}^\mathbf{T} \mathbf{B} \mathbf{X} \mathbf{w} - 2\mathbf{y}^\mathbf{T} \mathbf{B} \mathbf{X} \mathbf{w} + \mathbf{y}^\mathbf{T} \mathbf{y} \end{aligned}$$

Gradient

$$\nabla_\mathbf{w} J(\mathbf{w}) = 2\mathbf{X}^\mathbf{T} \mathbf{B} \mathbf{X} \mathbf{w} - 2\mathbf{X}^\mathbf{T} \mathbf{B} \mathbf{y}$$

Solution

$$\mathbf{w}^* = \left(\mathbf{X}^\mathbf{T} \mathbf{B} \mathbf{X}\right)^{-1} \mathbf{X}^\mathbf{T} \mathbf{B} \mathbf{y}$$

How to compute the gradient

The cost function is $J(\mathbf{w}) = |\mathbf{X}\mathbf{w} - \mathbf{y}|^2$. We could write this as a dot product and multiply out:

$$\begin{aligned} J(\mathbf{w}) &= (\mathbf{X}\mathbf{w} - \mathbf{y}) \cdot (\mathbf{X}\mathbf{w} - \mathbf{y}) \\ &= \mathbf{X}\mathbf{w} \cdot \mathbf{X}\mathbf{w} - 2\mathbf{X}\mathbf{w} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} \\ &= (\mathbf{X}\mathbf{w})^\mathbf{T} \mathbf{X}\mathbf{w} - 2(\mathbf{X}\mathbf{w})^\mathbf{T} \mathbf{y} + \mathbf{y}^\mathbf{T} \mathbf{y} \\ &= \mathbf{w}^T \mathbf{X}^\mathbf{T} \mathbf{X} \mathbf{w} - 2\mathbf{w}^\mathbf{T} \mathbf{X}^\mathbf{T} \mathbf{y} + \mathbf{y}^\mathbf{T} \mathbf{y}, \end{aligned}$$

and then we'd need to differentiate those terms w.r.t. $\mathbf{w}$. However, a better way is to use the chain rule. Define f and g such that $J : \mathbb{R}^d \to \mathbb{R}$ is their composition $J = g \circ f$:

$$\begin{aligned} f &: \mathbb{R}^d \to \mathbb{R}^n & f(\mathbf{w}) &= \mathbf{X}\mathbf{w} - \mathbf{y} \\ g &: \mathbb{R}^n \to \mathbb{R} & g(\mathbf{z}) &= |\mathbf{z}|^2. \end{aligned}$$

The chain rule says that $\nabla(g \circ f) = (Df)^\mathbf{T} \nabla g$, where Df is the derivative of f, i.e. the Jacobian matrix of first partial derivatives[12]. We have $Df(\mathbf{w}) = \mathbf{X}$ and $\nabla g(\mathbf{z}) = 2\mathbf{z}$, so

$$\begin{aligned} \nabla J(\mathbf{w}) &= 2\mathbf{X}^\mathbf{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) \\ &= 2\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{w} - 2\mathbf{X}^\mathbf{T}\mathbf{y}. \end{aligned}$$

10.10.2 Penalized Regression

TODO

10.10.3 Logistic Regression

- Two classes.
- The observations y_i are class labels (or probabilities thereof).

[12] The gradient ∇ applies only to scalar-valued functions.

- The model states that the probability of being in class 1 is given by the usual linear model, mapped onto $(0, 1)$ by the logistic function s:

$$y_i \sim \text{Bern}(s(\mathbf{x}_i^\mathbf{T}\mathbf{w})),$$

$$s(z) = \frac{1}{1+e^{-z}}$$

Note that $s'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = s(z)(1-s(z))$.

Likelihood

Let $s_i = s(\mathbf{x}_i^\mathbf{T}\mathbf{w})$.

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= \prod_i s_i^{y_i} \left(1-s_i\right)^{(1-y_i)} \\
\ell(\mathbf{w}) &= \sum_i y_i \log s_i + (1-y_i)\log\left(1-s_i\right) \\
\nabla \ell(\mathbf{w}) &= \sum_i \frac{y_i}{s_i}(s_i)(1-s_i)\mathbf{x}_i + \frac{1-y_i}{1-s_i}(-1)(s_i)(1-s_i)\mathbf{x}_i \\
&= \sum_i \mathbf{x}_i \left(y_i(1-s_i) - (1-y_i)s_i\right) \\
&= \sum_i \mathbf{x}_i \left(y_i - s_i\right) \\
&= \mathbf{X}^\mathbf{T}\left(\mathbf{y} - \mathbf{s}(\mathbf{X}\mathbf{w})\right) \qquad (d \times 1)
\end{aligned}$$

where $\mathbf{s} : \mathbb{R}^n \to \mathbb{R}^n$ applies s componentwise to the rows.

Optimization problem: Find $\mathbf{w}$ that minimizes the cost function $J(\mathbf{w}) = -\ell(\mathbf{w})$.

Because the weights $\mathbf{w}$ are tied up inside $s_i = s(\mathbf{x}_i^\mathbf{T}\mathbf{w})$ it's not possible to find the minimum $\mathbf{w}^*$ by setting the gradient equal to zero (i.e. by solving a linear system). We can use gradient descent, or Newton's method.

For Newton's method, we need the Hessian of the objective function. This is the $d \times$ d matrix of partial derivatives of the gradient, i.e. $\mathbf{X}^\mathbf{T}$ multiplied by the derivative (Jacobian matrix) of $\mathbf{s}(\mathbf{X}\mathbf{w})$. Define $\mathbf{f}(\mathbf{w}) = \mathbf{X}\mathbf{w}$ so now $\mathbf{s}(\mathbf{X}\mathbf{w}) = (\mathbf{s} \circ \mathbf{f})(\mathbf{w})$.

Function	domain $\to$ range	Jacobian	dim Jacobian
$\mathbf{f}(\mathbf{w}) = \mathbf{X}\mathbf{w}$	$\mathbb{R}^d \to \mathbb{R}^n$	$\mathrm{D}\,\mathbf{f} = \mathbf{X}$	$n \times d$
$\mathbf{s}(\mathbf{z})$	$\mathbb{R}^n \to \mathbb{R}^n$	$\mathrm{D}\,\mathbf{s}(\mathbf{z}) = \mathbf{S}$	$n \times n$

where $\mathbf{S}$ is a diagonal matrix with $\mathbf{S}_{ii} = \mathbf{s}(\mathbf{x}_i^\mathbf{T}\mathbf{w})(1 - \mathbf{s}(\mathbf{x}_i^\mathbf{T}\mathbf{w}))$. Now by the chain rule,

$$\begin{aligned}
\nabla^2 J(\mathbf{w}) &= \mathbf{X}^\mathbf{T}\,\mathrm{D}_\mathbf{w}\,\mathbf{s}(\mathbf{X}\mathbf{w}) \\
&= \mathbf{X}^\mathbf{T}(\mathrm{D}_\mathbf{f}\,\mathbf{s})(\mathrm{D}_w\,\mathbf{f}) \\
&= \mathbf{X}^\mathbf{T}\mathbf{S}\mathbf{X}.
\end{aligned}$$

10.11 Homework 2

10.11.1 Conditional Probability

In the following questions, **show your work**, not just the final answer.

(a) The probability that an archer hits her target when it is windy is 0.4; when it is not windy, her probability of hitting the target is 0.7. On any shot, the probability of a gust of wind is 0.3. Find the probability that

Let the random variables involved be $W \in \{0,1\}$ (wind no/yes) and $H \in \{0,1\}$ (hit no/yes).

(i) on a given shot there is a gust of wind and she hits her target.

$$\Pr(W=1,H=1) = \Pr(W=1)\Pr(H=1|W=1) = 0.3 \cdot 0.4 = 0.12$$

(ii) she hits the target with her first shot.

$$\Pr(H=1) = \textstyle\sum_{w\in\{0,1\}} \Pr(W=w)\Pr(H=1|W=w) = 0.7\cdot 0.7 + 0.3\cdot 0.4 = 0.61$$

(iii) she hits the target exactly once in two shots.

Each shot may be viewed as an independent draw of (W,H). Therefore we use $\Pr(H=1)$ from part (ii) as the success probability in a binomial distribution:

$$\Pr(\text{one hit in two trials}) = \binom{2}{1} \Pr(H=1)^1 \left(1-\Pr(H=1)\right)^1 = 2\cdot 0.61 \cdot 0.39 = 0.4758.$$

(iv) there was no gust of wind on an occasion when she missed.

$$\begin{aligned}
\Pr(W=0|H=0) &= \frac{\Pr(W=0,H=0)}{\Pr(H=0)} \\
&= \frac{\Pr(W=0)\Pr(H=0|W=0)}{\sum_{w\in\{0,1\}}\Pr(W=w)\Pr(H=0|W=w)} \\
&= \frac{0.7\cdot 0.3}{0.7\cdot 0.3 + 0.3\cdot 0.6} \\
&= 0.5385 \quad (4 \text{ d.p.})
\end{aligned}$$

(b) Let A, B, C be events. Show that if

$$P(A|B,C) > P(A|B)$$

then

$$P(A|B,C^c) < P(A|B),$$

where C^c denotes the complement of C. Assume that each event on which we are conditioning has positive probability.

First, we expand the conditional probabilities involved in the given inequality:

$$\Pr(A|B,C) = \frac{\Pr(A,B)\Pr(C|A,B)}{\Pr(B)\Pr(C|B)} > \Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}.$$

Multiplying both sides by $\frac{\Pr(B)}{\Pr(A,B)}$ shows that

$$\frac{\Pr(C|A,B)}{\Pr(C|B)} > 1,$$

i.e. $\Pr(C|A,B) > \Pr(C|B)$.

We can transform that into a statement about C^c by subtracting both sides from 1:

$$\Pr(C^c|A,B) = 1 - \Pr(C|A,B) < 1 - \Pr(C|B) = \Pr(C^c|B),$$

i.e.

$$\frac{\Pr(C^c|A,B)}{\Pr(C^c|B)} < 1.$$

Now, we want to show that $\Pr(A|B,C^c) < \Pr(A|B)$. The left hand side is

$$\Pr(A|B,C^c) = \frac{\Pr(A,B)\Pr(C^c|A,B)}{\Pr(B)\Pr(C^c|B)} < \frac{\Pr(A,B)}{\Pr(B)} = \Pr(A|B),$$

as required.

10.11.2 Positive Definiteness (2016)

3.

(a) Give an explicit formula for $x^{\mathbf{T}}Ax$. Write your answer as a sum involving the elements of A and x.

$$x^{\mathbf{T}}Ax = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$$

(b) Show that if A is positive definite, then the entries on the diagonal of A are positive (that is, $a_{ii} > 0$ for all $1 \leq i \leq n$.)

We prove the contrapositive: suppose $a_{ii} \leq 0$ for some $1 \leq i \leq n$. Now consider a particular x containing zeros everywhere except for $x_i = 1$. Then $x^{\mathbf{T}}Ax = a_{ii}x_i^2 = a_{ii} \leq 0$, so A is not positive definite.

4.

(b) Let A be positive definite. Prove that all eigenvalues of A are greater than zero.

Let λ be an eigenvalue of A and let $v \neq \mathbf{0}$ be an eigenvector for this eigenvalue, so that $Av = \lambda v$. Since A is positive definite, we have $v^{\mathbf{T}}Av = \lambda|v|^2 > 0$. Since $|v|^2 > 0$, we conclude $\lambda > 0$.

(c) Let A be positive definite. Prove that A is invertible.

$\det A$ is equal to the product of the eigenvalues. Since these are all positive $\det A > 0$ and so A is invertible.

(d) Let A be positive definite. Prove that there exist n linearly independent vectors $x_1, x_2, ..., x_n$ such that $A_{ij} = x_i^{\mathbf{T}}x_j$. (Hint: Use the spectral theorem and what you proved in (b) to find a matrix B such that $A = B^{\mathbf{T}}B$.)

The spectral theorem states that

10.11.3 Positive Definiteness

Definition. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix.

- We say that A is **positive definite** if $\forall x \in \mathbb{R}^n - \{0\}$, $x^\top Ax > 0$. We denote this with $A \succ 0$.
- Similarly, we say that A is **positive semidefinite** if $\forall x \in \mathbb{R}^n$, $x^\top Ax \geq 0$. We denote this with $A \succeq 0$.

(a) For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, prove that all of the following are equivalent.

(i) $A \succeq 0$.

(ii) $B^\top AB \succeq 0$, for some invertible matrix $B \in \mathbb{R}^{n \times n}$.

(iii) All the eigenvalues of A are nonnegative.

(iv) There exists a matrix $U \in \mathbb{R}^{n \times n}$ such that $A = UU^\top$.

(Suggested road map: (i) $\Leftrightarrow$ (ii), (i) $\Rightarrow$ (iii) $\Rightarrow$ (iv)$\Rightarrow$ (i). For the implication (iii) $\Rightarrow$ (iv) use the *Spectral Theorem for Symmetric Matrices.*

(i) $\iff$ (ii)

Let $B = A^{\mathbf{T}} = A^{-1}$. Then B is invertible and $B^{\mathbf{T}}AB = A$. Therefore $A \succeq 0 \iff B^{\mathbf{T}}AB \succeq 0$.

(i) $\implies$ (iii)

Let λ be an eigenvalue of A and let $v \neq \mathbf{0}$ be an eigenvector for this eigenvalue, so that $Av = \lambda v$. Since A is positive semidefinite, we have $v^{\mathbf{T}}Av = \lambda |v|^2 \geq 0$. Since $|v|^2 > 0$, we conclude $\lambda \geq 0$.

(iii) $\implies$ (iv)

We're asked to show that there exists a matrix U such that $A = UU^{\mathbf{T}}$.

Since A is symmetric, by the Spectral Theorem for Symmetric Matrices its eigenvectors are orthonormal and it can be "diagonalized" as $A = U^*\Lambda U^{*-1}$ where the columns of U^* are the eigenvectors of A and Λ is a diagonal matrix containing the eigenvalues. Since the inverse of an orthogonal matrix is its transpose, we have

$$A = U^*\Lambda U^{*-1} = U^*\Lambda U^{*\mathbf{T}}.$$

Now define $U = U^*\Lambda^{1/2}$, where $\Lambda^{1/2}$ is a diagonal matrix containing the square roots of the eigenvalues: $\left(\Lambda^{1/2}\right)_{jj} = \sqrt{\lambda_j}$. Note that $U^{\mathbf{T}} = (U^*\Lambda^{1/2})^{\mathbf{T}} = \Lambda^{1/2}U^{*\mathbf{T}}$. Then

$$A = U^*\Lambda U^{*\mathbf{T}} = U^*\Lambda^{1/2}\Lambda^{1/2}U^{*\mathbf{T}} = UU^{\mathbf{T}}.$$

(iv) $\implies$ (i)

Let $x \in \mathbb{R}^n$. We see that $x^{\mathbf{T}}Ax$ is equal to the squared l_2-norm of a vector and hence non-negative:

$$x^{\mathbf{T}}Ax = x^{\mathbf{T}}UU^{\mathbf{T}}x = (U^{\mathbf{T}}x)^{\mathbf{T}}U^{\mathbf{T}}x = |U^{\mathbf{T}}x|^2 \geq 0.$$

Incidentlly, $A = UU^{\mathbf{T}}$ implies that A is symmetric, since the following quantities are the same:

(a) i, j-th element of $UU^{\mathbf{T}}$

(b) dot product of U row i and U^T column j

(c) dot product of U row i and U row j

(d) dot product of U row j and U row i

(e) dot product of U row j and $U^{\mathbf{T}}$ column i

(f) j, i-th element of $UU^{\mathbf{T}}$.

(b) For a symmetric positive definite matrix $A \succ 0 \in \mathbb{R}^{n \times n}$, prove the following.

(i) For every $\lambda > 0$, we have that $A + \lambda I \succ 0$.

We want to show that $x^{\mathbf{T}}(A + \lambda I)x > 0$ for all $x \in \mathbb{R}^n$. We have

$$\begin{aligned} x^{\mathbf{T}}(A + \lambda I)x &= x^{\mathbf{T}}(Ax + \lambda Ix) \\ &= x^{\mathbf{T}}Ax + \lambda x^{\mathbf{T}}x > 0 \end{aligned}$$

where the inequality is true because $x^{\mathbf{T}}Ax > 0$ due to the positive definiteness of A, and $\lambda x^{\mathbf{T}}x > 0$ because $\lambda > 0$ and $x^{\mathbf{T}}x > 0$ because it is the square of the 2-norm of x.

(ii) There exists a $\gamma > 0$ such that $A - \gamma I \succ 0$.

We want to show that a $\gamma > 0$ exists such that

$$\begin{aligned} x^{\mathbf{T}}(A - \gamma I)x &= x^{\mathbf{T}}(Ax - \gamma Ix) \\ &= x^{\mathbf{T}}Ax - \gamma x^{\mathbf{T}}x > 0 \end{aligned}$$

for all non-zero $x \in \mathbb{R}^n$. To satisfy this, we can choose any $\gamma < \frac{x^{\mathbf{T}}Ax}{x^{\mathbf{T}}x}$. Both the numerator and denominator here are strictly positive (due to positive definiteness of A and positivity of squared norm), so such a $\gamma > 0$ does exist.

(iii) All the diagonal entries of A are positive; i.e. $A_{ii} > 0$ for $i = 1, \ldots, n$.

Let $\mathbf{x}$ be a vector containing zeros except for a 1 in the i-th position. Then $x^{\mathbf{T}}Ax = \sum_{i,j} A_{ij}x_i x_j = A_{ii}$ so this must be positive for A to be PD.

(iv) $\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} > 0$, where A_{ij} is the element at the i-th row and j-th column of A.

Consider $x = [1, 1, \ldots, 1]^{\mathbf{T}}$.

Since A is PD we require $x^{\mathbf{T}}Ax > 0$. But $x^{\mathbf{T}}Ax = \sum_j \sum_k A_{jk}x_j x_k = \sum_j \sum_k A_{jk}$.

10.11.4 Derivatives and Norms

In the following questions, **show your work**, not just the final answer.

(a) Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$. Compute $\nabla_{\mathbf{x}}(\mathbf{a}^\top \mathbf{x})$.

We view $\mathbf{a}$ as a constant vector and $\mathbf{a}^\mathbf{T}\mathbf{x}$ as a function $f : \mathbb{R}^n \to \mathbb{R}$ with $f(\mathbf{x}) = \mathbf{a}^\mathbf{T}\mathbf{x} = \sum_{i=1}^n a_i x_i$. The requested gradient is the column vector of first partial derivatives

$$\nabla_{\mathbf{x}}(\mathbf{a}^\mathbf{T}\mathbf{x}) = \begin{bmatrix} f_{x_1} \\ f_{x_2} \\ \vdots \\ f_{x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial a^\mathbf{T} x}{\partial x_1} \\ \frac{\partial a^\mathbf{T} x}{\partial x_2} \\ \vdots \\ \frac{\partial a^\mathbf{T} x}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{a}.$$

(b) Let $A \in \mathbb{R}^{n \times n}$, $\mathbf{x} \in \mathbb{R}^n$. Compute $\nabla_{\mathbf{x}}(\mathbf{x}^\top A\mathbf{x})$.
How does the expression you derived simplify in the case that A is symmetric?

(Hint: to get a feeling for the problem, explicitly write down a 2×2 or 3×3 matrix A with components A_{11}, A_{12}, etc., explicitly expand $x^\top Ax$ as a polynomial without matrix notation, calculate the gradient in the usual way, and put the result back into matrix form. Then generalize the result to the $n \times n$ case.)

2×2 **symmetric**

$$\begin{aligned} \mathbf{x}^\mathbf{T} A\mathbf{x} &= A_{11}x_1^2 + 2A_{12}x_1x_2 + A_{22}x_2^2 \\ &= \sum_{jk} A_{jk}x_j x_k \end{aligned}$$

$$\nabla_{\mathbf{x}}(\mathbf{x}^\top A\mathbf{x}) = \begin{bmatrix} 2A_{11}x_1 + 2A_{12}x_2 \\ 2A_{12}x_1 + 2A_{22}x_2 \end{bmatrix} = 2A\mathbf{x}$$

2×2

$$x^\mathbf{T} Ax = A_{11}x_1^2 + (A_{12} + A_{21})x_1x_2 + A_{22}x_2^2$$

$$\nabla_x(x^\top Ax) = \begin{bmatrix} 2A_{11}x_1 + (A_{12} + A_{21})x_2 \\ (A_{12} + A_{21})x_1 + 2A_{22}x_2 \end{bmatrix} = (A + A^\mathbf{T})\mathbf{x}$$

(c) Let $A, X \in \mathbb{R}^{n \times n}$. Compute $\nabla_X(\operatorname{trace}(A^\top X))$.

We view A as a constant matrix and trace $A^\mathbf{T}X$ as a function $f : \mathbb{R}^{n \times n} \to \mathbb{R}$ with

$$f(X) = \operatorname{trace} A^\mathbf{T}X = \sum_{j=1}^n A_{\cdot j} \cdot X_{\cdot j} = \sum_{j=1}^n \sum_{i=1}^n A_{ij}X_{ij},$$

where $B_{\cdot j}$ represents the j-th column of the matrix B.

The requested gradient is the matrix of first partial derivatives

$$\nabla_X(\text{trace}(A^\top X)) = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{12}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \frac{\partial f}{\partial X_{21}} & \frac{\partial f}{\partial X_{22}} & \cdots & \frac{\partial f}{\partial X_{2n}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f}{\partial X_{n1}} & \frac{\partial f}{\partial X_{n2}} & \cdots & \frac{\partial f}{\partial X_{nn}} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} = A.$$

(d) For a function $f : \mathbb{R}^d \to \mathbb{R}$ to be a norm, the distance metric $\delta(x, y) = f(x-y)$ must satisfy the triangle inequality. Is the function $f(x) = (\sqrt{|x_1|} + \sqrt{|x_2|})^2$ a norm for vectors $x \in \mathbb{R}^2$? Prove it or give a counterexample.

Consider $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. For f to be a valid norm we require $f(x) + f(y) \geq f(x + y)$. But $f(x) = f(y) = 1$ whereas $f(x + y) = 4$ so the triangle inequality does not hold.

(e) Let $x \in \mathbb{R}^n$. Prove that $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$.

Solution:

(f) Let $x \in \mathbb{R}^n$. Prove that $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$.
(Hint: The Cauchy–Schwarz inequality may come in handy.)

Solution:

10.11.5 Eigenvalues

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with $A \succeq 0$.

(a) Prove that the largest eigenvalue of A is

$$\lambda_{\max}(A) = \max_{\|x\|_2 = 1} x^\top A x.$$

(Hint: Use the *Spectral Theorem for Symmetric Matrices* to reduce the problem to the diagonal case.)

Solution:

(b) Similarly, prove that the smallest eigenvalue of A is

$$\lambda_{\min}(A) = \min_{\|x\|_2 = 1} x^\top A x.$$

Solution:

(c) Is either of the optimization problems described in parts (a) and (b) a convex program? Justify your answer.

Solution:

(d) Show that if λ is an eigenvalue of A then λ^2 is an eigenvalue of A^2, and deduce that

$$\lambda_{\max}(A^2) = \lambda_{\max}(A)^2 \text{ and } \lambda_{\min}(A^2) = \lambda_{\min}(A)^2.$$

Solution:

(e) From parts (a), (b), and (d), show that for any vector $x \in \mathbb{R}^n$ such that $\|x\|_2 = 1$,

$$\lambda_{\min}(A) \leq \|Ax\|_2 \leq \lambda_{\max}(A).$$

Solution:

(f) From part (e), deduce that for any vector $x \in \mathbb{R}^n$,

$$\lambda_{\min}(A)\|x\|_2 \leq \|Ax\|_2 \leq \lambda_{\max}(A)\|x\|_2.$$

Solution:

10.11.6 Gradient Descent

Consider the optimization problem $\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Ax - b^\top x$, where A is a symmetric matrix with $0 < \lambda_{\min}(A)$ and $\lambda_{\max}(A) < 1$.

(a) Using the first order optimality conditions, derive a closed-form solution for the minimum possible value of x, which we denote x^*.

Let $f(\mathbf{x}) = \frac{1}{2}x^\mathbf{T}Ax - b^\top x$. Since $x^\mathbf{T}Ax = \sum_{j,k} A_{jk}x_jx_k$, the gradient in the x_j direction is

$$(\nabla_x f)_j = \sum_k A_{jk}x_k - b_j$$

(the factor of 1/2 cancels the 2s deriving from differentiating x_j^2 and $2x_jx_k$).

In other words,

$$\nabla_x f = A\mathbf{x} - \mathbf{b}.$$

Setting this equal to zero gives $x^* = A^{-1}b$.

Compare the 1D version: $f(x) = \frac{1}{2}ax^2 - bx \implies f'(x) = ax - b \implies x^* = b/a$.

(b) Solving a linear system directly using Gaussian elimination takes $O(n^3)$ time, which may be wasteful if the matrix A is sparse. For this reason, we will use gradient descent to compute an approximation to the optimal point x^*. Write down the update rule for gradient descent with a step size of 1.

for j in $1 \dots d$

$$x_j^{(i)} \leftarrow x_j^{(i-1)} - \sum_k A_{jk}x_k^{(i-1)} + b_j$$

Or in other words,

$$\begin{aligned} x^{(i)} &\leftarrow x^{(i-1)} - Ax^{(i-1)} + b \\ &= (I - A)x^{(i-1)} + b \end{aligned}$$

(c) Show that the iterates $x^{(i)}$ satisfy the recursion

$$x^{(i)} - x^* = (I - A)(x^{(i-1)} - x^*).$$

$$\begin{aligned} x^{(i)} - x^* &= (I - A)x^{(i-1)} + b - x^* \\ &= (I - A)x^{(i-1)} + Ax^* - x^* \\ &= (I - A)x^{(i-1)} + (A - I)x^* \\ &= (I - A)(x^{(i-1)} - x^*) \end{aligned}$$

(d) Show that for some $0 < \rho < 1$,

$$\|x^{(i)} - x^*\|_2 \le \rho\|x^{(i-1)} - x^*\|_2.$$

Solution:

(e) Let $x^{(0)} \in \mathbb{R}^n$ be a starting value for our gradient descent iterations. If we want our solution $x^{(i)}$ to be $\epsilon > 0$ close to x^*, i.e. $\|x^{(i)} - x^*\|_2 \leq \epsilon$, then how many iterations of gradient descent should we perform? In other words, how large should k be? Give your answer in terms of $\rho, \|x^{(0)} - x^*\|_2$, and ϵ. Note that $0 < \rho < 1$, so $\log \rho < 0$.

Solution:

(f) Observe that the running time of each iteration of gradient descent is dominated by a matrix-vector product. What is the overall running time of gradient descent to achieve a solution $x^{(i)}$ which is ϵ-close to x^*? Give your answer in terms of $\rho, \|x^{(0)} - x^*\|_2, \epsilon$, and n.

Solution:

10.11.7 Classification

Suppose we have a classification problem with classes labeled $1, \ldots, c$ and an additional "doubt" category labeled $c+1$. Let $f : \mathbb{R}^d \to \{1, \ldots, c+1\}$ be a decision rule. Define the loss function

$$R(f(x) = i|x) = \begin{cases} 0 & \text{if } i = j \quad i, j \in \{1, \ldots, c\} \\ \lambda_r & \text{if } i = c+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

where $\lambda_r \geq 0$ is the loss incurred for choosing doubt and $\lambda_s \geq 0$ is the loss incurred for making a misclassification. Hence the risk of classifying a new data point x as class $i \in \{1, 2, \ldots, c+1\}$ is

$$R(f(x) = i|x) = \sum_{j=1}^{c} L(f(x) = i, y = j) P(Y = j|x).$$

(a) Show that the following policy obtains the minimum risk. (1) Choose class i if $P(Y = i|x) \geq P(Y = j|x)$ for all j and $P(Y = i|x) \geq 1 - \lambda_r/\lambda_s$; (2) choose doubt otherwise.

Solution:

(b) What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$? Explain why this is consistent with what one would expect intuitively.

Solution:

10.11.8 Gaussian Classification

Let $P(x|\omega_i) \sim N(\mu_i, \sigma^2)$ for a two-category, one-dimensional classification problem with classes ω_1 and ω_2, $P(\omega_1) = P(\omega_2) = 1/2$, and $\mu_2 > \mu_1$.

(a) Find the Bayes optimal decision boundary and the corresponding Bayes decision rule.

A Bayes optimal decision boundary for a one-dimensional, two-class problem is a point x^* at which the two class posterior probabilities are equal. Since the variances and priors are equal the problem is symmetric and it seems intuitively clear that the decision boundary must be $x^* = \frac{\mu_1+\mu_2}{2}$, with rule

$$f(x) = \begin{cases} \omega_1, & x < x^* \\ \omega_2, & x > x^* \end{cases}$$

(undefined classification exactly at the boundary).

To prove this, first note that the posterior probability of membership of a point x in class ω_i is

$$\begin{aligned} \mathbf{p}(\omega_i|x) &= \frac{\mathbf{p}(\omega_i)\mathbf{p}(\omega_i|x)}{\mathbf{p}(x)} \\ &= \frac{1}{2\mathbf{p}(x)} \frac{1}{(\sqrt{2\pi}\sigma)} \exp(-\frac{(x-\mu_i)^2}{2\sigma^2}) \end{aligned}$$

Viewed as a function of ω_i, the log posterior is

$$\log \mathbf{p}(\omega_i|x) = -\frac{(x-\mu_i)^2}{2\sigma^2} + \text{constant},$$

so the decision boundary x^* satisfies

$$\begin{aligned} -\frac{(x^*-\mu_1)^2}{2\sigma^2} &= -\frac{(x^*-\mu_2)^2}{2\sigma^2} \\ \implies (x^*-\mu_1)^2 &= (x^*-\mu_2)^2 \\ \implies x^* &= \frac{\mu_2^2-\mu_1^2}{2(\mu_2-\mu_1)} = \frac{\mu_2+\mu_1}{2}. \end{aligned}$$

(b) The Bayes error is the probability of misclassification,

$$P_e = P((\text{misclassified as } \omega_1)|\omega_2)P(\omega_2) + P((\text{misclassified as } \omega_2)|\omega_1)P(\omega_1).$$

Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-z^2/2} dz$$

where $a = \frac{\mu_2-\mu_1}{2\sigma}$.

Let the random variables X and Y represent the sample point and its class respectively. The

probability of misclassification is

$$\begin{aligned} P_e &= P((\text{misclassified as } \omega_1)|Y=\omega_2)P(Y=\omega_2) + P((\text{misclassified as } \omega_2)|Y=\omega_1)P(Y=\omega_1) \\ &= \frac{1}{2}\left(\mathbf{p}(X<x^*|Y=\omega_2) + \mathbf{p}(X>x^*|Y=\omega_1)\right). \end{aligned}$$

These two probability distributions are 1D Gaussians with variance σ^2 and means μ_2 and μ_1 respectively. Now change the parameterization of these Gaussians so that they both have variance 1 and mean 0. The above probability becomes

$$\begin{aligned} P_e &= \frac{1}{2}\left(\mathbf{p}(X<\frac{x^*-\mu_2}{\sigma}|Y=\omega_2) + \mathbf{p}(X>\frac{x^*-\mu_1}{\sigma}|Y=\omega_1)\right) \\ &= \frac{1}{2\sqrt{2\pi}}\left(\int_{-\infty}^{(x^*-\mu_2)/\sigma} e^{-z^2}\,\mathrm{d}\,z + \int_{(x^*-\mu_1)/\sigma}^{\infty} e^{-z^2}\,\mathrm{d}\,z\right) \end{aligned}$$

10.11.9 Maximum Likelihood Estimation

Let X be a discrete random variable which takes values in $\{1, 2, 3\}$ with probabilities $P(X = 1) = p_1, P(X = 2) = p_2$, and $P(X = 3) = p_3$, where $p_1 + p_2 + p_3 = 1$. Show how to use the method of maximum likelihood to estimate p_1, p_2, and p_3 from n observations of $X : x_1, \ldots, x_n$. Express your answer in terms of the counts

$$k_1 = \sum_{i=1}^{n} \mathbb{1}(x_i = 1), k_2 = \sum_{i=1}^{n} \mathbb{1}(x_i = 2), \text{ and } k_3 = \sum_{i=1}^{n} \mathbb{1}(x_i = 3),$$

where

$$\mathbb{1}(x = a) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{if } x \neq a. \end{cases}$$

Let the observed data vector be $\mathbf{k} = [k_1, k_2, k_3]^\mathbf{T}$ and the parameter vector be $\mathbf{p} = [p_1, p_2, p_3]^\mathbf{T}$. The sampling model is $\mathbf{k} \sim \text{Multinomial}(\mathbf{p})$, so the probability of the observed data vector is

$$\Pr(\mathbf{k}|\mathbf{p}) = \frac{n!}{k_1!k_2!k_3!} \prod_{j=1}^{3} p_j^{k_j},$$

giving the following log-likelihood function:

$$l(\mathbf{p}) = \log \Pr(\mathbf{k}|\mathbf{p}) = \sum_{j=1}^{3} k_j \log p_j.$$

We want to maximize this log-likelihood subject to the constraint that $\sum_j p_j = 1$. To do so, we maximize the Lagrangian

$$\mathcal{L}(\mathbf{p}, \lambda) = \sum_{j=1}^{3} k_j \log p_j - \lambda \left(\sum_{j=1}^{3} p_j - 1 \right).$$

The gradient of the Lagrangian is

$$\nabla_{\mathcal{L}} = \begin{bmatrix} \partial \mathcal{L} / \partial p_1 \\ \partial \mathcal{L} / \partial p_2 \\ \partial \mathcal{L} / \partial p_3 \\ \partial \mathcal{L} / \partial \lambda \end{bmatrix} = \begin{bmatrix} k_1/p_1 - \lambda \\ k_2/p_2 - \lambda \\ k_3/p_3 - \lambda \\ 1 - \sum_{j=1}^{3} p_j \end{bmatrix}.$$

Solving $\nabla_{\mathcal{L}} = 0$ yields $k_j = \widehat{\lambda} \widehat{p}_j$ and $\sum_j \widehat{p}_j = 1$. Therefore $n = \sum_j k_j = \widehat{\lambda} \sum_j \widehat{p}_j = \widehat{\lambda}$, giving the maximum likelihood parameter estimates

$$\widehat{p}_j = \frac{k_j}{n}.$$

(Confirm that this point is a maximum.)

10.12 Homework 3

10.12.1 Independence vs. Correlation

(a) Consider the random variables $X, Y \in \mathbb{R}$ with the following conditions.

(i) X and Y can take values $\{-1, 0, 1\}$.

(ii) Either X is 0 with probability $(\frac{1}{2})$, or Y is 0 with probability $(\frac{1}{2})$.

(iii) When X is 0, Y takes values 1 and -1 with equal probability $(\frac{1}{2})$. When Y is 0, X takes values 1 and -1 with equal probability $(\frac{1}{2})$.

Are X and Y uncorrelated? Are X and Y independent? Prove your assertions. *Hint:* Graph these points in the plane. What's each point's joint probability?

The information we are given corresponds to the following entries in a joint probability distribution table.

			Y		
		-1	0	1	
X	-1		1/4		
	0	1/4		1/4	1/2
	1		1/4		
			1/2		

Using the fact that the rows and columns must sum to the marginal totals, and that each margin must sum to one, we can fill out the full joint distribution:

			Y		
		-1	0	1	
X	-1	0	1/4	0	1/4
	0	1/4	0	1/4	1/2
	1	0	1/4	0	1/4
		1/4	1/2	1/4	1

We have

$$\mathbb{E}[X] = \mu_X = -1 \times \frac{1}{4} + 0 \times \frac{1}{2} + 1 \times \frac{1}{4} = 0,$$

and $\mathbb{E}[Y] = \mu_Y = \mu_X$ because the marginal distributions of X and Y are identical.

Are X and Y uncorrelated? Yes. The definition of "uncorrelated" is that their covariance is zero. Their covariance is

$$\text{Cov}(X, Y) = \mathbb{E}\ (X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) = \mathbb{E}[XY] - \mu_X \mu_Y.$$

But note that $\mu_X \mu_Y = 0 \cdot 0 = 0$, and for every sample point with non-zero probability, it is true that either $X = 0$ or $Y = 0$. Therefore $\text{Cov}(X, Y) = 0$; X and Y are uncorrelated.

Are X and Y independent? No. The definition of "independent" is that Y contributes no information about X (and equivalently, X contributes no information about Y). More formally, X and Y are independent if and only if

$$\mathbf{p}(X = x | Y = y) = \mathbf{p}(X = x)$$

for every pair (x, y).

But this means that the columns of the joint probability distribution are identical (and hence the rows also). Since that is not the case, X and Y are not independent.

(b) Consider three Bernoulli random variables B, C, and D which take values $\{0, 1\}$ with equal probability. Construct three more random variables X, Y, Z such that $X = B \oplus C, Y = C \oplus D$, and $Z = B \oplus D$, where $\oplus$ is the XOR (exclusive or) operator. Are X, Y, and Z pairwise independent? Mutually independent? Prove it.

B	C	D	X	Y	Z	Probability
0	0	0	0	0	0	1/8
0	0	1	0	1	1	1/8
0	1	0	1	1	0	1/8
0	1	1	1	0	1	1/8
1	0	0	1	0	1	1/8
1	0	1	1	1	0	1/8
1	1	0	0	1	1	1/8
1	1	1	0	0	0	1/8

Are X, Y, and Z pairwise independent? Yes.

We have $\mathbf{p}(X = 1) = \mathbf{p}(Y = 1) = \mathbf{p}(Z = 1) = 1/2$. Since all three have non-zero probability there's no risk on conditioning on an impossible event, and we can take the definition of pairwise independence to be: X, Y, and Z are pairwise independent if and only if

$$\mathbf{p}(X = 1|Y) = \mathbf{p}(X = 1)$$
$$\mathbf{p}(X = 1|Z) = \mathbf{p}(X = 1)$$
$$\mathbf{p}(Y = 1|Z) = \mathbf{p}(Y = 1).$$

Consider X conditioned on Y. Of the events for which $Y = 0$, half have $X = 0$ and half have $X = 1$. Similarly, of the events (rows) for which $Y = 1$, half have $X = 0$ and half have $X = 1$. Therefore $\mathbf{p}(X = 1|Y) = \mathbf{p}(X = 1) = 1/2$. By the symmetry of the problem, the same is true for $\mathbf{p}(X = 1|Z)$ and $\mathbf{p}(Y = 1|Z)$. Therefore X, Y, and Z are pairwise independent.

Are X, Y, and Z mutually independent? No.

We can take the definition of mutual independence to be: X, Y, and Z are mutually independent if and only if

$$\mathbf{p}(X = 1|Y, Z) = \mathbf{p}(X = 1)$$
$$\mathbf{p}(Y = 1|X, Z) = \mathbf{p}(Y = 1)$$
$$\mathbf{p}(Z = 1|X, Y) = \mathbf{p}(Z = 1).$$

It suffices to exhibit one counter-example. Consider conditioning on $Y = 1, Z = 1$. Of the events (rows) for which that is true, X is always 0. Therefore $\mathbf{p}(X = 1|Y, Z) = 0 \neq \mathbf{p}(X = 1)$.

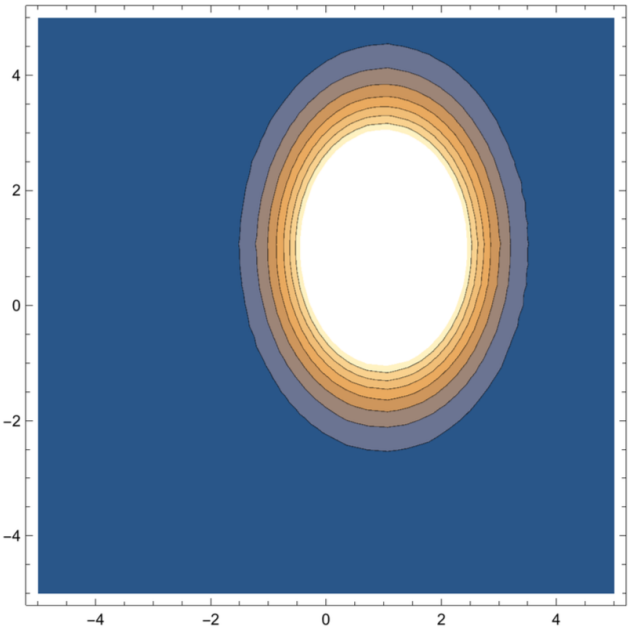
10.12.2 Isocontours of Normal Distributions

Let $f(\mu, \Sigma)$ be the density function of a normally distributed random variable in $\mathbb{R}^2$. Plot isocontours of the following functions.

(a) $f(\mu, \Sigma)$, where $\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.

```
In[23]:= ContourPlot[
          PDF[
           MultinormalDistribution[{1, 1}, {{1, 0}, {0, 2}}],
           {x, y}
          ],
          {x, -5, 5},
          {y, -5, 5}
         ]
```

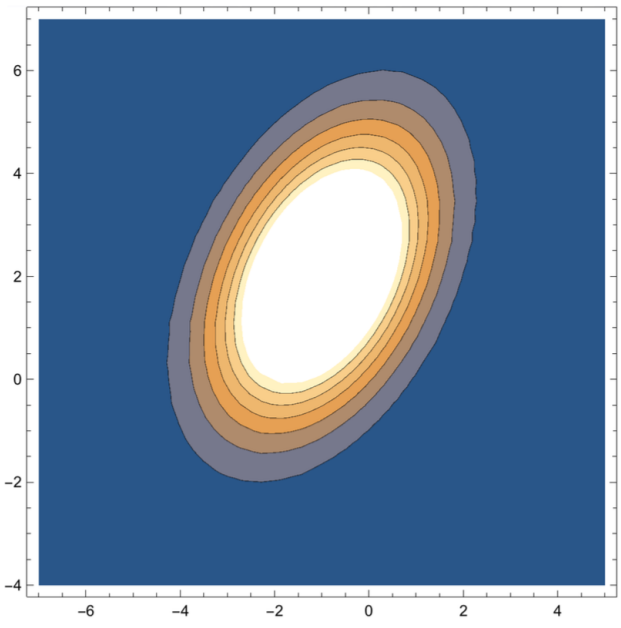
Out[23]=



(b) $f(\mu, \Sigma)$, where $\mu = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$.

In[25]:=
```
ContourPlot[
 PDF[
  MultinormalDistribution[{-1, 2}, {{2, 1}, {1, 3}}],
  {x, y}
 ],
 {x, -7, 5},
 {y, -4, 7}
]
```

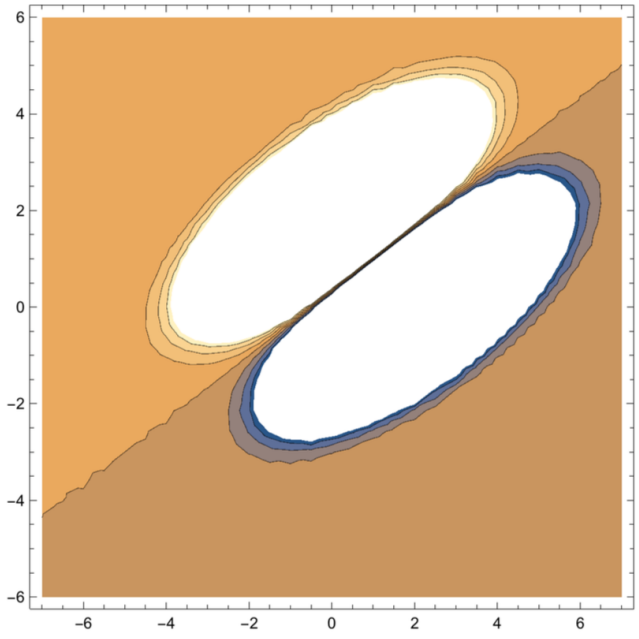
Out[25]=



(c) $f(\mu_1, \Sigma_1) - f(\mu_2, \Sigma_2)$, where $\mu_1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\Sigma_1 = \Sigma_2 = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$.

In[28]:=
```
ContourPlot[
 PDF[
   MultinormalDistribution[{0, 2}, {{2, 1}, {1, 1}}],
   {x, y}
  ]
  -
  PDF[
   MultinormalDistribution[{2, 0}, {{2, 1}, {1, 1}}],
   {x, y}
  ] ,
 {x, -7, 7},
 {y, -6, 6}
]
```

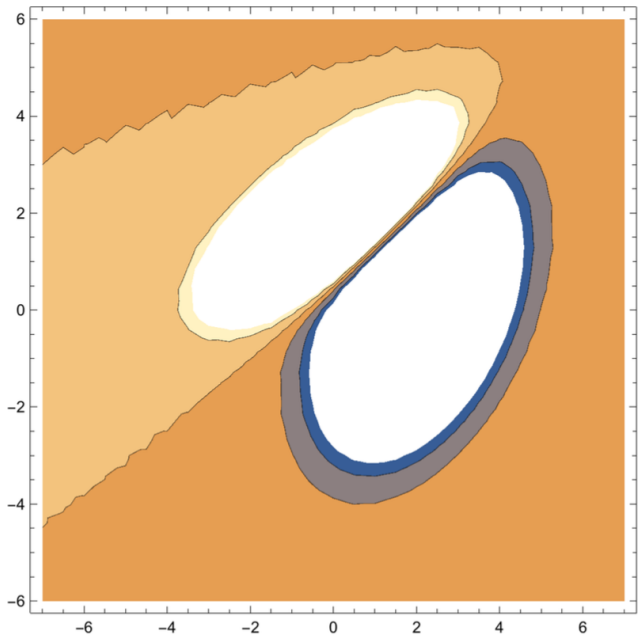
Out[28]=



(d) $f(\mu_1, \Sigma_1) - f(\mu_2, \Sigma_2)$, where $\mu_1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$.

In[29]:=
```
ContourPlot[
 PDF[
   MultinormalDistribution[{0, 2}, {{2, 1}, {1, 1}}],
   {x, y}
  ]
  -
  PDF[
   MultinormalDistribution[{2, 0}, {{2, 1}, {1, 3}}],
   {x, y}
  ] ,
 {x, -7, 7},
 {y, -6, 6}
]
```

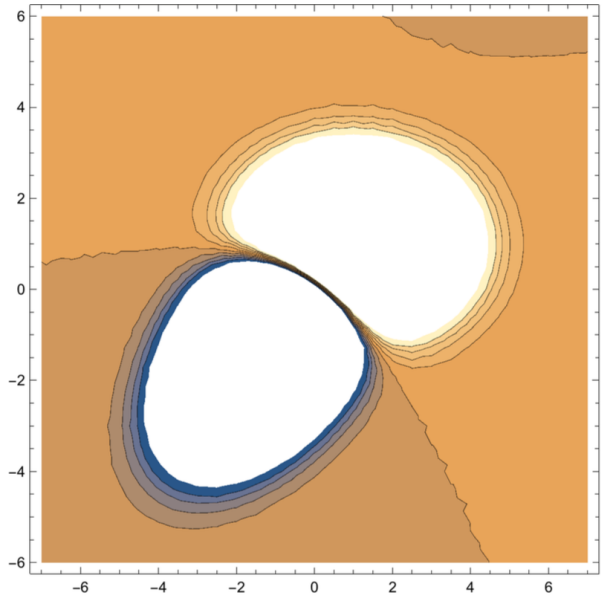
Out[29]=



(e) $f(\mu_1, \Sigma_1) - f(\mu_2, \Sigma_2)$, where $\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mu_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

In[30]:=
```
ContourPlot[
 PDF[
   MultinormalDistribution[{1, 1}, {{2, 0}, {0, 1}}],
   {x, y}
  ]
  -
  PDF[
   MultinormalDistribution[{-1, -1}, {{2, 1}, {1, 2}}],
   {x, y}
  ] ,
 {x, -7, 7},
 {y, -6, 6}
]
```

Out[30]=



10.12.3 Eigenvectors of the Gaussian Covariance Matrix

Consider two one-dimensional random variables $X_1 \sim \mathcal{N}(3, 9)$ and $X_2 \sim \frac{1}{2}X_1 + \mathcal{N}(4, 4)$, where $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. In software, draw $N = 100$ random two-dimensional sample points from (X_1, X_2) such that the ith value sampled from X_2 is calculated based on the ith value sampled from X_i.

(a)
```
from numpy.random import normal

X1 = normal(3, 3, 100)
X2 = X1/2 + normal(4, 2, 100)
X = np.stack([X1, X2], axis=1)
n, d = X.shape
```

(b) Compute the mean (in $\mathbb{R}^2$) of the sample.

```
mu = X.mean(axis=0)
```

(c) Compute the 2×2 covariance matrix of the sample.

```
Sigma = (X - mu).T @ (X - mu) / (n * d)
```

(d) Compute the eigenvectors and eigenvalues of this covariance matrix.

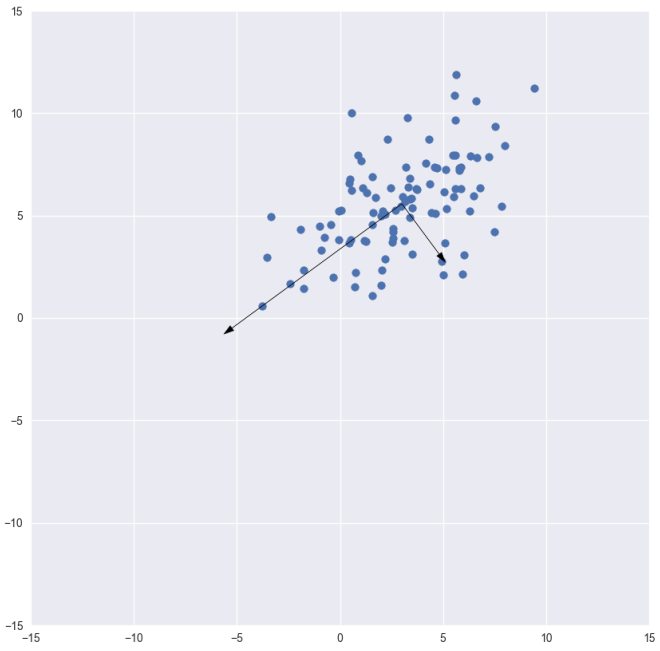
```
from numpy.linalg import eigh
evals, evecs = eigh(Sigma)
```

(e) On a two-dimensional grid with a horizontal axis for X_1 with range $[-15, 15]$ and a vertical axis for X_2 with range $[-15, 15]$, plot

(i) all $N = 100$ data points, and

(ii) arrows representing both covariance eigenvectors. The eigenvector arrows should originate at the mean and have magnitudes equal to their corresponding eigenvalues.

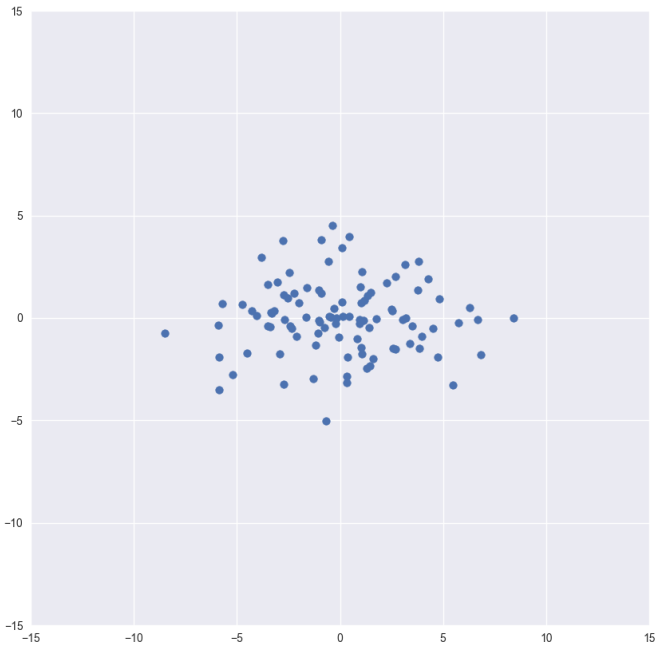
```
fig = plt.figure(figsize=(10,10))
plt.xlim(-15,15)
plt.ylim(-15,15)
plt.scatter(X[:,0], X[:,1])
arrow_kwargs = dict(fc="k", ec="k", head_width=0.3, head_length=0.5)
plt.arrow(mu[0], mu[1],
evecs[:,0][0] * evals[0],
evecs[:,0][1] * evals[0],
**arrow_kwargs)
plt.arrow(mu[0], mu[1],
evecs[:,1][0] * evals[1],
evecs[:,1][1] * evals[1],
**arrow_kwargs)
```



(f) Let $U = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$ be a 2×2 matrix whose columns are the eigenvectors of the covariance matrix, where v_1 is the eigenvector with the larger eigenvalue. We use $U^\top$ as a rotation matrix to rotate each sample point from the (X_1, X_2) coordinate system to a coordinate system aligned with the eigenvectors. (As $U^\top = U^{-1}$, the matrix U reverses this rotation, moving back from the eigenvector coordinate system to the original coordinate system). Center your sample points by subtracting the mean μ from each point; then rotate each point by $U^\top$, giving $x_{\text{rotated}} = U^\top(x - \mu)$. Plot these rotated points on a new two dimensional-grid, again with both axes having range $[-15, 15]$.

```
U = evecs[:,::-1]
X_centered = X - mu
X_centered_rotated = (U.T @ X_centered.T).T

fig = plt.figure(figsize=(10,10))
plt.xlim(-15,15)
plt.ylim(-15,15)
plt.scatter(X_centered_rotated[:,0], X_centered_rotated[:,1])
```



10.12.4 Maximum Likelihood Estimation

Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be n sample points drawn independently from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$.

(a) Suppose the normal distribution has an unknown diagonal covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & & & \\ & \sigma_2^2 & & & \\ & & \sigma_3^2 & & \\ & & & \ddots & \\ & & & & \sigma_d^2 \end{bmatrix}$$

and an unknown mean μ. Derive the maximum likelihood estimates, denoted $\widehat{\mu}$ and $\widehat{\sigma_i}$ for μ and σ_i. Show all your work.

(Answer starts on next page)

First, let's get some intuition for the situation: the covariance matrix is diagonal, so the isocontours of the PDF of the Gaussian are axis-aligned. That means that the PDF can be factored into a product of one-dimensional marginal densities: i.e. we can compute the density of a sample vector $\mathbf{x}$ as the product of densities of its scalar components (individual features): $\mathbf{p}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \prod_{j=1}^{d} \mathbf{p}(\mathbf{x}_j; \mu_j, \sigma_j^2)$. We therefore expect the estimation problem to be fairly straightforward, essentially involving fitting d one-dimensional Gaussians independently.

The likelihood function is

$$\begin{aligned} \mathcal{L}(\mu, \Sigma) &= \prod_{i=1}^{n} \mathbf{p}(X_i; \mu, \Sigma) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{d} \mathbf{p}(\mathbf{X}_{ij}; \mu_j, \sigma_j^2) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{(\sqrt{2\pi})^d \sigma_j^d} \exp\left(\frac{-(X_{ij} - \mu_j)^2}{2\sigma_j^2}\right), \end{aligned}$$

giving the log-likelihood function

$$\begin{aligned} \ell(\mu, \Sigma) &= \sum_{i=1}^{n} \sum_{j=1}^{d} -d \log \sigma_j - \frac{(X_{ij} - \mu_j)^2}{2\sigma_j^2} + \text{constant} \\ &= \sum_{j=1}^{d} -nd \log \sigma_j - \frac{1}{2\sigma_j^2} \sum_{i=1}^{n} (X_{ij} - \mu_j)^2 + \text{constant}. \end{aligned}$$

Fix a particular feature j. The partial derivatives with respect to the mean and variance parameter

for that feature are

$$\frac{\partial \ell}{\partial \mu_j} = \frac{1}{\sigma_j^2}\sum_{i=1}^{n}(X_{ij} - \mu_j) = \frac{1}{\sigma_j^2}\left(-n\mu_j + \sum_{i=1}^{n} X_{ij}\right)$$
$$\frac{\partial \ell}{\partial \sigma_j} = -\frac{nd}{\sigma_j} + \frac{1}{\sigma_j^3}\sum_{i=1}^{n}(X_{ij} - \mu_j)^2.$$

To find the MLE $\widehat{\mu}_j$ we set the partial derivative equal to zero and solve for μ:

$$\frac{\partial \ell}{\partial \mu_j} = 0 \implies -n\widehat{\mu}_j + \sum_{i=1}^{n} X_{ij} = 0$$
$$\implies \widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}.$$

To find the MLE $\widehat{\sigma}_j$ we set the partial derivative equal to zero, set $\mu_j = \widehat{\mu}_j$, and solve for σ:

$$\frac{\partial \ell}{\partial \sigma_j} = 0 \implies -nd + \frac{1}{\widehat{\sigma}_j^2}\sum_{i=1}^{n}(X_{ij} - \widehat{\mu}_j)^2 = 0$$
$$\implies \widehat{\sigma}_j^2 = \frac{1}{nd}\sum_{i=1}^{n}(X_{ij} - \widehat{\mu}_j)^2.$$

(Why is it valid to substitute $\widehat{\mu}_j$ for μ_j?)

To verify that these critical points are indeed maxima, we note first that $\ell(\mu, \Sigma)$ is a quadratic in μ, in which the sign of μ_j is negative. Therefore it is a concave-down quadratic in μ_j and has only a maximum; no minimum.

For σ we compute the second partial derivative,

$$\frac{\partial^2 \ell}{\partial \sigma_j^2} = \frac{nd}{\sigma_j^2} - \frac{3}{\sigma_j^4}\,{}_j,$$

where ${}_j = \sum_{i=1}^{n}(X_{ij} - \mu_j)^2$, and evaluate it at the critical point:

$$\frac{\partial^2 \ell}{\partial \sigma_j^2}(\sigma_j) = \frac{(nd)^2}{{}_j} - \frac{3(nd)^4}{\left({}_j\right)^4}\,{}_j$$
$$= \frac{(nd)^2}{{}_j} - \frac{3(nd)^4}{\left({}_j\right)^3}.$$

(I was expecting to be able to show that $\frac{\partial^2 \ell}{\partial \sigma_j^2}(\sigma_j)$ is negative but I don't seem to be managing to do so.)

(b) Suppose the normal distribution has a known covariance matrix Σ and an unknown mean $A\mu$, where Σ and A are known $d \times d$ matrices, Σ is positive definite, and A is invertible. Derive the maximum likelihood estimate, denoted $\widehat{\mu}$, for μ.

Let $\eta = A\mu$. Then $\widehat{\eta}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ as above. There is a theorem (the "invariance property") regarding MLEs which states that $\widehat{g(\theta)}$ is $g(\widehat{\theta})$, where $\widehat{\ }$ denotes MLE.

We know $\widehat{A\mu}$. We want $\widehat{\mu} = \widehat{A^{-1}A\mu}$. By the invariance property for MLEs this is

$$\widehat{\mu} = A^{-1}\widehat{A\mu} = A^{-1}\frac{1}{n}\sum_{i=1}^{n} X_{ij}.$$

I'm not entirely sure what requirements this theorem makes of g but I am confident that the invertible linear transformation A^{-1} satisfies them.

Alternatively, without relying on this theorem, I tried to argue from first principles, but currently am confused when attempting to compute the gradient for μ:

The likelihood function for μ is

$$\mathcal{L}(\mu) = \prod_{i=1}^{n} \frac{1}{\left(\sqrt{2\pi}\right)^{d}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(X_i - A\mu)^{\mathbf{T}}\Sigma^{-1}(X_i - A\mu)\right),$$

and the log-likelihood function is

$$\ell(\mu) = -\frac{1}{2}\sum_{i=1}^{n}(X_i - A\mu)^{\mathbf{T}}\Sigma^{-1}(X_i - A\mu) + \text{constant}.$$

Let $\dot{X}_i = X_i - A\mu$. Then the log-likelihood is the following quadratic form (up to an additive constant):

$$\begin{aligned}
\ell(\mu) &= -\frac{1}{2}\sum_{i=1}^{n} \dot{X}_i^{\mathbf{T}}\Sigma^{-1}\dot{X}_i \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d}\left(\Sigma^{-1}\right)_{jk}\dot{X}_{ij}\dot{X}_{ik} \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d}\left(\Sigma^{-1}\right)_{jk}\left(X_{ij} - (A\mu)_j\right)\left(X_{ik} - (A\mu)_k\right) \\
&= -\frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d}\left(\Sigma^{-1}\right)_{jk}\sum_{i=1}^{n}(X_{ij} - (A\mu)_j)(X_{ik} - (A\mu)_k) \\
&= -\frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d}\left(\Sigma^{-1}\right)_{jk}\sum_{i=1}^{n}\left(X_{ij}X_{ik} - X_{ik}(A\mu)_j - X_{ij}(A\mu)_k + (A\mu)_j(A\mu)_k\right)
\end{aligned}$$

10.12.5 Covariance Matrices and Decompositions

As described in lecture, the covariance matrix $\text{Var}(R) \in \mathbb{R}^{d \times d}$ for a random variable $R \in \mathbb{R}^d$ with mean μ is

$$\text{Var}(R) = \text{Cov}(R, R) = \mathbb{E}[(R-\mu)(R-\mu)^\top] = \begin{bmatrix} \text{Var}(R_1) & \text{Cov}(R_1, R_2) & \dots & \text{Cov}(R_1, R_d) \\ \text{Cov}(R_2, R_1) & \text{Var}(R_2) & & \text{Cov}(R_2, R_d) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(R_d, R_1) & \text{Cov}(R_d, R_2) & \dots & \text{Var}(R_d) \end{bmatrix}$$

where $\text{Cov}(R_i, R_j) = \mathbb{E}[(R_i - \mu_i)(R_j - \mu_j)]$ and $\text{Var}(R_i) = \text{Cov}(R_i, R_i)$.

If the random variable R is sampled from the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with the PDF

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{((x-\mu)^\top \Sigma^{-1}(x-\mu))/2},$$

then $\text{Var}(R) = \Sigma$.

Given n points $X_1, X_2, \dots, X_n$ sampled from $\mathcal{N}(\mu, \Sigma)$, we can estimate Σ with the maximum likelihood estimator

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^\top,$$

which is also known as the covariance matrix of the sample

(a) The estimate $\widehat{\Sigma}$ makes sense as an approximation of Σ only if $\widehat{\Sigma}$ is invertible. Under what circumstances is $\widehat{\Sigma}$ not invertible? Make sure your answer is complete; i.e., it includes all cases in which the covariance matrix of the sample is singular. Express you answer in terms of the geometric arrangement of the sample points Xi.

Let $\dot{X}$ represent the centered data, i.e. $\dot{X}_i = X_i - \mu$.

Note that $\widehat{\Sigma}$ is the mean of a collection of n outer product matrices $\dot{X}_i \dot{X}_i^{\mathbf{T}}$, where each outer product matrix is contributed by a single sample point. Also note that the columns of $\dot{X}_i \dot{X}_i^{\mathbf{T}}$ are all scalar multiples of $\dot{X}_i$.

$\widehat{\Sigma}$ is invertible if and only if it is full-rank. Full-rank means that its columns are linearly independent.

From this point of view, the following circumstances will lead to $\widehat{\Sigma}$ being singular:

(a) **There is only one point.** If $n = 1$ then $\mu = X_1$ and $\dot{X}_1 = \mathbf{0}$, and the outer product is the zero matrix. This is singular (e.g. determinant is zero).

(b) $d > 1$ **and there are only two points.** If $n = 2$ then μ lies on the line connecting the two points, so $\dot{X}_1 = a\dot{X}_2$ for some scalar a. Therefore the columns of the sum of the two outer product matrices differ only by a scalar multiple.

In general, the centered sample vectors $\{\dot{X}_i : 1 < i \leq n\}$ must span $\mathbb{R}^d$. I.e. if the sample vectors lie in an affine hyperplane of $\mathbb{R}^d$ then $\widehat{\Sigma}$ will be singular.

(b) Suggest a way to fix a singular covariance matrix estimator $\widehat{\Sigma}$ by replacing it with a similar but invertible matrix. Your suggestion may be a kludge, but it should not change the covariance matrix too much. Note that infinitesimal numbers do not exist; if your solution uses a very small number, explain how to calculate a number that is sufficiently small for your purposes.

In my code I have used the pseudoinverse function `numpy.linalg.pinv`.

(c) Consider the normal distribution $\mathcal{N}(0, \Sigma)$ with mean $\mu = 0$. Consider all vectors of length 1; i.e., any vector x for which $|x| = 1$. Which vector(s) x of length 1 maximizes the PDF $f(x)$? Which vector(s) x of length 1 minimizes $f(x)$? (Your answers should depend on the properties of Σ.) Explain your answer.

Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be unit-length eigenvectors of Σ, arranged in order of decreasing eigenvalue.

Note that $f(x)$ is maximum at the mean $\mathbf{0}$ and decreases with increasing distance from $\mathbf{0}$. The exact form of this decrease is determined by the quadratic form $(\mathbf{x} - \mu)^{\mathbf{T}} \Sigma^{-1} (\mathbf{x} - \mu)$.

Then the unit vector x that maximizes $f(\mathbf{x})$ is $\mathbf{v}_n$. This is because the eigenvector with smallest eigenvalue points in the direction of least slope of the quadratic form. Similarly, $\mathbf{v}_1$ is the unit vector that minimizes $f(\mathbf{x})$ because the eigenvector with largest eigenvalue points in the direction of greatest slope of the quadratic form.

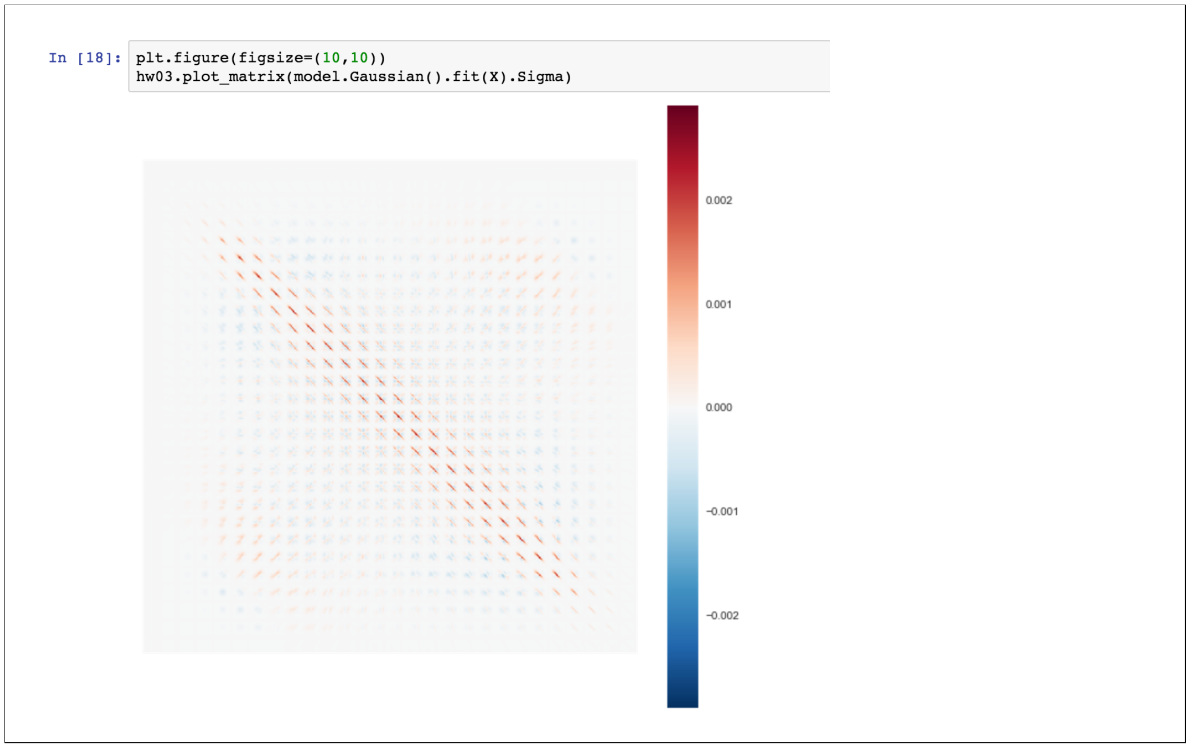
10.12.6 Gaussian Classifiers for Digits and Spam

In this problem, you will build classifiers based on Gaussian discriminant analysis. Unlike Homework 1, you are NOT allowed to use any libraries for out-of-the-box classification (e.g `sklearn`). You may use anything in `numpy` and `scipy`.

The training and test data can be found on Piazza in the post corresponding to this homework. Don't use the training/test data from Homework 1, as they have changed for this homework. Submit your predicted class labels for the test data on the Kaggle competition website and be sure to include your Kaggle display name and scores in your writeup. Also be sure to include an appendix of your code at the end of your writeup.

(a) Taking pixel values as features (no new features yet, please), fit a Gaussian distribution to each digit class using maximum likelihood estimation. This involves computing a mean and a covariance matrix for each digit class, as discussed in lecture. *Tip*: You may, and probably should, contrast-normalize the images before using their pixel values. One way to normalize is to divide the pixel values of an image by the l_2 norm of its pixel values.

(b) (Written answer) Visualize the covariance matrix for a particular class (digit). How do the diagonal terms compare with the off-diagonal terms? What do you conclude from this?



(c) Classify the digits in the test set on the basis of posterior probabilities with two different approaches.

(i) Linear discriminant analysis (LDA). Model the class conditional probabilities as Gaussians $\mathcal{N}(\mu_C, \Sigma)$ with different means μ_C (for class C) and the same covariance matrix Σ, the average covariance matrix of the 10 classes.

Hold out 10,000 randomly chosen training points for a validation set. Classify each image in the validation set into one of the 10 classes (with a 0-1 loss function). Compute the error rate and

plot it over the following numbers of randomly chosen training points:

$$[100, 200, 500, 1,000, 2,000, 5,000, 10,000, 30,000, 50,000].$$

(Expect some variance in your error rate when few training points are used.)

(ii) Quadratic discriminant analysis (QDA). Model the class conditionals as Gaussians $\mathcal{N}(\mu_C, \Sigma_C)$, where Σ_C is the estimated covariance matrix for class C. (If any of these covariance matrices turn out singular, implement the trick you described in Q5.(b). You are welcome to use k-fold cross validation to choose the right constant(s) for that trick.) Repeat the same tests and error rate calculations you did for LDA.

(iii) (Written answer.) Which of LDA and QDA performed better? Why?

My QDA implementation is currently incorrect. The unit tests pass for 1 and 2D cases but when the sample points are higher-dimensional QDA is classifying points essentially uniformly at random (around 90% error rate).

Here are the error rates for my LDA implementation using the provided data, with contrast normalization.

#training points	Error rate
100	0.51
200	0.91
500	0.78
1000	0.36
2000	0.34
5000	0.31
10000	0.34
30000	0.29
50000	0.28

(iv) Train your best classifier with `train.mat` and classify the images in `test.mat`. Submit your labels to the online Kaggle competition. Record your optimum prediction rate in your submission. You are welcome to compute extra features for the Kaggle competition. If you do so, please describe your implementation in your assignment. Please use extra features **only** for this portion of the assignment. In your submission, include plots of error rate versus number of training examples for both LDA and QDA. Also include tables giving the error rates (as percentages) for each number of training examples for both LDA and QDA. Include written answers where indicated.

0.77 accuracy score for digits; LDA.

(d) Next, apply LDA or QDA (your choice) to spam. Submit your test results to the online Kaggle competition. Record your optimum prediction rate in your submission. If you use additional features (or omit features), please describe them.

Optional: If you use the defaults, expect relatively low classification rates. The TAs suggest using a bag-of-words model. You may use third-party packages to implement that if you wish. Also, normalizing your vectors might help.

0.71 accuracy score for spam; LDA.

(e) *Extra for Experts:* Using the `training_data` and `training_labels` in `spam.mat`, identify 10 words in your features set corresponding to the maximum and minimum variances. Use k-fold cross validation to train your classifier using only 10 variance-maximum words and record your average error rate. Do the same with the 10 minimum-variance words. What do you notice?

Solution:

10.13 Homework 4 - Regression

10.13.1 Logistic Regression with Newton's Method

Consider sample points $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ and associated values $y_1, y_2, \ldots, y_n \in \{0, 1\}$, an $n \times d$ design matrix $X = [X_1 \quad \ldots \quad X_n]^\mathbf{T}$ and an n-vector $y = [y_1 \quad \ldots \quad y_n]^\mathbf{T}$.

If we add ℓ_2-regularization to logistic regression, the cost function is

$$J(w) = \lambda \left|w\right|_2^2 - \sum_{i=1}^{n} \left(y_i \ln s_i + (1 - y_i) \ln(1 - s_i) \right)$$

where $s_i = s(X_i \cdot w)$, $s(\gamma) = 1/(1+e^{-\gamma})$, and $\lambda > 0$ is the regularization parameter. As in lecture, the vector $s = [s_1 \quad \ldots \quad s_n]^\mathbf{T}$ is a useful shorthand.

In this problem, you will use Newton's method to minimize this cost function on the four-point, two dimensional training set

$$X = \begin{bmatrix} 0 & 3 \\ 1 & 3 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad\qquad y = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

You may want to draw these points on paper to see what they look like. The y-vector implies that the first two sample points are in class 1, and the last two are in class 0.

These sample points cannot be separated by a decision boundary that passes through the origin. As described in lecture, append a 1 to each X_i vector and use a weight vector $w \in \mathbb{R}^3$ whose last component is the bias term (the term we call α in lecture).

1. Derive the gradient of the cost function $J(w)$. Your answer should be a simple matrix-vector expression. Do NOT write your answer in terms of the individual components of the gradient vector.

   Note that $s'(\gamma) = \frac{e^{-\gamma}}{(1+e^{-\gamma})^2} = s(\gamma)(1 - s(\gamma))$.

   Let $s_i = s(\mathbf{x}_i^\mathbf{T}\mathbf{w})$, so that $\nabla_\mathbf{w}\, s_i = \mathbf{x}_i$. We have

   $$J(\mathbf{w}) = \lambda|\mathbf{w}|^2 - \sum_i y_i \log s_i + (1 - y_i) \log\left(1 - s_i\right),$$

so

$$\begin{aligned}\nabla J(\mathbf{w}) &= 2\lambda\mathbf{w} - \sum_i \frac{y_i}{s_i}(s_i)(1-s_i)\mathbf{x}_i + \frac{1-y_i}{1-s_i}(-1)(s_i)(1-s_i)\mathbf{x}_i \\ &= 2\lambda\mathbf{w} - \sum_i \mathbf{x}_i \left(y_i(1-s_i) - (1-y_i)s_i\right) \\ &= 2\lambda\mathbf{w} - \sum_i \mathbf{x}_i \left(y_i - s_i\right) \\ &= 2\lambda\mathbf{w} - \mathbf{X}^\mathbf{T}\left(\mathbf{y} - \mathbf{s}(\mathbf{X}\mathbf{w})\right) \qquad (d\times 1)\end{aligned}$$

where $\mathbf{s} : \mathbb{R}^n \to \mathbb{R}^n$ applies s componentwise to the rows.

We can interpret this expression a bit. $s(\mathbf{X}\mathbf{w})$ is an n-vector containing the predicted values for each sample point, so $\mathbf{y} - s(\mathbf{X}\mathbf{w})$ is the error in the current predicted values, and $\mathbf{X}^\mathbf{T}(\mathbf{y} - s(\mathbf{X}\mathbf{w}))$ is a d-vector whose j-th component is large if feature j is correlated with (has a large dot product with) the current errors. So the steepest direction downhill will tend to put more weight on features that are correlated with the current error in the predictions.

2. Derive the Hessian of $J(w)$. Again, your answer should be a simple matrix-vector expression.

The Hessian is the $d \times$ d matrix of partial derivatives of the gradient, so we have

$$\nabla^2 J(\mathbf{w}) = 2\lambda\mathbf{I} + \mathbf{X}^\mathbf{T} \operatorname{Jac}\mathbf{s}(\mathbf{X}\mathbf{w}),$$

where $\operatorname{Jac}\mathbf{s}$ is the Jacobian matrix of the vector-valued function $\mathbf{s}$.

We can compute the Jacobian using the chain rule. Define $\mathbf{f}(\mathbf{w}) = \mathbf{X}\mathbf{w}$ so now $\mathbf{s}(\mathbf{X}\mathbf{w}) = (\mathbf{s}\circ\mathbf{f})(\mathbf{w})$:

Function	domain $\to$ range	Jacobian	dim Jacobian
$\mathbf{f}(\mathbf{w}) = \mathbf{X}\mathbf{w}$	$\mathbb{R}^d \to \mathbb{R}^n$	$\mathrm{D}\,\mathbf{f} = \mathbf{X}$	$n \times d$
$\mathbf{s}(\mathbf{z})$	$\mathbb{R}^n \to \mathbb{R}^n$	$\mathrm{D}\,\mathbf{s}(\mathbf{z}) = \mathbf{S}$	$n \times n$

where $\mathbf{S}$ is a $n \times n$ diagonal matrix with $\mathbf{S}_{ii} = s_i(1-s_i)$. Now by the chain rule,

$$\begin{aligned}\nabla^2 J(\mathbf{w}) &= 2\lambda\mathbf{I} + \mathbf{X}^\mathbf{T}\,\mathrm{D}_\mathbf{w}\,\mathbf{s}(\mathbf{X}\mathbf{w}) \\ &= 2\lambda\mathbf{I} + \mathbf{X}^\mathbf{T}(\mathrm{D}_\mathbf{f}\,\mathbf{s})(\mathrm{D}_w\,\mathbf{f}) \\ &= 2\lambda\mathbf{I} + \mathbf{X}^\mathbf{T}\mathbf{S}\mathbf{X}. \qquad (d\times d)\end{aligned}$$

3. State the update equation for one iteration of Newton's method for this problem.

The quadratic approximation to the cost function at $\mathbf{v}$ is

$$q(\mathbf{w}) = J(\mathbf{v}) + (\mathbf{w}-\mathbf{v})^\mathbf{T}\left(\nabla J(\mathbf{v})\right) + \frac{1}{2}(\mathbf{w}-\mathbf{v})^\mathbf{T}\left(\nabla^2 J(\mathbf{v})\right)(\mathbf{w}-\mathbf{v}).$$

We want to find the $\mathbf{w}$ that minimizes this. The gradient of this is something like

$$\nabla q(\mathbf{w}) = \nabla J(\mathbf{v}) + \left(\nabla^2 J(\mathbf{v})\right)\mathbf{w},$$

but that's not quite right. Anyway, from the lecture notes, setting the gradient equal to zero gives

$$\mathbf{w} = \mathbf{v} - \left(\nabla^2 J(\mathbf{v})\right)^{-1} \nabla J(\mathbf{v}).$$

For our problem, this is (writing $\mathbf{w}^{(l)}$ instead of $\mathbf{v}$ for the value of $\mathbf{w}$ at iteration l.)

$$\mathbf{w}^{(l+1)} = \mathbf{w}^{(l)} - \left(2\lambda\mathbf{I} + \mathbf{X}^\mathbf{T}\mathbf{S}\mathbf{X}\right)^{-1}\left(2\lambda\mathbf{w}^{(l)} - \mathbf{X}^\mathbf{T}\left(\mathbf{y} - s(\mathbf{X}\mathbf{w}^{(l)})\right)\right).$$

4. We are given a regularization parameter of $\lambda = 0.07$ and a starting point of $\mathbf{w}^{(0)} = \begin{bmatrix}-2 & 1 & 0\end{bmatrix}^\mathbf{T}$.

```
from numpy import array
from numpy import diag
from numpy import exp
from numpy.linalg import inv


def q1_4():
    X = array([[0, 3, 1],
               [1, 3, 1],
               [0, 1, 1],
               [1, 1, 1]])
    y = array([[1],
               [1],
               [0],
               [0]])
    lambda_ = 0.07

    w0 = array([[-2],
                [ 1],
                [ 0]])

    s0 = logistic(X @ w0)

    w1 = logistic_regression_newton_update(w0, X, y, lambda_)

    s1 = logistic(X @ w1)

    w2 = logistic_regression_newton_update(w1, X, y, lambda_)


def logistic_regression_newton_update(w, X, y, lambda_):
    s = logistic(X @ w)
    gradient = 2 * lambda_ * w - X.T @ (y - s)
    B = diag((s * (1 - s) + 2 * lambda_).ravel())
    hessian = X.T @ B @ X
    return w - inv(hessian) @ gradient


def logistic(z):
    return 1 / (1 + exp(-z))
```

(a) State the value of $s^{(0)}$ (the value of s before any iterations).

```
[[ 0.95257413]
 [ 0.73105858]
 [ 0.73105858]
 [ 0.26894142]]
```

(b) State the value of $w^{(1)}$ (the value of w after one iteration).

```
[[ 0.03660748]
```

```
 [ 1.77901816]
 [-3.1787346 ]]
```

(c) State the value of $s^{(1)}$.

```
[[ 0.89644368]
 [ 0.89979306]
 [ 0.19786111]
 [ 0.20373548]]
```

(d) State the value of $w^{(2)}$ (the value of w after two iterations).

```
[[-0.84243273]
 [ 1.2968546 ]
 [-1.60471569]]
```

10.13.2 ℓ_1- and ℓ_2-Regularization

Consider sample points $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ and associated values $y_1, y_2, \ldots, y_n \in \mathbb{R}$, an $n \times d$ design matrix $X = [X_1 \quad \ldots \quad X_n]^{\mathbf{T}}$ and an n-vector $y = [y_1 \quad \ldots \quad y_n]^{\mathbf{T}}$. For the sake of simplicity, assume that the sample data has been centered and whitened so that each feature has mean 0 and variance 1 and the features are uncorrelated; i.e., $X^{\mathbf{T}}X = nI$. For this question, we will not use a fictitious dimension nor a bias term; our linear regression function will be zero for $x = 0$.

Consider linear least-squares regression with regularization in the ℓ_1-norm, also known as Lasso. The Lasso cost function is

$$J(w) = |Xw - y|^2 + \lambda \|w\|_{\ell_1}$$

where $w \in \mathbb{R}^d$ and $\lambda > 0$ is the regularization parameter. Let $w^* = \text{argmin}_{w \in \mathbb{R}^d} J(w)$ denote the weights that minimize the cost function.

In the following steps, we will show that whitened training data decouples the features, so that w_i^* is determined by the i^{th} feature alone (i.e., column i of the design matrix X), regardless of the other features. This is true for both Lasso and ridge regression.

1. We use the notation $X_{*1}, X_{*2}, \ldots, X_{*d}$ to denote column i of the design matrix X, which represents the i^{th} feature. (Not to be confused with row i of X, the sample point $X_i^{\mathbf{T}}$.) Write $J(w)$ in the following form for appropriate functions g and f.

$$J(w) = g(y) + \sum_{i=1}^{d} f(X_{*i}, w_i, y, \lambda)$$

The cost function is

$$\begin{aligned} J(\mathbf{w}) =& |\mathbf{X}\mathbf{w} - \mathbf{y}|^2 + \lambda ||\mathbf{w}||_1 \\ =& \mathbf{w}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{w} - 2y^{\mathbf{T}}\mathbf{X}\mathbf{w} + \mathbf{y}^{\mathbf{T}}\mathbf{y} + \lambda ||\mathbf{w}||_1 \\ =& n\mathbf{w}^{\mathbf{T}}\mathbf{w} - 2y^{\mathbf{T}}\mathbf{X}\mathbf{w} + \mathbf{y}^{\mathbf{T}}\mathbf{y} + \lambda ||\mathbf{w}||_1 \qquad \text{(because } \mathbf{X}^{\mathbf{T}}\mathbf{X} = n\mathbf{I}\text{).} \end{aligned}$$

Now $\mathbf{w}^{\mathbf{T}}\mathbf{w} = \sum_{i=1}^{d} w_i^2$, and $||\mathbf{w}||_1 = \sum_{i=1}^{d} |w_i|$, and

$$y^{\mathbf{T}}\mathbf{X}\mathbf{w} = (\mathbf{y}^{\mathbf{T}}\mathbf{X})\mathbf{w} = \sum_{i=1}^{d} \mathbf{X}_{*i}^{\mathbf{T}}\, \mathbf{y} w_i,$$

so

$$\begin{aligned} J(\mathbf{w}) &= g(\mathbf{y}) + \sum_{i=1}^{d} f(\mathbf{X}_{*i}, w_i, y, \lambda), \qquad \text{where} \\ g(\mathbf{y}) &= \mathbf{y}^{\mathbf{T}}\mathbf{y} \qquad \text{and} \\ f(\mathbf{X}_{*i}, w_i, \mathbf{y}, \lambda) &= nw_i^2 + \lambda |w_i| - 2\mathbf{X}_{*i}^{\mathbf{T}}\, \mathbf{y} w_i. \end{aligned}$$

2. If $w_i^* > 0$, what is the value of w_i^*?

For $w_i \geq 0$, the i-th component of $J(\mathbf{w})$ is

$$J(\mathbf{w})_i = nw_i^2 + w_i(\lambda - 2\mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y}) + \text{constant}.$$

so

$$\frac{\partial J}{\partial w_i} = 2nw_i + \lambda - 2\mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y},$$

and setting the gradient equal to zero gives

$$w_i^* = \begin{cases} \frac{2\mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y} - \lambda}{2n}, & \mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y} > \frac{\lambda}{2} \\ 0, & \text{otherwise.} \end{cases}$$

3. If $w_i^* < 0$, what is the value of w_i^*?

For $w_i \leq 0$, the i-th component of $J(\mathbf{w})$ is

$$J(\mathbf{w})_i = nw_i^2 - w_i(\lambda + 2\mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y}) + \text{constant}.$$

so

$$\frac{\partial J}{\partial w_i} = 2nw_i - \lambda - 2\mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y},$$

and setting the gradient equal to zero gives

$$w_i^* = \begin{cases} \frac{\lambda + 2\mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y}}{2n}, & \mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y} < -\frac{\lambda}{2} \\ 0, & \text{otherwise.} \end{cases}$$

4. Considering parts 2 and 3, what is the condition for w_i^* to be zero?

$|\mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y}| \leq \frac{\lambda}{2}$.

5. Now consider ridge regression, which uses the ℓ_2 regularization term $\lambda\,|w|^2$. How does this change the function $f(\cdot)$ from part 1? What is the new condition in which $w_i^* = 0$? How does it differ from the condition you obtained in part 4?

For ridge regression we have

$$J(\mathbf{w}) = g(\mathbf{y}) + \sum_{i=1}^{d} f(\mathbf{X}_{*i}, w_i, y, \lambda), \qquad \text{where}$$
$$f(\mathbf{X}_{*i}, w_i, \mathbf{y}, \lambda) = (n + \lambda)w_i^2 - 2\mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y}w_i,$$

and g is as above. So

$$\frac{\partial J}{\partial w_i} = 2(n + \lambda)w_i - 2\mathbf{X}_{*i}^{\mathbf{T}}\ \mathbf{y},$$

and

$$w_i^* = \frac{\mathbf{X}_{*i}^{\mathbf{T}}\,\mathbf{y}}{n+\lambda}.$$

So the weight for the i-th feature is zero if and only if $\mathbf{X}_{*i}^{\mathbf{T}}\,\mathbf{y} = 0$, i.e. the n-vector containing the i-th feature is orthogonal to the observed training values $\mathbf{y}$.

This is in contrast to Lasso, for which the i-th feature receives a weight of zero if $|\mathbf{X}_{*i}^{\mathbf{T}}\,\mathbf{y}| \leq \frac{\lambda}{2}$, i.e. if the dot product of the i-th feature with the training values $\mathbf{y}$ falls below $\lambda/2$.

This result is consistent with the general notion that Lasso tends to set some weights to exactly zero whereas ridge regression would set them to a small but usually non-zero value.

10.13.3 Regression and Dual Solutions

a) For a vector w, derive $\nabla |w|^4$. Then derive $\nabla_w |Xw - y|^4$.

Suppose $\mathbf{w} \in \mathbb{R}^d$. Then $|\mathbf{w}|^4 \in \mathbb{R}$ is

$$|\mathbf{w}|^4 = \left(\sum_{j=1}^{d} w_j^2 \right)^2 = \sum_{j=1}^{d} \sum_{k=1}^{d} w_j^2 w_k^2.$$

Now consider the j-th component. Viewed as a function of $\mathbf{w}_j$, we have

$$|\mathbf{w}|^4 = w_j^4 + 2w_j^2 \sum_{k \neq j} w_k^2 + \text{constant}$$

therefore

$$\begin{aligned} \frac{\partial |\mathbf{w}|^4}{\partial w_j} &= 4w_j^3 + 4w_j \sum_{k \neq j} w_k^2 \\ &= 4|\mathbf{w}|^2 w_j \end{aligned}$$

so

$$\nabla_{\mathbf{w}} |\mathbf{w}|^4 = 4|\mathbf{w}|^2 \mathbf{w}.$$

Now let $|\mathbf{Xw} - \mathbf{y}|^4 = (g \circ f)(\mathbf{w})$, where

$$\begin{aligned} f &: \mathbb{R}^d \to \mathbb{R}^n & f(\mathbf{w}) &= \mathbf{Xw} - \mathbf{y} \\ g &: \mathbb{R}^n \to \mathbb{R} & g(\mathbf{z}) &= |\mathbf{z}|^4. \end{aligned}$$

The chain rule states that $\nabla(g \circ f) = (Df)^{\mathbf{T}} \nabla g$, where Df is the Jacobian matrix of first partial derivatives of f. We have $\nabla g(\mathbf{z}) = 4|\mathbf{z}|^2 \mathbf{z}$ and $Df = \mathbf{X}$, so

$$\begin{aligned} \nabla_w |\mathbf{Xw} - \mathbf{y}|^4 &= (Df)^{\mathbf{T}} \nabla g \\ &= 4|\mathbf{Xw} - \mathbf{y}|^2 \mathbf{X}^{\mathbf{T}} (\mathbf{Xw} - \mathbf{y}) \\ &= 4|\mathbf{Xw} - \mathbf{y}|^2 \mathbf{X}^{\mathbf{T}} \mathbf{Xw} - \mathbf{X}^{\mathbf{T}} \mathbf{y} \end{aligned}$$

b) Consider sample points $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ and associated values $y_1, y_2, \ldots, y_n \in \mathbb{R}$, an $n \times d$ design matrix $X = [X_1 \quad \ldots \quad X_n]^{\mathbf{T}}$ and an n-vector $y = [y_1 \quad \ldots \quad y_n]^{\mathbf{T}}$, and the regularized regression problem

$$w^* = \operatorname*{argmin}_{w \in \mathbb{R}^d} |Xw - y|^4 + \lambda |w|^2,$$

which is similar to ridge regression, but we take the fourth power of the error instead of the squared error. (It is not possible to write the optimal solution w^* as the solution of a system of linear equations, but it can be found by gradient descent or Newton's method.)

Show that the optimum w^* is unique. By setting the gradient of the objective function to zero, show that w^* can be written as a linear combination $w^* = \sum_{i=1}^{n} a_i X_i$ for some scalars $a_1, \ldots, a_n$. Write the vector a of dual coefficients in terms of X, y, and the optimal solution w^*.

The objective function $J(\mathbf{w})$ is

$$\begin{aligned}|\mathbf{X}\mathbf{w}-\mathbf{y}|^4+\lambda|\mathbf{w}|^2 &= \big((\mathbf{X}\mathbf{w}-\mathbf{y})\cdot(\mathbf{X}\mathbf{w}-\mathbf{y})\big)^2+\lambda|\mathbf{w}|^2 \\ &= \Big(\mathbf{w}^T\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{w}-2\mathbf{w}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y}+\mathbf{y}^\mathbf{T}\mathbf{y}\Big)^2+\lambda|\mathbf{w}|^2\end{aligned}$$

I think this objective function is convex in $\mathbf{w}$, but I'm not sure how to show that. Basically, I think $\mathbf{w}^T\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{w}$ is a convex function of $\mathbf{w}$ since $\mathbf{X}^\mathbf{T}\mathbf{X}$ is positive definite, and $2\mathbf{w}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y}$ is also a convex function of $\mathbf{w}$, and the sum of convex functions is convex, and the square of a convex function is convex, so the whole expression is convex. Being convex in $\mathbf{w}$ means that there is a unique minimum at $\mathbf{w}^*$.

The gradient of the objective function $J(\mathbf{w})$ is

$$\nabla_\mathbf{w} J(\mathbf{w}) = 4|\mathbf{X}\mathbf{w}-\mathbf{y}|^2\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{w}-\mathbf{X}^\mathbf{T}\mathbf{y}+2\lambda\mathbf{w}.$$

Setting this equal to zero gives TODO: LaTeX error but I don't see how to simplify this to show that w^* can be written as a linear combination $w^*=\sum_{i=1}^n a_iX_i$ for some scalars $a_1,\ldots,a_n$.

c) Consider the regularized regression problem

$$w^* = \underset{w\in\mathbb{R}^d}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^n L(w^\mathbf{T}X_i,y_i)+\lambda\,|w|^2$$

where the loss function L is convex in its first argument. Prove that the optimal solution has the form $w^*=\sum_{i=1}^n a_iX_i$. If the loss function is not convex, does the optimal solution always have the form $w^*=\sum_{i=1}^n a_iX_i$? Justify your answer.

10.13.4 Classification + Logistic Regression

Daylen is planning the frat party of the semester. He's completely stocked up on Franzia. Unfortunately, the labels for 497 boxes (test set) have been scratched off, and he needs to quickly find out which boxes contain Red wine (label 1) and White wine (label 0). Fortunately, for him the boxes still have their Nutrition Facts (features) intact and detail the chemical composition of the wine inside the boxes (the description of these features and the features themselves are provided in `data.mat`). He also has 6,000 boxes with Nutrition Facts and labels intact (train set). Help Daylen figure out what the labels should be for the 497 mystery boxes.

1. Derive and write down the batch gradient descent update equation for logistic regression with ℓ_2 regularization.

   From Q1, the gradient of the cost function is

   $$\nabla J(\mathbf{w}) = 2\lambda\mathbf{w} - \mathbf{X}^\mathbf{T}\left(\mathbf{y} - s(\mathbf{X}\mathbf{w})\right),$$

   where $s(\mathbf{X}\mathbf{w}) = \begin{bmatrix} 1/(1+e^{-\mathbf{X}_1\cdot\mathbf{w}}) \\ \vdots \\ 1/(1+e^{-\mathbf{X}_n\cdot\mathbf{w}}) \end{bmatrix}$ contains the predicted values (class probability) for each sample point, given parameters $\mathbf{w}$.

   Therefore the batch gradient descent update equation with learning rate ϵ at iteration k is

   $$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \epsilon\left(2\lambda\mathbf{w}^{(k)} - \mathbf{X}^\mathbf{T}\left(\mathbf{y} - s(\mathbf{X}\mathbf{w}^{(k)})\right)\right).$$

   Choose a reasonable regularization parameter value and a reasonable learning rate. Run your algorithm and plot the cost function as a function of the number of iterations. (As this is batch descent, one "iteration" should use every sample point once.)

   My implementation of logistic regression with regularization passes the simple 1D test case, but has some numerical problems when used on the real data, causing the cost function to not always decrease under gradient "descent". Therefore I failed to make a Kaggle submission.

2. Derive and write down the stochastic gradient descent update equation for logistic regression with ℓ_2 regularization. Choose a suitable learning rate. Run your algorithm and plot the cost function as a function of the number of iterations—where now each "iteration" uses *just one* sample point.

   The stochastic gradient descent update equation is:

   On each iteration:

   (a) Sample i from a discrete Uniform distribution on $1, \ldots, n$.

   (b) Update w according to

   $$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \epsilon\left(2\lambda\mathbf{w}^{(k)} - \mathbf{x}_i^\mathbf{T}\left(\mathbf{y} - s(\mathbf{x}_i^\mathbf{T}\mathbf{w}^{(k)})\right)\right).$$

   Comment on the differences between the convergence of batch and stochastic gradient descent.

3. Instead of a constant learning rate ϵ, repeat part 2 where the learning rate decreases as $\epsilon \propto 1/t$ for the t^{th} iteration. Plot the cost function vs. the number of iterations. Is this strategy better than having a constant ϵ?

4. Finally, train your classifier on the entire training set. Submit your predictions for the test set to Kaggle. You can only submit twice per day, so get started early! In your writeup, include your Kaggle display name and score and describe the process you used to decide which parameters to use for your best classifier.

10.13.5 Real World Spam Classification

Motivation: After taking CS 189 or CS 289A, students should be able to wrestle with "real-world" data and problems. These issues might be deeply technical and require a theoretical background, or might demand specific domain knowledge. Here is an example that a past TA encountered.

Daniel (a past CS 189 TA) interned as an anti-spam product manager for an email service provider. His company uses a linear SVM to predict whether an incoming spam message is spam or ham. He notices that the number of spam messages received tends to spike upwards a few minutes before and after midnight. Eager to obtain a return offer, he adds the timestamp of the received message, stored as number of milliseconds since the previous midnight, to each feature vector for the SVM to train on, in hopes that the ML model will identify the abnormal spike in spam volume at night. To his dismay, after testing with the new feature, Daniel discovers that the linear SVM's success rate barely improves.

Why can't the linear SVM utilize the new feature well, and what can Daniel do to improve his results? Daniel is unfortunately limited to a quadratic kernel i.e. the features are at most polynomials of degree 2 over the original variables. This is an actual interview question Daniel received for a machine learning engineering position!

Write a short explanation. This question is open ended, and there can be many correct answers.

The way the new feature was defined means that both small and large values are associated with being spam. I wonder if it would perform better if instead he defined it as:

Absolute value of (timestamp at noon) minus (timestamp of email)

although, perhaps the quadratic kernel would already be handling this.

10.14 Homework 6 - Neural Networks

10.14.1 Model specification

K possible output categories; one hidden layer of H units; tanh activation in the hidden layer; logistic activation in the output layer. Notation:

		indices	dimensions
Input layer	$\mathbf{x}$	x_j	$d \times 1$
Weights	$\mathbf{V}$	V_{hj}	$H \times d$
Hidden layer	$\mathbf{z} = \tanh(\mathbf{Vx})$	z_h	$H \times 1$
Weights	$\mathbf{W}$	W_{kh}	$K \times H$
Ouput layer	$\widehat{\mathbf{y}} = \sigma(\mathbf{Wz})$	$\widehat{y}_k$	$K \times 1$
Loss	$L(\widehat{\mathbf{y}}, \mathbf{y})$		scalar

where σ is the logistic function $\sigma(x) = (1 - e^{-x})^{-1}$, and tanh and σ act elementwise.

The loss (cost) function is the cross-entropy (log likelihood of training labels given predictions)

$$-L(\widehat{y}, \mathbf{y}) = \sum_k y_k \log(\widehat{y}_k) + (1 - y_k) \log(1 - \widehat{y}_k).$$

10.14.2 Gradient descent algorithm

We want to do gradient descent on the full set $(\mathbf{V}, \mathbf{W})$ of parameters. This involves computing gradients of the loss function $\nabla_V L$ and $\nabla_W L$. We derive the gradients with respect to one row of these matrices at a time, and give code fragments showing how to compute the matrix of derivatives efficiently.

10.14.3 Gradient with respect to weight matrix W

$\mathbf{W}_k$ is one row of $\mathbf{W}$, of length $H + 1$. We have

$$\nabla_{\mathbf{W}_k} L = \frac{\partial L}{\partial \widehat{y}_k} \nabla_{\mathbf{W}_k} \widehat{y}_k.$$

Now, $\widehat{y}_k = \sigma(\mathbf{W}_k \mathbf{z})$, so

$$\nabla_{\mathbf{W}_k} \widehat{y}_k = \mathbf{z} \widehat{y}_k (1 - \widehat{y}_k).$$

This expression is still correct if the offset is implemented as an additional "dimension", in which case the last element of $\mathbf{W}_k$ is the offset and the last element of $\mathbf{z}$ is 1.

The derivative of the loss with respect to $\widehat{y}_k$ is

$$\frac{\partial L}{\partial \widehat{y}_k} = -\frac{y_k}{\widehat{y}_k} + \frac{1 - y_k}{1 - \widehat{y}_k} = \frac{\widehat{y}_k - y_k}{\widehat{y}_k (1 - \widehat{y}_k)}.$$

Multiplying these quantities gives

$$\nabla_{\mathbf{W}_k} L = \mathbf{z}(\widehat{y}_k - y_k).$$

In code we can compute the full matrix of derivatives $\nabla_{\mathbf{W}}$ using vector/matrix primitives as

$$\text{diag}(\widehat{\mathbf{y}} - \mathbf{y})\ \mathbf{Z},$$

where the rows of $\mathbf{Z}$ are each equal to $\mathbf{z}$:

```
grad__L__z = (W.T * (yhat - y)).sum(axis=1)
zz = z.reshape((1, H + 1)).repeat(K, 0)
grad__L__W = diag(yhat - y) @ zz
```

10.14.4 Gradient with respect to weight matrix V

$\mathbf{V}_h$ is one row of $\mathbf{V}$, of length $d + 1$. We have

$$\nabla_{\mathbf{V}_h} L = \frac{\partial L}{\partial \mathbf{z}_h} \nabla_{\mathbf{V}_h} \mathbf{z}_h.$$

Now, $\frac{\partial L}{\partial z_h} = \sum_k \frac{\partial L}{\partial \widehat{y}_k} \frac{\partial \widehat{y}_k}{\partial z_h}$. We've already found $\frac{\partial L}{\partial \widehat{y}_k}$ above, and $\frac{\partial \widehat{y}_k}{\partial z_h} = W_{kh}\widehat{y}_k(1-\widehat{y}_k)$, giving

$$\frac{\partial L}{\partial z_h} = \sum_k W_{kh}(\widehat{y}_k - y_k).$$

$\mathbf{z}_h = \tanh(\mathbf{V}_h\mathbf{x})$, so $\nabla_{\mathbf{V}_h}\mathbf{z}_h = \mathbf{x}(1-z_h^2)$, and multiplying the two quantities gives

$$\nabla_{\mathbf{V}_h} L = \mathbf{x}(1-z_h^2)\sum_k W_{kh}(\widehat{y}_k - y_k).$$

Again, in code we can compute the full matrix of derivatives $\nabla_{\mathbf{V}} L$ using vector/matrix primitives:

```
grad__L__z = (W.T * (yhat - y)).sum(axis=1)
xx = x.reshape((1, d + 1)).repeat(H + 1, 0)
grad__L__V = diag((1 - z ** 2) * grad__L__z) @ xx
```

kaggle: `dandavison7` 0.88577

No submission for this question (I'm auditing the class, and just had time for the derivations and implementation, but do appreciate the grading on my derivations!)

kaggle: `dandavison7` 0.88577

No submission for this question (I'm auditing the class, and just had time for the derivations and implementation, but do appreciate the grading on my derivations!)

kaggle: `dandavison7` 0.88577

No submission for this question (I'm auditing the class, and just had time for the derivations and implementation, but do appreciate the grading on my derivations!)

kaggle: `dandavison7` 0.88577