Undergraduate Topics in Mathematics 3

Proof writing, The theory of sets, Axiomatic geometry, Numerical analysis

Nathan Warner

Northern Illinois University

Computer Science
Northern Illinois University
United States

Contents

Proofs

1.1 Intro to proof writing, intuitive proofs

- **Intro to definitions, propositions and proofs: the chessboard problem**: Suppose you have a chessboard (8×8 grid of squares) and a bunch of dominoes (2×1 block of squares), so each domino can perfectly cover two squares of the chessboard.

  Note that with 32 dominoes you can cover all 64 squares of the chessboard. There are many different ways you can place the dominoes to do this, but one way is to cover the first column by 4 dominoes end-to-end, cover the second column by 4 dominoes, and so on

  Math runs on definitions, so let's give a name to this idea of covering all the squares. Moreover, let's not define it just for 8×8 boards — let's allow the definition to apply to boards of other dimensions

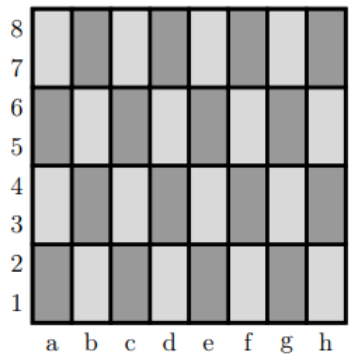
  Definition. A perfect cover of an $m \times n$ board with 2×1 dominoes is an arrangement of those dominoes on the chessboard with no squares left uncovered, and no dominoes stacked or left hanging off the end.

  As we demonstrated above, there exist perfect covers of the 8×8 chessboard. This is a book about proofs, so let's write this out as a proposition (something which is true and requires proof) and then let's write out a formal proof of this fact.

  Proposition. There exists a perfect cover of an 8×8 chessboard.

  This proposition is asserting that "there exists" a perfect cover. To say "there exists" something means that there is at least one example of it. Therefore, any proposition like this can be proven by simply presenting an example which satisfies the statement.

  Proof. Observe that the following is a perfect cover.



  We have shown by example that a perfect cover exists, completing the proof. ■

We typically put a small box at the end of a proof, indicating that we have completed our argument. This practice was brought into mathematics by Paul Halmos, and it is sometimes called the Halmos tombstone

One apocryphal story is that Halmos regarded proofs as living until proven. Once proven, they have been defeated — killed. And so he wrote a little tombstone to conclude his proof

What if I cross out the bottom-left and top-left squares, can we still perfectly cover the 62 remaining squares?

As you can probably already see, the answer is yes. For example, the first column can now be covered by 3 dominoes and the other columns can be covered by 4 dominoes each.

What if I cross out just one square, like the top-left square? Can this be perfectly covered?

The answer is no

Proposition. If one crosses out the top-left square of an 8×8 chessboard, the remaining squares can not be perfectly covered by dominoes.

Proof Idea. The idea behind this proof is that one domino, wherever it is placed, covers two squares. And two dominoes must cover four squares. And three cover six. In general, the number of squares covered — 2, 4, 6, 8, 10, etc. — is always an even number. This insight is the key, because the number of squares left on this chessboard is 63— an odd number

Proof. Since each domino covers 2 squares and the dominoes are non-overlapping, if one places our k dominoes on the board, then they will cover $2k$ squares, which is always an even number. Therefore, a perfect cover can only cover an even number of squares. Notice, though, that the board has 63 remaining squares, which is an odd number. Thus, it can not be perfectly covered.

What if I take an 8×8 chessboard and cross out the top-left and the bottom-right squares? Then can it be covered by dominoes?

Proposition. If one crosses out the top-left and bottom-right squares of an 8×8 chessboard, the remaining squares can not be perfectly covered by dominoes.

Proof. Observe that the chessboard has 62 remaining squares, and since every domino covers two squares, if a perfect cover did exist it would require

$$\frac{62}{2} = 31 \text{ dominoes.}$$

Also observe that every domino on the chessboard covers exactly one white square and exactly one black square

Thus, whenever you place 31 non-overlapping dominoes on a chessboard, they will collectively cover 31 white squares and 31 black squares.

Next observe that since both of the crossed-out squares are white squares, the remaining squares consist of 30 white squares and 32 black squares. Therefore, it is impossible to have 31 dominoes cover these 62 squares. ■

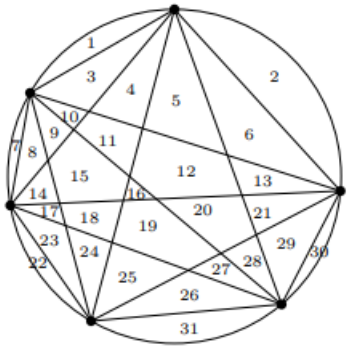
- **Naming Results**: So far, all of our results have been called "propositions." Here's the run-down on the naming of results:
  - A theorem is an important result that has been proved.
  - A proposition is a result that is less important than a theorem. It has also been proved.
  - A lemma is typically a small result that is proved before a proposition or a theorem, and is used to prove the following proposition or theorem.
  - A corollary is a result that is proved after a proposition or a theorem, and which follows quickly from the proposition or theorem. It is often a special case of the proposition or theorem.

  All of the above are results that have been proved — a conjecture, though, has not.
  - A conjecture is a statement that someone guesses to be true, although they are not yet able to prove or disprove it.
- **Conjectures and counterexamples**: As an example of a conjecture, suppose you were investigating how many regions are formed if one places n dots randomly on a circle and then connects them with lines.

$n = 1$ 1 region | $n = 2$ 2 regions | $n = 3$ 4 regions | $n = 4$ 8 regions | $n = 5$ 16 regions

At this point, if you were to conjecture how many regions there will be for the $n = 6$ case, your guess would probably be 32 regions — the number of regions certainly seems to be doubling at every step. In fact, if it kept doubling, then with a little more thought you might even conjecture a general answer: that n randomly placed dots form 2^{n-1} regions;

Surprisingly, this conjecture would be incorrect. One way to disprove a conjecture is to find a counterexample to it. And as it turns out, the $n = 6$ case is such a counterexample



$n = 6$
31 regions

This counterexample also underscores the reason why we prove things in math. Sometimes math is surprising. We need proofs to ensure that we aren't just guessing at what seems reasonable. Proofs ensure we are always on solid ground. Further, proofs help us understand why something is true — and that understanding is what makes math so fun

Lastly, we study proofs because they are what mathematicians do

- **The pingeonhole principal**

principal. The principal has a simple form and a general form. Assume k and n are positive integers

Simple form: If $n + 1$ objects are placed into n boxes, then at least one box has at least two objects in it.

General form: If $kn + 1$ objects are placed into n boxes, then at least one box has at least $k + 1$ objects in it.

Birthday example: If there are 330 million people in the united states, how many U.S. residents are guaranteed to have the same birthday according to the pigeonhole principal?

To determine this, let's see what would happen if each date of the year had exactly the same number of people born on it

$$\frac{330 \times 10^6}{366} = 901,639.344.$$

Since 901,639.344 people are born on an average day of the year, we should be able round up and say that at least one day of the year has had at least 901,640 people born on it. That is, with the pigeonhole principal we should be able to prove that there are at least 901,640 people in the USA with the same birthday

Solution. Imagine you have one box for each of the 366 dates of the (leap) year, and each person in the U.S. is considered an object. Put each person in the box corresponding to their birthday. By the general form of the pigeonhole principal (with $n = 366$ and $k = 901,639$ and thus $k + 1 = 901,640$), any group of

$$(901,639)(366) + 1.$$

people is guaranteed to contain 901,640 people which have the same birthday.

- **Another pingeonhole example**:

  Proposition. Given any five numbers from the set $\{1, 2, 3, 4, 5, 6, 7, 8\}$, two of the chosen numbers will add up to 9.

  We may think to start by listing the pairs that sum to 9. We have

  $$1 + 8$$
  $$2 + 7$$
  $$3 + 6$$
  $$4 + 5.$$

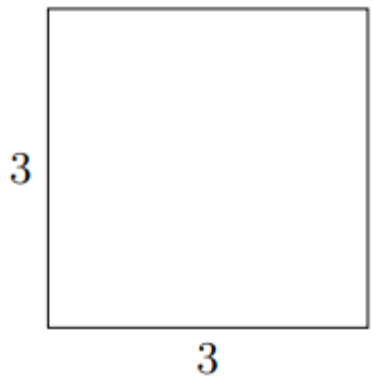
  And of course $8 + 1, 7 + 2, ..$ etc. We see we have four sums, we choose these sums as our boxes. If each of the four sums is a box, and each number is an object, then we are placing five objects into four boxes

  Proof. Let one box correspond to the numbers 1 and 8, a second box correspond to 2 and 7, another to 3 and 6, and a final box to 4 and 5. Notice that each of these pairs adds up to 9.

  Given any five numbers from $\{1, 2, 3, 4, 5, 6, 7, 8\}$, place each of these five numbers in the box to which it corresponds; for example, if your first number is a 6, then place it in the box labeled "3 and 6." Notice that we just placed five numbers into four boxes. Thus, by the simple form of the pigeonhole principal, there must be some box which contains two numbers in it. These two numbers add up to 9, as desired

- **Another pingeonhole example**:

  Proposition. Given any collection of 10 points from inside the following square (of side-length 3), there must be at least two of these points which are of distance at most $\sqrt{2}$



Proof. Divide the 3×3 square into nine 1×1 boxes. Placing 10 arbitrary points amongst the boxes gaurantees that at least one box will have at least two points. We observe that the farthest these two points can be from each other is when they sit in two corners such that a diagonal line through the box hits both points. The length of this line is given by

$$\sqrt{1^2 + 1^2} = \sqrt{2}.$$

Thus, we observe that the maximum distance of these two points is $\sqrt{2}$ ■

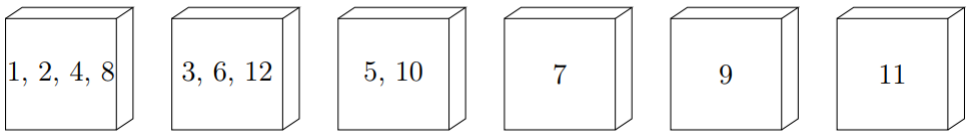
- **Another pingeonhole example**:

Proposition. Given any 101 integers from $\{1, 2, 3, ..., 200\}$, at least one of these numbers will divide another

Solution. As we ponder about how to construct 100 boxes from the properties of the set, we may wonder how the even and odd members partition this set. Call $S = \{1, 2, 3, ..., 200\}$, $E = \{2, 4, 6, ..., 200\}$, and $O = \{1, 3, 5, ..., 199\}$. Note that $E \cup O = S$. We notice that these two sets are arithmetic sequences, each with difference two. If $a_n = a_1 + (n-1)d$, then

$$\begin{aligned} n &= \frac{a_n - a_1}{2} + 1 \\ \implies n &= 100. \end{aligned}$$

Let's make the odd numbers are boxes. We note that any even number ℓ can be written as $\ell = 2^k m$, where m is odd, and k is the highest power of two that divides ℓ. Thus, in box m, we place any number of the form $2^k m$



For any pair of numbers in the same box, the smaller divides the larger. Picking 101 numbers from the set S, and only 100 boxes... by the pigeonhole principal we must have atleast two numbers in the same box, and thus the smaller divides the larger. ■.

Formal proof. Proof. For each number n from the set $\{1, 2, 3, \ldots, 200\}$, factor out as many 2's as possible, and then write it as $n = 2^k \cdot m$, where m is an odd number. So, for example, $56 = 2^3 \cdot 7$, and $25 = 2^0 \cdot 25$. Now, create a box for each odd number from 1 to 199; there are 100 such boxes.

Remember that we are given 101 integers and we want to find a pair for which one divides the other. Place each of these 101 integers into boxes based on this rule:

If the integer is n, then place it in Box m if $n = 2^k \cdot m$ for some k.

For example, $72 = 2^3 \cdot 9$ would go into Box 9, because that's the largest odd number inside it.

Since 101 integers are placed in 100 boxes, by the pigeonhole principal (principal 1.5) some box must have at least 2 integers placed into it; suppose it is Box m. And suppose these two numbers are $n_1 = 2^k \cdot m$ and $n_2 = 2^\ell \cdot m$, and let's assume the second one is the larger one, meaning $\ell > k$. Then we have now found two integers where one divides the other; in particular n_1 divides n_2, because:

$$\frac{n_2}{n_1} = \frac{2^\ell \cdot m}{2^k \cdot m} = 2^{\ell - k}.$$

This completes the proof.■

- **Another pigeonhole example**

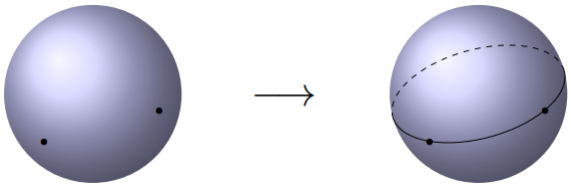
  Proposition. Suppose G is a graph with $n \geqslant 2$ vertices. Then G contains two vertices which have the same degree.

  We start by observing that the minimum degree is zero, and the maxmium is $n - 1$. It could happen that a vertex is connected to no other vertices, and a vertex could be connected to all other vertices. If a vertex is connected to all other vertices, than it has degree $n - 1$, because it has an edge going to all vertices but itself. Thus, we have our boxes. But you may notice that we have n boxes for n vertices. This may seem like a problem, but after some thought you may see that it is not possible for the zero box and the $n - 1$ box to both be used for a specific graph G. Thus, we have only $n - 1$ boxes for n vertices.

  The rest of the proof is left as an exercise for the reader.

- **Classic Geometry Theorem**. Given any two points on the sphere, there is a great circle that passes through those two points.

  Given a sphere, there are infinitely many ways to cut it in half, and each of these paths of the knife is called a great circle



- **Final pigeonhole example**

  Proposition. If you draw five points on the surface of an orange in marker, then there is always a way to cut the orange in half so that four points (or some part of the point) all lie on one of the halves.

Proof. Consider an orange with five points drawn on it. Pick any two of these points, and call them p and q. By the Classic Geometry Theorem, there exists a great circle passing through these points; angle your knife to cut along this great circle. Because the points are drawn in marker, they are wide enough so that part of these two points appear on both halves.

Now consider the remaining three points and the two halves that you just cut the orange into. Consider these three points to be objects and the halves to be boxes; by the simple form of the pigeonhole principal, at least two of these three points are on the same orange half. These two, as well a portion of p and of q, give four points or partial points, as desired ■

1.2 Direct proofs

- **Fact about integers**: The sum of integers is an integer, the difference of integers is an integer, and the product of integers is an integer. Also, every integer is either even or odd.

  We are calling these facts because, while they are true and one could prove them, we will not be proving them here

- **Even and odd integers**: An integer n is *even* if $n = 2k$ for some integer k

  An integer n is *odd* if $n = 2k + 1$ for some integer k

- **Sum of two even integers**

  Proposition. The sum of two even integers is even

  Proof. Assume n and m are even integers, then $n = 2a$, and $m = 2b$ for some integers a and b. Furthermore,

  $$n + m = 2a + 2b = 2(a + b).$$

  Since the sum of two integers is itself an integer, then we have two times an integer, which satisfies the definition of an even number. Hence, the sum $n + m$ is even, where n and m are even. $\int$

- **More on propositions**: We can rewrite our propositions to take the form

  > if *statement* is true, then *other statement* is also true

  For example,

  > if m and n are even, then $m + n$ is also even

  Another way to summarize such statements is this:

  > *some statement* is true implies *some other statement* is true.

  Which allows us to use the implies symbol $\implies$. For example,

  > m and n being even $\implies$ $m + n$ is even

  We have the general form $P \implies Q$, where P and Q are statements

  However, when writing formally, like when writing up the final draft of your homework, these symbols are rarely used. You should write out solutions with words, complete sentences, and proper grammar. Pick up any of your math textbooks, or look online at math research articles, and you will find that such practices are standard.

- **The structure of direct proofs**: A direct proof is a way to prove a "$P \Rightarrow Q$" proposition by starting with P and working your way to Q. The "working your way to Q" stage often involves applying definitions, previous results, algebra, logic, and techniques. Here is the general structure of a direct proof:

Proposition. $P \implies Q$

Proof. Assume P

Explain what P means by applying definitions and/or other results

$\vdots$ Apply algebra,

$\vdots$ logic techniques.

Hey look, that's what Q means

Therefore Q ■

- **Proof by cases**: A related proof strategy is proof by cases. This is a "divide and conquer" strategy where one breaks up their work into two or more cases

The below example of proof by cases will also give us more practice with direct proofs involving definitions. Indeed, when you break up a problem in two parts, those two parts still need to be proven, and a direct proof is often the way to tackle each of those parts

Proposition. If n is an integer, then $n^2 + n + 6$ is even.

Proof. Assume n is an integer, then either n is even or it is odd.

Case I. Assume n is even, then $n = 2m$ for some integer m. Thus, we have

$$\begin{aligned} n^2 + n + 6 &= (2m)^2 + 2m + 6 \\ &= 4m^2 + 2m + 6 \\ &= 2(2m^2 + m + 3). \end{aligned}$$

Observe that $2m^2 + m + 3 \in \mathbb{Z}$. Thus, we have two times an integer, which satisfies the definition of an even number.

Case 2. Assume n is odd, then $n = 2m + 1$ for some integer m. Thus,

$$\begin{aligned} n^2 + n + 6 &= (2m + 1)^2 + 2m + 1 + 6 \\ &= 4m^2 + 4m + 1 + 2m + 7 \\ &= 4m^2 + 6m + 8 \\ &= 2(2m^2 + 3m + 4). \end{aligned}$$

Since m is an integer, $2m^2 + 3m + 4$ is an integer, and we again have two times an integer, which is an even integer.

We have shown that $n^2 + n + 6$ is even whether n is even or odd. Combined, this shows that $n^2 + n + 6$ is even for all integers n ■

- **Proof by exhaustion (brute force proof)**: A proof by cases cuts up the possibilities into more manageable chunks. If the theorem refers to a collection of elements and your proof is simply checking each element individually, then it is called a *proof by exhaustion* or a *brute force proof*

- **Divisibility**: An integer a is said to divide an integer b if $b = ak$ for some integer k. When a does divide b, we write $a \mid b$, and when a does not divide b, we write $a \nmid b$.

  Note: A common mistake is to see something like "$2 \mid 8$" and think that this equals 4. The expression "$a \mid b$" is either true or false

  Remark. $a \mid 0$ for any integer a, because $0 = a \cdot 0$ for every such a

  $0 \nmid b$ for any nonzero integer b, because for any such b, we have $b \neq 0 \cdot k$ for any integer k

- **The transitive property of divisibility**:

  Proposition. Let a, b, and c be integers, if $a \mid b$ and $b \mid c$, then $a \mid c$

  Proof. Assume a, b, and c are integers. Further assume that $a \mid b$, and $b \mid c$

  > By the definition of divisibility, $a \mid b$ and $b \mid c$ implies $b = ak$ for some integer k, and $c = bs$ for some integer s
  >
  > If $a \mid c$, we require that $c = ar$ for some integer r

  $$\begin{aligned} & b = ak \\ \implies & c = (ak)s \\ \implies & c = a(ks). \end{aligned}$$

  Since k and s are integers, then their product ks is itself an integer. Let $r = ks$. Then $c = ar$, which is precisely the definition of divisiblity, and we conclude that $a \mid c$. ■

- **The division algorithm**:

  Theorem. For all integers a and m with $m > 0$, there exist unique integers q and r such that

  $$a = mq + r.$$

  Where $0 \leqslant r < m$. We call q the *quotient* and r the *remainder*

- **Common divisor, greatest common divisor**: Let a and b be integers. If $c \mid a$ and $c \mid b$, then c is said to be a common divisor of a and b.

  The greatest common divisor of a and b is the largest integer d such that $d \mid a$ and $d \mid b$. This number is denoted $\gcd(a, b)$.

Note that there is one pair of integers that does not have a greatest common divisor; if $a = 0$ and $b = 0$, then every positive integer d is a common divisor of a and b. This means that no divisor is the greatest divisor, since you can always find a bigger one. Thus, in this one case, $gcd(a, b)$ does not exist

- **Bezout's identity**: If a and b are positive integers, then there exist integers k and ℓ such that

$$\gcd(a, b) = ak + b\ell.$$

As an example, suppose $a = 12$ and $b = 20$, then $\gcd(12, 20) = 4$, and we have

$$\begin{aligned} 4 &= 12k + 20\ell \\ \implies \ell &= \frac{1}{5} - \frac{3}{5}k. \end{aligned}$$

Let $k = 2$, then we see $\ell = -1$. We see that there are infinitely many solutions, $k = 2, \ell = -1$ is just one of them. Nevertheless, this theorem simply says that at least one solution must exist.

Proof. Assume a and b are fixed positive integers, notice that the expression $ax + by$ can take many values for integers x and y. Let d be the *smallest positive integer* that $ax + by$ can be equal. Let k and ℓ be the x and y that obtain this d. That is,

$$d = ak + b\ell.$$

We now must show that d is a common divisor of a and b, and then that it is the *greatest common divisor*

Part 1 (common divisor). d is a common divisor of a and b if $d \mid a$ and $d \mid b$. To see that $d \mid a$, we examine the division algorithm. We know that there exsits unique integers q and r such that

$$a = dq + r.$$

With $0 \leqslant r < d$. We have

$$\begin{aligned} r &= a - dq \\ &= a - (ak + b\ell)q \\ &= a - akq - b\ell q \\ &= a(1 - kq) + b(-\ell q). \end{aligned}$$

Observe that $1 - kq$, and $-\ell q$ are both integers, Since r is written in the form $ax + by$, $0 \leqslant r < d$, and d is the smallest positive integer that this form can produce (with the given a, b), it must be that $r = 0$. Thus,

$$a = dq + 0 = dq.$$

And we see that $d \mid a$. A similar argument will show that $d \mid b$ as well. This proves that d is a common divisor of a and b.

Part 2 (gcd). Assume that d' is some other common divisor of a and b. We must show that $d' \leqslant d$. If d' is a common divisor of a and b, then $d' \mid a$ and $d' \mid b$, which implies $a = d'n$, and $b = d'm$, for some integers n and m. If $d = ak + b\ell$, then

$$\begin{aligned} d &= d'nk + d'm\ell \\ &= d'(nk + m\ell) \\ \implies d' &= \frac{d}{nk + m\ell}. \end{aligned}$$

Since $n, k, m, \ell \in \mathbb{Z}$, it follows that $nk + m\ell \in \mathbb{Z}$. Thus, $d' \leqslant d$.

Therefore, we have shown that d is not only a common divisor of a and b, but that it is also the largest, and hence the *gcd*. Thus,

$$\gcd(a, b) = d = ak + b\ell.$$

■

A corollary from this result is that $\gcd(ma, mb) = m \gcd(a, b)$. If $\gcd(a, b) = ak + b\ell$, we have

$$\begin{aligned} \gcd(ma, mb) &= mak + mb\ell \\ &= m(ak + b\ell) \\ &= m \gcd(a, b). \end{aligned}$$

- **Modulo and congruence**: For integers a, r, and m, we say that a is congruent to r modulo m and we write $a \equiv r \pmod{m}$ if $m \mid (a - r)$.

  For example, $18 \equiv 4 \pmod{7}$ because $18 = 7(2) + 4$, we see that $7 \mid (18 - 4)$

  If a divided by m leaves a remainder of r, then $a \equiv r \pmod{m}$. However, this is not the only way to have $a \equiv r \pmod{m}$ — it is not required that r be the remainder when a is divided by m; all that is required is that a and r have the same remainder when divided by m. For example:

  $$18 = 11 \pmod{7}.$$

- **Properties of modular congruence**: Assume that a, b, c, d and m are integers, $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$. Then

  (i) $a + c \equiv b + d \pmod{m}$
  (ii) $a - c \equiv b - d \pmod{m}$
  (iii) $a \cdot c \equiv b \cdot d \pmod{m}$

  Proof of property *i*. Assume that $a \equiv b \pmod{m}$, and $c \equiv d \pmod{m}$, we must show that $a + c \equiv b + d \pmod{m}$

  If $a \equiv b \pmod{m}$, then $m \mid a - b$, which implies $a - b = mk$ for some $k \in \mathbb{Z}$. Similarly, $c \equiv d \pmod{m} \implies m \mid c - d \implies c - d = m\ell$, for some $\ell \in \mathbb{Z}$. Adding these two equations yields

  $$\begin{aligned} &(a - b) + (c - d) = mk + m\ell \\ \implies &(a + c) - (b + d) = m(k + \ell). \end{aligned}$$

Since $k + \ell \in \mathbb{Z}$, then by the definition of divisibility

$$m \mid (a + c) - (b + d).$$

Which then by the definition of congruence

$$a + c \equiv b + d \pmod{m}.$$

■

Proof of property *iii*. Assume $a \equiv b \pmod{m}$, and $c \equiv d \pmod{m}$

From above we know it follows that $a - b = mk$, and $c - d = m\ell$, for $k, \ell \in \mathbb{Z}$. If $ac \equiv bd \pmod{m}$, it must be that $ac - bd = ms$, for some $s \in \mathbb{Z}$. Let's see if we can derive $ac - bd$ in terms of what we know, namely $a - b$ and $c - d$. Amazingly,

$$\begin{aligned} ac - bd &= (a - b)c + (c - d)b \\ &= mkc + m\ell b \\ &= m(kc + \ell b). \end{aligned}$$

It then follows that

$$m \mid ac - bd.$$

Thus,

$$ac \equiv bd \pmod{m}.$$

■

- **Prime and composite integers**: An integer $p \geqslant 2$ is prime if its only positive divisors are 1 and p. An integer $n \geqslant 2$ is composite if it is not prime. Equivalently, n is composite if it can be written as $n = st$, where s and t are integers and $1 < s, t < n$.

  Note: To be clear, "$1 < s, t < n$" means that both s and t are between 1 and n.

- **Properties of primes and divisibility**:

  Lemma. Let a, b and c be integers, and let p be a prime:

  (i) If $p \nmid a$, then $\gcd(p, a) = 1$.

  (ii) If $a \mid bc$ and $\gcd(a, b) = 1$, then $a \mid c$.

  (iii) If $p \mid bc$, then $p \mid b$ or $p \mid c$ (or both).

  Proof of property *i*. Assume that p does not divide a, then p cannot possibly be a common divisor of a and p, because it is not a divisor of a.

  Since $p \in \mathbb{P}$[1], then the only divisors of p are one and itself. Thus, the only option left is one. Hence, the greatest common divisor is one. ■

[1] Where $\mathbb{P}$ is the family of primes

***Proof of property** ii.* Assume $a \mid bc$, and $\gcd(a,b) = 1$. Then, $bc = ar$ for some integer r, and by Bezout's identity, there exist some integers k, ℓ such that

$$\begin{aligned} \gcd(a,b) &= ak + b\ell \\ \implies 1 &= ak + b\ell. \end{aligned}$$

If $a \mid c$, we require $c = as$, for some integer s. If we multiply the above expression by c, we get

$$c = cak + cb\ell.$$

Since we assumed $a \mid bc$, then it must be that $bc = ar$, for $r \in \mathbb{Z}$. Thus, we have

$$\begin{aligned} c &= cak + ar\ell \\ &= a(ck + r\ell). \end{aligned}$$

Since $c, k, r, \ell \in \mathbb{Z}$, the expression $ck + r\ell$ is also an integer, and by the definition of divisibility, it must be that $a \mid c$ ■

***Proof of property** iii.* Assume that $p \mid bc$. Then there are two cases, either $p \mid b$, or $p \nmid b$.

Case I. If $p \mid b$, then the statement is true and we are done

Case II. If $p \nmid b$, then by property i, it must be that $\gcd(p,b) = 1$. By property ii, if $p \mid bc$, and $\gcd(p,b) = 1$, then it must be that $p \mid c$. ■.

- **More on properties of congruence**: We return to congruence to examine the statement

$$ak \equiv bk \pmod{m} \stackrel{?}{\implies} a \equiv b \pmod{m}.$$

Proposition (*modular cancellation law*). Let a, b, k, m be integers. If $ak \equiv bk \pmod{m}$, and $\gcd(m,k) = 1$, then $a \equiv b \pmod{m}$

Proof. Assume $ak \equiv bk \pmod{m}$, and $\gcd(m,k) = 1$, then $m \mid ak - bk$, and $ak - bk = m\ell$, for some integer ℓ.

If $a \equiv b \pmod{m}$, then $m \mid a - b$, and $a - b = mr$, for some integer r. Since $ak \equiv bk \pmod{m}$, then it must be that

$$\begin{aligned} ak - bk &= m\ell \\ \implies k(a-b) &= m\ell \\ \implies a - b &= \frac{m\ell}{k}. \end{aligned}$$

Thus, we require $\frac{\ell}{k}$ to be an integer, it then follows that the proposition holds true.

We know that if $a \mid bc$, and $\gcd(a,b) = 1$, then $a \mid c$. Thus, since $k \mid m\ell$, and $\gcd(m,k) = 1$, it must be that $k \mid \ell$. Hence, $\frac{\ell}{k} \in \mathbb{Z}$, and

$$a - b = m\left(\frac{\ell}{k}\right).$$

And by the definition of divisibility, $m \mid a - b$, which implies $a \equiv b \pmod{m}$ ■.

- **Fermat's little theorem**: If a is an integer and p is a prime which does not divide a, then

$$a^{p-1} \equiv 1 \pmod{p}.$$

Proof. Assume that a is an integer and p is a prime which does not divide a. We begin by proving that when taken modulo p,

$$\{a, 2a, 3a, \ldots, (p-1)a\} \equiv \{1, 2, 3, \ldots, p-1\}.$$

To do this, observe that the set on the right has every residue modulo p except 0, and each such residue appears exactly once. Therefore, since both sets have $p-1$ elements listed, in order to prove that the left set is the same as the right set, it suffices to prove this:

1. No element in the left set is congruent to 0, and
2. Each element in the left set appears exactly once.

In doing so, we will twice use the modular cancellation law (Proposition 2.18) to cancel out an a, and so we note at the start that by Lemma 2.17 part (i) we have $\gcd(p, a) = 1$.

Step 1. First we show that none of the terms in $\{a, 2a, 3a, \ldots, (p-1)a\}$, when considered modulo p, are congruent to 0. To do this, we will consider an arbitrary term ia, where i is anything in $\{1, 2, 3, \ldots, p-1\}$. Indeed, if we did have some

$$ia \equiv 0 \pmod{p},$$

which is equivalent to

$$ia \equiv 0a \pmod{p},$$

then by the modular cancellation law (Proposition 2.18) we would have

$$i \equiv 0 \pmod{p}.$$

That is, in order to have $ia \equiv 0 \pmod{p}$, that would have to have $i \equiv 0 \pmod{p}$. Therefore we are done with Step 1, since no i from $\{1, 2, 3, \ldots, p-1\}$ is congruent to 0 modulo p.

Step 2. Next we show that every term in $\{a, 2a, 3a, \ldots, (p-1)a\}$, when considered modulo p, does not appear more than once in that set. Indeed, if we did have

$$ia \equiv ja \pmod{p},$$

for i and j from $\{1, 2, 3, \ldots, p-1\}$, then by the modular cancellation law (Proposition 2.18) we have

$$i \equiv j \pmod{p}.$$

And since i and j are both from the set $\{1, 2, 3, \ldots, p-1\}$, this means that $i = j$. In other words, each term in $\{a, 2a, 3a, \ldots, (p-1)a\}$ is not congruent to any other term from that set — it is only congruent to itself. This completes Step 2.

We have succeeded in proving that when taken modulo p,

$$\{a, 2a, 3a, \ldots, (p-1)a\} \equiv \{1, 2, 3, \ldots, p-1\},$$

even though the numbers in these sets may be in a different order. But since the order does not matter when multiplying numbers, we see that

$$a \cdot 2a \cdot 3a \cdot 4a \cdot \cdots \cdot (p-1)a \equiv 1 \cdot 2 \cdot 3 \cdot 4 \cdot \cdots \cdot (p-1) \pmod{p}.$$

Then, since $\gcd(2, p) = 1$ by Lemma 2.17 part (i), by the modular cancellation law (Proposition 2.18) we may cancel a 2 from both sides:

$$a \cdot 3a \cdot 4a \cdot \cdots \cdot (p-1)a \equiv 1 \cdot 3 \cdot 4 \cdot \cdots \cdot (p-1) \pmod{p}.$$

Then, since $\gcd(3, p) = 1$ by Lemma 2.17 part (i), by the modular cancellation law (Proposition 2.18) we may cancel a 3 from both sides:

$$a \cdot a \cdot 4a \cdot \cdots \cdot (p-1)a \equiv 1 \cdot 4 \cdot \cdots \cdot (p-1) \pmod{p}.$$

Continuing to do this for the $4, 5, \ldots, (p-1)$ on each side (each of which has a greatest common divisor of 1 with p, by Lemma 2.17 part (i)), by the modular cancellation law (Proposition 2.18) we obtain

$$\underbrace{a \cdot a \cdot a \cdot \cdots \cdot a}_{p-1 \text{ copies}} \equiv 1 \pmod{p},$$

which is equivalent to what we sought to prove:

$$a^{p-1} \equiv 1 \pmod{p}.$$

- **Bonus proof**:

  Proposition. If x and y are positive integers, and $x \geqslant y$, then $\sqrt{x} \geqslant \sqrt{y}$

  Proof. Assume x and y are positive integers, and $x \geqslant y$. Then

  $$\begin{aligned} & x \geqslant y \\ \implies & x - y \geqslant 0 \\ & . \end{aligned}$$

  Since $x, y \geqslant 0$, $\sqrt{x^2} = |x| = x$, and $\sqrt{y^2} = |y| = y$. Thus,

  $$\begin{aligned} & x - y \geqslant 0 \\ \implies & \sqrt{x^2} - \sqrt{y^2} \geqslant 0 \\ \implies & (\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y}) \geqslant 0 \\ \implies & \sqrt{x} - \sqrt{y} \geqslant 0 \qquad \blacksquare. \end{aligned}$$

- **The AM-GM inequality**:

  Theorem (*AM-GM inequality*). If $x, y \geqslant 0 \in \mathbb{Z}$, then $\sqrt{xy} \leqslant \frac{x+y}{2}$

Proof. Assume $x, y \geqslant 0 \in \mathbb{Z}$. Consider

$$0 \leqslant (x-y)^2.$$

Which we know to be true, squaring an integer is always positive, and we know $x - y$ to be an integer. It then follows that

$$0 \leqslant x^2 - 2xy + y^2.$$

If we add $4xy$ to both sides, we get

$$\begin{aligned} & 4xy \leqslant x^2 + 2xy + y^2 \\ \implies & 4xy \leqslant (x+y)^2 \\ & . \end{aligned}$$

Now let's take the square root of both sides

$$2\sqrt{xy} \leqslant |x+y|.$$

Since $x, y \geqslant 0$, $|x+y| = x+y$. Thus,

$$\begin{aligned} & 2\sqrt{xy} \leqslant x+y \\ \therefore & \sqrt{xy} \leqslant \frac{x+y}{2}. \end{aligned}$$

Note: Some of the steps taken in this proof may seem a bit random, but if we start at the proposition $\sqrt{xy} \leqslant \frac{x+y}{2}$ and work backwards algebraically, we see

$$\begin{aligned} \sqrt{xy} &\leqslant \frac{x+y}{2} \\ 2\sqrt{xy} &\leqslant x+y \\ 4xy &\leqslant (x+y)^2 \\ 4xy &\leqslant x^2 + 2xy + y^2 \\ 0 &\leqslant x^2 + 2xy + y^2 - 4xy \\ 0 &\leqslant x^2 - 2xy + y^2 \\ 0 &\leqslant (x-y)^2. \end{aligned}$$

We see that we have derived a starting point, and were just working backwards in the proof.

1.3 Sets

- **Vacuous truth**: a vacuous truth is a conditional or universal statement (a universal statement that can be converted to a conditional statement) that is true because the antecedent cannot be satisfied.[1] It is sometimes said that a statement is vacuously true because it does not really say anything. For example, the statement "all cell phones in the room are turned off" will be true when no cell phones are present in the room. In this case, the statement "all cell phones in the room are turned on" would also be vacuously true, as would the conjunction of the two: "all cell phones in the room are turned on and turned off", which would otherwise be incoherent and false.
- **Review: Proper subset**: If $A = B$, then $A \subseteq B$. In the case that $A \subseteq B$ and $A \neq B$, we say that A is a proper subset of B. the correct notation for this is "$A \subset B$."
- **Proving** $A \subseteq B$

  Definition. Suppose A and B are sets. If every element in A is also an element of B, then A is a subset of B, which is denoted $A \subseteq B$

  Note: For every set B, it is true that $\varnothing \subseteq B$. To see it, first note that, because there are no elements in $\varnothing$, it would be true to say "for any $x \in \varnothing$, x is a purple elephant that speaks German." It's vacuously[2] true! You certainly can't disprove it, right? You can't present to me any element in $\varnothing$ that is not a purple elephant that speaks German.

  By this reasoning, I could switch out "is a purple elephant that speaks German" for any other statement, and it would still be true! And this includes the subset criteria: if $x \in \varnothing$, then $x \in B$, which by definition means that $\varnothing \subseteq B$. Again, you certainly can not present to me any $x \in \varnothing$ which is not also an element of B, can you?

  in order to prove that $A \subseteq B$, what we would have to show is this:

  $$\text{If } x \in A, \text{ then } x \in B.$$

  In other words, for any arbitrary element in A, that same element is also in B

  Proposition. It is the case that

  $$\{n \in \mathbb{Z} : \ 12 \mid n\} \subseteq \{n \in \mathbb{Z} : \ 3 \mid n\}.$$

  Proof. Let $A = \{n \in \mathbb{Z} : \ 12 \mid n\}$, and $B = \{n \in \mathbb{Z} : \ 3 \mid n\}$. Assume $a \in A$

  Since $a \in A$, then $12 \mid a$, which implies $a = 12k$, for some $k \in \mathbb{Z}$. If $a \in B$, then $3 \mid a \implies a = 3\ell$

  Since $a = 12k$, and $a = 3\ell$, then $12k = 3\ell \implies \ell = 4k$. Thus, we have

  $$a = 3(4k).$$

  Which by the definition of divisiblity, and since $4k \in \mathbb{Z}$, we have $3 \mid a$.

  Therefore, $a \in B$ ■

[2] A statement is vacuously true if it asserts something about all elements of the empty set.

- **Proving** $A = B$: Recall that, for sets A and B, to say that "$A = B$" is to say that these two sets contain *exactly* the same elements. Said differently, it means these two things:

  1. Every element in A is also in B (which means $A \subseteq B$), and
  2. Every element in B is also in A (which means $B \subseteq A$).

  Indeed, a slick way to prove that $A = B$ is to prove both $A \subseteq B$ and $B \subseteq A$, both of which can be done using the approach discussed above.

- **Review of set operations**:

  - The *union* of sets A and B is the set $A \cup B = \{x : x \in A \text{ or } x \in B\}$.
  - The *intersection* of sets A and B is the set $A \cap B = \{x : x \in A \text{ and } x \in B\}$.
  - Likewise, if $A_1, A_2, A_3, \ldots, A_n$ are all sets, then the union of all of them is the set
    $$A_1 \cup A_2 \cup \cdots \cup A_n = \{x : x \in A_i \text{ for some } i\}.$$
    This set is also denoted
    $$\bigcup_{i=1}^{n} A_i.$$
  - Likewise, if $A_1, A_2, A_3, \ldots, A_n$ are all sets, then the intersection of all of them is the set
    $$A_1 \cap A_2 \cap \cdots \cap A_n = \{x : x \in A_i \text{ for all } i\}.$$
    This set is also denoted
    $$\bigcap_{i=1}^{n} A_i.$$

  Assume A and B are sets and "$x \notin B$" means that x is not an element of B.

  - The *subtraction* of B from A is $A \setminus B = \{x : x \in A \text{ and } x \notin B\}$.
  - If $A \subseteq U$, then U is called a *universal set* of A. The *complement* of A in U is $A^c = U \setminus A$.

  Furthermore,

  - The *power set* of a set A is $\mathcal{P}(A) = \{X : X \subseteq A\}$.
  - The *cardinality* of a set A is the number of elements in the set, and it is denoted $|A|$.

  Assume A and B are sets, The Cartesian product of A and B is

  $$A \times B = \{(a, b) : \ a \in Aandb \in B\}..$$

- **More on power sets**:

  Proposition. Suppose A and B are sets. If $\mathcal{P}(A) \subseteq \mathcal{P}(B)$, then $A \subseteq B$.

  Proof. Assume A and B are sets, and $\mathcal{P}(A) \subseteq \mathcal{P}(B)$.

  > Choose $x \in \mathcal{P}(A)$, which means $x \subseteq A$. Since $\mathcal{P}(A) \subseteq \mathcal{P}(B)$, it follows that $x \in \mathcal{P}(B)$, which means $x \subseteq B$. Let $x = A$, since $A \in \mathcal{P}(A)$. Since $x \subseteq B$, then $A \subseteq B$

  Therefore, $A \subseteq B$ ■

- **De Morgan's law**:

  Theorem. Suppose A and B are subsets of a universal set U. Then,

  $$(A \cup B)^C = A^C \cap B^C. \tag{1}$$

  And

  $$(A \cap B)^C = A^C \cup B^C. \tag{2}$$

  Proof (1). Assume A and B are subsets of a universal set U, since $(A \cup B)^C$, and $A^C \cap B^C$ are sets, we show equality by showing $(A \cup B)^C \subseteq A^C \cap B^C$, and $A^C \cap B^C \subseteq (A \cup B)^C$. It then follows that $(A \cup B)^C = A^C \cap B^C$

  > Choose $x \in (A \cup B)^C$, by the definition of the complement, we have $x \notin (A \cup B)$, which by the definition of the union means x cannot be in A, and it cannot be in B. In other words, $x \notin A$ and $x \notin B \implies x \in A^C$ and $x \in B^C$. Therefore,
  >
  > $$x \in A^C \cap B^C.$$
  >
  > Which by the definition of the subset, means $(A \cup B)^C \subseteq A^C \cap B^C$
  >
  > Next, let $x \in A^C \cap B^C$, then $x \in A^C$ and $x \in B^C$, which means $x \notin A$ and $x \notin B$, which implies $x \notin (A \cup B) \implies x \in (A \cup B)^C$.
  >
  > Therefore, since $x \in A^C \cap B^C \implies x \in (A \cup B)^C$, by the definition of a subset, we have $A^C \cap B^C \subseteq (A \cup B)^C$

  Since both $(A \cup B)^C \subseteq A^C \cap B^C$, and $A^C \cap B^C \subseteq (A \cup B)^C$, it must be the case that $(A \cup B)^C = A^C \cap B^C$ ■

  It should be addressed that this proof can be done by simply manipulating the set builder notation. We have

  $$\begin{aligned} A^C \cap B^C &= \{x \in \mathbb{R} : \ x \in A^C \text{ and } x \in B^C\} \\ &= \{x \in \mathbb{R} : \ x \notin A \text{ and } x \notin B\} \\ &= \{x \in \mathbb{R} : \ x \notin (A \cup B)\} \\ &= \{x \in \mathbb{R} : \ x \in (A \cup B)^C\}. \end{aligned}$$

  ■

- **Proving** $a \in A$: Consider the set $\{x \in S : \ P(x)\}$, where $P(x)$ is some condition on x

Given a set of this form, if you are presented with a specific a and you wish to prove that $a \in A$, then you must show that

1. $a \in S$
2. $P(a)$ is true

For example, Let $A = \{(x, y) \in \mathbb{Z} \times \mathbb{N} : \ x \equiv y \pmod 5\}$, then $(17, 2) \in A$

Proof. First, note that $(17, 2) \in \mathbb{Z} \times \mathbb{N}$ because $17 \in \mathbb{Z}$, and $2 \in \mathbb{N}$, Next, observe that

$$17 \equiv 2 \pmod 5.$$

Because $5 \mid (17 - 2)$

- **Indexed Families of Sets**: Consider a set $\mathcal{F}$, If every element of $\mathcal{F}$ is itself a set, then $\mathcal{F}$ is called a *family of sets*. Then, one can ask questions about such a family, — like, what is the union of all of the sets in $\mathcal{F}$. That is,

$$\bigcup_{S \in \mathcal{F}} S = \{x : \ x \in S \text{ for some } S \in \mathcal{F}\}.$$

Likewise,

$$\bigcap_{S \in \mathcal{F}} S = \{x : \ x \in S \text{ for every } S \in \mathcal{F}\}.$$

- **Bonus example I**.

Proposition. It is the case that

$$\{n \in \mathbb{Z} : \ 12 \mid n\} = \{n \in \mathbb{Z} : \ 3 \mid n\} \cap \{n \in \mathbb{Z} : \ 4 \mid n\}.$$

Proof. Let $A = \{n \in \mathbb{Z} : \ 12 \mid n\}$, $B = \{n \in \mathbb{Z} : \ 3 \mid n\}$, and $C = \{n \in \mathbb{Z} : \ 4 \mid n\}$

> *Part i.)* Choose $x \in A$, we then have $12 \mid x$, and $x = 12k$, for some $k \in \mathbb{Z}$. Thus,
>
> $$x = 12k = 3(4k) = 4(3k).$$
>
> Which by the definition of divisibility implies both $3 \mid x$ and $4 \mid x$, since both $4k$ and $3k \in \mathbb{Z}$. Hence, $x \in B \cap C$
>
> *Part ii.)* Choose $x \in B \cap C$, then both $x = 3r$ and $x = 4s$, for $r, s \in \mathbb{Z}$. We have
>
> $$3r = 4s.$$
>
> Which implies $3 \mid 4s$, since $r \in \mathbb{Z}$. Because $3 \in \mathbb{P}$, we know that either $3 \mid 4$ or $3 \mid s$. Since it is clear that $3 \nmid 4$, it must be the case that $3 \mid s$, and thus $s = 3\ell$ for an integer ℓ. It then follows that
>
> $$x = 4s = 4(3\ell) = 12\ell.$$
>
> Which by the definition of divisibility implies $12 \mid x$, and thus $x \in A$

Since choosing an $x \in A \implies x \in B \cap C$, it must be that $A \subseteq B \cap C$, and choosing an $x \in B \cap C \implies x \in A$, it must also be that $B \cap C \subseteq A$. With these two facts, we can assert that $A = B \cap C$ ■

- **The Cardinality of the Power Set**: Suppose A is a set with n elements. How many subsets of A are there? Said differently, what is $|P(A)|$?

We could check the first few cases by hand

A	$\lvert A\rvert = n$	$\lvert\mathcal{P}(A)\rvert$
$\{1\}$	1	2
$\{1,2\}$	2	4
$\{1,2,3\}$	3	8
$\{1,2,3,4\}$	4	16

It sure looks like if $|A| = n$, then $|P(A)| = 2^n$. Why would this be true? There is actually a pretty slick way to see it. Every subset of $\{1, 2, 3\}$ can be thought of by asking whether or not each element is included in the subset. For example, $\{1, 3\}$ can be thought of as $\langle$yes, no, yes$\rangle$, since 1 was included, 2 was not, and 3 was.

Suppose you're trying to generate a subset of $\{1, 2, 3\}$. You could think about doing so by asking three yes/no questions, the answers to which uniquely determine your set. With 2 options for the first element, 2 for the second, and 2 for the third, in total there are $2 \times 2 \times 2 = 8$ ways to answer the three questions, and hence 8 subsets!

With n straight yes/no questions, there are $2 \times 2 \times \cdots \times 2 = 2^n$ ways to answer the questions, each corresponding uniquely to a subset of A. Thus, if $|A| = n$, then $|P(A)| = 2^n$.

- **A consequence of the above fact**:

Proposition. Given any $A \subseteq \{1, 2, 3, \ldots, 100\}$ for which $|A| = 10$, there exist two different subsets $X \subseteq A$ and $Y \subseteq A$ for which the sum of the elements in X is equal to the sum of the elements in Y.

For example, consider the set $\{6, 23, 30, 39, 44, 46, 62, 73, 90, 91\}$, If we let

$$X = \{6, 23, 46, 73, 90\} \text{ and } Y = \{30, 44, 73, 91\}.$$

then the elements in both sets sum to 238:

Proof. We prove this fact using the pigeonhole principal. Consider the smallest and largest possible subset sums. If $A = \varnothing \subseteq \{1, 2, 3, ..., 100\}$, then the sum is 0. If $A = \{91, 92, 93, 94, 95, 96, 97, 98, 99, 100\}$, then the subset sum is 955. Thus, there are no more than 956 possible subset sums for the set $A \subseteq \{1, 2, 3, ..., 100\}$, for which $|A| = 10$.

Consider 956 boxes, each representing a unique subset sum. Since we have $2^{|A|} = 2^{10} = 1024$ subsets and only 956 boxes to place each subset in, there must be a box containing two subsets A, which means they must have the same sum ■.

- **The symmetric difference of sets**. The *symmetric difference* of two sets A and B, denoted $A\Delta B$, or $A \ominus B$, is the set which contains the elements which are either in set A or in set B but not in both

1.4 Induction

- **Dominoes**: Consider a line of dominoes, perfectly arranged, just waiting to be knocked over. Dominoes stacked up like this have the following properties:

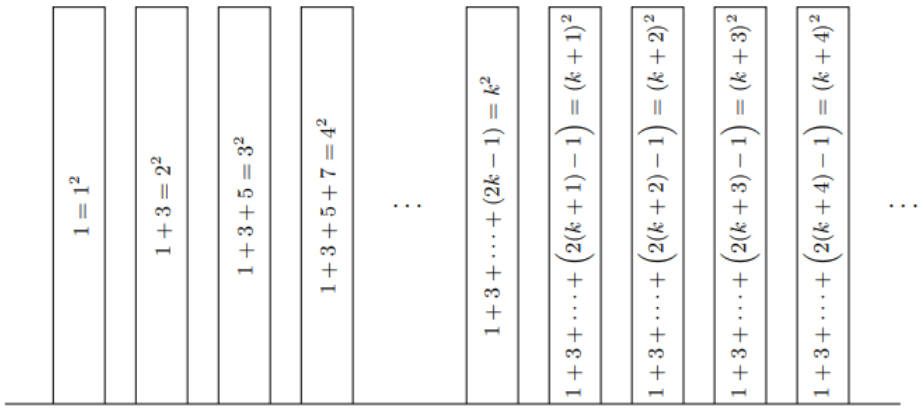
  1. If you give the first domino a push, it will fall (in particular, it will fall into the second domino, knocking it over).
  2. Moreover, every domino, when it's knocked over, falls into the next one and knocks it over.

  Given these two properties, it must be the case that if you knock over the first domino, then every domino will eventually fall. The first premise gets the process going, as it implies that the first domino will fall. And then the second premise keeps it going: Applying the second premise means that the falling first domino will cause the second domino to fall. Applying the second premise again means that the second falling domino will cause the third domino to fall. Applying the second premise again means that the third falling domino will cause the fourth domino to fall. And so on.

- **Sum of the first n odd numbers**: Take a look at the following

$$\begin{aligned} 1 &= 1 = 1^2 \\ 1+3 &= 4 = 2^2 \\ 1+3+5 &= 9 = 3^2 \\ 1+3+5+7 &= 16 = 4^2 \\ 1+3+5+7+9 &= 25 = 5^2 \\ 1+3+5+7+9+11 &= 36 = 6^2 \\ 1+3+5+7+9+11+13 &= 49 = 7^2. \end{aligned}$$

  It sure looks like the sum of the first n odd numbers is n^2. But how can we prove that it's true for every one of the infinitely many n? The trick is to use the domino idea. Imagine one domino for each of the above statements.



  Suppose we do the following:

  – Show that the first domino is true (this is trivial, since obviously $1 = 1^2$).

  – Show that any domino, if true, implies that the following domino is true too

Given these two, we may conclude that all the dominoes are true. It's exactly the same as noting that all the dominoes from earlier will fall. This is a slick way to prove infinitely many statements all at once, and it is called the *principal of mathematical induction*, or, when among friends, it is simply called *induction*.

- **Induction**: Consider a sequence of mathematical statements, $S_1, S_2, S_3, ...$
  - Suppose S_1 is true, and
  - Suppose, for each $k \in \mathbb{N}$, if S_k is true then S_{k+1} is true.

  Then, S_n is true for every $n \in \mathbb{N}$.

- **Induction framework**:

  Proposition. $S_1, S_2, S_3, ...$ are all true

  Proof. *General setup or assumptions if needed*

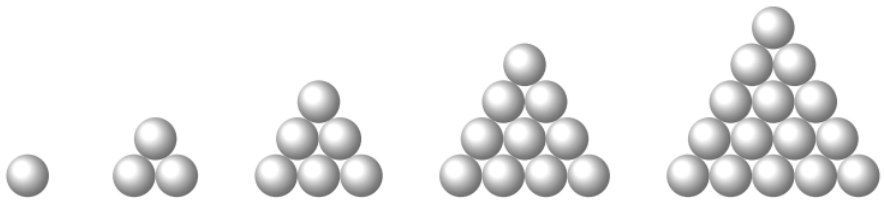
  Base case. $\langle\langle$Demonstration that S_1 is true$\rangle\rangle$

  Inductive hypothesis. Assume that S_k is true

  Induction step. $\langle\langle$Proof that S_k implies $S_{k+1}\rangle\rangle$

  Conclusion. Therefore, by induction, all the S_n are true. ■

- **Induction example 1**: Let's simply sum the first n natural numbers: $1+2+3+4+$ ůůů $+ n$. These sums are called the triangular numbers since they can be pictured as the number of balls in the following triangles.



Proposition. For any $n \in \mathbb{N}$, $\sum_{i=1}^{n} i = 1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$

Proof. We proceed by induction

Base case: The base case is when $n = 1$, and

$$1 = \frac{1(1+1)}{2} = 1.$$

Inductive hypothesis: Let $k \in \mathbb{N}$, assume

$$1 + 2 + 3 + ... + k = \frac{k(k+1)}{2}.$$

Inductive step: We aim to show that the result holds for $k + 1$. Thus,

$$1 + 2 + 3 + ... + k + k + 1 = \frac{(k+1)((k+1)+1)}{2}.$$

We have

$$\begin{aligned} 1 + 2 + 3 + ... + k + k + 1 &= \frac{(k+1)(k+2)}{2} \\ \implies \frac{k(k+1)}{2} + k + 1 &= \frac{(k+1)(k+2)}{2} \\ \implies \frac{k^2 + k + 2k + 1}{2} &= \frac{k^2 + 2k + k + 2}{2}. \end{aligned}$$

Therefore, by induction, $1 + 2 + 3 + ... + n = \frac{n(n+1)}{2}$ for all $n \in \mathbb{N}$ ■

- **Induction example 2**:

Proposition. Let S_n be the sum of the first n natural numbers. Then, for any $n \in \mathbb{N}$,

$$S_n + S_{n+1} = (n+1)^2.$$

We will prove this proposition twice. The first proof is a direct proof, the second will be by induction.

Direct proof. We have

$$\begin{aligned} S_n + S_{n+1} &= \frac{n(n+1)}{2} + \frac{(n+1)((n+1)+1)}{2} \\ &= \frac{n^2 + n}{2} + \frac{n^2 + 2n + n + 2}{2} \\ &= \frac{n^2 + n + n^2 + 3n + 2}{2} \\ &= \frac{2n^2 + 4n + 2}{2} \\ &= \frac{2(n^2 + 2n + 1)}{2} \\ &= n^2 + 2n + 1 \\ &= (n+1)^2 \quad \blacksquare. \end{aligned}$$

Proof by induction. We proceed by induction

Base case: The base case is when $n = 1$, and

$$S_1 + S_2 = 1 + 3 = 4 = (1 + 1)^2.$$

as desired

Inductive hypothesis. Let $k \in \mathbb{N}$, and assume that

$$S_k + S_{k+1} = (k + 1)^2.$$

Inductive step. We aim to prove that the result holds for $k + 1$. That is,

$$S_{k+1} + S_{k+2} = (k + 2)^2.$$

For this, we use the fact that S_{k+1} is the sum of the first $k + 1$ natural numbers, thus we can write it as $S_k + (k+1)$. Likewise, $S_{k+2} = S_{k+1} + (k + 2)$. Thus,

$$\begin{aligned} S_{k+1} + S_{k+2} &= S_k + (k + 1) + S_{k+1} + (k + 2) \\ &= S_k + S_{k+1} + 2k + 3 \\ &= (k + 1)^2 + 2k + 3 \\ &= k^2 + 2k + 1 + 2k + 3 \\ &= k^2 + 4k + 4 \\ &= (k + 2)^2. \end{aligned}$$

Conclusion. Therefore, by induction, the proposition holds for all $n \in \mathbb{N}$ ■

- **A quick note about induction**: For some proof techniques, adding a sentence at the end of your proof is nice but not required. For induction, though, it really is required. You can prove that the first domino will fall, and you can prove that each domino — if fallen— will knock over the next domino, but why does this mean they all fall? Because induction says so! Until you say "by induction. . . " your work will not officially prove the result

- **Induction example 3.**

Proposition. For every $n \in N$, the product of the first n odd natural numbers equals $\frac{(2n)!}{2^n n!}$. That is,

$$1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n - 1) = \frac{(2n)!}{2^n n!}.$$

Proof. We proceed by induction.

Base case: The base case occurs when $n = 1$,

$$1 = \frac{(2(1))!}{2^1 1!} = 1.$$

As desired

Inductive hypothesis. Let $k \in \mathbb{N}$, assume

$$1 \cdot 3 \cdot 5 \cdot ... \cdot (2k-1) = \frac{(2k)!}{2^k k!}.$$

Inductive step. We aim to prove that the result holds for $k+1$. Thus, we wish to show

$$\begin{aligned} 1 \cdot 3 \cdot 5 \cdot ... \cdot (2k-1) \cdot (2(k+1)-1) &= \frac{(2(k+1))!}{2^{k+1}(k+1)!} \\ &= \frac{(2k+2)!}{2^{k+1}(k+1)!}. \end{aligned}$$

By the inductive hypothesis, we have

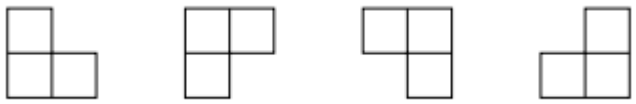
$$\begin{aligned} 1 \cdot 3 \cdot 5 \cdot ... \cdot (2k-1) \cdot (2k+1) &= \frac{(2k)!}{2^k k!}(2k+1) \\ &= \frac{(2k)!(2k+1)}{2^k k!} \\ &= \frac{(2k+1)!}{2^k k!} \\ &= \frac{(2k+1)!}{2^k k!} \cdot \frac{(2k+2)}{(2k+2)} \\ &= \frac{(2k+2)!}{2^k k!(2k+2)} \\ &= \frac{(2k+2)!}{2^k k! \cdot 2(k+1)} \\ &= \frac{(2k+2)!}{2^{k+1}(k+1)!}. \end{aligned}$$

Therefore, by induction, the proposition holds for all $n \in \mathbb{N}$ ■

- **Induction example 4**.

Proposition. For every $n \in \mathbb{N}$, if any one square is removed from a $2^n \times 2^n$ chessboard, the result can be perfectly covered with L-shaped tiles.

The tiles cover three squares and look like this:

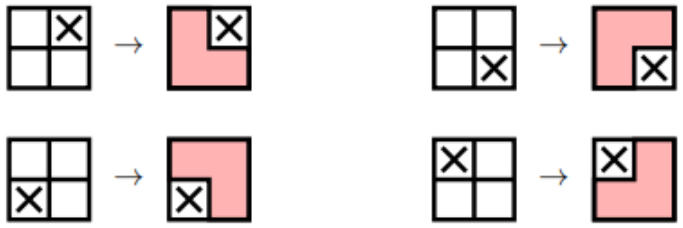


Since the proposition refers to something being true "for every $n \in \mathbb{N}$," that's a pretty good indication that induction is the way to proceed. The base case (when $n = 1$) will be fine. For the inductive hypothesis, we will be assuming that any $2^k \times 2^k$ board, with one square removed, can be perfectly covered by L-shaped tiles.

In the induction step we are going to consider a $2^{k+1} \times 2^{k+1}$ board — a board that is twice as big in each dimension— with one square missing.

Proof. We proceed by induction

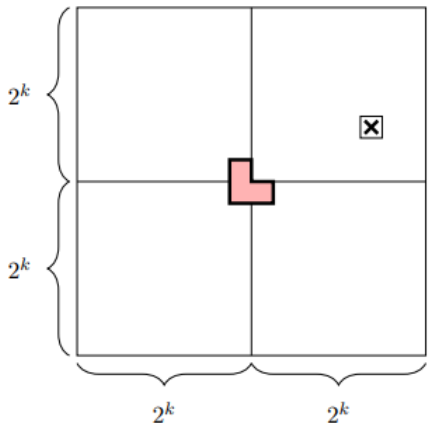
Base Case. The base case is when $n = 1$, and among the four possible squares that one can remove from a 2×2 chessboard, each leaves a chessboard which can be perfectly covered by a single L-shaped tile:



Inductive Hypothesis. Let $k \in \mathbf{N}$, and assume that if any one square is removed from a $2^k \times 2^k$ chessboard, the result can be perfectly covered with L–shaped tiles.

Induction Step. Consider a $2^{k+1} \times 2^{k+1}$ chessboard with any one square removed. Cut this chessboard in half vertically and horizontally to form four $2^k \times 2^k$ chessboards. One of these four will have a square removed, and hence, by the induction hypothesis, can be perfectly covered.

Next, place a single L-shaped tile so that it covers one square from each of the other three 2^k»2^k chessboards, as shown in the picture below.



Each of these other three $2^k \times 2^k$ chessboards can be perfectly covered by the inductive hypothesis, and hence the entire $2^{k+1} \times 2^{k+1}$ chessboard can be perfectly covered.

Conclusion. By induction, for every $n \in \mathbb{N}$, if any one square is removed from a $2^n \times 2^n$ chessboard, the result can be perfectly covered with L-shaped tiles.

- **Another note about induction**: So far, in all of our examples we proved that a statement holds from all $n \in \mathbb{N}$. The base case was $n = 1$ and in the inductive hypothesis we assumed that the result holds for some $k \in \mathbb{N}$.

  There are times where one instead wants to prove that a statement holds for only the natural numbers past some point. For example, it is possible to prove the p-test by induction, a result that you might remember from your calculus class:

  $$\sum_{i=1}^{\infty} \frac{1}{i^n} \text{ converges for all integers } n \geqslant 2.$$

  To prove this result, the base case would be $n = 2$ and in the inductive hypothesis we would assume that the result holds for some $k \in \{2, 3, 4, 5, \ldots\}$.

  At other times, you may want to prove that a result holds for more than just the natural numbers. For example, a result from combinatorics is that

  $$\sum_{i=1}^{n} \binom{n}{i} = 2^n \text{ holds for all integers } n \geqslant 0.$$

  Here, the base case is $n = 0$, and the inductive hypothesis is the assumption that this holds for some $k \in \{0, 1, 2, 3, \ldots\}$.

- **Strong induction idea**: The idea behind strong induction is that at the point when the 100th domino is the next to get knocked down, you know for sure that all of the first 99 dominoes have fallen, not just the 99th. Likewise, when you are proving some sequence of statements $S_1, S_2, S_3, S_4, \ldots$, instead of just assuming that S_k is true in order to prove S_{k+1}, why not just assume that $S_1, S_2, \ldots, S_k$ are all true in order to prove S_{k+1} — because by the time you are proving S_{k+1}, you have shown them all to be true!

- **Strong induction**: Consider a sequence of mathematical statements, $S_1, S_2, S_3, ...$
  - Suppose S_1 is true, and
  - Suppose, for any $k \in \mathbb{N}$, if $S_1, S_2, \ldots, S_k$ are all true, then S_{k+1} is true.

  Then S_n is true for every $n \in \mathbb{N}$.

  Note: In regular induction, you essentially use S_1 to prove S_2, and then S_2 to prove S_3, and then S_3 to prove S_4, and so on. With strong induction, you use S_1 to prove S_2, and then S_1 and S_2 to prove S_3, and then S_1, S_2, and S_3 to prove S_4, and so on.

- **Fundemental theorem of arithmetic**: If n is an integer and $n \geqslant 2$, then n is either prime or composite. An integer p is prime if $p \geqslant 2$ and its only positive divisors are 1 and p. A positive integer $n \geqslant 2$ that is not prime is called composite, and is therefore one that can be written as $n = st$, where s and t are integers smaller than n but larger than 1. And with that, it is time for a really big and important result.

  Theorem 4.8 (Fundamental Theorem of Arithmetic). Every integer $n \geqslant 2$ is either prime or a product of primes.

  Proof. We proceed by strong induction

Base case. The base case occurs when $n = 2$. Observe that $2 \in \mathbb{P}$

Inductive hypothesis. Let $k \in \mathbb{N}$ such that $k \geqslant 2$. Assume that the integers $2, 3, 4, ..., k$ are either prime or a product of primes.

Induction step. Next, we consider $k + 1$. We aim to show that $k + 1$ is either prime or a product of primes. Since $k + 1$ is larger than one, it is either prime or composite. Consider these two cases separately. Case 1 is that $k + 1$ is prime. In this case, our goal is achieved.

Case 2 is that $k+1$ is composite; that is, $k+1$ has positive factors other than one and itself. Say, $k + 1 = st$, where s, t are positive integers greater than zero, and

$$1 < s < k + 1 \quad 1 < t < k + 1.$$

By the inductive hypothesis, both s and t can be written as a product of primes, say

$$s = p_1 \cdot p_2 \cdot ... \cdot p_m$$
$$t = q_1 \cdot q_2 \cdot ... \cdot q_\ell.$$

Where each $p_i, q_j \in \mathbb{P}$, then

$$k + 1 = st = (p_1 \cdot p_2 \cdot ... \cdot p_m)(q_1 \cdot q_2 \cdot ... \cdot q_\ell).$$

Is written as a product of primes

Note that if s or t where prime, then m or ℓ would be one. Say s was prime, then $s = p_1$

Conclusion. By strong induction, every positive integer larger than 2 can be written as a product of primes.

- **Chocolate bar example**:

Proposition. Suppose you have a chocolate bar that is an $m \times n$ grid of squares. The entire bar, or any smaller rectangular piece of that bar, can be broken along the vertical or horizontal lines separating the squares.

The number of breaks to break up that chocolate bar into individual squares is precisely $mn - 1$.

Proof. We proceed by strong induction

Base case: The base case occurs when $n = 1$, which is an 1×1 chocolate bar. Since the number of breaks needed to break the bar into individual squares is clearly zero, we have

$$0 = 1(1) - 1 = 0.$$

As desired

Inductive hypothesis: Let $k \in \mathbb{N}$, assume that all bars with at most k squares satisfy the proposition.

Induction step: Consider now any bar with $k + 1$ squares, suppose this bar has dimensions $m \times n$. Consider an arbitrary first break, and suppose the two smaller bars have a squares and b squares, respectively. Note that we must have $a + b = mn$, because the number of squares in the smaller bars must add up to the number of squares in the original $m \times n$ bar.

By the inductive hypothesis, the bar with a squares will require $a - 1$ breaks to completely break it up, and the bar with b breaks will require $b - 1$ breaks. Therefore, to break up the $m \times n$ bar, we must make a first break, followed by $(a - 1) + (b - 1)$ additional breaks. The total number of breaks is then

$$\begin{aligned} 1 + (a - 1) + (b - 1) &= a + b - 1 \\ &= mn - 1. \end{aligned}$$

And $mn - 1$ is indeed one less than the number of squares in the $m \times n$ bar.

Conclusion. By strong induction, a chocolate bar of any size requires one break less than its number of squares to break it up into individual squares ■

Note: What if the pieces were in the shape of a triangle? If it had T squares would it still require $T - 1$ breaks?

What about other shapes? What if there are pieces missing in the middle? Interestingly, the answer is $T - 1$ no matter the bar's shape, and even if pieces are missing! As long as each of your "breaks" divides one chunk into two, that's the answer.

Here is some intuition for that: No matter the shape, the bar starts out as a single "chunk" of chocolate, and after your sequence of breaks the bar is broken into T chunks of chocolate — the T individual squares. How many breaks does it take to move from 1 chunk to T chunks? Notice that every break increases the number of chunks by 1. So after 1 break, there will be 2 chunks. After 2 breaks, there will be 3 chunks. And so on. Thus, after $T - 1$ breaks there will be T chunks, which is why $T - 1$ breaks is guaranteed to be the answer, no matter which shape you started with.

- **Multiple base cases**: When proving the $(k + 1)$st case within the induction step, strong induction allows you to apply not just the kth step, but any of the steps $1, 2, 3, \ldots, k$. In the previous two examples, you had no idea which earlier steps you will need, so it was vital that you assumed them all. At times, though, you really only need, say, the previous two steps. The kth step is perhaps not enough, but the $(k - 1)$st step and the kth step is guaranteed to be enough.

If you rely on the two previous steps, then that is analogous to saying that it takes the previous two dominoes to knock over the next one. Thus, if you knock over dominoes 1 and 2, then they will collectively knock over the third. Then, since the second and third have fallen, those two will collectively knock over the fourth. Then the third and fourth will knock over the fifth. And so on. Thus, the induction relies on two base cases, because without knocking over the first two the third won't fall and the process won't begin

Example:

Proposition. Every $n \in N$ with $n \geqslant 11$ can be written as $2a + 5b$ for some natural numbers a and b.

Base Cases. In the induction step, we will need two cases prior, so we show two base cases here: $n = 11$ and $n = 12$. Both of these can be written as asserted:

$$11 = 2 \cdot 3 + 5 \cdot 112 = 2 \cdot 1 + 5 \cdot 2.$$

Inductive Hypothesis. Assume that for some integer $k \geqslant 12$, the results hold for

$$n = 11, 12, 13, \ldots, k.$$

Induction Step. We aim to prove the result for $k + 1$. By the inductive hypothesis,

$$k - 1 = 2a + 5b$$

for some $a, b \in \mathbb{N}$. Adding 2 to both sides,

$$k + 1 = 2(a + 1) + 5b.$$

Observe that $(a + 1) \in \mathbb{N}$ and $b \in \mathbb{N}$, proving that this is indeed a representation of $(k + 1)$ in the desired form.

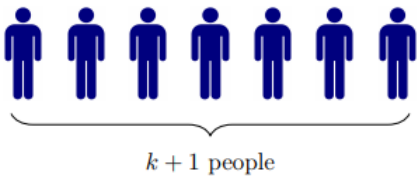
Conclusion. Therefore, by strong induction, every integer $n \geqslant 11$ can be written as the proposition asserts. ■

- **False proofs with induction**:

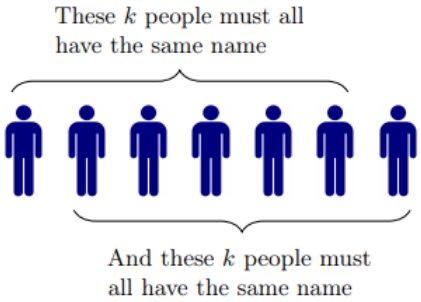
  Proposition. Everyone on Earth has the same name

  Fake Proof. We will consider groups of n people at a time, and by induction we will "prove" that for every $n \in \mathbb{N}$, every group of n people must have everyone with the same name.

  > **Base Case.** If $n = 1$, then of course everyone in the group has the same name, since there's only one person in the group!
  >
  > **Inductive Hypothesis.** Let $k \in \mathbb{N}$, and assume that any group of k people all have the same name.
  >
  > **Induction Step.** Consider a group of $k + 1$ people.

  > But notice that we can look at the first k of these people and then the last k of these people, and to each of these groups we can apply the inductive hypothesis:

  > And the only way that this can all happen, is if all $k + 1$ people have the same name.

  Conclusion. This "proves" by induction that for every $n \in N$, every group of n people must have the same name. So if you let n be equal to the number of people on Earth, this "proves" that everyone has the same name.

  For $k + 1$ people, the proof assumes that you can take the first k people and the last k people, and both of these subsets must have the same name because the induction hypothesis applies to them individually.

However, this reasoning fails when $k+1=2$. For $k+1=2$, the first subset has one person, and the second subset also has one person. These subsets do not overlap, so there is no logical connection ensuring that these two people share the same name.

The induction relies on overlapping subsets of k people to conclude that all $k+1$ people must have the same name. However, this overlap only works if $k+1>2$, meaning the proof doesn't actually establish the result for $k+1=2$, which breaks the induction chain. Without the foundation for $n=2$, the argument fails for all larger n.

- **Induction bonus example 1**.

  Lemma 4.13. For every $n \in \mathbb{N}_0$,

  $$1+2+4+8+...+2^n = 2^{n+1}-1.$$

  For example,

  $$\begin{aligned} 1 &= 2^1-1 \\ 1+2 &= 2^2-1 \\ 1+2+4 &= 2^3-1 \\ 1+2+4+8 &= 2^4-1. \end{aligned}$$

  Base case. The base case occurs when $n=1$, we have

  $$1 = 2^1-1 = 1.$$

  As desired

  Inductive hypothesis. Let $k \in \mathbb{N}_0$, assume that

  $$1+2+4+...+2^k = 2^{k+1}-1.$$

  Induction step. We wish to show that the result holds for $k+1$. That is,

  $$1+2+4+...+2^k+2^{k+1} = 2^{(k+1)+1}-1 = 2^{k+2}-1.$$

  By the inductive hypothesis, we have

  $$\begin{aligned} 1+2+4+...+2^k+2^{k+1} &= 2^{k+1}-1+2^{k+1} \\ &= 2(2^{k+1})-1 \\ &= 2^{k+2}-1. \end{aligned}$$

  As desired

  Therefore, by induction, the proposition holds for all $n \in \mathbb{N}_0$

- **Induction bonus example 2**. **Proof.** We proceed by strong induction. **Base Case.** Our base case is when $n=1$. Note that 1 can be written as 2^0, and this is the only way to write 1 as a sum of distinct powers of 2, because all other powers of 2 are larger than 1.

Inductive Hypothesis. Let $k \in \mathbb{N}$, and assume that each of the integers $1, 2, 3, \ldots, k$ can be expressed as a sum of distinct powers of 2 in precisely one way.

Induction Step. We now aim to show that $k+1$ can be expressed as a sum of distinct powers of 2 in precisely one way.

Let 2^m be the largest power of 2 such that $2^m \leqslant k+1$. We now consider two cases: the first is if $2^m = k+1$, and the second is if $2^m < k+1$.

Case 1: $2^m = k+1$. If this occurs, then 2^m itself is a way to express $k+1$ as a (one-term) sum of distinct powers of 2. Moreover, there is no other way to express $k+1$ as a sum of distinct powers of 2, because by Lemma 4.13 all smaller powers of 2 sum to $2^m - 1 = k$. Thus, even by including all smaller powers of 2, we are unable to reach $k+1$. So, in Case 1, there is precisely one such expression for $k+1$.

Case 2: $2^m < k+1$. In order to apply the inductive hypothesis, we will consider $(k+1) - 2^m$. First, note that $(k+1) - 2^m$ is less than 2^m, because otherwise $k+1$ would have two copies of 2^m within it, implying that $2^m + 2^m \leqslant k+1$. However, since $2^m + 2^m = 2 \cdot 2^m = 2^{m+1}$, this would mean $2^{m+1} \leqslant k+1$. This can't be, since 2^m was chosen to be the largest power of 2 that is at most $k+1$. Thus, it must be the case that $(k+1) - 2^m < 2^m$.

Next, by the inductive hypothesis, $(k+1) - 2^m$ can be expressed as a sum of distinct powers of 2 in precisely one way, and since $(k+1) - 2^m < 2^m$, this unique expression for $(k+1) - 2^m$ will not contain a 2^m. Thus, by adding a 2^m to it, we obtain an expression for $k+1$ as a sum of powers of 2. And this expression is unique because $(k+1) - 2^m$ is unique according to the inductive hypothesis, and the 2^m portion is unique because, again by Lemma 4.13, even if you summed all of the smaller powers of 2, you will not reach 2^m.

Conclusion. By strong induction, every $n \in \mathbb{N}$ can be expressed as a sum of distinct powers of 2 in precisely one way. □

- **Induction bonus example 3.**

Theorem 4.15 (*The binomial theorem*). For $x, y \in \mathbb{R}$, and $n \in \mathbb{N}_0$

$$(x+y)^n = \sum_{m=0}^{n} \binom{n}{m} x^{n-m} y^m.$$

Here, when $n \geqslant m$, the binomial coefficient $\binom{n}{m}$ is defined to be

$$\binom{n}{m} = \frac{n!}{m!(n-m)!},$$

which one can show is always an integer. The binomial coefficients can also be defined combinatorially: $\binom{n}{m}$ is equal to the number of ways to choose m elements from an n-element set; in fact, $\binom{n}{m}$ is read "n choose m." For example,

$$\binom{4}{2} = 6$$

because there are six subsets of the set $\{1, 2, 3, 4\}$ containing two elements:

$$\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}.$$

Binomial coefficients can be computed iteratively using *Pascal's rule*, which says that

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r},$$

as well as the fact that

$$\binom{n}{0} = 1 \quad \text{and} \quad \binom{n}{n} = 1 \quad \text{for all } n \in \mathbb{N}_0.$$

A beautiful way to combine these facts is called *Pascal's triangle*:

$$\begin{array}{ccccccccccc}
 & & & & & \binom{0}{0} & & & & & \\
 & & & & \binom{1}{0} & & \binom{1}{1} & & & & \\
 & & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & \\
 & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & \\
 & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & \\
\binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5}
\end{array}
=
\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}$$

Indeed, we can even prove the binomial theorem by induction, by making use of Pascal's rule. Here is a sketch of that proof:

Proof sketch. The base case is when $n = 0$, and indeed $(x + y)^0 = 1$. The next couple cases are more interesting, and you can check that $(x + y)^1 = x + y$ and $(x+y)^2 = x^2 + 2xy + y^2$ do indeed match the theorem. The inductive hypothesis will be

$$(x+y)^k = x^k + \binom{k}{1}x^{k-1}y + \binom{k}{2}x^{k-2}y^2 + \cdots + \binom{k}{k-1}xy^{k-1} + y^k.$$

For the induction step, we perform easy algebra, then apply the inductive hypothesis, then perform hard algebra, then apply Pascal's rule:

$$\begin{aligned}
(x+y)^{k+1} &= (x+y)(x+y)^k \\
&= (x+y)\left[x^k + \binom{k}{1}x^{k-1}y + \binom{k}{2}x^{k-2}y^2 + \cdots + \binom{k}{k-1}xy^{k-1} + y^k\right] \\
&= x^{k+1} + \left[\binom{k}{0}\right]x^k y + \left[\binom{k}{1}\right]x^{k-1}y^2 + \cdots + \left[\binom{k}{k}\right]xy^k + y^{k+1} \\
&= x^{k+1} + \binom{k+1}{1}x^k y + \binom{k+1}{2}x^{k-1}y^2 + \cdots + \binom{k+1}{k}xy^k + y^{k+1}.
\end{aligned}$$

And that—a few boring algebraic details omitted—is the proof.

The binomial theorem tells us that in order to expand $(x+y)^5$ you can just look at the 5th row of Pascal's triangle (where the top element counts as the 0th row, so the 5th row is 1 5 10 10 5 1):

$$(x+y)^5 = 1x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + 1y^5.$$

Moreover, by plugging in special values for x and y, all sorts of neat identities pop out. There are loads of examples of this, but here are just three:

- By plugging in $x = 1$, $y = 1$, we prove $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.
- By plugging in $x = 2$, $y = 1$, we prove $3^n = \sum_{k=0}^{n} \binom{n}{k} 2^k$.
- By plugging in $x = -1$, $y = 1$, we prove $0 = \sum_{k=0}^{n} (-1)^k \binom{n}{k}$.

1.5 Logic

- **Statements**: A statement is a sentence or mathematical expression that is either true or false. If the logic is valid and the statements are true, then it is called sound

  Every theorem/proposition/lemma/corollary is a (true) statement; Every conjecture is a statement (of unknown truth value); and Every incorrect calculation is a (false) statement.

- **Open sentence**: A related notion is that of an *open sentence*, which refers to sentences or mathematical expressions that:

  1. do not have a truth value,
  2. depend on some unknown, like a variable x or an arbitrary function f, and
  3. when the unknown is specified, the open sentence becomes a statement (and thus has a truth value).

  Their truth value depends on the specific value of x or f that is chosen.

  Typically, we use capital letters for statements, like P, Q and R. Open sentences are often written the same, or perhaps like $P(x)$, $Q(x)$ or $R(x)$ when one wishes to emphasize the variabl

- **And, or, not**: Let P and Q be statements or open sentences.

  1. $P \wedge Q$ means "P and Q".
  2. $P \vee Q$ means "P or Q (or both)".
  3. $\sim P$ means "not P".

- **Implies, iff**: Let P and Q be statements or open sentences.

  1. $P \implies Q$ means "P implies Q".
  2. $P \iff Q$ means "P if and only if Q".

  Let's now discuss a subtle aspect of implications: Translating them to and from English. Language can be complicated,[3] and we in fact have many different ways in English to say "P implies Q." Here are some examples:

  - If P, then Q
  - Q if P
  - P only if Q
  - Q whenever P
  - Q, provided that P
  - Whenever P, then also Q
  - P is a sufficient condition for Q
  - For Q, it is sufficient that P
  - For P, it is necessary that Q

  For example, "If it is raining, then the grass is wet" has the same meaning as "The grass is wet if it is raining." These also mean the same as "The grass is wet whenever it is raining" or "For the grass to be wet, it is sufficient that it is raining."

[3] Language nuances can make logical translation challenging.

Next, here are some ways to say "P if and only if Q":

- P is a necessary and sufficient condition for Q.
- For P, it is necessary and sufficient that Q.
- P is equivalent to Q.
- If P, then Q, and conversely.
- P implies Q and Q implies P.
- Shorthand: P iff Q.
- Symbolically: $(P \implies Q) \wedge (Q \implies P)$.

The fact that "P implies Q" is the same as "If P, then Q" or "Q if P" is sometimes intuitive to students. But the fact that these are all the same as "P only if Q" is often confusing. Most people's guts tell them that "P implies Q" should be the same as "Q only if P."

The answer is "P only if Q", and the way to think about it is that "P implies Q" means that whenever P is true, Q must also be true. And "P only if Q" means that P can only be true if Q is true...that is, whenever P is true, it must be the case that Q is also true...that is, $P \implies Q$.

- **Conditional, biconditional statements**: Now, if P and Q are statements, then "$P \implies Q$" and "$P \iff Q$" are also statements, meaning they must also be either true or false. The statement $P \implies Q$ is called a conditional statement, whereas $P \iff Q$ is called a biconditional statement. These are minor definitions, but the following is an important definition.
- **Converse**: The *converse* of $P \implies Q$ is $Q \implies P$

  Note: If $P \implies Q$, it is not necessarily the case that $Q \implies P$
- **Truth tables for and, or, and not**: A truth table models the relationship between the truth values of one or more statements, and that of another

P	Q	$P \wedge Q$
True	True	True
True	False	False
False	True	False
False	False	False

For for "P and Q" to be a true statement, both P and Q must be independently true

Here's how the truth values for P and for Q affect the truth value for $P \vee Q$.

P	Q	$P \vee Q$
True	True	True
True	False	True
False	True	True
False	False	False

It is sufficient that either P is true or that Q is true (or both).

Finally, here is how the truth values for P affects that of $\neg P$.

P	$\neg P$
True	False
False	True

In order for "not P" to be true, it is required that P be false. By applying this reasoning twice, this also implies that $\sim\sim P$ and P always have the same truth value.

One last example shows how we proceed with more complicated statements

P	Q	$P \vee Q$	$P \wedge Q$	$\neg(P \wedge Q)$	$(P \vee Q) \wedge \neg(P \wedge Q)$
True	True	True	True	False	False
True	False	True	False	True	True
False	True	True	False	True	True
False	False	False	False	True	False

- **De Morgan's Logic Laws**: Take a loot at the truth tables for $\neg(P \wedge Q)$ and $\neg P \vee \neg Q$, side by side:

P	Q	$P \wedge Q$	$\neg(P \wedge Q)$
True	True	True	False
True	False	False	True
False	True	False	True
False	False	False	True

P	Q	$\neg P$	$\neg Q$	$\neg P \vee \neg Q$
True	True	False	False	False
True	False	False	True	True
False	True	True	False	True
False	False	True	True	True

Since the final columns are the same, if one is true, the other is true; if one is false, the other is false; that is, there is no way to select P and Q without these two agreeing. When two statements have the same final column in their truth tables, like in the example above, they are said to be logically equivalent (one is true if and only if the other is true), which we denote with an " $\iff$ " symbol. De Morgan's logic law, for example, can be written like this:

$$\neg(P \wedge Q) \iff (\neg P \vee \neg Q)$$

"P and Q are not both true" is the same as "P is false or Q is false."

Theorem: If P and Q are statements, then

$$\neg(P \wedge Q) \iff \neg P \vee \neg Q \quad \text{and} \quad \neg(P \vee Q) \iff \neg P \wedge \neg Q.$$

- **P, Q, and their names**: In logical statements involving P and Q, the terms P and Q are referred to as propositions or statements. Depending on the logical operator used, they may also have more specific names:

  1. **In a conjunction ($P \wedge Q$):**
     - P and Q are called **conjuncts**.
  2. **In a disjunction ($P \vee Q$):**
     - P and Q are called **disjuncts**.
  3. **In an implication ($P \implies Q$):**
     - P is called the **antecedent** (or **hypothesis**, **premise**).

- Q is called the **consequent** (or **conclusion**).

4. **In a biconditional ($P \iff Q$):**
   - P and Q are called **equivalents** (since $P \iff Q$ means P and Q are logically equivalent).
5. **In negation ($\neg P$):**
   - P is simply the proposition being negated.

- **Implications**: We call the conditional statemests, $P \implies Q$ *implications*. They are called implications because they express a logical relationship where one statement (the premise, P) "implies" or leads to another statement (the conclusion, Q). The word "implication" comes from the Latin root *implicare*, meaning "to entwine" or "to involve," reflecting the idea that P is connected to Q.

  A biconditional statement combines two implications, $P \implies Q$ AND $Q \implies P$

- **Truth Tables with Implications**: Consider the truth table for the implication $P \implies Q$

P	Q	$P \implies Q$
True	True	True
True	False	False
False	True	True
False	False	True

  The results of the first two rows are trivial, but the last two may be hard to grasp.

  Why is the implication true if the assumption, P, is false? It's kind of like how we said that this is true: "If $x \in \varnothing$, then x is a purple elephant that speaks German." Since there is nothing in the empty set, if you suppose $x \in \varnothing$, you can then claim anything you want about x and it is inherently true — you certainly cannot present to me any element in the empty set that is not a purple elephant that speaks German. In the set theory chapter, we called such a claim *vacuously true*.

  Likewise, in a universe where P is true, the statement $P \implies Q$ has some real meaning that needs to be proven or disproven: Does P being true imply Q is true, or not? But in a universe where P is not true, it claims nothing, and hence $P \implies Q$ is *vacuously true*.

  "If unicorns exist, then they can fly" can certainly not be considered false, because unicorns do not exist, so any claim about them is considered vacuously true. Indeed, the way to falsify that proposition would be to locate a unicorn that cannot fly, which is impossible to do. Every unicorn in existence can indeed fly! Also, every unicorn in existence cannot fly! Neither can be disproven!

  Let's now consider the truth table for the statement $P \iff Q$

P	Q	$P \iff Q$
True	True	True
True	False	False
False	True	False
False	False	True

  We can see this by writing $P \iff Q$ as $(P \implies Q) \wedge (Q \implies P)$

- **Quantifiers**: Consider the sentence

$$n \text{ is even}$$

Which is not a statement because it is neither true nor false. One way to turn a sentence like this into a statement is to give n a value. For example,

$$\text{If } n = 5, \text{ then } n \text{ is even}$$

What I'd like to discuss now are two other basic ways to turn "n is even" into a statement: add quantifiers. A quantifier is an expression which indicates the number (or quantity) of our objects

$$\forall\ n \in \mathbb{N},\ n \text{ is even}$$
$$\exists\ n \in \mathbb{N} \text{ such that } n \text{ is even}$$

Where $\forall$ means "for all", and $\exists$ means "there exists". The symbol $\forall$ is known as the *universal quantifer*. Whereas $\exists$ is known as the *existential quantifier.*

Note: We also have $\nexists$ "there does not exist", and $\exists!$ "there exists a unique"

- **Rules of negating**: We have the following rules for negating statements
  - $\neg\wedge = \vee$
  - $\neg\vee = \wedge$
  - $\neg\forall = \exists$
  - $\neg\exists = \forall$

Consider the statement, R : for every real number x, there is some real number y such that $y^3 = x$. Symbolically, we have

$$\forall\ x \in \mathbb{R},\ \exists\ y \in \mathbb{R} \text{ such that } y^3 = x.$$

Then,

$$\neg(\forall\ x \in \mathbb{R},\ \exists\ y \in \mathbb{R} \text{ such that } y^3 = x).$$

Is equivalent to the statement

$$\exists\ x \in \mathbb{R},\ \text{such that } \forall\ y \in \mathbb{R},\ y^3 \neq x.$$

- **Negations with implications**: First, recall the truth table for $P \implies Q$

P	Q	$P \implies Q$
True	True	True
True	False	False
False	True	True
False	False	True

The only way for $P \implies Q$ to be false is for both P to be true and for Q to be false. This shows that

$$\neg(P \implies Q) \Leftrightarrow P \wedge \neg Q.$$

Consider the statement

$$S: \ \forall\, n \in \mathbb{N}, (3 \mid n) \implies (6 \mid n).$$

Then,

$$\begin{aligned} \neg S : \neg(\forall\, n \in \mathbb{N}, (3 \mid n) &\implies (6 \mid n)) \\ &\Leftrightarrow \exists\, n \in \mathbb{N} \text{ such that } (3 \mid n) \wedge (6 \nmid n). \end{aligned}$$

- **The contrapositive (and the inverse)**: The *contrapositive* of $P \implies Q$ is $\neg Q \implies \neg P$

  Note: The *inverse* of $P \implies Q$ is $\neg P \implies \neg Q$

  Theorem: An implication is logically equivalent to its contrapositive. That is,

  $$P \implies Q \Leftrightarrow \neg Q \implies \neg P.$$

  The truth table easily verifys this

- **Proving quantified statements: Existential proofs**: To prove an existence statement, it suffices to exhibit an example satisfying the criteria. The above strategy is called a constructive proof — you literally construct an example. There are also non-constructive ways to prove something exists. Often (but not always!) non-constructive proofs make use of some other theorem.

- **Proving quantified statements: Universal proofs**: To prove a universal statement, it suffices to choose an arbitrary case and prove it works there. We have seen several examples of this. For example, if you were asked to prove that "For every odd number n, it follows that $n+1$ is even," your proof wouldn't explicitly check 1 and 3 and 5 and so on. Rather, you would say "Since n is odd, $n = 2a+1$ for some $a \in \mathbb{Z}$." Then you would note that

  $$n + 1 = (2a + 1) + 1 = 2(a + 1)$$

  is even. The point here is that by letting $n = 2a + 1$, you were essentially selecting an arbitrary odd number, and operating on that. Every odd number can be written in that form, and every odd number can have 1 added to it and then factored like we did. Since our n was completely arbitrary, everything we did could be applied to any particular odd number. Proving something holds for an arbitrary element of a set, proves that it in turn holds for every element in that set.

- **Proving biconditional statements**: In order to prove a statement in the form $P \implies Q$, we must prove both directions. That is, $P \implies Q$ and $Q \implies P$

1.6 Proof using the contrapositive

- **Proof outline**:

  Proposition. $P \implies Q$

  Proof. We will use the contrapositive. Assume not-Q

  $\langle\langle$ An explanation of what not-Q means $\rangle\rangle$, use definitions, and/or other results

  $\vdots$ Apply algebra,

  $\vdots$ logic, techniques.

  $\langle\langle$ Hey look, that's what not-P means $\rangle\rangle$

  Therefore not-P

  Since not-$Q \implies$ not-P, by the contrapositive $P \implies Q$ ■

- **Contrapositive proof 1.**

  Proposition. Suppose $n \in \mathbb{N}$, if n^2 is odd, then n is odd.

  Proof. We will use the contrapositive. The statement, $\forall\ n \in \mathbb{N},\ n^2 = 2k+1 \implies n = 2\ell+1,\ k, \ell \in \mathbb{Z}$ has the logically equivalent contrapositive $\forall n \in \mathbb{N},\ n \neq 2\ell+1 \implies n^2 \neq 2k+1$. Since $n \in \mathbb{N}$, if n, n^2 is not odd, then it must be even. Thus, the statement becomes $\forall n \in \mathbb{N},\ n = 2\ell \implies n^2 = 2k,\ k, \ell \in \mathbb{N}$ which becomes much easier to proof. For some extra practice negating statements, here is the negation

  $$\begin{aligned}&\neg(\forall n \in \mathbb{N},\ n^2 = 2k+1 \implies n = 2\ell+1,\ k, \ell \in \mathbb{N})\\ &= \exists n \in \mathbb{N} \text{ such that } n^2 = 2k+1 \wedge n \neq 2\ell+1.\end{aligned}$$

  Recall $\neg(P \implies Q) = P \wedge \neg Q$

  Assume $n \in \mathbb{N}$, and that n is even. Since n is even, it must be that $n = 2\ell$, for some integer ℓ. Squaring both sides, we get

  $$\begin{aligned}n^2 &= (2\ell)^2\\ &= 4\ell^2 = 2(2\ell^2).\end{aligned}$$

  Since $\ell \in \mathbb{Z}$, we know $2\ell^2 \in \mathbb{Z}$, and thus n^2 is even.

  Therefore, since n not being odd implies n^2 is also not odd, we have shown by the contrapositive that if n^2 is odd, n is also odd ■

- **Contrapositive proof 2.**

  Proposition. Suppose $n \in N$. Then, n is odd if and only if $3n + 5$ is even

Proof. We will prove this in two parts

Part 1: If n is odd then $3n+5$ is even. Assume $n \in \mathbb{N}$ is odd, then $n = 2k+1$, for $k \in \mathbb{N}_0$. Thus,

$$\begin{aligned} 3n+5 &= 3(2k+1)+5 \\ &= 6k+3+5 = 6k+8 \\ &= 2(3k+4). \end{aligned}$$

Thus even.

Part 2: $3n+5$ being even implies n is odd. We prove this by use of the contrapositive. The given statement has the following contrapositive...

$$n = 2k \implies 3n+5 = 2\ell+1,\ k, \ell \in \mathbb{N}_0.$$

Thus,

$$\begin{aligned} 3n+5 &= 3(2k)+5 \\ &= 6k+5 = 6k+4+1 \\ &= 2(3k+2)+1. \end{aligned}$$

Thus odd.

Since $P \implies Q$, and $Q \implies P$, it must be that $P \iff Q$ is true. Thus, we assert for $n \in \mathbb{N}$, n is odd if and only if $3n+5$ is even.

- **Contrapositive proof 3.**:

Proposition. Let $a, b \in \mathbb{Z}$, and $p \in \mathbb{P}$. If $p \nmid ab$, then $p \nmid a$ and $p \nmid b$ **Proof.** Suppose $a, b \in \mathbb{Z}$ and p is a prime. We will use the contrapositive. Suppose that it is not true that $p \nmid a$ and $p \nmid b$. By the logic form of De Morgan's law (Theorem 5.9), this is equivalent to saying it is not true that $p \nmid a$ *or* it is not true that $p \nmid b$. That is, $p \mid a$ *or* $p \mid b$. Let's consider these two cases separately.

Case 1. Suppose $p \mid a$, which by the definition of divisibility (Definition 2.8) means that $a = pk$ for some $k \in \mathbb{Z}$. Thus,

$$ab = (pk)b = p(kb).$$

Since $k, b \in \mathbb{Z}$, also $(kb) \in \mathbb{Z}$. And so, by the definition of divisibility (Definition 2.8), $p \mid ab$.

Case 2. Suppose $p \mid b$, which by the definition of divisibility (Definition 2.8) means that $b = p\ell$ for some $\ell \in \mathbb{Z}$. Thus,

$$ab = a(p\ell) = b(a\ell).$$

Since $a, \ell \in \mathbb{Z}$, also $(a\ell) \in \mathbb{Z}$. And so, by the definition of divisibility (Definition 2.8), $p \mid ab$.

In either case, we concluded that $p \mid ab$, which is equivalent to saying that it is not true that $p \nmid ab$.

We proved that if it is not true that $p \nmid a$ and $p \nmid b$, then it is not true that $p \nmid ab$. Hence, by the contrapositive, this implies that if $p \mid ab$, then $p \mid a$ and $p \mid b$. □

Note: Mathematicians have agreed that we should be allowed to skip essentially-identical cases

If you have two cases, like $p \mid a$ and $p \mid b$, and there is literally no mathematical distinction between them, then you are allowed to say "without loss of generality, assume $p \mid a$." This allows you to skip the "$p \mid b$" case entirely.

Condensed, Elder-Approved Proof. Suppose $a, b \in \mathbb{Z}$ and p is a prime. We will use the contrapositive. Suppose that it is not true that $p \nmid a$ and $p \nmid b$. By the logic form of De Morgan's law (Theorem 5.9), this is equivalent to saying it is not true that $p \nmid a$ *or* it is not true that $p \nmid b$. That is, $p \mid a$ *or* $p \mid b$. Without loss of generality, assume $p \mid a$.

By the definition of divisibility (Definition 2.8), this means that $a = pk$ for some $k \in \mathbb{Z}$. Thus,

$$ab = (pk)b = p(kb).$$

Since $k, b \in \mathbb{Z}$, also $(kb) \in \mathbb{Z}$. And so, by the definition of divisibility (Definition 2.8), $p \mid ab$.

We proved that if it is not true that $p \nmid a$ and $p \nmid b$, then it is not true that $p \nmid ab$. Hence, by the contrapositive, this implies that if $p \mid ab$, then $p \mid a$ and $p \mid b$. □

- **Contrapositive proof 4.**

Proposition. Let $a, b, n \in \mathbb{N}$. If $36a \not\equiv 36b \pmod{n}$, then $n \nmid 36$

Proof idea. The fact that this proposition says a lot of things are not happening is one indication that the contrapositive could be worthwhile. The contrapositive states For $a, b, n \in \mathbb{N}$, If $n \mid 36$, then $36a \equiv 36b \pmod{n}$

Proof. Assume $a, b, n \in \mathbb{N}$, and $n \mid 36$. In this case, we have $36 = nk$, for $k \in \mathbb{Z}$. We require $36a - 36b = n\ell$, for $\ell \in \mathbb{Z}$. We then examine the quantity $36a - 36b$. Since $36 = nk$, we have

$$\begin{aligned} 36a - 36b &= nka - nkb \\ &= n(ka - kb). \end{aligned}$$

Which is precisely the definition of divisibility, since it is clear that $ka - kb \in \mathbb{Z}$. Thus, we have $n \mid 36a - 36b$, and by the definition of modular congruence $36a \equiv 36b \pmod{n}$.

Therefore, by the contrapositive, $36a \not\equiv 36b \pmod{n}$ implies that $n \nmid 36$ ■

- **Lemma 6.6** This lemma has two parts

  (i) If $m \in \mathbb{Z}$, then $m^2 + m$ is even

  (ii) If $a \in \mathbb{Z}$, and a^2 is even, then a is even

  This proof is trivial and will not be shown. Proving i is simply a proof by cases. To prove ii, we can use the contrapositive, instead proving that if a is odd, then a^2 is odd. Which, by the contrapositive shows that if a^2 is even, then a must also be even.

- **Contrapositive proof 5.**

  Proposition. If a is an odd integer, then $x^2 + x - a^2 = 0$ has no integer solution.

  Proof idea. We will use the contrapositive, which states if $x^2 + x - a^2 = 0$ has an integer solution, then a is even.

  Note: Negating Q in this case ($x^2 + x - a^2 = 0$ has no integer solution) does not given $x^2 + x - a^2 \neq 0$... It is important to question what it means for the given statement to be false in order to properly negate. The negation of the statement is "it is false that $x^2 + x - a^2 = 0$ has no integer solutions", which must mean that some integer m exists such that $m^2 + m - a^2 = 0$.

  Proof. Suppose that a is an odd integer. We will use the contrapositive. Assume that it is false that $x^2 + x - a^2 = 0$ has no integer solutions; that is, assume that there is some integer m such that

  $$m^2 + m - a^2 = 0.$$

  By the quadratic formula[9] and then some algebra,

  $$\begin{aligned} m &= \frac{-1 \pm \sqrt{1^2 - 4(1)(-a^2)}}{2(1)} \\ m &= \frac{-1 \pm \sqrt{1 + 4a^2}}{2} \\ 2m &= -1 \pm \sqrt{1 + 4a^2} \\ 2m + 1 &= \pm\sqrt{1 + 4a^2} \\ 4m^2 + 4m + 1 &= 1 + 4a^2 \\ m^2 + m &= a^2 .. \end{aligned}$$

Next, observe that $m^2 + m$ is guaranteed to be even, by Lemma 6.6 part (i). Thus, since we just deduced that $m^2 + m = a^2$, this means that a^2 must be even. And since a is an integer, a^2 being even implies that a is even, by Lemma 6.6 part (ii). In particular, this means that a is not odd.

We have shown that if it is false that $x^2 + x - a^2 = 0$ has no integer solutions, then it is also false that a is an odd integer. By the contrapositive, if a is an odd integer, then $x^2 + x - a^2 = 0$ has no integer solution. $\square$

1.7 Contradiction

- **The idea**: The big idea is this: If you start with something true and apply correct logic to it, you will never arrive at something false. So it can't be true that Carmen stole the bag, if that would imply the falsity that she can be in two places at once. Indeed, if your assumptions imply something false, then something you assumed had to be false as well.

  Suppose we had a theorem $P \implies Q$. Throughout the problem, we assume P to be true. The goal is to show that Q is also true. By the truth tables, either Q is true or $\neg Q$ is true, not both. This gives two options.

  1. P is true and Q is true ($P \wedge Q$)
  2. P is true and $\neg Q$ is true ($P \wedge \neg Q$)

  If $P \wedge \neg Q$ implies anything false, that can't be the correct option. That is, it must be $P \wedge Q$. Thus, we have shown $P \implies Q$

  Notice that the only way that $P \implies Q$ can be false is if P is true and Q is false.

P	Q	$P \implies Q$
True	True	True
True	False	False
False	True	True
False	False	True

  Thus, this is the only case we have to rule out in order to prove our theorem: that $P \implies Q$ is true. So, if you assume that P is true and Q is false, and manage to use that to deduce a contradiction, then you will have ruled out the one and only bad case, which in turn means that the theorem must be true!

  In other words, if $P \wedge \neg Q$ cannot be, then it must be that $P \implies Q$

- **Contradiction example 1.**

  Proposition. There does not exist a largest natural number

  Proof Idea. One quick note: This proposition is not phrased explicitly as "$P \implies Q$," but you are probably starting to see how to rephrase propositions in this form. For example, this proposition could instead be stated as: "If N is the set of natural numbers, then N does not have a largest element." Or, equivalently: "If N is larger than every natural number, then $N \notin \mathbb{N}$" Or, equivalently: "If N is a natural number, then there exists a natural number larger than N."

  For our proof by contradiction, we will assume that there *is* a largest natural number, and then deduce a contradiction. There are several ways to do this, but one way is to assume that N is the largest and then show that $N + 1$ must be larger—if it weren't, we could deduce that $0 \geqslant 1$, which is clearly a contradiction. Here's that:

  Proof. Assume for a contradiction that there is a largest element of $\mathbb{N}$, and call this number N. Being larger than every other natural number, N has the property that $N \geqslant m$ for all $m \in \mathbb{N}$.

Observe that since $N \in \mathbb{N}$, also $(N+1) \in \mathbb{N}$. And so, by assumption,

$$N \geqslant N + 1.$$

Subtracting N from both sides,

$$0 \geqslant 1.$$

This is a contradiction[1] since we know that $0 < 1$, and therefore there must not be a largest element of $\mathbb{N}$. $\square$

- **Contradiction example 2.**

  Proposition. There does not exist a smallest positive rational number.

  Proof. Assume for the sake of contradiction that there does exist a smallest positive rational number. Call this number q. Since $q \in \mathbb{Q}$, we have

  $$q = \frac{a}{b}.$$

  Where $a, b \in \mathbb{Z}$, and $a, b > 0$. Since q is the smallest, than for all $r \in \mathbb{Q}$, we have $q \leqslant r$. Let $r = \frac{a}{2b}$. Then,

  $$\begin{aligned} \frac{a}{b} &\leqslant \frac{a}{2b} \\ \implies 2ab &\leqslant ab \\ \implies 2 &\leqslant 1. \end{aligned}$$

  This is a contradiction, since we know $2 > 1$. It must be that there is no smallest positive rational number.

- **Proof by contradiction general form**:

  Proposition. $P \implies Q$

  Proof. Assume for the sake of contradiction P and $\neg Q$

  $\langle\langle$ An explanation of what these mean $\rangle\rangle$

  $\vdots$ Apply algebra,

  $\vdots$ logic, techniques.

  $\langle\langle$ Hey look, that contradicts something we know to be true $\rangle\rangle$

  We obtained a contradiction, therefore $P \implies Q$ $\blacksquare$

- **Proof by contradiction example 3.**

  Proposition. If A, B are sets, then $A \cap (B \setminus A) = \varnothing$

Proof. Assume for the sake of contradiction, that $A \cap (B \setminus A) \neq \varnothing$

> Since $A \cap (B \setminus A) \neq \varnothing$, then $\exists x \in A \cap (B \setminus A)$. Thus, $x \in A \wedge x \in (B \setminus A)$. Rewrite $B \setminus A$ as $B \cap A^C$. Thus, $x \in B \ \wedge \ x \in A^C$. Since $x \in A^C$, it must be that $x \notin A$. Thus, we have $x \in A$, $x \in B$, and $x \notin A$

Therefore, since $x \in A$ and $x \notin A$ is a contradiction, it must be that if A, and B are sets, then $A \cap (B \setminus A) = \varnothing$ ■.

- **Proof by contradiction example 4.**

Proposition. There does not exists integers m, n such that $15m + 35n = 1$

Proof. Assume for the sake of contradiction there does exist integers m, n such that $15m + 35n = 1$, since $m, n \in \mathbb{Z}$, $3m + 7n \in \mathbb{Z}$, but

$$\begin{aligned} & 15m + 35n = 1 \\ \implies & 3m + 7n = \frac{1}{5}. \end{aligned}$$

Since $3m + 7n \notin \mathbb{Z}$, we have a contradiction. Thus, it must be that there does not exist integers m, n such that $15m + 35n = 1$.

Alternatively, we could have done

$$\begin{aligned} & 15m + 35n = 1 \\ \implies & 5(3m + 7n) = 1. \end{aligned}$$

Which implies $5 \mid 1$. But it is clearly the case that $5 \nmid 1$, since there exists no $k \in \mathbb{Z}$ such that $1 = 5k$. Thus, another way to arrive at a contradiction. ■

- **Proof by contradiction example 5.**

Proposition. There are infinitely many primes.

Proof. Suppose for the sake of contradiction that there are finitely many primes, say k in total. Let $p_1, p_2, p_3, ..., p_k$ be the complete list. Consider the number $N = p_1 \cdot p_2 \cdot p_3 \cdot ... \cdot p_k$. Next, consider $N + 1$. That is, $p_1 p_2 p_3 ... p_k + 1$. Either $N + 1$ is prime or it is composite, we consider both cases separately

Case 1: $N + 1$ is prime. In this case, $N + 1$ is prime and greater than all the p_is we have previously considered. Thus, we have found a new prime.

Case 2: $N + 1$ is composite. We begin by showing that no such p_i divides $N + 1$. Because we know that $p_i \mid N$, we have

$$N \equiv 0 \pmod{p_i}.$$

Adding one to both sides, we get

$$N + 1 \equiv 1 \pmod{p_i}.$$

Hence, it must be that $p_i \nmid N + 1$. Since p_i was arbitrary, this shows that none of our k primes divide $N + 1$

We assumed that $p_1, p_2, \ldots, p_k$ was the complete list of prime numbers. And recall that $N+1$ is assumed to be composite, which means it is a product of primes. But since none of the p_i divide $N+1$, there must be some other prime number, q, which divides $N+1$. And hence, we have again found a new prime.

In either case, we have contradicted the claim that $p_1, p_2, \ldots, p_k$ was an exhaustive list of the prime numbers. Therefore, there must be infinitely many primes. ■

- **Proof by contradiction example 6.**

Proposition The number $\sqrt{2}$ is irrational

Proof. Assume for a contradiction that $\sqrt{2}$ is rational. Then there must be some non-zero integers p and q where

$$\sqrt{2} = \frac{p}{q}.$$

Moreover, we may assume that this fraction is written in *lowest terms*, meaning that p and q have no common divisors. Then,

$$\sqrt{2}q = p.$$

By squaring both sides,

$$2q^2 = p^2.$$

Since $q^2 \in \mathbb{Z}$, by the definition of divisibility, this implies that $2 \mid p^2$, and hence $2 \mid p$ by Lemma 2.17 part (iii). By a second application of the definition of divisibility, this means that $p = 2k$ for some non-zero integer k. Plugging this in:

$$\begin{aligned} 2q^2 &= p^2, \\ 2q^2 &= (2k)^2, \\ 2q^2 &= 4k^2, \\ q^2 &= 2k^2 \end{aligned}$$

Therefore, $2 \mid q^2$, and hence $2 \mid q$, again by Lemma 2.17 part (iii). But this is a contradiction: We had assumed that p and q had no common factors, and yet we proved that 2 divides each. Therefore, $\sqrt{2}$ cannot be rational, meaning it is irrational.

The following is a geometric proof that $\sqrt{2} \in \bar{\mathbb{Q}}$. Recall that $\bar{\mathbf{Q}}$ is the set of irrational numbers.

Assume for a contradiction that $\sqrt{2} = \frac{p}{q}$ where $p, q \in \mathbb{N}$ and the fraction is written in lowest terms. This implies that

$$2q^2 = p^2,$$

but this time let's think about this as

$$p^2 = 2q^2.$$

Or, better yet,

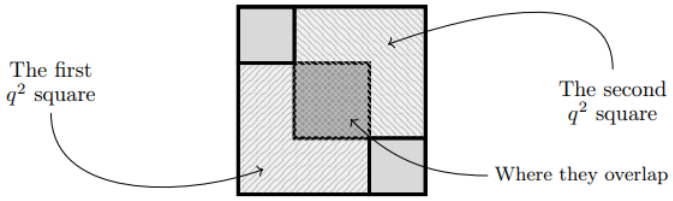
$$p^2 = q^2 + q^2.$$

Since p and q are integers, p^2 represents the area of a square with side length p, and each q^2 represents the area of a square with side length q.

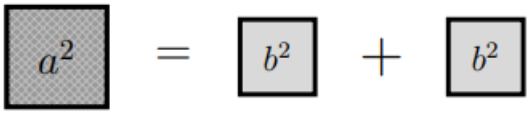
$$\begin{array}{ccccccccccc}
 & & & & \binom{0}{0} & & & & & & \\
 & & & \binom{1}{0} & & \binom{1}{1} & & & & & \\
 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & & \\
 & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & & \\
\binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & & \\
\binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5}
\end{array} = \begin{array}{ccccccccccc}
 & & & & 1 & & & & & & \\
 & & & 1 & & 1 & & & & & \\
 & & 1 & & 2 & & 1 & & & & \\
 & 1 & & 3 & & 3 & & 1 & & & \\
1 & & 4 & & 6 & & 4 & & 1 & & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}$$

Recall that $\sqrt{2} = \frac{p}{q}$ was written in lowest terms. In particular, this means that there do not exist any smaller integers a and b for which $\sqrt{2} = \frac{a}{b}$. Our contradiction will be to find such a and b.

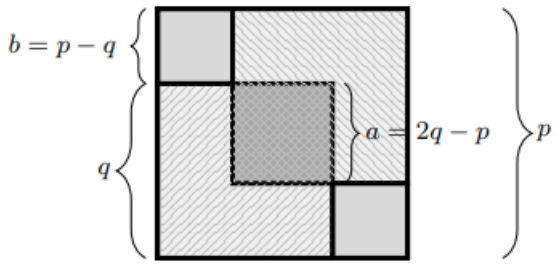
Getting back to the squares above, we are now going to imagine each square is a piece of paper and we are going to place the two q^2 squares on top of the p^2 square. If one q^2 square is placed in the lower-left, and the other is placed in the upper-right, this happens



Notice that there is one square region in the middle that was covered twice, and two small squares in the upper-left and lower-right that were not covered at all. And remember: The amount of area in the p^2 square is equal to the amount of area in the two q^2 squares. Therefore, the area that was covered twice must equal the area that was not covered at all! Let's suppose the middle square has dimensions $a \times a$, and the two corner squares have dimensions $b \times b$. Then, this reasoning shows that



And those a and b must also be integers, since they are the difference of integers from the overlap picture:



We had assumed that p and q were the smallest integers for which $\sqrt{2} = \frac{p}{q}$, and yet the above image shows that a and b are also integers, and since $a^2 = b^2 + b^2$, which implies $2b^2 = a^2$, we have $2 = \frac{a^2}{b^2}$. And so, finally, by taking the square root of each side, we see that

$$\sqrt{2} = \frac{a}{b}.$$

We have shown that a and b are integers with the above property. The picture above also shows that a is smaller than p, and b is smaller than q. Combined, this contradicts our assumption that p and q are the smallest integers where $\sqrt{2} = \frac{p}{q}$.

- **The irrational numbers**: The fact that irrational numbers exist explains why we need the real numbers $\mathbb{R}$—the rational numbers $\mathbb{Q}$ are clearly not enough! Next, note that while $\sqrt{2}$ is not a ratio of integers, it is a root of $x^2 - 2 = 0$, which is a polynomial with integer coefficients.

  Big Question: Is every irrational number a root of a polynomial with integer coefficients?

  Big Answer: Nope! In 1844, Joseph Liouville proved that

  $$\sum_{k=1}^{\infty} \frac{1}{10^{k!}} = 0.11000100000000000000000100\ldots$$

  is not the root of any polynomial with integer coefficients.

  The irrational numbers were thus partitioned into *algebraic numbers*, which are the roots of such polynomials, and *transcendental numbers*, which are not. Today, π and e are the most famous numbers which have been proved to be transcendental.

- **Proof of the halting problem**:

  Theorem. Assume that P is an arbitrary program and i is a possible input of P; we write $P(i)$ to be the result of plugging input i into the program P. There does not exist a program $H(P(i))$ which determines whether $P(i)$ will eventually halt.

  Proof. Assume for a contradiction that such a program H did exist. Create a new program $T(x)$; its input, x, is itself a program with some input. Now, we define the program $T(x)$ as follows:

```
0  Input: A program x, with its own input
1  Run H(x)
2  if H(x) answers Program x will halt then
3      begin an infinite loop
4  else halt
```

The program T is designed to run counter to x: If the input program x was going to halt, then T begins an infinite loop. And if the input program was going to run forever, then T says to halt

The program T accepts as input any program. And since T is itself a program, we are allowed to *plug T into itself!* What is the result? Well, since $T(T)$ is a program, like any program either $T(T)$ contains an infinite loop or it does not. Let's consider each of these two cases.

Case 1: Observe that if $T(T)$ has an infinite loop, then like all programs with infinite loops, it will not halt — but by looking at the above pseudocode for T, it is clear that if $T(T)$ has an infinite loop, then it will halt! This is a contradiction.

Case 2: Conversely, if $T(T)$ does not have an infinite loop, then like all programs without an infinite loop it must eventually halt — but by looking at the above pseudocode for T, it is clear that if $T(T)$ will eventually halt, then it will begin an infinite loop which will prevent it from halting! This is again a contradiction.

Whether T does or does not have an infinite loop, we have reached a contradiction. And since T was built from H, our assumption that there exists a halting program H must have been incorrect. This concludes the proof. ■

- **Proof by contradiction example 7**:

  Proposition. Every natural number is interesting

  Proof. Assume for a contradiction that not every natural number is interesting. Then, there must be a smallest uninteresting number, which we call n. But being the smallest uninteresting number is a very interesting property for a number to have! So n is both uninteresting and interesting, which gives the contradiction. Therefore, every natural number must be interesting. ■

- **Proof by minimal counterexample**: We proved that every natural number is interesting. The way we did this was by assuming for a contradiction that not every number is interesting. Under this assumption, there exist uninteresting natural numbers, and so there must exist a smallest uninteresting natural number.

  Despite it being a silly example, there is an important idea behind it which is sometimes called *proof by minimal counterexample*. Consider a theorem which asserts something is true for every natural number, and you are attempting to prove it by contradiction. Then you would assume for a contradiction not every natural number satisfies the result — that is, you're assuming there is at least one counterexample. Well, among all of the counterexamples, one of them must be the smallest. And thinking about that smallest counterexample — such as the smallest uninteresting number — can at times be a powerful variant of proof by contradiction.

  We used strong induction to prove the fundamental theorem of arithmetic. But there's another slick proof of this theorem that uses a proof by minimal counterexample

  Theorem (*Fundemental theorem of arithmetic*). Every integer $n \geqslant 2$ is either prime or a product of primes.

Recall that every integer $n \geqslant 2$ is either prime or composite, and being composite means it is a product of smaller integers

Proof. Assume for a contradiction that this is not true. Then there must be a minimal counterexample; let's say N is the smallest natural number at least 2 which is neither prime nor the product of primes. The fact that it is not prime means that it is composite: $N = ab$ for some $a, b \in \{2, 3, \ldots, N-1\}$.

We now make use of the fact that N is assumed to be the minimal counterexample to this result — which means that everything smaller than N must satisfy the result. In particular, since a and b are smaller than this smallest counterexample, a and b must each be prime or a product of primes.

And this gives us a contradiction: Since $N = ab$, if a and b are each prime or a product of primes, then their product — which equals N — must be as well. This contradicts our assumption that N was a counterexample, completing the proof.

Another way to think about this proof is that it argues that if N were a counterexample, then since $N = ab$, it can't possibly be that both a and b are primes or a product of primes, since as we just saw, that would produce a contradiction. And therefore, it must be the case that either a or b is also a counterexample. This implies that every counterexample produces a smaller counterexample — every N produces an a or a b. But this is a contradiction, since you can not repeatedly find smaller and smaller natural numbers — at some point you reach the bottom.

- **Proof of the division algorithm**

Theorem (*The division algorithm*): For all integers a and m with $m > 0$, there exist unique integers q and r such that

$$a = mq + r,$$

where $0 \leqslant r < m$.

Proof. **Existence.** First, note that if $a = 0$, then by simply choosing $q = 0$ and $r = 0$, the theorem follows. Thus, we may assume that $a \neq 0$.

Next, we will argue that if the theorem holds for all positive a, then it also holds for all negative a. Indeed, assume that $a > 0$, and suppose a and m can be expressed as

$$a = mq + r,$$

where $0 \leqslant r < m$. Then, $-a$ has an expression as well. In particular, if we let $q' = -q - 1$ and $r' = m - r$, then

$$mq' + r' = m(-q-1) + (m-r) = -mq - m + m - r = -(mq + r) = -a.$$

Therefore, for these integers q' and r',

$$-a = mq' + r',$$

where $0 \leqslant r' < m$. Because of this, any expression for $a > 0$ immediately produces one for $-a$. Thus, we need only prove the case where a is a positive integer.

We will implement a proof by minimal counterexample in order to prove the case where a is positive. Fix any $m > 0$, and assume for a contradiction that not every $a \in \mathbb{N}$ satisfies the theorem, which in turn means that there is a smallest a for which the theorem fails. Consider three cases.

Case 1: $a < m$. In this case, we can simply let $q = 0$ and $r = a$, and we have obtained

$$a = m \cdot q + r,$$

with $0 \leqslant r < m$, and the theorem is satisfied.

Case 2: $a = m$. In this case, we can simply let $q = 1$ and $r = 0$, and we have obtained

$$a = m \cdot q + r,$$

with $0 \leqslant r < m$, and the theorem is satisfied.

Case 3: $a > m$. Recall that the theorem assumes that $m > 0$, and so in this case we have $a > m > 0$. In particular, note that $a > a - m$ and also $a - m > 0$.

Since a is the smallest positive counterexample to this theorem, and $a - m$ is both positive and less than a, the integer $a' = a - m$ must satisfy this theorem! That is, there must exist integers d and s for which

$$(a - m) = m \cdot d + s,$$

with $0 \leqslant s < m$. By moving the m on the left side over,

$$a = m \cdot d + s + m.$$

By factoring,

$$a = m \cdot (d + 1) + s.$$

Thus, by letting $q = d+1$ and $r = s$, we have shown that our smallest counterexample is not a counterexample at all:

$$a = m \cdot q + r,$$

with $0 \leqslant r < m$. Since there cannot exist a smallest counterexample, there cannot exist any counterexample. Thus, for each a and m, there must exist a q and r as the theorem asserts.

Uniqueness. Assume for a contradiction that for our fixed a and m, the q and r are not unique. That is, assume there exist two different representations of a:

$$a = mq + r \quad \text{and} \quad a = mq' + r',$$

where $q, r, q', r' \in \mathbb{Z}$ and $0 \leqslant r, r' < m$. Then,

$$mq + r = mq' + r'.$$

By some algebra, we find:

$$r - r' = mq' - mq,$$

which means

$$r - r' = m(q' - q).$$

Since q and q' are integers, so is $q-q'$ (by Fact 2.1), which means the above expression matches the definition of divisibility (Definition 2.8)! That is, $m \mid (r - r')$.

Notice that since $0 \leqslant r, r' < m$, the difference $r - r'$ would have these restrictions:

$$-m < r - r' < m.$$

And the only number in this range which is divisible by m is zero. That is, $r - r' = 0$, or $r = r'$.

Next, since $r = r'$, the fact that $r - r' = m(q - q')$ implies that

$$0 = m(q - q').$$

Since $m > 0$, we may divide both sides by m, which means $0 = q - q'$, or $q = q'$.

We assumed that

$$a = mq + r \quad \text{and} \quad a = mq' + r'$$

were two different representations of a and m, but we have proven that $q = q'$ and $r = r'$, proving that they are in fact the same representation, giving the contradiction and concluding the proof.

1.8 Functions

- **The definition of a function**: Given a pair of sets A and B, suppose that each element $x \in A$ is associated, in some way, to a unique element of B, which we denote $f(x)$. Then f is said to be a function from A to B. This is often denoted $f : A \to B$.

  Furthermore, A is called the **domain** of f, and B is called the **codomain** of f.

  The set $\{f(x) : x \in A\}$ is called the **range** of f.

- **The *Existence*, and *uniqueness* property of functions**: When discussing functions, the ideas of existence and uniqueness will come up repeatedly. We defined a function $f : A \to B$ to be a rule which sends each $x \in A$ to some $f(x) \in B$. What this means is that $f(x)$ must exist (it must be equal to some $b \in B$), and it must be unique (it must be equal to only one $b \in B$).

  For example, defining $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \ln(x)$ fails the *existence* requirement of functions, because the natural logarithm function $\ln(x)$ is not defined for negative values of x or $x = 0$. his means that the function $\ln(x)$ would fail the requirement of existence for all elements in the domain $\mathbb{R}$.

  To make $f(x) = \ln(x)$ a valid function, we must adjust the domain to only include values for which $\ln(x)$ is defined. The correct domain is $(0, \infty)$, the set of positive real numbers. Thus, we would write

  $$f : (0, \infty) \to \mathbb{R}.$$

  A "function" that fails the uniqueness requirement of functions would assign a single element in the domain to more than one element in the codomain.

  Consider a rule $f : A \to B$ defined as

  $$f(x) = \begin{cases} b_1 & \text{if } x = a \\ b_2 & \text{if } x = a \end{cases}.$$

  Where $b_1 \neq b_2$, and $a \in A$. This rule clearly violates the *uniqueness* criterion, and is therefore not a function.

  In high school you were probably taught the *vertical line test* to check whether a graph corresponds to a function. The vertical line test says that if every vertical line hits the graph in one (existence) and only one (uniqueness) spot, then the graph corresponds to a function

- **Injections, Surjections and Bijections**: A function $f : A \to B$ is injective (or one-to-one) if $f(a_1) = f(a_2)$ implies that $a_1 = a_2$.

  The contrapositive of the second half states, A function $f : A \to B$ is *injective* if $a_1 \neq a_2$ implies that $f(a_1) \neq f(a_2)$

  A function $f : A \to B$ is surjective (or onto) if, for every $b \in B$, there exists some $a \in A$ such that $f(a) = b$

  Let's take a look at another way to define this same idea, by again applying the contrapositive (and doing a little rearranging).

A function $f : A \to B$ is surjective (or onto) if there does not exist any $b \in B$ for which $f(a) \neq b$ for all $a \in A$.

When defining a function $f : A \to B$, the ideas of existence and uniqueness were focused on A — for every $x \in A$, we demanded that $f(x)$ exist and be unique. To be injective and surjective, the attention shifts to B. To be surjective means that B has an existence criterion (for every $b \in B$, there exists some $a \in A$ that maps to it). And to be injective means that B has a uniqueness-type criterion (for every $b \in B$, there is at most one $a \in A$ that maps to it).

A function $f : A \to B$ is *bijective* if it is both injective and surjective.

Defining a function $f : A \to B$ placed existence and uniqueness criteria on A. If f is both injective and surjective, then this adds existence and uniqueness criteria to B. Thus, if f is a bijection, then it has these criteria on both sides: Every $a \in A$ is mapped to precisely one $b \in B$, and every $b \in B$ is mapped to by precisely one $a \in A$. In effect, this pairs up each element of A with an element of B; namely, a is paired with $f(a)$ in this way.

- **Proving $x$jectiveness for $x \in$ {in,sur,bi}**: Based on its definition, this is the outline to prove a function is injective.

Proposition. $f : A \to B$ is an injection

Proof. Assume $x, y \in A$, and $f(x) = f(y)$

$\vdots$ Apply algebra,

$\vdots$ logic, techniques.

Therefore, $x = y$

Since $f(x) = f(y)$ implies $x = y$, f is injective ■

Alternatively, one could use the contrapositive, which would mean one starts by assuming $x \neq y$, and then concludes that $f(x) \neq f(y)$.

Next, here's the outline for a surjective proof.

Proposition. $f : A \to B$ is a surjection

Proof. Assume $b \in B$

$\vdots$ Magic to find an $a \in A$

$\vdots$ where $f(a) = b$.

Since every $b \in B$ has an $a \in A$ where $f(a) = b$, f is surjective ■

- **Proving jectiveness examples**

- $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = x^2$ is not injective, surjective, or bijective.
- $g : \mathbb{R}^+ \to \mathbb{R}$ where $g(x) = x^2$ is injective, but not surjective or bijective.
- $h : \mathbb{R} \to \mathbb{R}^+$ where $h(x) = x^2$ is surjective, but not injective or bijective.
- $k : \mathbb{R}^+ \to \mathbb{R}^+$ where $k(x) = x^2$ is injective, surjective, and bijective.

Proof (part a). Observe that $f(-2) = f(2) = 4$, while $-2 \neq 2$. Thus, f is not injective. Next, notice that $f(x) = x^2 > 0$. Thus, there is no such $a \in \mathbb{R}$ such that $f(a) = -4$. Since -4 is in the codomain and is not hit, f is not surjective. Since f is not both injective and surjective, it is therefore not bijective.

Part b. Let $a_1, a_2 \in \mathbb{R}^+$, assume $g(a_1) = g(a_2)$. Thus,

$$\begin{aligned} & a_1^2 = a_2^2 \\ \implies & a_1 = \pm a_2. \end{aligned}$$

But, for all $a \in \mathbb{R}^+$, $a > 0$. Thus, $a_1 = a_2$ and g is injective. Observe that again there is no such value in the domain of g such that $g(x) = -4$. Since -4 is in the codomain of g, it is not surjective, and is therefore not bijective.

Part c. Observe that $h(-2) = h(2) = 4$, while $-2 \neq 2$. Thus, h is not injective. Further, let $b \in \mathbb{R}^+$, then

$$\begin{aligned} & h(a) = b \\ \implies & a^2 = b \\ \implies & a = \pm b. \end{aligned}$$

But, the codomain is restricted to positive values, thus $a = b$ and h is surjective. Since h is not injective, it is not bijective.

Part d. Let $a_1, a_2 \in \mathbb{R}^+$, assume $f(a_1) = f(a_2)$, which implies

$$\begin{aligned} & a_1^2 = a_2^2 \\ \implies & a_1 = \pm a_2. \end{aligned}$$

Again, since the domain is restricted to positive values, we have $a_1 = a_2$ and f is injective. Next, let $b \in \mathbb{R}^+$, then

$$\begin{aligned} & f(a) = b \\ \implies & a^2 = b \\ \implies & a = \pm b. \end{aligned}$$

But since the codomain is restricted to positive values, $a = b$ and the function is surjective. Since the function is both onto and one-to-one, the function is bijective (invertible). ■

- **Proving jectiveness example 2.** Show $f : (\mathbb{Z} \times \mathbb{Z}) \to (\mathbb{Z} \times \mathbb{Z})$, with $f(x, y) = (x + 2y, 2x + 3y)$ is a bijection.

Proof. First, we show injectiveness. Let $(a,b),(c,d) \in \mathbb{Z}^2$. Assume $f(a,b) = f(c,d)$. Thus,

$$\begin{aligned} &(a+2b, 2a+3b) = (c+2d, 2c+3d) \\ \implies &\begin{cases} a+2b &= c+2d \\ 2a+3b &= 2c+3d \end{cases} \\ \implies &\begin{cases} a+2b-2a-3b &= 0 \\ 2a+3b-2c-3d &= 0 \end{cases}. \end{aligned}$$

We then solve this system,

$$\left.\begin{array}{cccc|c} 1 & 2 & -1 & -2 & 0 \\ 2 & 3 & -2 & -3 & 0 \end{array}\right. \implies \left.\begin{array}{cccc|c} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{array}\right..$$

Which implies

$$\begin{cases} a &= c \\ b &= d \end{cases}$$

As desired. Thus, f is injective. Next, let $(c,d) \in \mathbb{Z}^2$. Require $f(a,b) = (c,d)$ for some $(a,b) \in \mathbb{Z}^2$. Thus,

$$\begin{aligned} &(a+2b, 2a+3b) = (c,d) \\ \implies &\begin{cases} a+2b &= c \\ 2a+3b &= d \end{cases}. \end{aligned}$$

Solving this system yields

$$\left.\begin{array}{cc|c} 1 & 2 & c \\ 2 & 3 & d \end{array}\right. \implies \left.\begin{array}{cc|c} 1 & 0 & -3c+2d \\ 0 & 1 & 2c-d \end{array}\right..$$

Thus, $(a,b) = (-3c+2d, 2c-d)$ and the function is surjective. Because the function is both injective and surjective, it is therefore bijective.

Alternatively, observe that $f:\ \mathbb{Z}^2 \to \mathbb{Z}^2,\ f(x,y) = (x+2y, 2x+3y)$ is given by the matrix representation $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$

$$\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Thus, since A is square, we can simply check its determinant. [4]

$$\det \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} = 1(2) - 2(3) = -1.$$

Since $\det(A) \neq 0$, the function is invertible

- **The func-y pigeonhole principal**:

  Theorem 8.10 (The func-y pigeonhole principal): Suppose A and B are finite sets and $f:\ A \to B$ is any function.

  (a) If $|A| > |B|$, then f is not injective.

  (b) If $|A| < |B|$, then f is not surjective.

[4] Common linear algebra W

Proof. **Part (a).** Consider each element in A to be an object and each element of B to be a box. Given an $a \in A$, place object a into box b if $f(a) = b$. Since there are more objects than boxes, by the pigeonhole principal at least one box has at least two objects in it. That is, $f(a_1) = f(a_2)$ for some distinct a_1 and a_2, implying that f is not injective.

Part (b). Since f is a function, each $a \in A$ is mapped to only one $b \in B$. Thus, k elements in A can map to at most k elements of B. And so the $|A|$ elements in A can map to at most $|A|$ elements in B. However, since $|A| < |B|$, there must be some elements not hit, meaning that f is not surjective.

It is again useful to think about what the contrapositive tells us:

(a) If f is injective, then $|A| \leqslant |B|$.
(b) If f is surjective, then $|A| \geqslant |B|$.

Viewing the statements this way is beneficial for another reason: It demonstrates clearly that in order for f to be a bijection—meaning an injection and a surjection—we would need $|A| = |B|$.

It is also worth mentioning that this theorem still holds true in the case that $|A|$ and/or $|B|$ are infinite.[5]

- **The Composition**: Let A, B, and C be sets, $g : A \to B$, and $f : B \to C$. Then the composition function is denoted $f \circ g$ and is defined as follows:

$$(f \circ g) : A \to C \quad \text{where} \quad (f \circ g)(a) = f(g(a)).$$

Suppose

$$g : \mathbb{R} \to \mathbb{R},\ g(x) = x + 1$$
$$f : \mathbb{R} \to \mathbb{R}^+,\ f(x) = x^2.$$

Then,

$$(f \circ g) : \mathbb{R} \to \mathbb{R}^+,\ (f \circ g)(x) = (x + 1)^2.$$

- **Property of injective functions under composition**:

Theorem 8.13. Suppose A, B and C are sets, $g : A \to B$ is injective, and $f : B \to C$ is injective. Then $f \circ g$ is injective

Proof. Since $(f \circ g) : A \to C$, to show that is an injection we must show that for all $a_1, a_2 \in A$, $(f \circ g)(a_1) = (f \circ g)(a_2)$ implies $a_1 = a_2$. Assume $a_1, a_2 \in A$, and $(f \circ g)(a_1) = (f \circ g)(a_2)$. Using the definition of the composition, we have

$$f(g(a_1)) = f(g(a_2)).$$

Since f is injective, we know that for any $b_1, b_2 \in B$, $f(b_1) = f(b_2)$ implies $b_1 = b_2$. Since $g(a_1), g(a_2) \in B$, we have

$$g(a_1) = g(a_2).$$

Likewise, since g is injective, it must be that $a_1 = a_2$

[5]But proving this to be the case would take us too far afield.

Thus, we have shown that for any $a_1, a_2 \in A$, if $(f \circ g)(a_1) = (f \circ g)(a_2)$, then $a_1 = a_2$. Therefore, $(f \circ g)$ is an injection. ■

- **Property of surjective functions under composition**:

  Theorem 8.14: Suppose A, B and C are sets, $g : A \to B$ is surjective, and $f : B \to C$ is surjective. Then $f \circ g$ is surjective.

  Proof. Since $(f \circ g) : A \to C$, to show that $f \circ g$ is surjective, we must show that for all $c \in C$, there exists some $a \in A$ such that $(f \circ g)(a) = c$. To start, since f is surjective, then for all $c \in C$, there exists some $b \in B$ such that $f(b) = c$. Further, we know that g is surjective. Thus, for all $b \in B$, there exists some $a \in A$ such that $g(a) = b$.

  Thus, for an arbitrary $c \in C$, we have found an $a \in A$ such that

  $$(f \circ g)(a) = f(g(a)) = f(b) = c.$$

  Completing the proof ■

- **A corollary from the above two results**: Suppose A, B and C are sets, $g : A \to B$ is bijective, and $f : B \to C$ is bijective. Then $f \circ g$ is bijective.

  Proof. By Theorem 8.13, $f \circ g$ is an injection. By Theorem 8.14, $f \circ g$ is a surjection. Thus, by the definition of a bijection (Definition 8.7), $f \circ g$ is a bijection.

- **Note about compositions**: Notice that in our definition of function composition (Definition 8.11) we had functions g and f where $g : A \to B$, and $f : B \to C$. Notice that we don't really need the codomain of g to equal the domain of f. If we had $g : A \to B$ and $f : D \to C$ where $B \subseteq D$, that would be enough (for the definition, and for these last two theorems). As long as $g(a)$ is a part of f's domain, then $f(g(a))$ will make sense, which is all we need.

- **Identity function and invertibility**: For a set A, the identity function on A is the function

  $$i_A : A \to A \text{ where } i_A(x) = x \text{ for every } x \in A$$

  The inverse of a function $f : A \to B$, if it exists, is the function $f^{-1} : B \to A$ such that $f^{-1} \circ f = i_A$ and $f \circ f^{-1} = i_B$.

  For example, if $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = x + 1$, then $f^{-1} : \mathbb{R} \to \mathbb{R}$ is the function $f^{-1}(x) = x - 1$. To see this, simply note that

  $$(f \circ f^{-1})(x) = f(f^{-1}(x)) = f(x - 1) = (x - 1) + 1 = x$$

  and

  $$(f^{-1} \circ f)(x) = f^{-1}(f(x)) = f^{-1}(x + 1) = (x + 1) - 1 = x.$$

- **Arctan and the natural logarithm**: this is a great opportunity to mention a couple important functions — $\arctan(x)$ and $\ln(x)$ — which are defined as the inverses to other important function.

  - If $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$ is the tangent function, then its inverse is defined to be $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$, and is called the arctangent function.[6]
  - If $\exp : \mathbb{R} \to \mathbb{R}^+$ is the exponential function (that is, $\exp(x) = e^x$), then its inverse is defined to be $\ln : \mathbb{R}^+ \to \mathbb{R}$, and is called the natural logarithm function.

- **When does an inverse exist**:

[6]

Theorem: A function $f: A \to B$ is invertible if and only if f is a bijection.

Proof. First, suppose that $f : A \to B$ is invertible. We will prove that f is both an injection and a surjection, which will prove that f is a bijection. To see that f is a surjection, choose any $b \in B$. We aim to find an $a \in A$ such that $f(a) = b$. To this end, let $a = f^{-1}(b)$, which exists and is in A because $f^{-1} : B \to A$. Now simply observe that the definition of an invertible function (Definition 8.16) implies

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

This proves that f is a surjection.

To see that f is an injection, let $a_1, a_2 \in A$ and assume $f(a_1) = f(a_2)$. Note that $f(a_1)$ (and hence $f(a_2)$, since they're equal) is an element of B due to the fact that $f : A \to B$. And so, since $f^{-1} : B \to A$, we may apply f^{-1} to both sides:

$$f(a_1) = f(a_2)$$

$$f^{-1}(f(a_1)) = f^{-1}(f(a_2))$$

$$a_1 = a_2,$$

by the definition of the inverse. Thus, f is an injection. And since we already showed that f is a surjection, it must be a bijection. This concludes the forward direction of the theorem.

As for the backwards direction, assume that f is a bijection. For $b \in B$, we will now define $f^{-1}(b)$ like this:

$$f^{-1}(b) = a \quad \text{if} \quad f(a) = b.$$

That is, we are defining f^{-1} to act as an inverse from B to A should act, without yet claiming that f^{-1} is a function. Our goal now is to demonstrate that this definition of f^{-1} satisfies the conditions to be a function, which would prove that f is invertible. To do so, recall that to be a function there is an existence condition ($f^{-1}(b)$ must be equal to some $a \in A$) and a uniqueness condition ($f^{-1}(b)$ must be equal to only one $a \in A$). We will check these separately.

Existence: Let $b \in B$. Since f is surjective, there must be some $a \in A$ such that $f(a) = b$. Hence, by our definition of f^{-1}, we have $f^{-1}(b) = a$. We have shown that for every $b \in B$ there exists at least one $a \in A$ for which $f^{-1}(b) = a$, which concludes the existence portion of this argument.

Uniqueness: Suppose $f^{-1}(b) = a_1$ and $f^{-1}(b) = a_2$, for some $b \in B$ and $a_1, a_2 \in A$. By the definition of f^{-1}, this means that $f(a_1) = b$ and $f(a_2) = b$. But since f is injective, this means that $a_1 = a_2$. We have shown that $f^{-1}(b)$ can not be equal to two different elements of A, which concludes the uniqueness portion of this argument.

Combined, these two parts show that $f^{-1} : B \to A$ is a function, hence proving that f is invertible.

We have proved the forwards and backwards directions of Theorem 8.17, which completes its proof. □

- **The image and inverse image**: Let $f : A \to B$ be a function, and assume $X \subseteq A$ and $Y \subseteq B$. The *image* of A is

$$f(X) = \{y \in B : y = f(x) \text{ for some } x \in X\},$$

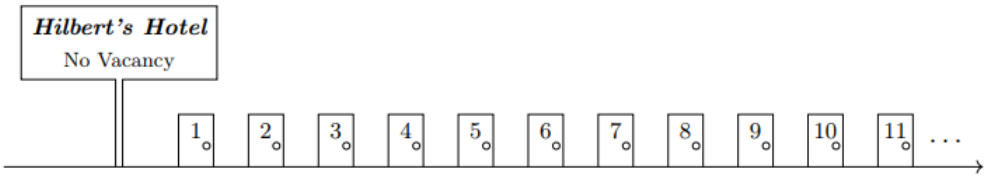
and the *inverse image* of Y is

$$f^{-1}(Y) = \{x \in A : f(x) \in Y\}.$$

- **The bijection principal**:

principal (*The bijection principal.*) Two sets have the same size if and only if there is a bijection between them.

- **Hilbert's hotel**: We begin by talking about the set of problems related to the so-called Hilbert's Hotel. Assume that there is a hotel, called Hilbert's Hotel, which has infinitely many rooms in a row.



  - Assume every room has someone in it, and so the "No Vacancy" sign has been turned on. With most hotels, this would mean that if someone else arrives at the hotel, they will not be given a room. But this isn't the case with Hilbert's Hotel. If, for $n \in \mathbb{N}$, the patron in room n moves to room $n + 1$, then nobody is left without a room and suddenly room 1 is completely open! So the new customer can go to room 1. We created a room out of nothing!
  - Now imagine 2 people arrived at the hotel. Can we accommodate them? Certainly! Now, just have everyone move from room n to room $n + 2$. This leaves rooms 1 and 2 open to the newcomers, and we are again good-to-go.
  - What if, however, we have infinitely many people lined up wanting a room? Can we accommodate all of them? Yes! We still can! Just have the person in room n move to room $2n$. Then all of the odd-numbered rooms are vacant and the infinite line of people can take these rooms.

  The first point of this exercise is to simply realize that weird stuff can happen when dealing with the infinite. The second point, though, is to realize that each time the people switched rooms, those same exact people got new rooms. So in the first example when they each just moved one room down, that should mean that there are just as many rooms from 1 to ∞ as there are from 2 to ∞. . . And likewise for the others.

- **Cardinality and infinite sets**:

Example There are the same number of natural numbers as there are natural numbers larger than 1 (that is, $|\mathbb{N}| = |\{2, 3, 4, \dots\}|$). What's the bijection that shows this? Let

$$f : \mathbb{N} \to \{2, 3, 4, \dots\} \quad \text{where} \quad f(n) = n + 1.$$

In other (non-)words, this is the pairing

$$1 \leftrightarrow 2 \quad 2 \leftrightarrow 3 \quad 3 \leftrightarrow 4 \quad 4 \leftrightarrow 5 \quad \dots$$

The Moral. Two sets can have the same size even though one is a proper subset of the other.

Example. There are the same number of natural numbers as even natural numbers (that is, $|\mathbb{N}| = |2\mathbb{N}|$). What's the bijection that shows this? Let

$$f : \mathbb{N} \to \{2, 4, 6, 8, \dots\} \quad \text{where} \quad f(n) = 2n.$$

In other (non-)words, this is the pairing

$$1 \leftrightarrow 2 \quad 2 \leftrightarrow 4 \quad 3 \leftrightarrow 6 \quad 4 \leftrightarrow 8 \quad \dots$$

The Moral. Two sets can have the same size even though one is a proper subset of the other and the larger one even has *infinitely many more elements* than the smaller one.

And in a similar way, one can prove that $|\mathbb{N}| = |\mathbb{Z}|$. Indeed, a bijection $f : \mathbb{N} \to \mathbb{Z}$ can be given by following this pattern:

$$f(1) = 0, \quad f(2) = 1, \quad f(3) = -1, \quad f(4) = 2, \quad f(5) = -2, \quad f(6) = 3, \quad \dots$$

One way to write such a function is this:

$$f : \mathbb{N} \to \mathbb{Z} \quad \text{where} \quad f(n) = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even;} \\ -\frac{(n-1)}{2} & \text{if } n \text{ is odd.} \end{cases}$$

1.9 Relations

- **Set partitions**: A partition of a set A is a collection of non-empty subsets of A for which each element of A is in one and only one of the subsets.

  Formally, a partition is a collection of non-empty sets $\{P_i\}_{i \in S}$ such that

  1. $P_i \subseteq A$ for all i
  2. $\bigcup_{i \in S} P_i = A$
  3. $P_i \cap P_j = \varnothing$ for all $i \neq j$

  A partition of $\mathbb{Z}$ is the set of evens and the set of odds. Another partition of $\mathbb{Z}$ is the positive integers, the negative integers, and $\{0\}$. Another is the non-17 integers and $\{17\}$. Another is the five sets in the Mod-5 Property section on the previous page. And the simplest partition of $\mathbb{Z}$ is simply $\mathbb{Z}$ — a partition with only one part.

- **Index sets**: In the formal definition of a partition, S is the index set that labels or indexes the subsets P_i in the partition.

  S can be any set (e.g., $N, \{1, 2, ..., n\}$, or any other index set), as long as it provides unique labels for each subset P_i

- **Equivalence Relations**: An *equivalence relation* on a set A is an ordered relationship between pairs of elements of A for which the pair is either *related* or is *not related*. If $a, b \in A$, we denote $a \sim b$ if a is related to b, and $a \not\sim b$ if a is not related to b.

  For $\sim$ to be an equivalence relation, it also must satisfy the following three properties:

  - **Reflexive:** $a \sim a$ for all $a \in A$;
  - **Symmetric:** If $a \sim b$, then $b \sim a$ for all $a, b \in A$; and
  - **Transitive:** If $a \sim b$ and $b \sim c$, then $a \sim c$ for all $a, b, c \in A$.

  Lastly, if $\sim$ is an equivalence relation and $a \in A$, define the *equivalence class* containing a to be the set

  $$\{b \in A : a \sim b\}.$$

- **Relations**: A relation on a set A is any ordered relationship between pairs of elements of A for which the pair is either *related* or is *not related*. If $a, b \in A$, we denote $a \sim b$ if a is related to b, and $a \not\sim b$ if a is not related to b.

  Lastly, if $\sim$ is a relation and $a \in A$, define the class containing a to be the set

  $$\{b \in A : a \sim b\}.$$

- **Equivalence relations and partitions**:

  Theorem 9.5. Assume $\sim$ is a relation on A. The relation $\sim$ partitions the elements of A into classes if and only if $\sim$ is an equivalence relation.

  Before we prove this theorem, we first define some notation. We denote the equivalence class of an element $a \in A$, $\{x \in A : \ a \sim x\}$ by $[a]$.

  Next, a lemma.

Lemma 9.10. Suppose $\sim$ is an equivalence relation on a set A, and let $a, b \in A$. Then,

$$[a] = [b] \text{ if and only if } a \sim b$$

Proof of lemma 9.10. For the (straight)forward direction, assume that $[a] = [b]$. Observe that since $\sim$ is reflexive, $b \sim b$ and so $b \in [b]$. And since $[a] = [b]$, this in turn means that $b \in [a]$, which by Notation 9.9 implies $a \sim b$. This concludes the forward direction.

As for the backward direction, we begin by assuming $a \sim b$, and we aim to prove that $[a] = [b]$. This will be accomplished by demonstrating that $[a] \subseteq [b]$ and $[b] \subseteq [a]$. To prove the former, choose any $x \in [a]$; we will show that $x \in [b]$. By assumption we have $a \sim b$, and because $x \in [a]$ we have $a \sim x$. That is,

$$a \sim b \quad \text{and} \quad a \sim x.$$

By the symmetry property of $\sim$,

$$b \sim a \quad \text{and} \quad a \sim x.$$

By the transitivity property of $\sim$,

$$b \sim x.$$

And so, by Notation 9.9,

$$x \in [b].$$

We have shown that $x \in [a]$ implies $x \in [b]$, and hence $[a] \subseteq [b]$.

The reverse direction is nearly the same. Let $x \in [b]$, which means $b \sim x$. Combining this, the transitivity of $\sim$, and our assumption that $a \sim b$, we get $a \sim x$, which means $x \in [a]$. And since $x \in [b]$ implies $x \in [a]$, we have $[b] \subseteq [a]$.

We have shown that $[a] \subseteq [b]$ and $[b] \subseteq [a]$, which proves that $[a] = [b]$. This concludes the backward direction, and hence the proof. 😊

We now proceed to the proof of theorem 9.5

- **Equivalence relation example 1**: Let $\sim$ be the relation on $\mathbb{R}$ where

$$a \sim b \text{ if } \lfloor a \rfloor = \lfloor b \rfloor$$

We can verify that $\sim$ is an equivalence relation by checking that it satisfies the three criteria. It is reflexive because certainly $\lfloor a \rfloor = \lfloor a \rfloor$ for any $a \in \mathbb{R}$; it is symmetric because if $\lfloor a \rfloor = \lfloor b \rfloor$, then certainly $\lfloor b \rfloor = \lfloor a \rfloor$; and it is transitive because if $\lfloor a \rfloor = \lfloor b \rfloor$ and $\lfloor b \rfloor = \lfloor c \rfloor$, then $\lfloor a \rfloor = \lfloor c \rfloor$. Each of these is immediate because the equal sign already has these properties.

This means that the equivalence classes must then partition all of $\mathbb{R}$, and indeed they do. The class of all numbers that are equivalent to 12.4 is the set of numbers in the interval $[12, 13)$; that is, all numbers x such that $12 \leqslant x < 13$. Indeed, the equivalence classes for $\sim$ are all intervals of the form $[n, n+1)$ for $n \in \mathbb{Z}$.

Moreover, by Theorem 9.5 this means that the equivalence classes must then partition all of $\mathbb{R}$, and they do: every $x \in \mathbb{R}$ is in precisely one of these intervals:

$$\ldots, [2, 3), [3, 4), [4, 5), [5, 6), [6, 7), \ldots.$$

😊

Elementary fields, groups, and rings

- **Modular congruence and congruence classes**: Recall that two integers a and b are said to be congruent modulo n if they leave the same remainder when divided by n. Mathematically, this is written as

$$a \equiv b \pmod{n}.$$

Which means

$$n \mid a - b.$$

When an integer a is divided by n

$$a = q_1 n + r_1 \quad \text{with} 0 \leqslant r_1 < n.$$

Similarly, for an integer b divided by n

$$b = q_2 n + r \quad \text{with} 0 \leqslant r_2 < n.$$

Subtracting b from a

$$a - b = (q_1 - q_2)n + (r_1 - r_2). \tag{1}$$

If $n \mid (a - b)$,

$$a - b = nk, \quad k \in \mathbb{Z}.$$

By (1) above, we have

$$(q_1 - q_2)n + (r_1 - r_2) = kn.$$

For this to hold, we require $r_1 - r_2$ to be a multiple of n, since $q_1 - q_2$ is already a multiple of n. Since r_1, r_2 satisfy $0 \leqslant r_1, r_2 < n$. It must be that $-n < r_1 - r_2 < n$. In this case, for n to divide $r_1 - r_2$. It must be that

$$r_1 - r_2 = 0.$$

Which implies $r_1 = r_2$. Hence, a and b have the same remainder when divided by n when $n \mid a - b$.

A congruence class modulo n is the set of all integers that are congruent to a particular integer a modulo n. This set is denoted as

$$[a]_n = \{x \in \mathbb{Z} \mid x \equiv a \pmod{n}\}.$$

For example, $[0]_3$ is

$$\{x \in \mathbb{Z} : \ x \equiv 0 \pmod{3}\}$$

.

Which is the integers x such that $3 \mid x - 0$. In other words, it describes the set of integers that are divisible by 3.

The set $[1]_3$ is the set

$$[1]_3 = \{x \in \mathbb{Z} : \ x \equiv 1 \pmod{3}\}.$$

Which implies $3 \mid x - 1$, and thus $x = 3k + 1$, for $k \in \mathbb{Z}$. In words, it is the set of integers that leave a remainder of one when divided by three.

The modulus n partitions the integers into n distinct congruence classes:

$$[0]_n, [1]_n, ..., [n-1]_n.$$

Every integer belongs to exactly one of these classes.

Arithmetic operations can be performed within the framework of congruence classes

- **Addition**: If $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then

$$a + c \equiv b + d \pmod{n}.$$

- **Multiplication**: If $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then

$$ac \equiv bd \pmod{n}.$$

- **Groups**: A group is a collection of objects G, together with one operation $\oplus$, which has the following properties:
  - **Associativity**: $a \oplus (b \oplus c) = (a \oplus b) \oplus c$
  - **Identity**: There is an element $e \in G$ such that $e \oplus g = g \oplus e = g$ for all $g \in G$
  - **Inverse**: For every $g \in G$, there exists $g^{-1} \in G$ such that $g \oplus g^{-1} = g^{-1} \oplus g = e$

  For example, $\mathbb{Z}$ is a group under addition.
  - **Associativity**: Two integers a, b are associative, $a + (b + c) = (a + b) + c$
  - **Identity**: Zero is the identity element, since $0 \in \mathbb{Z}$ and $0 + a = a + 0 = a$
  - **Inverse**: $a + (-a) = (-a) + a = 0$

  Note: A group is said to be *abelian* if it is commutative under its operation. In other words, $x \oplus y = y \oplus x$ for all $x, y \in G$

- **Rings**: A ring is a set R, together with two operations $\oplus$ and $*$, which has the following properties
  - R is an abelian group under $\oplus$
  - R is associative under $*$
  - The operation $*$ distributes over $\oplus$

$$a * (b \oplus c) = (a * b) \oplus a * c$$
$$(a \oplus b) * c = (a * c) \oplus (b * c).$$

  For example, $\mathbb{Z}$ is a ring under addition and multiplication. First note that $\mathbb{Z}$ is an abelian group under addition. Further, for $a, b \in \mathbb{Z}$, $a \cdot b = b \cdot a$.

  $1 \in \mathbb{Z}$ is the identity, $1 \cdot a = a \cdot 1 = a$ for all $a \in \mathbb{Z}$, and we know that multiplication distributes over addition

$$a \cdot (b + c) = a \cdot b + a \cdot c$$
$$(a + b) \cdot c = a \cdot c + b \cdot c.$$

- **Fields**: A field is a set F, together with two operations $\oplus$ and $*$, which has the following properties
  - F is a commutative ring under $\oplus$ and $*$

- Every nonzero $f \in F$ has a multplicative inverse, that is, some element $g \in F$ for which

$$f * g = g * f = 1.$$

The sets $\mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$ under addition and multiplication are examples of fields. The set of integers $\mathbb{Z}$ is not. Although it is a commutative ring under addition and multiplication, not every element has a multiplicative inverse. For example, there is no such $a \in \mathbb{Z}$ such that $2 \cdot g = 1$

- **Vector spaces**: A vector space is a set of vectors V, together with a set of scalars F, with the following properties
  - V is a abelian group under vector addition
  - F is a field under multiplication
  - For each $s \in F$, and $\mathbf{v} \in V$, scalar multiplication gives a unique element $s \cdot \mathbf{v} \in V$
  - Additional properties

$$\begin{aligned} 1\mathbf{v} &= \mathbf{v} \\ a(b\mathbf{v}) &= (ab)\mathbf{v} \\ a(\mathbf{u} + \mathbf{v}) &= a\mathbf{u} + a\mathbf{v} \\ (a + b)\mathbf{v} &= a\mathbf{v} + b\mathbf{v}. \end{aligned}$$

Combinatorics

3.1 Introduction

- **What is combinatorics?**: Combinatorics is a collection of techniques and a language for the study of finite or countably infinite discrete structures. Given a set of elements and possibly some structure on that set, typical questions are
  - Does a specific arrangement of the elements exists?
  - How many such arrangemets are there?
  - What properties do these arrangements have?
  - Which one of the arrangemetns is maximal, minimal, or optimal according to some criterion?
- **Counting the number of subsets for a set**: Let $[n] = \{1, 2, ..., n\}$, and let $f(n)$ be the number of subsets of $[n]$. Then $f(n) = 2^n$. For any particular subset of $[n]$, each element is either in that subset or not. Thus, to construct a subset, we have to make one of two choices for each element of $[n]$. Furthermore, these choices are independent of each other. Hence, the total number of choices, and consequently the total number of subsets is

$$\underbrace{2 \times 2 \times ... \times 2}_{n} = 2^n.$$

- **Number of subsets without consecutive integers**: For a sequence $[n] = \{1, ..., n\}$ we can count the number of subsets given by $f(n)$, that do not contain consecutive integers with the recurrence relation

$$f(n) = f(n-1) + f(n-2).$$

We consider two cases

1. n in not included in the subsets
2. n is included in the subsets. In this case, we build the subsets considering the subsequence $[n-2] = \{1, ..., n-2\}$. Note that if we include n, we must exclude $n-1$, because $n-1$ and n are consecutive, this will become cleare in the upcoming example.

Consider the sequence $[n] = \{1, 2, 3, 4\}$. By the relation above,

$$f(4) = f(3) + f(2).$$

Before we are able to compute this, we must define our base cases.

$$f(n) = \begin{cases} 3 & \text{if } n = 2 \\ 2 & \text{if } n = 1 \end{cases}.$$

If $n = 2$, we have $\{1, 2\}$, and the allowed subsets are $\varnothing, \{1\}, \{2\}$. If we have $n = 1$, the subsets are $\{\varnothing, \{1\}\}$. Thus

$$\begin{aligned} f(4) &= f(3) + f(2) = f(2) + f(1) + f(2) \\ &= 3 + 2 + 3 = 8. \end{aligned}$$

Let's explicitly break up the given sequence so we can see whats going on. In the first case, n is excluded, thus the sequence becomes $\{1,2,3\}$. If n is included, the sequence becomes $\{1,2\}$, where we build the subsets of $\{1,2\}$, and then add 4 to each one. Thus,

$$\begin{aligned}\{1,2,3\}+\{1,2\} &= \{1,2,3\}+\varnothing+\{1\}+\{2\}\\ &= \{1,2,3\}+\{4\}+\{1,4\}+\{2,4\}.\end{aligned}$$

Since the sequence $\{1,2,3\}$ in not a base case, we must split this one up aswell, we have

$$\begin{aligned}\{1,2,3\} &= \{1,2\}+\{1\}+\{4\}+\{1,4\}+\{2,4\}\\ &= \varnothing+\{1\}+\{2\}+\varnothing+\{1\}+\{4\}+\{1,4\}+\{2,4\}\\ &= \varnothing+\{1\}+\{2\}+\{3\}+\{1,3\}+\{4\}+\{1,4\}+\{2,4\}.\end{aligned}$$

Thus, we conclude all "good" subsets of $[n]$ either have n or don't have n. The ones that don't have n are exactly the "good" subsets of $[n-1]$. The "good" subsets of $[n]$ that include n are exactly the "good" subsets of $[n-2]$ together with n. Thus $f(n)=f(n-1)+f(n-2)$ ■

3.2 Induction and recurrence relations

- **Principal of Mathematical Induction**: Given an infinite sequence of propositions

$$P_1, P_2, P_3, ...P_n, ..., .$$

In order to prove that all of them are true, it is enough to show two things

1. **The base case:** P_1 is true
2. **The inductive step**: For all positive integers k, if P_k is true, then so is P_{k+1}

Example: Show that

$$1 + 2 + 3 + ... + n = \frac{n(n+1)}{2}.$$

Base case:

$$1 = \frac{1(1+1)}{2} = \frac{2}{2} = 1.$$

Inductive step: P_k is given by

$$1 + 2 + 3 + ... + k = \frac{k(k+1)}{2}.$$

P_{k+1} is given by

$$1 + 2 + 3 + ... + k + k + 1 = \frac{k+1(k+2)}{2}.$$

If $1 + 2 + 3 + ... + k = \frac{k(k+1)}{2}$, then

$$\begin{aligned} 1 + 2 + 3 + ... + k + k + 1 &= \frac{k+1(k+2)}{2} \\ \frac{k(k+1)}{2} + k + 1 &= \frac{k+1(k+2)}{2} \\ \frac{k(k+1) + 2k + 2}{2} &= \frac{k^2 + 3k + 2}{2} \\ \frac{k^2 + 3k + 2}{2} &= \frac{k^2 + 3k + 2}{2}. \end{aligned}$$

Thus, we have showed that $P_k \implies P_{k+1}$ ■.

Note: Our aim is not to directly prove P_{k+1}, but to prove that P_k implies P_{k+1}. In the inductive step we assume P_k to be true, then show under this assumption, P_{k+1} is also true.

- **Understanding gauss's formula for the sum of the first n natural numbers**: Suppose we want to find the sum $1+2+3+...+(n-1)+n$. We could have discovered the formula that we proved above by first writing the sum twice

$$\begin{aligned} &1 + 2 + 3 + ... + (n-1) + n \\ &n + (n-1) + (n-2) + ... + 2 + 1. \end{aligned}$$

The sum of the two numbres in each column is $n+1$, and there are n columns, so the total sum is $n(n+1)$, it then follows that the actual sum is $\frac{1}{2}n(n+1)$

- **Trianglular numbers**: The sequence of integers

  $$1 \qquad\qquad 3 = 1 + 2$$
  $$6 = 1 + 2 + 3$$
  $$10 = 1 + 2 + 3 + 4$$
  $$15 = 1 + 2 + 3 + 4 + 5$$
  $$....$$

  Are called *triangular numbers*. If you were to make a triangle of dots out of the sum, where the highest number is the base, the second highest is the layer ontop of the base, etc, you would form a triangle.

- **Strong induction**: Given an infinite sequence of propositions

  $$P_1, P_2, P_3, ..., P_n.$$

  In order to demonstrate that all of them are true, it is enough to know two things.

  1. **The base case**: P_1 is true
  2. **The inductive step**: For all integers $k \geqslant 1$, if $P_1, P_2, P_3, ..., P_k$ are true, then so is P_{k+1}

- **Pingala-fibonacci numbers**: Define a sequence of positive integers as follows: $F_0 = 0, F_1 = 1$, and for $n = 2, 3, ...$ we have

  $$F_n = F_{n-2} + F_{n-1}.$$

  This sequence is also known as *the fibonacci sequence*.

- **Lucas numbers**: Change the initial values on the fibonacci sequence. Let $L_0 = 2, L_1 = 1$, and $L_n = L_{n-2} + L_{n-1}$. Then, we get the *Lucas numbers*

  $$2, 1, 3, 4, 7, 11, 18, 29, 47, ...$$

$$\mathcal{L}.$$

Axiomatic geometry

4.1 Euclids elements and the question of parallels

- **Mathematical axioms and postulates**: Axioms are general truths or statements accepted without proof. Postulates are assumptions specific to a particular mathematical framework, often geometry. They serve as starting points for reasoning within that system.

  In short, axioms are universal truths in mathematics. Postulates are subject-specific assumptions.

- **Euclids definitions**:

  1. **Point:** That which has no part.
  2. **Line:** Breadthless length.
  3. The ends of a line are points.
  4. **Straight line:** A line which lies evenly with the points on itself.
  5. **Surface:** That which has length and breadth only.
  6. The edges of a surface are lines.
  7. **Plane surface:** A surface which lies evenly with the straight lines on itself.
  8. **Angle:** The inclination to one another of two lines in a plane which meet one another and do not lie in a straight line.
  9. **Right angle:** When a straight line set up on another straight line makes the adjacent angles equal to one another, each of the equal angles is a right angle.
  10. **Perpendicular:** A straight line standing on another straight line to form right angles with it.
  11. **Obtuse angle:** An angle greater than a right angle.
  12. **Acute angle:** An angle less than a right angle.
  13. **Boundary:** That which is the extremity of anything.
  14. **Figure:** That which is contained by any boundary or boundaries.
  15. **Circle:** A plane figure contained by one line (the circumference) such that all straight lines falling upon it from one point among those lying within the figure are equal to one another.
  16. **Center of a circle:** The point from which all straight lines drawn to the circumference are equal.
  17. **Diameter of a circle:** Any straight line drawn through the center and terminated in both directions by the circumference.
  18. **Semicircle:** The figure contained by the diameter and the circumference cut off by it. The center of the semicircle is the same as that of the circle.
  19. **Segment of a circle:** The figure contained by a straight line and the circumference it cuts off.
  20. **Rectilineal figure:** A figure contained by straight lines.
  21. **Trilateral figure:** A rectilineal figure contained by three straight lines (a triangle).
  22. **Quadrilateral figure:** A rectilineal figure contained by four straight lines.
  23. **Multilateral figure (polygon):** A rectilineal figure contained by more than four straight lines.

24. **Equilateral triangle:** A triangle with three equal sides.
25. **Isosceles triangle:** A triangle with two equal sides.
26. **Scalene triangle:** A triangle with three unequal sides.
27. **Right-angled triangle:** A triangle with one right angle.
28. **Obtuse-angled triangle:** A triangle with one obtuse angle.
29. **Acute-angled triangle:** A triangle with three acute angles.
30. **Parallel lines:** Straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

- **Euclids postulates**

1. To draw a straight line from any point to any point
2. To produce a finite straight line continuously in a straight line
3. To describe a circle with any center and distance
4. That all right angles are equal to one another
5. That, if a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angels

- **Euclids axioms**:

1. Things which are equal to the same thing are also equal to one another
2. If equals be added to equals, the wholes are equal
3. If equals be subtracted from equals, the remainders are equal.
4. Things which coincide with one another are equal to one another
5. The whole is greater than the part

- **Definitions rephrased**:

1. **Point:** A location that has no size or dimension.
2. **Line:** A one-dimensional object that has length but no width.
3. **Endpoints of a line:** The points where a line begins or ends.
4. **Straight line:** A line that does not curve and lies evenly between its endpoints.
5. **Surface:** A two-dimensional object that has length and width but no thickness.
6. **Edges of a surface:** The boundaries of a surface are lines.
7. **Plane surface:** A flat surface where any straight line connecting two points on it lies entirely on the surface.
8. **Angle:** The measure of the inclination or separation between two lines that meet at a point but are not aligned.
9. **Right angle:** An angle formed when one line meets another to create two equal angles (90 degrees each).
10. **Perpendicular lines:** Two lines that meet to form a right angle.
11. **Obtuse angle:** An angle larger than a right angle (greater than 90 degrees).
12. **Acute angle:** An angle smaller than a right angle (less than 90 degrees).
13. **Boundary:** The edge or limit of an object.
14. **Figure:** A shape that is enclosed by boundaries.

15. **Circle:** A shape where all points on the boundary (the circumference) are the same distance from a central point.
16. **Center of a circle:** The point that is equidistant from every point on the circle's boundary.
17. **Diameter of a circle:** A straight line passing through the center of a circle that touches the boundary on both sides.
18. **Semicircle:** Half of a circle, defined by dividing a circle along its diameter.
19. **Segment of a circle:** A region of a circle bounded by a chord (a straight line) and the arc it cuts off.
20. **Polygon (rectilinear figure):** A shape enclosed by straight lines.
21. **Triangle:** A polygon with three sides.
22. **Quadrilateral:** A polygon with four sides.
23. **Polygon (multilateral figure):** A shape with more than four sides.
24. **Equilateral triangle:** A triangle where all three sides are equal in length.
25. **Isosceles triangle:** A triangle where two sides are equal in length.
26. **Scalene triangle:** A triangle where all three sides are of different lengths.
27. **Right triangle:** A triangle with one right angle (90 degrees).
28. **Obtuse triangle:** A triangle with one obtuse angle (greater than 90 degrees).
29. **Acute triangle:** A triangle where all angles are acute (less than 90 degrees).
30. **Parallel lines:** Two straight lines in the same plane that, no matter how far extended, will never meet

- **Postulates rephrased**

1. It is possible to draw a straight line connecting any two points.
2. A finite straight line can be extended indefinitely in a straight line.
3. A circle can be drawn with any center and any radius.
4. All right angles are equal to each other.
5. If a straight line intersects two straight lines such that the interior angles on one side add up to less than two right angles, then the two straight lines, if extended indefinitely, will meet on the side where the angles are less than two right angles.

- **Axioms rephrased**:

1. Things equal to the same thing are equal to each other.
2. If equals are added to equals, the results are equal.
3. If equals are subtracted from equals, the remainders are equal.
4. Things that overlap or coincide exactly are equal.
5. The whole is greater than any of its parts.

- **More on Euclids 5th postulate**: Unlike the other four postulates, the 5th postulate is more complex and less intuitive. It essentially describes the behavior of parallel lines, but its wording led mathematicians to wonder if it could be derived from the other postulates.

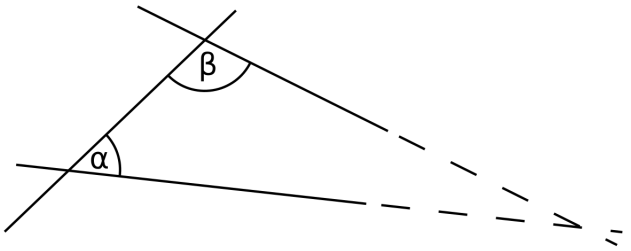
  For centuries, mathematicians like Proclus, Ptolemy, and others tried to prove the 5th postulate as a theorem based on the other four postulates. These attempts were unsuccessful, as the postulate is independent.

In the 19th century, mathematicians like Lobachevsky, Bolyai, and Gauss explored what happens if the 5th postulate is replaced with different assumptions. This led to the development of non-Euclidean geometries:

- **Hyperbolic geometry:** There are infinitely many parallel lines through a point not on a given line.
- **Elliptic geometry:** No parallel lines exist.

The questioning of the 5th postulate revolutionized mathematics, leading to a broader understanding of geometry and the realization that Euclidean geometry is just one of many possible systems.

Observe Euclids 5th postulate



- **Playfair's Postulate**: Is an equivalent form of Euclid's 5th postulate which states

  "Through a given point not on a line, there is exactly one line parallel to the given line"

4.2 Five examples

- **The Euclidean plane**: The Euclidean plane is a two-dimensional geometric space that forms the foundation of Euclidean geometry, as described in Euclid's Elements. It is characterized by the following properties

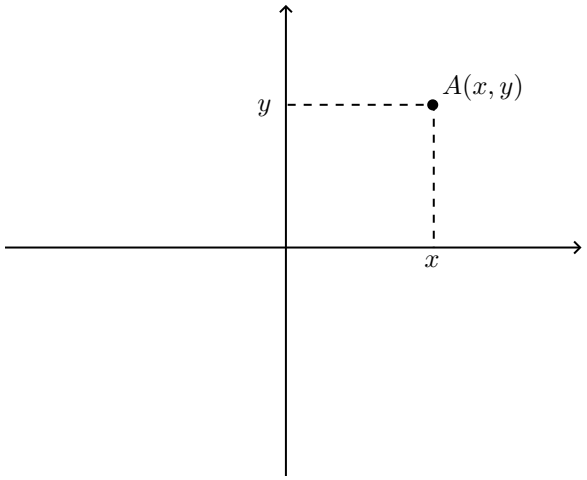
  1. **Flat Surface:** The Euclidean plane is flat, meaning it has no curvature.
  2. **Points and Lines:** It consists of an infinite set of points. Straight lines can be drawn to connect any two points, and these lines extend infinitely in both directions.
  3. **Distance and Angles:** Distance between points is measured using the Euclidean distance formula. Angles are measured in degrees or radians.
  4. **Postulates:** The plane follows Euclid's postulates, including the 5th (parallel postulate), which ensures the uniqueness of parallel lines.
  5. **Coordinate Representation:** Often represented using the Cartesian coordinate system, where every point is defined by an ordered pair (x, y)
  6. **Dimensions:** It has two dimensions: length and width.

  Note that the 2-dimensional cartesian plane is a mathematical representation of the Euclidean plane using a coordinate system. The Euclidean plane is a more general geometric concept, while the Cartesian plane provides a numerical framework (coordinates) for working with Euclidean geometry. In practical applications, the Cartesian plane is often used to model the Euclidean plane

  Let $\mathbb{E}$ denote the Euclidean plane.

  Coordinates: The points in $\mathbb{E}$ are in one-to-one correspondence with the ordered pairs of real numbers. Each point A corresponds to a pair of real numbers (x, y), called the *coordinates* of A, where the pair is assigned in the familiar way

  We often identify A with its pair of coordinates (x, y)



  Equations of lines: Each *nonvertical line* ℓ in $\mathbb{E}$ consists of all points (x, y), where $y = mx + b$ for some fixed m and b. each *vertical line* ℓ consists of all (x, y), where $x = a$ for some fixed a

For any two points $A(x_1, y_1)$ and $B(x_2, y_2)$, the *slope* of the line ℓ through A and B is

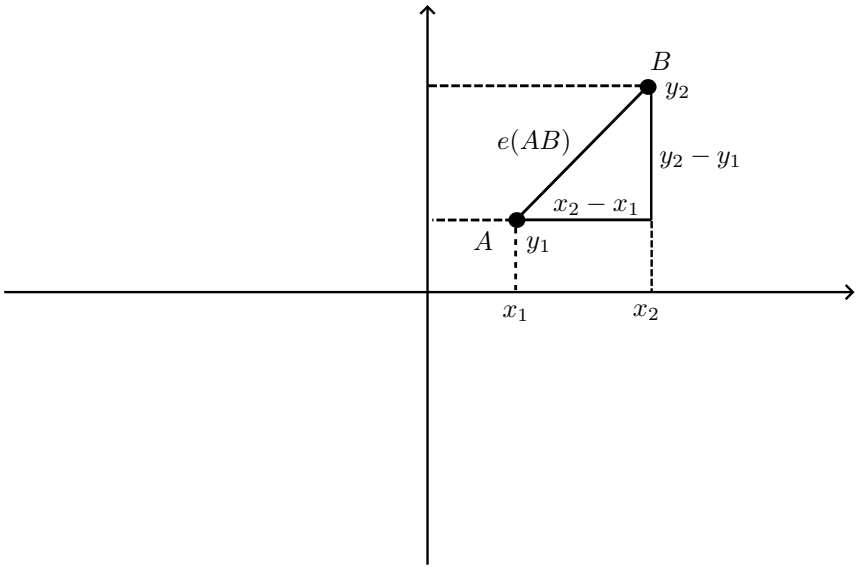
$$m = \frac{y_2 - y_1}{x_2 - x_1} \quad (\text{if } x_2 \neq x_1)$$

And an equation for ℓ is given by

$$y - y_1 = m(x - x_1) \quad (\text{if } x_2 \neq x_1)$$

The *Euclidean distance* $e(AB)$ between A and B satisfies the formula

$$e(AB) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Then

$$(e(AB))^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$$
$$e(AB) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- **More on Euclidean distance**

Proposition 1.1 If $A(x_1, y_1)$ and $B(x_2, y_2)$ are on the line $y = mx + b$, then $e(AB) = |x_1 - x_2|\sqrt{m^2 + 1}$

Proof. Assume $A(x_1, y_1)$ and $B(x_2, y_2)$ are on the line $y = mx + b$, and $e(AB) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. Observe that the slope m of the line is given by

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

Which implies

$$y_2 - y_1 = m(x_2 - x_1)$$

Plugging this expression for $y_2 - y_1$ into $e(AB)$ yields

$$\begin{aligned} e(AB) &= \sqrt{(x_2 - x_1)^2 + (m(x_2 - x_1))^2} \\ &= \sqrt{(x_2 - x_1)^2 + (m^2(x_2 - x_1)^2)} \\ &= \sqrt{(x_2 - x_1)^2[1 + m^2]} \\ &= \sqrt{(x_2 - x_1)^2} \cdot \sqrt{m^2 + 1} \\ &= |x_2 - x_1| \sqrt{m^2 + 1} \end{aligned}$$

As desired ■

- **The Minkowski plane, or taxicab plane**: Let $\mathbb{M}$ denote the Minkowski plane. $\mathbb{M}$ has the same points, lines, and coordinates as $\mathbb{E}$, but distance is different. For any $A(x_1, y_1)$ and $B(x_2, y_2)$, define the *Minkowski distance* $d_{\mathbb{M}}$ as

$$d_{\mathbb{M}} = |x_2 - x_1| + |y_2 - y_1|$$

Thus, the *Minkowski distance* $d_{\mathbb{M}}(AB)$ is defined as the sum of the horizontal and vertical "ordinary distances"

For example, consider $A(1, 2), B(-1, -3)$, then

$$d_{\mathbb{M}}(AB) = |-1 - 1| + |-3 - 2| = 7$$

- **More on Minkowski distance**:

Proposition 1.2 If $A(x_1, y_1)$ and $B(x_2, y_2)$ are on the line $y = mx+b$, then $d_{\mathbb{M}}(AB) = |x_1 - x_2|(1 + |m|)$

Proof. Assume $A(x_1, y_1)$ and $B(x_2, y_2)$ are on the line $y = mx + b$, and $d_{\mathbb{M}} = |x_2 - x_1| + |y_2 - y_1|$. Observe that the slope m of the line is given by

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

Which implies

$$y_2 - y_1 = m(x_2 - x_1)$$

Plugging this expression for $y_2 - y_1$ into $d_{\mathbb{M}}$ yields

$$\begin{aligned} d_{\mathbb{M}} &= |x_2 - x_1| + |y_2 - y_1| \\ &= |x_2 - x_1| + |m(x_2 - x_1)| \\ &= |x_2 - x_1| + |m|\,|x_2 - x_1| \\ &= |x_2 - x_1|(1 + |m|) \\ &= |-(x_1 - x_2)|(1 + |m|) \\ &= |-1|\,|x_1 - x_2|(1 + |m|) \\ &= |x_1 - x_2|(1 + |m|) \end{aligned}$$

As desired ■

- **The spherical plane**: Let $\mathbb{S}(r)$ denote the surface of the sphere of radius r; that is, the *spherical plane*.

Once r is fixed, we shorten the notation to $\mathbb{S}$. We shall assume that our spheres are centered at the origin $(0,0,0)$ in three-dimensional space. Then $\mathbb{S}$ is the set of all (x,y,z) such that $x^2+y^2+z^2=r^2$. Points are as usual, and lines on $\mathbb{S}$ are defined to be the *great circles*. A great circle is the intersection of the sphere with a plane that cuts the sphere in half. Then, any two points have a unique line joining them, unless they are oppoosite (antipodes). In this case, they have infinitely many lines joining them.

Distance in $\mathbb{S}$: For points A, B on $\mathbb{S}$, define distance

$$d_{\mathbb{S}}(AB) = \text{length of the minor (shorter) arc of the great circle (line) through } A \text{ and } B$$

To compute $d_{\mathbb{S}}(AB)$ more easily, we must recall the forumula for the *arc length in a circle of radius* r. Let θ be the radian measure of $\angle POQ$. The angle that sweeps out the full circle has measure 2π, and the circumference is $2\pi r$. The sector formed by $\angle POQ$ makes up $\frac{\theta}{2\pi}$ of the full circle, so

$$\text{arc length } PQ = \frac{\theta}{2\pi} \cdot 2\pi r = \theta r$$

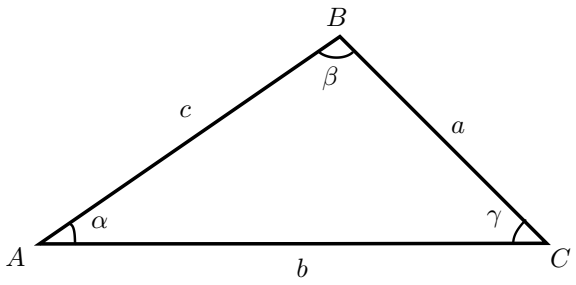
An explicit formula for the spherical distance between two points, in terms of their coordinates, is given next. It follows from the distance formula for three-dimensional space and the Law of Cosines.

If $P(a,b,c)$ and $Q(x,y,z)$ are points on the surface of the sphere of radius r centered at $(0,0,0)$ then

$$d_{\mathbb{S}} = r\cos^{-1}\left(\frac{ax+by+cz}{r^2}\right)$$

First, recall the law of cosines

Remark. *(Law of Cosines.)*



In trigonometry, the **law of cosines** (also known as the *cosine formula* or *cosine rule*) relates the lengths of the sides of a triangle to the cosine of one of its angles. For a triangle with sides a, b, and c, opposite respective angles α, β, and γ (see Fig. 1), the law of cosines states:

$$c^2 = a^2 + b^2 - 2ab\cos\gamma,$$

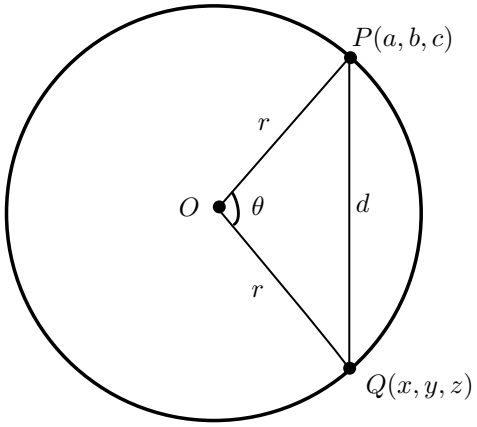
$$a^2 = b^2 + c^2 - 2bc\cos\alpha,$$
$$b^2 = a^2 + c^2 - 2ac\cos\beta.$$

The law of cosines generalizes the Pythagorean theorem, which holds only for **right triangles**: if γ is a right angle then $\cos\gamma = 0$, and the law of cosines reduces to:

$$c^2 = a^2 + b^2.$$

The law of cosines is useful for solving a triangle when all three sides or two sides and their included angle are given. 😊

Consider the points $P(a, b, c)$ and $Q(x, y, z)$ and the line (great circle) connecting them



Let d be the Euclidean distance PQ and θ be the radian measure of $\angle POQ$. By the law of cosines,

$$d^2 = r^2 + r^2 - 2r^2\cos(\theta)$$
$$\implies \cos(\theta) = \frac{d^2 - r^2 - r^2}{-2r^2} = \frac{d^2 - 2r^2}{-2r^2} = \frac{2r^2 - d^2}{2r^2}$$

The Euclidean distance d is given by

$$\begin{aligned} d &= \sqrt{(x-a)^2 + (y-b)^2 + (z-c)^2} \\ &= \sqrt{x^2 + y^2 + z^2 + a^2 + b^2 + c^2 - 2ax - 2by - 2cz} \end{aligned}$$

Thus,

$$\begin{aligned} \cos(\theta) &= \frac{2r^2 - \left(\sqrt{x^2 + y^2 + z^2 + a^2 + b^2 + c^2 - 2ax - 2by - 2cz}\right)^2}{2r^2} \\ &= \frac{2r^2 - x^2 - y^2 - z^2 - a^2 - b^2 - c^2 + 2ax + 2by + 2cz}{2r^2} \end{aligned}$$

Observe that since points P, Q lie on a sphere, they must obey the equations

$$x^2 + y^2 + z^2 = r^2$$

Thus, since P is given by the pair (a, b, c), and Q is given by (x, y, z), we have

$$a^2 + b^2 + c^2 = r^2$$
$$x^2 + y^2 + z^2 = r^2$$

Thus,

$$\begin{aligned}
\cos(\theta) &= \frac{2r^2 - x^2 - y^2 - z^2 - a^2 - b^2 - c^2 + 2ax + 2by + 2cz}{2r^2} \\
&= \frac{r^2 + r^2 - x^2 - y^2 - z^2 - a^2 - b^2 - c^2 + 2ax + 2by + 2cz}{2r^2} \\
&= \frac{a^2 + b^2 + c^2 + x^2 + y^2 + z^2 - x^2 - y^2 - z^2 - a^2 - b^2 - c^2 + 2ax + 2by + 2cz}{2r^2} \\
&= \frac{2ax + 2by + 2cz}{2r^2} \\
&= \frac{2(ax + by + cz)}{2r^2} \\
&= \frac{ax + by + cz}{r^2}
\end{aligned}$$

Since $d_{\mathbb{S}} = r\theta$, we finally arrive at the expression

$$d_{\mathbb{S}} = r\theta = r\cos^{-1}\left(\frac{ax + by + cz}{r^2}\right)$$

As desired ■

Note: There are no parallel lines in $\mathbb{S}$, any two great circles meet at a pair of antipodes.

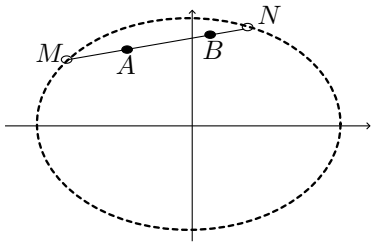
- **The chords of a circle**: A chord of a circle is a straight line segment whose endpoints lie on the circle. In other words, it is a line segment that connects two points on the circumference of a circle
- **The hyperbolic plane (Poincare disk model)**: Let $\mathbb{H}$ denote the hypebolic plane, which is the set of all points inside (but not on) the unit circle in $\mathbb{E}$. That is, all (x, y) with $x^2 + y^2 < 1$

Lines in $\mathbb{H}$ are defined to be the chords of the circle.

Distance: If A, B are two points in $\mathbb{H}$, define $d_{\mathbb{H}}(AB)$, the distance between them in $\mathbb{H}$ as follows: Draw the chord AB, and let M, N be the points where the chord meets the unit circle (M, N are in $\mathbb{E}$ but not $\mathbb{H}$). label so that B separates A and N.

Let $e(PQ)$ denote the usual Euclidean distance between points, and define

$$d_{\mathbb{H}}(AB) = \ln\left(\frac{e(AN)e(BM)}{e(AM)e(BN)}\right)$$



Since $e(AN) > e(BN)$ and $e(BM) > e(AM)$, we have $\frac{e(AN)}{e(BN)} > 1$ and $\frac{e(BM)}{e(AM)} > 1$. Hence $\frac{e(AN)e(BM)}{e(AM)e(BN)} = \frac{e(AN)}{e(BN)} \cdot \frac{e(BM)}{e(AM)} > 1$. It follows from a property of ln that $d_{\mathbb{H}}(AB) > 0$. Note that $d_{\mathbb{H}}(AB) = d_{\mathbb{H}(BA)}$. Also,

$$d_{\mathbb{H}}(AB) = \left| \ln \left(\frac{e(AN)e(BM)}{e(AM)e(BN)} \right) \right| = \left| \ln \left(\frac{e(AM)e(BN)}{e(AN)e(BM)} \right) \right|$$

So if absolute value is used in this way, then we need not worry about which point on the unit circle is marked M and which is marked N.

If $A = B$ in $\mathbb{H}$, take any chord through A and let M, N be as previously. Since $\frac{e(AN)e(AM)}{e(AM)e(AN)} = 1$, it is consistent with the preceding definition to set $d_{\mathbb{H}}(AA) = 0$.

We note that N using the distance formula above is always the point from A through B, and the point M is the point from B through A. With this in mind, it is clear that $\frac{e(AN)e(BM)}{e(AM)e(BN)} \to \infty$ as we move A and B closer to the opposing sides of the unit circle. Since $\ln : (0, \infty) \to \mathbb{R}$, and we noted earlier that $\frac{e(AN)e(BM)}{e(AM)e(BN)} > 1$, distances in the hyperbolic plane can get arbitrary large or small, without bound.

Further, Euclids 5th postulate/Playfairs postulate is false on the hyperbolic plane. Observe

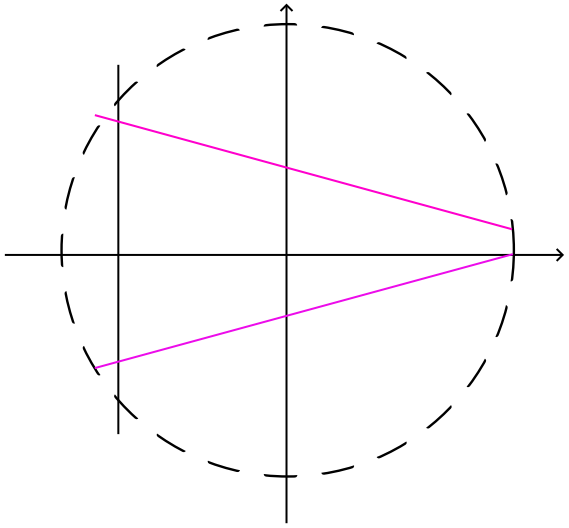
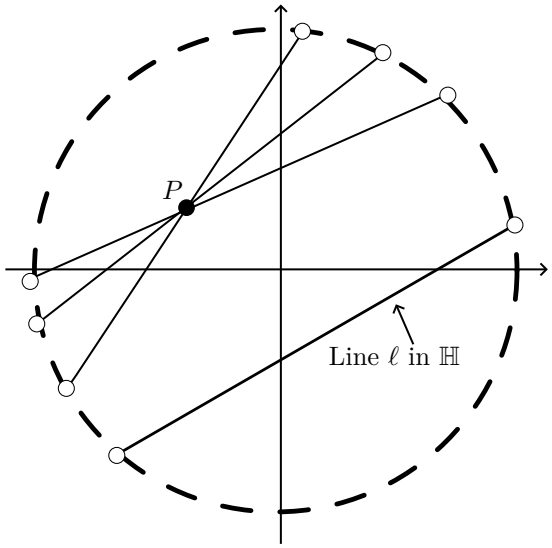


Figure 1: *Euclids fifth postulate does not hold on the hyperbolic plane*

These lines will never meet, because they are stopped by the unit circle boundary. Further, they will in a sense continue on forever, because distances can get arbitrarily large

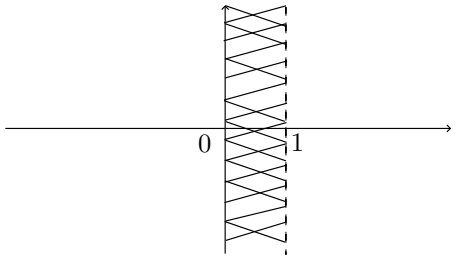
Also,



We see that given a point P not on the line ℓ, there are many lines through P that are parallel to ℓ. All of these lines are parallel to ℓ, because they will never intersect with ℓ

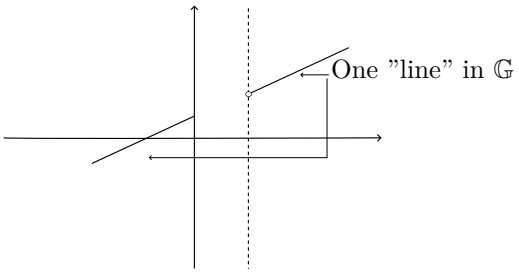
- **The gap plane**: Let $\mathbb{G}$ denote the *gap*, or *missing strip* plane. The points of $\mathbb{G}$ are all those of $\mathbb{E}$ except those (x, y) with $0 < x \leqslant 1$

So the y-axis is part of $\mathbb{G}$, but the line $x = 1$ is not (and neither is any vertical line



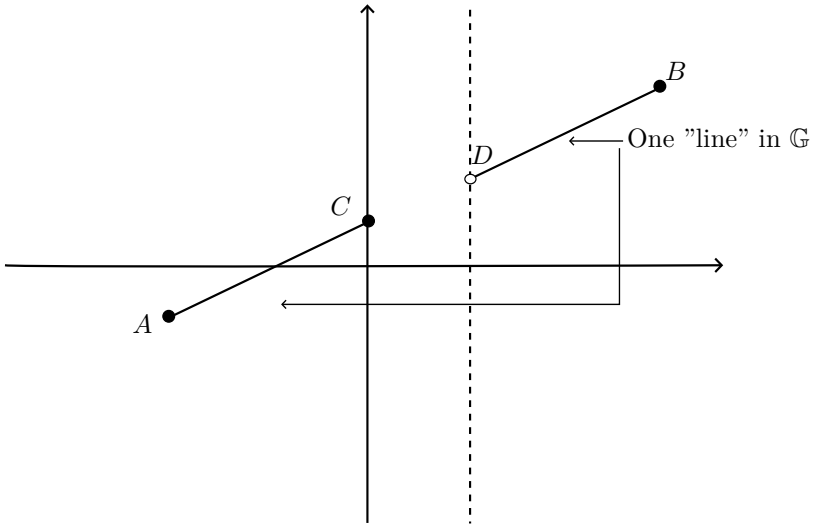
$x = a$ for $0 < a < 1$)

Lines in $\mathbb{G}$ are defined to be the same as in $\mathbb{E}$, except that for any nonvertical line $y = mx + b$, the part in the missing strip is deleted. So a typical nonvertical line ℓ consists of all (x, y) with $y = mx + b$ (m, b fixed) and with $x \leqslant 0$ or $x > 1$



Behold a line in $\mathbb{G}$

Distance: For points A, B in $\mathbb{G}$, we define $d_{\mathbb{G}}(AB)$ as follows. First; if A and B lie on opposite sides of the gap, let C be the point where segment $\overline{AB}$ meets the y-axis, and D the point where $\overline{AB}$ meets the vertical line $x = 1$ (D is not in $\mathbb{G}$)



Now define

$$d_{\mathbb{G}}(AB) = \begin{cases} e(AB) & \text{for } A, B \text{ on the same side of the gap} \\ e(AB) - e(CD) & \text{for } A, B \text{ on the opposite sides of the gap} \end{cases}$$

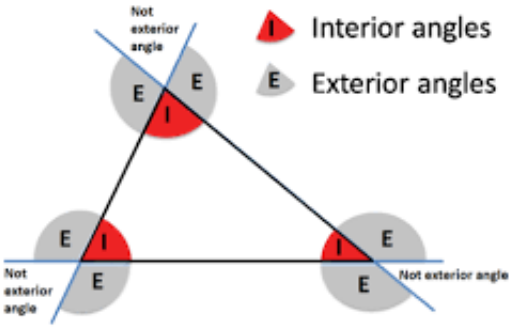
- **Interior and Exterior angles**: Interior angles are the angles inside the triangle. Each vertex of the triangle has one interior angle. The sum of the interior angles of a triangle is always 180°

  Exterior angles are the angles formed outside the triangle when one side of the triangle is extended. At each vertex, an exterior angle is supplementary to the interior angle (they add up to 180°

If an interior angle at a vertex is A, the corresponding exterior angle E is:

$$E = 180° - A$$

The sum of the exterior angles of a triangle (one at each vertex) is always $360°$, regardless of the shape of the triangle.



- **Remote angles**: Remote angles refer to the interior angles of a triangle that are not adjacent to a given exterior angle
- **More on points**:
  - **Collinear points**: Points that lie on the same straight line.
  - **Noncollinear points**: Points that do not lie on the same straight line.
  - **Coplanar points**: Points that lie on the same plane.
  - **Concurrent Points**: Points where three or more lines intersect.
  - **Equidistant Points**: Points that are all the same distance from a particular point or object.
  - **Lattice Points**: Points with integer coordinates.
  - **Interior points:** Points that lie inside a given shape.
  - **Exterior points:** Points that lie outside a given shape.
- **Congruent triangles**: Congruent triangles are triangles that are exactly the same in shape and size. This means that all corresponding sides and angles of one triangle are equal to those of the other triangle.
- **Vertical (opposite) angles**: Vertical angles (also called opposite angles) are the angles that are formed by two intersecting lines and are opposite to each other
- **Reading angle notation**: Suppose you have an angle $\angle ABC$. This angle refers to the angle formed at vertex B by the two line segments or rays:

  One extending from B to A, the other extending from B to C. The middle letter, B, always represents the vertex of the angle (the point where the two lines meet).

  Note: If there's no ambiguity about which angle is being referred to, the angle might simply be denoted as $\angle B$.
- **Potential dangers and the exterior angle inequality**:

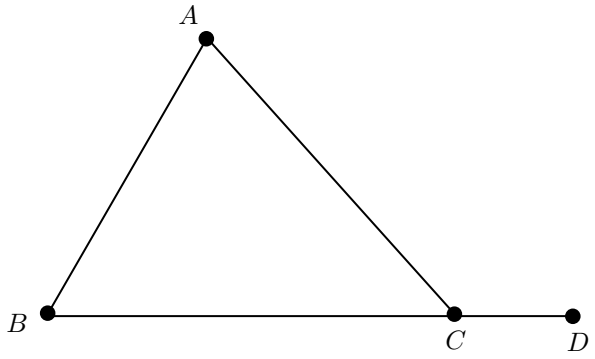
Theorem (*Exterior angle inequality*): An exterior angle of a triangle is greater than either remote interior angle. That is, if $\triangle ABC$ is a any triangle, and point D is on the extension of segment $\overline{BC}$ through C, then

$$\angle ACD > \text{ both } \angle A \text{ and } \angle B$$

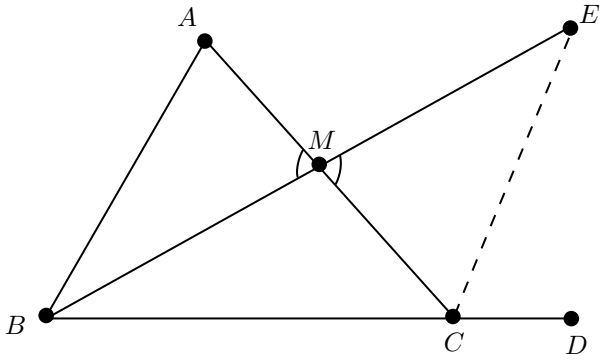
Background facts that are ok for both $\mathbb{E}$ and $\mathbb{S}$

1. **Triangles:** Line segments that join any three noncollinear points
2. **Angle measures**: Are defined for every angle
3. **Vertical angles**: Have equal measure
4. **side-angle-side**: Criterion for congruent triangles, If two sides and the angle between them in one triangle are equal to the corresponding parts in another triangle, the triangles are congruent.

Consider the triangle



Euclid's proof of EAI: Let M be the midpoint of $\overline{AC}$ so $\overline{AM} = \overline{CM}$. Next, extend $\overline{BM}$ through M to point E such that $\overline{MB} = \overline{ME}$



Notice that since $\angle AMB$ and $\angle CME$ are vertical, they must be equal. That is, $\angle AMB = \angle CME$. Since

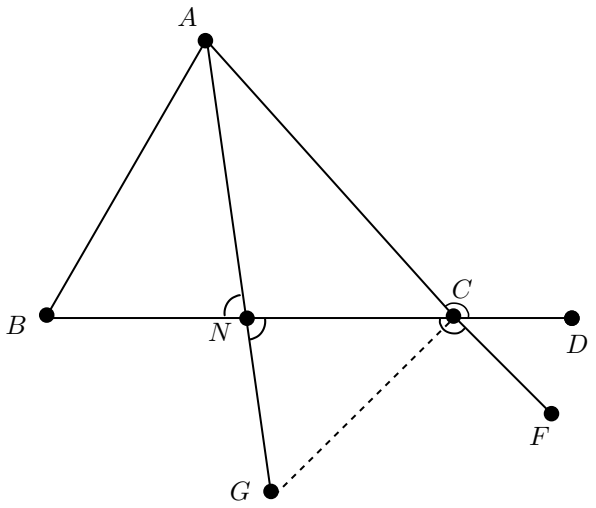
1. $AM = CM$
2. $MB = ME$
3. $\angle AMB = \angle CME$

We have met the side-angle-side criterion for congruent triangles. Thus, $\triangle AMB \cong \triangle CME$. Consequently, we have $\angle BAM = \angle ECM$. Further, notice that

$$\begin{aligned}\angle ACD &= \angle ACE + \angle ECD \\ &= \angle ECM + \angle ECD \\ &= \angle BAM + \angle ECD > \angle BAM = \angle A\end{aligned}$$

Thus, $\angle ACD > \angle A$. To show $\angle ACD > \angle B$, first, extend AC through C to point F, forming $\angle BCF$. Notice that since $\angle ACD$ and $\angle BCF$ are vertical, they must be equal. That is, $\angle ACD = \angle BCF$

Next, let N be the midpoint of BC such that $BN = CN$. Extend A through N to point G such that $AN = GN$.



Note that since $\angle ANB$ and $\angle CNG$ are vertical, they are equal. That is, $\angle ANB = \angle CNG$. Further, since we have

1. $\angle ANB = \angle CNG$
2. $AN = GN$
3. $BN = CN$

We have congruence, $\triangle ANB \cong \triangle CNG$. Thus, $\angle ABN = \angle NCG$. Therefore,

$$\begin{aligned}\angle ACD = \angle BCF &= \angle BCG + \angle GCF \\ &= \angle NCG + \angle GCF \\ &= \angle ABN + \angle GCF > \angle ABN = \angle B\end{aligned}$$

Thus, we have shown that $\angle ACD > \angle A$ and $\angle B$ ■

4.3 Intro to geometric proofs and some set theory

- **Transversal**: a transversal is a line that passes through two lines in the same plane at two distinct points.
- **Relationship of angles**: Consider the transversal configuration

  We see that we get eight formed angles.

  - **Interior angles**: Interior angles are the angles that are inside the transversal configuration. Angles a, b, c, d are interior
  - **Exterior angles**: Exterior angles are the angles that are inside the transversal configuration. Angles e, f, g, h are exterior
  - **Consecutive interior angles**: Pairs of interior angles that are on the same side of the transversal. Angles c, d are consecutive interior, and a, b are consecutive interior
  - **Consecutive exterior angles**: Pairs of exterior angles that are on the outside of the transversal configuration. Angles e, g are consecutive exterior, angles f, h are consecutive exterior
  - **Alternate interior angles**: Pairs of interior angles that are on opposite sides but not complementary, angles b, d and a, c are alternate interior
  - **Alternate exterior angles**: Pairs of exterior angles that are on opposite sides but not complementary, angles e, h, and f, g are alternate exterior
  - **Vertical angles**: Angles that are opposite each other, formed when two lines intersect. Vertical angles are of equal measure. Pairs d, h - a, g - e, b - and f, c are vertical
  - **Supplementary angles**: Angle pairs that sum to 180, pairs a, h - d, g - f, b - and e, c are supplementary
  - **Complementary angles**: Angle pairs that sum to 90, none in the transversal configuration

  Proposition (Equal alternate interior angles). Suppose $a + b = 180$, then $b = d$, and $c = a$.

  Proof. Consider the transversal configuration shown above. Assume $a + b = 180$, then $a = 180 - b$. Since vertical angles are equal, we have $d = h$. But since a, h are supplementary, we have $a + h = 180$, which implies $h = 180 - a$. Thus,

  $$d = h = 180 - a$$

  Since $a + b = 180$ implies $b = 180 - a$, we have

  $$d = h = 180 - a = b$$

  Thus, $d = b$. Next, we show that $c = a$. Since c and f are vertical, we have $c = f$. Further, since $a+b = 180$, we have $a = 180-b$. Notice that b and f are supplementary, which implies $b + f = 180$, or $f = 180 - b$. So, since $c = f = 180 - b$, and $a = 180 - b$, we have $c = f = 180 - b = a$. Thus, $c = a$

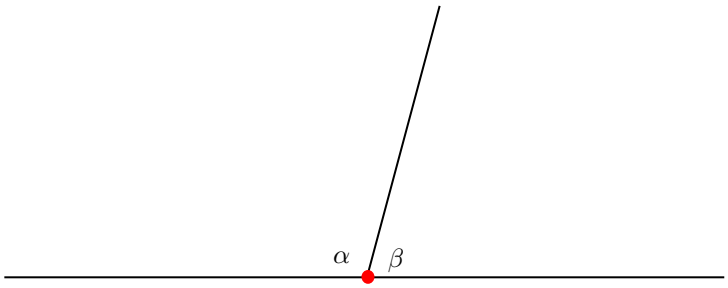
  Therefore, we conclude that if $a + b = 180$, $b = d$ and $c = a$ ■

- **Background on Euclid's plane without the fifth postulate**:

Assumptions:

- Two points determine a unique line
- Distances between points on a line include all positive real numbers
- Angles are measured

We say that α and β are *supplementary* because $\alpha + \beta = 180°$. Note that two



angles that are supplementary to each other do not have to be next to each other, only the sums of their angles must be 180°.

As a side note, recall that *complementary* angles are angles that sum to 90°

Definitions:

- **Angles**: An angle is formed when two rays meet at a common endpoint, called the vertex.
- **Vertical angles**: Vertical angles (or opposite angles) are the angles formed when two lines intersect.
- **Triangle**: A triangle is a polygon with three sides, three vertices, and three angles.

  The sum of the interior angles of a triangle is always 180°

  A triangle is a closed geometric figure formed by three line segments connecting three non-collinear points
- **Congruent**: Congruent refers to figures or shapes that are identical in size and shape.

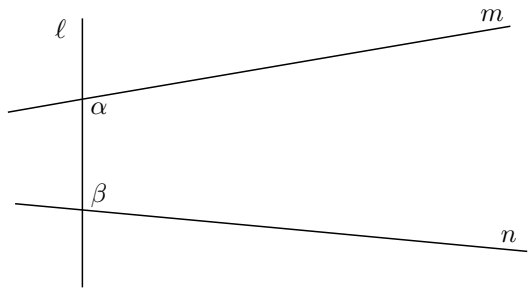
  Two triangles are congruent if their corresponding sides and angles are equal (e.g., by SSS, SAS, ASA, or AAS congruence criteria).

  We generally use the side-angle-side criterion to determine congruent triangles.

Also, recall the exterior angle theorem proved above.

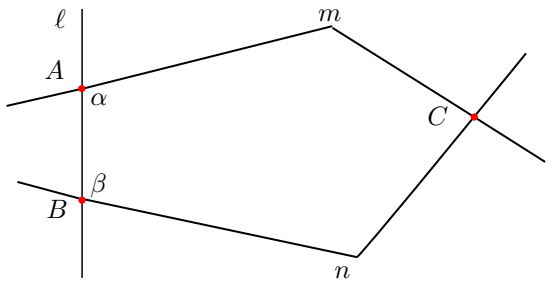
- **Example of proof by contradiction**: Suppose we are on Euclid's plane without the fifth postulate

Proposition 1. Suppose that line ℓ crosses m and n so that the interior angles on one side of ℓ add to more than $180°$



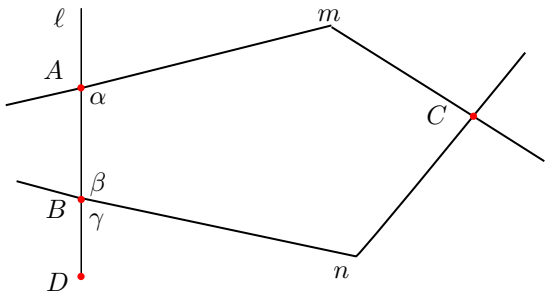
Then, m, n do not meet on that side of ℓ

Proof. Assume for the sake of contradiction that the statement is false. That is, suppose m, n meet on that side of ℓ. Then, we must have



Call the point where they meet C, since we have three noncollinear points A, B, C, $\triangle ABC$ is formed.

Define $\angle CBD$ as the exterior angle for $\triangle ABC$, call it measure γ



β and γ are supplementary, so $\beta + \gamma = 180°$. Thus, $\gamma = 180° - \beta$. By the EAI, $\gamma > \alpha$, which means $180° - \beta > \alpha$. Thus, we have $180° > \alpha + \beta$. But, we stated that $\alpha + \beta > 180°$, which is a contradiction.

Therefore, by contradiction, are assumption that m, n meet on that side is false, and therefore m, n must not meet on that side. ■

- **Upper bounds**: Suppose S is a set of real numbers, we define $b \in \mathbb{R}$ as an *upper bound* for S if for all $x \in S, x \leqslant b$

  The negation of this definition is, there exists $x \in S$ such that $x \nleq b$, or $x > b$. Thus, to prove some b is not an upper bound for S, we can show that some element of S is greater than b

  There are of course sets that do not have any upper bounds. Consider the set $S = \{n : \ n \in \mathbb{N} \text{ and } n > 0\}$. This set has no upper bound.

  If $S = \varnothing$, then every $b \in \mathbb{R}$ is an upper bound for S. This statement is vacuously true.

- **Least upper bound (supremum)**: $c \in \mathbb{R}$ is a *least upper bound* of a set S of real numbers if

  1. c is an upper bound for S
  2. $c \leqslant b$ for all upper bounds b of S

  Note: The supremum of a set S is denoted $b = \sup(S)$, where b is the supremum of the set

- **Least upper bound property of** $\mathbb{R}$: If S is a nonempty set of real numbers that has an upper bound in $\mathbb{R}$, then S has a least upper bound (l.u.b) in $\mathbb{R}$

  This justifies, among other things, that infinite decimals exist as real numbers, since an infinite decimal can be defined as the least upper bound of the set of all its finite truncations. For example, suppose S is the set of all finite decimal expansions of π.

  $$S = \{3, 3.1, 3.14, 3.141, 3.1415, ...\}$$

  Then, S as an *l.u.b* π, and $\pi \notin S$

- **Least upper bound proposition**

Proposition. Let S be a nonempty set of real numbers that has a least upper bound $b \in \mathbb{R}$. Let $t \in \mathbb{R}$ such that $t < b$. Then, there exists some $s \in S$ such that $t < s \leqslant b$.

Proof. Assume S is a nonempty subset of the real numbers with a least upper bound b. Let $t \in \mathbb{R}$ such that $t < b$. Since b is a least upper bound of S, we have

$$\forall\ s \in S,\ s \leqslant b$$

Since $t < b$, t cannot be an upper bound for S. If it were, then that would contradict b being the least upper bound. Since t is not an upper bound of S, then this implies the existence of some $s \in S$ such that $t < s$. If this were not the case, then the negation which states, for all $s \in S$, $t \geqslant s$ would be true. Since the negation implies that t is an upper bound, which we know can't be the case, there must exist some $s \in S$ such that $t < s$.

Since $s \leqslant b$ for all $s \in S$, and we know that there exists some $s \in S$ such that $t < s$, there must be at least one s that satisfies

$$t < s \leqslant b$$

■

- **Lower bounds**: Let S be a nonempty set of real numbers. Then $g \in \mathbb{R}$ is a *lower bound* for S if $g \leqslant x$ for all $x \in S$.
- **Greater lower bounds (Infimum)**: $h \in \mathbb{R}$ is a *greatest lower bound*, also called the *infimum*, or *inf* for S if h is a lower bound for S and $h \geqslant g$ for all lower bounds g of S
- **Infimum proposition**

Proposition. Let S be a nonempty set or real numbers that has a lower bound in $\mathbb{R}$. Then S has a infimum in $\mathbb{R}$

Proof. Assume $S \subseteq \mathbb{R}$, $S \neq \varnothing$ that has a lower bound in $\mathbb{R}$.

Let B be the set of all lower bounds of S. Since S has a lower bound, B is nonempty. Define

$$B = \{b \in \mathbb{R} : b \leqslant s\ \forall\ s \in S\}$$

We first note that every $s \in S$ serves as an upper bound for B. This is because for any $b \in B,\ b \leqslant s$ for all $s \in S$, thus satisfying the definition of an upper bound

Since B is nonempty and bounded above by all elements of S, B has a least upper bound (supremum) in $\mathbb{R}$. Let λ be this supremum. That is, $\lambda = \sup B$. To show that this supremum is precisely the infimum for S is to show two things

1. $\lambda \in B$. That is, λ is a lower bound for S
2. $\lambda \geqslant b$ for all lower bounds b of S

We begin by showing that $\lambda \in B$. If $\lambda \in B$, then by definition of B, $\lambda \leqslant s\ \forall\ s \in S$. Assume for the sake of contradiction that there exists some $s \in S$ such that $\lambda > s$. This would contradict the fact that λ is the least upper bound for B because then s would be an upper bound for B smaller than λ. Thus, there are no such $s \in S$ such that $s < \lambda$, and λ must therefore be in B

Next, we show that λ is truly the greatest lower bound of S, that $\lambda \geqslant b$ for all lower bounds b of S. Assume for the sake of contradiction that there exists some $b \in B$ such that $\lambda < b$. This would mean λ is not actually an upper bound for B which again contradicts the fact that λ is the supremum of B

Thus, since $\lambda \in B$, and $\lambda \geqslant b$ for all $b \in B$. We have that λ is the greatest lower bound of S, or $\lambda = \inf S$ ■

4.4 An axiom system for geometry: First steps.

- **What is projective geometry?** Projective geometry is a branch of geometry where any two distinct lines intersect in exactly one point, meaning there are no parallel lines. It extends Euclidean geometry by adding "points at infinity" to ensure this property holds. Projective geometry focuses on incidence relations (how points and lines are related) rather than distances or angles.
- **What is incidence?** In geometry, "incident" means that a point lies on a line (or a plane, in higher dimensions), or that a line passes through a point. More generally, it describes a fundamental relationship between geometric objects in an incidence structure.

  For example:

  - A point is incident to a line if it lies on that line.
  - A line is incident to a point if it passes through that point.
  - In projective geometry, two lines are incident to the same point if they intersect at that point.

  It is a basic, undefined term in axiomatic geometry, meaning it is taken as a fundamental concept rather than being defined in terms of simpler notions.
- **What is incidence geometry**: Incidence geometry is the study of geometric structures based only on points, lines, and their incidence relations (which points lie on which lines). It focuses on which objects are connected rather than distances, angles, or measurements. The main rules are typically:

  1. Any two distinct points determine a unique line.
  2. Any two distinct lines intersect in at most one point.
  3. There exist at least four points, not all on the same line (to avoid trivial cases).

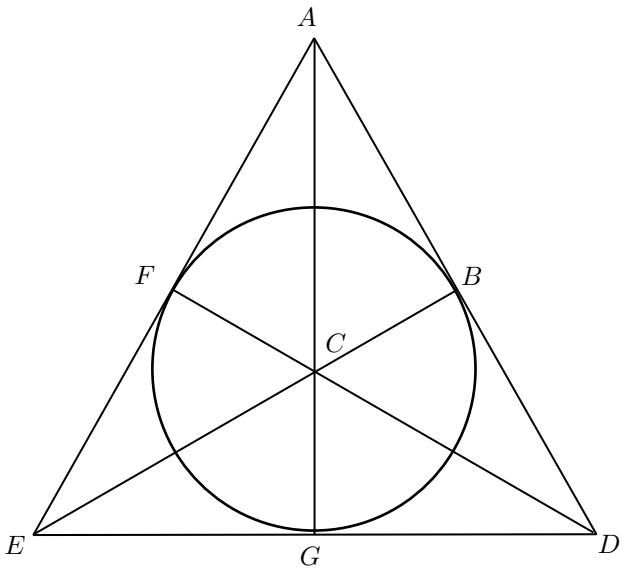
  It includes Euclidean, affine, and projective geometries as special cases.
- **The Fano plane**: The Fano plane is a *projective plane of order two.*

  When we say that the Fano Plane is a projective plane of order two, we mean that

  - **It is a finite projective plane**: – A projective plane is a type of incidence geometry satisfying specific axioms
    1. Any two distinct points determine a unique line.
    2. Any two distinct lines intersect in a unique point.
    3. There exist four points, no three of which are collinear (this ensures it is not a degenerate geometry).
  - **Order two ($q = 2$)**: The order of a finite projective plane is a parameter q that determines its structure. The order q is defined by the number of points on each line minus one. In the Fano Plane:
    1. Every line contains exactly $q + 1 = 3$ points
    2. Every point is on exactly $q + 1 = 3$ lines
    3. The total number of points is $q^2 + q + 1 = 2^2 + 2 + 1 = 7$
    4. The total number of lines is $q^2 + q + 1 = 2^2 + 2 + 1 = 7$

  Since the Fano Plane satisfies these properties for $q = 2$, it is called a projective plane of order two.
- **More on the Fano plane**: There are seven points $\{A, B, C, D, E, F, G\}$, and there are seven lines $\{A, B, D\}, \{C, D, F\}, \{A, F, E\}, \{A, C, G\}, \{B, C, E\}, \{B, F, G\}, \{D, E, G\}$

There are three points on each line, and three points through each line



Which points on which line? Write points in alphabetical order in three rows, start with A, then B, then with D

$$\begin{array}{c} A\ B\ C\ D\ E\ F \\ B\ C\ D\ E\ F\ A \\ D\ E\ F\ AB\ C \end{array}$$

Note that the columns give the lines

Note: The triangle picture is a good visual aid, but the Fano plane is not part of the Euclidean plane.

- **Coordinates for the Fano plane**: Each point is an ordered triple (x, y, z), where x, y, z are integers mod 2

$$\begin{cases} 0 & \text{stands for all even numbers} \\ 1 & \text{stands for all odd numbers} \end{cases}$$

We further note that odd + odd = even. Or, $1+1=0$. Other than that it is business as usual... $0+0=0,\ 1+0=0+1=1$

We have the points

$$\begin{array}{llll} A(1,0,0) & B(1,1,0) & D(0,1,0) & E(0,0,1) \\ C(1,1,1) & F(1,0,1) & G(0,1,1) & \text{No point : } (0,0,0) \end{array}$$

Given points P, Q, find the third point collinear with P, Q. We simply add the coordinate triples for P, Q. For example, suppose $A(1,0,0), B(1,1,0)$. Then,

$$(1,0,0) + (1,1,0) = (0,1,0) = D$$

- **Distance on the Fano plane**: We define distance for Fano points, but its not Euclidean distance

  Given points P, Q,

  $$d(PQ) = \text{ number of different respective coordinates}$$

  For example, $B(1,1,0), G(0,1,1)$ implies $d(BG) = 2$

- **General finite projective plane**:In general, for a finite projective plane of order q

  1. There are $q^2 + q + 1$ points
  2. There are $q^2 + q + 1$ lines
  3. Every line contains $q + 1$ points
  4. Every point is contained in $q + 1$ lines

  And satisfies

  1. Any two distinct points determine a unique line.
  2. Any two distinct lines intersect at a unique point.
  3. There exist at least four points, no three of which are collinear. (This ensures non-triviality.)

  Thus, the Fano Plane is the smallest projective plane, and it uniquely exists for order 2.

  Note: The "projective" part in the name projective plane comes from its connection to projective geometry, which generalizes Euclidean geometry by removing the notion of parallel lines.

- **Fine projective plane with order one?**: a finite projective plane cannot have order $q = 1$ because it would not satisfy the axioms of a projective plane.

  If $q = 1$:

  1. **Number of points**: $1^2 + 1 + 1 = 3$
  2. **Number of lines**: $1^2 + 1 + 1 = 2$
  3. **Each line has**: $1 + 1 = 2$ points
  4. **Each point is on**: $1 + 1 = 2$ lines

  This configuration forms a triangle, but therefore fails the requirement that a finite projective plane has at least four points (it only has three)

- **Some extra planes**
  - $\hat{\mathbb{E}}$**: The bumpy plane**: Which is $\mathbb{E}$, but warped. Has bumps and depressions, not always flat.
  - $\mathbb{R}^3$**:** Points, lines, distance of usual 3-dimensional space.

– $\varnothing$: Has the components necessary for a plane vacuously

- **Define a plane**: Let's define a plane called $*$,

$$\mathbb{P} = \{A, B, C, D\} \quad \text{(4 points)}$$
$$\mathbb{L} = \{A, B, C\}, \{A, C, D\}, \{B, D\} \quad \text{3 lines}$$

With distance function

	A	B	C	D
A	0	1	2	$\frac{1}{2}$
B	1	0	$\frac{3}{2}$	$\frac{1}{2}$
C	2	$\frac{3}{2}$	0	$\frac{3}{2}$
D	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	0

- **Axiom system for plane geometry**:

Undefined terms:

– $\mathbb{P}$**:** Set of elements, called **points.**
– $\mathbb{L}$**:** Collection of subsets of $\mathbb{P}$, called **lines**
– A function $d: \mathbb{P} \times \mathbb{P} \to \mathbb{R}$, called a **distance function**

Call anything with these components a **plane**

Notation, terminology

– A line is a set of points
– If P is on the line m ($P \in m$), we say that "P is on m", or "m goes through P"
– If two or more points are on the same line, we say they are **collinear**
– Denote distance $d(P, Q)$, or $d(PQ)$, or just PQ

Axiom of distance: For all points P, Q

1. $PQ \geqslant 0$
2. $PQ = 0 \iff P = Q$
3. $PQ = QP$

These are true for all planes mentioned so far, even $*$ and $\varnothing$

The distance set: Define $\mathbb{D} = \{PQ: P, Q \in \mathbb{R}\}$. This is the set of all distances that occur between points of $\mathbb{P}$, with respect to the given distance function.

The diameter of the plane $\mathbb{P}$, ω

$$\begin{cases} \omega = \sup \mathbb{D} & \text{if } \mathbb{D} \text{ has an upper bound in } \mathbb{R} \\ \omega = \infty & \text{if } \mathbb{D} \text{ has no an upper bound in } \mathbb{R} \end{cases}$$

Note that ∞ is not a real number, but we still say $r < \infty$ for all $r \in \mathbb{R}$

$\mathbb{P}$	$\mathbb{D}$	ω
$\mathbb{E}$	$[0, \infty)$	∞
$\mathbb{M}$	$[0, \infty)$	∞
$\mathbb{S}(r)$	$[0, \pi r]$	πr
$\mathbb{H}$	$[0, \infty)$	∞
$\mathbb{G}$	$[0, \infty)$	∞
Fano	$\{0, 1, 2, 3\}$	3
$\hat{\mathbb{E}}$	$[0, \infty)$	∞
$\mathbb{R}^3$	$[0, \infty)$	∞
$\varnothing$	$\varnothing$	$\times$
$(*)$	$\{0, \frac{1}{2}, 1, \frac{3}{2}, 2\}$	2

Note: Whether ω is a finite number or ∞, each distance PQ is a nonnegative, finite real number

Why not assume that two points determine a unique line? That two points are together in exactly one line? The sphere $\mathbb{S}$, which we want to include as a plane, has many lines through two points, when the points are antipodes. These are the points P, Q where $PQ = \pi r = \omega$.

Thus, our axioms will allow multiple lines through two points, but only if their distance apart is precisely ω, the diameter of the plane. Note that P, Q, with $PQ = \omega$ **may or may not** have more than one line through them.

Axioms of incidence

1. There are at least two different lines
2. Each line contains at least two different points
3. Each pair of points are together in at least one line
4. Each pair of points P, Q, with $PQ < \omega$ are together in at most one line

Note: These are true for all discussed planes except $\varnothing$. 1 and 2 are false for $\varnothing$

- **So what exactly is a plane?**: Based on the provided axioms, the definition of a plane in this system is simply a structure consisting of
  - A set of points $\mathbb{P}$
  - A collection of subsets of $\mathbb{P}$ called lines $\mathbb{L}$
  - A distance function $d : \mathbb{P} \times \mathbb{P} \to \mathbb{R}$.

Thus, a set $\mathbb{P}$ and a set of lines $\mathbb{L}$ can be called a plane as long as they fit this definition, regardless of whether they satisfy the axioms of distance or incidence.

However, for a plane to behave in a meaningful way in axiomatic geometry (i.e., to be one of the discussed geometric planes like $\mathbb{E}, \mathbb{M}, \mathbb{S}(r)$, etc...) it must satisfy the axioms of distance and incidence. These axioms impose necessary geometric structure, ensuring that distances behave as expected and that lines and points interact according to the incidence rules.

Thus, a plane can exist without satisfying the axioms, but to be a meaningful model of geometry, it is typically expected to satisfy them.

- **Plane example**: Consider the plane with $\mathbb{P}:$ all points inside the unit circle in $\mathbb{E}$, and $\mathbb{L}$ be the set of all chords inside the circle

  For points P, Q in $\mathbb{P}$, define $d(PQ) = PQ = e(PQ)$. Ie the Euclidean distance

  Note that the seven axioms are true statements for this example.

  We have $\mathbb{D} = [0, 2)$, so $\omega = 2$, but $PQ < 2$ for all $P, Q \in \mathbb{P}$

- **Trivial discrete model (TDM)**: Let $\mathbb{P}$ be any set of at least three elements. Let $\mathbb{L}$ be the collection of all two element subsets of $\mathbb{P}$

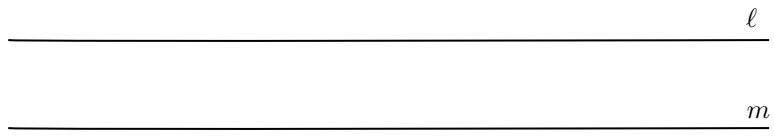
  Define distance as follows: For all $x \neq y \in \mathbb{P}$,

  $$\begin{cases} xy & = 1 \\ xx & = 0 \end{cases}$$

  The seven axioms are true for the TDM. We have $\mathbb{D} = \{0, 1\}$, thus $\omega = 1$

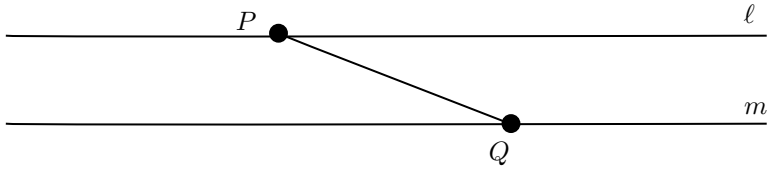
  For example, if $\mathbb{P} = \{A, B, C\}$, which implies $\mathbb{L} = \{A, B\}, \{A, C\}, \{B, C\}$, which forms a triangle where all sides are of length one.

- **White stripes model (ws)**: Let ℓ, m be two parallel lines in $\mathbb{E}$



  Define $\mathbb{P} = \{\text{all points on } \ell\} \cup \{\text{all points on } m\}$, and $\mathbb{L} = \ell, m$, and all two point sets $\{P, Q\}$ where P on ℓ, Q on m. Define distance $d =$ Euclidean distance $e(PQ)$

  Note that the seven axioms are true statements for ws, and $\mathbb{D} = [0, \infty),\ \omega = \infty$

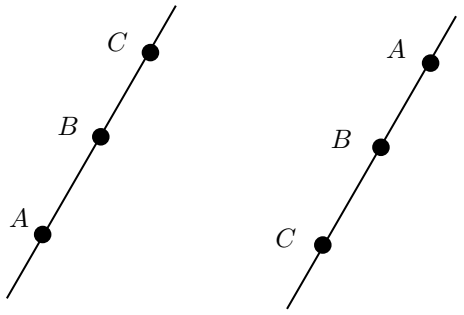


4.5 Betweenness, segements, and rays

- **Betweenness**: Let $\mathbb{P}$ be a plane with points, lines, distance, and satisfy the seven axioms (3 distance, 4 incidence). Define

  Definition. Point B lies **between** points A and C, denoted $A - B - C$ provide that

  1. A, B, and C are different and collinear
  2. $AB + BC = AC$



- **Betweenness example 1**:

$$P = \{A, B, C, D\}$$
$$L = \{\{A, B, C\}, \{A, C, D\}, \{B, D\}\}$$

Distance:

	A	B	C	D
A	0	1	2	$\frac{1}{2}$
B	1	0	$\frac{3}{2}$	$\frac{1}{2}$
C	2	$\frac{3}{2}$	0	$\frac{3}{2}$
D	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	0

On line $\{A, C, D\}$:

$$AC = 2, \quad AD = \frac{1}{2}, \quad DC = \frac{3}{2}$$

$AD + DC = AC$. Thus, $A - D - C$.

On line $\{A, B, C\}$:

$$AB = 1, \quad AC = 2, \quad BC = \frac{3}{2}$$

No two of these add to the third, so there is **no betweenness relation** among A, B, C.

- **Betweenness on the Fano plane**: We have the collinear points $A(1,0,0), B(1,1,0), D(0,1,0)$, with

$$AB = 1, \quad BD = 1, \quad AD = 2$$

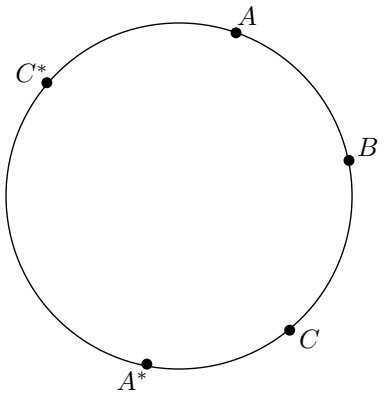
We see $AB + BD = AD$. Thus, $A - B - D$

- **Betweenness on the spherical plane**: Consider points A, C, with $A \neq C$, and distance $AC < \omega = \pi r$. So, A, C determine unique great circle (line) $\overleftrightarrow{AC}$. Let A^* be the antipode of A, and C^* be the antipode of C. We check all points B on $\overleftrightarrow{AC}$ and see in which locations there is betweenness $A - B - C$.

  First, consider B on minor arc $\widehat{AC}$. Notice that the minor arc $\widehat{AB}$ plus the minor arc $\widehat{BC}$ equals the minor arc $\widehat{AC}$. Thus,

$$d_{\mathbb{S}}(AB) + d_{\mathbb{S}}(BC) = d_{\mathbb{S}}(AC)$$

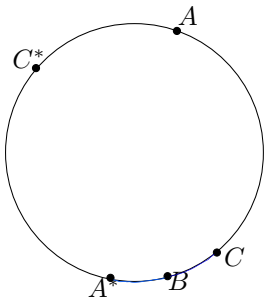
  Thus, $A - B - C$



  Next, let B be on the minor arc $\widehat{A^*C}$. Observe that

$$d_{\mathbb{S}}(AC) + d_{\mathbb{S}}(CB) = d_{\mathbb{S}}(AB)$$

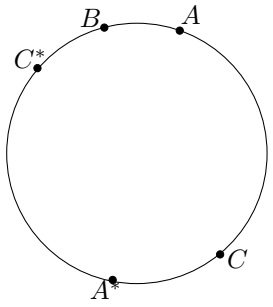
  Thus, $A - C - B$



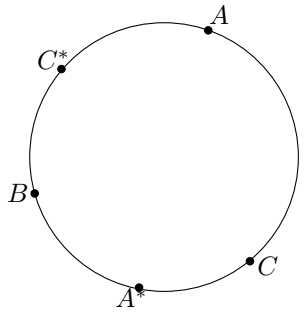
Next, let B be on the minor arc $\overset{\frown}{C^*A}$. Observe that

$$d_{\mathbb{S}}(BA) + d_{\mathbb{S}}(AC) = d_{\mathbb{S}}(BC)$$

Thus, $B - A - C$



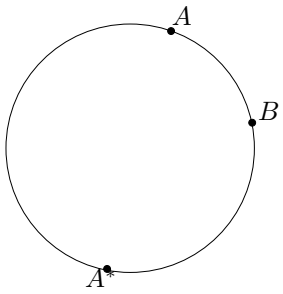
Next, let B be on the minor arc $\overset{\frown}{A^*C^*}$, any two of $d_{\mathbb{S}}(AB), d_{\mathbb{S}}(BC), d_{\mathbb{S}}(AC)$ add to more than πr, hence more than any distance on $\mathbb{S}$. Therefore, no two add to the third and A, B, C have no betweenness relation



Finally, consider two points A, A^*, where A^* is A's antipode. Let B be any point not equal to A or A^*. Then, B is collinear with A, A^*. Observe that

$$d_{\mathbb{S}}(AB) + d_{\mathbb{S}}(BA^*) = \pi r = d_{\mathbb{S}}(AA^*)$$

Thus, $A - B - A^*$



- **Betweenness theorem 1**:

  Theorem 6.1 (Symmetry of betweenness). For a general plane $\mathbb{P}$ with points, lines, distance, and satisfy the seven axioms, $A - B - C \iff C - B - A$

  Proof. Suppose that $A - B - C$, by definition, A, B, C are different and collinear. Hence, C, B, A are different and collinear, and $AB + BC = AC$

  By distance axiom three, $AB = BA$, $BC = CB$, and $AC = CA$. Thus,

  $$\begin{aligned} & AB + BC = AC \\ \implies & BA + CB = CA \end{aligned}$$

  But by the commutative property of $+$ in $\mathbb{R}$

  $$\begin{aligned} & BA + CB = CA \\ \implies & CB + BA = CA \end{aligned}$$

  Therefore, by the definition of betweenness, $C - B - A$. Thus, by similar steps, if $C - B - A$, then $A - B - C$ ■

- **Uniqueness Middle Theorem (UMT)**:

  Theorem 6.2 (UMT): If $A - B - C$ then $B - A - C$ and $A - C - B$ are false.

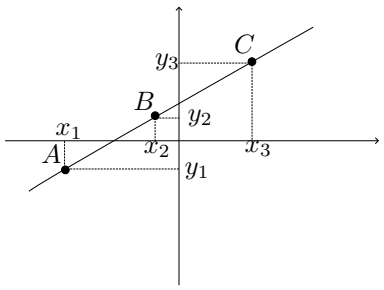
  Proof. Assume $A - B - C$, then A, B, C are different, collinear, and $AB + BC = AC$. We know by distance axioms 1 and 2 that each of AB, BC, and AC are greater than zero. Thus,

  $$AC > AB \quad \text{and } AC > BC$$

  Suppose for the sake of contradiction that $B - A - C$ is also true. Thus, BC would be larger than both $BA = AB$ and AC.

  Since this contradicts the fact that $AC > BC$, which must be true if $A - B - C$, it must be that $B - A - C$ is false. By similar steps, $A - C - B$ is also false. ■

- **Betweennes in** $\mathbb{M}$: Suppose $A - B - C$ is true in $\mathbb{E}$. Then, we have in the Minkowski plane



So we see

$$d_{\mathbb{M}}(AB) + d_{\mathbb{M}}(BC) = |x_1 - x_2| + |y_1 - y_2| + |x_2 - x_3| + |y_2 - y_3|$$

We can then drop the absolute value bars by examining the configuration and determining which order the subtraction needs to happen to yield a positive result. We have

$$\begin{aligned}&(x_2 - x_1) + (y_2 - y_1) + (x_3 - x_2) + (y_3 - y_2)\\ &= (x_3 - x_1) + (y_3 - y_1) = d_{\mathbb{M}}(AC)\end{aligned}$$

Thus, for $A - B - C$ in $\mathbb{E}$, $A - B - C$ in $\mathbb{M}$ holds true. Similarly, $B - A - C$ in $\mathbb{E}$ implies $B - A - C$ in $\mathbb{M}$, and $A - C - B$ in $\mathbb{E}$ implies $A - C - B$ in $\mathbb{M}$

So for three collinear points A, B, C in $\mathbb{E}$, exactly one (by the UMT) of $A - B - C$, $B - A - C$, $A - C - B$ occurs, and each relation implies the same relation happens in $\mathbb{M}$.

If $A-B-C$ happens in $\mathbb{M}$, then the other two do not by the UMT, so only $A-B-C$ will then be true in $\mathbb{E}$. We state

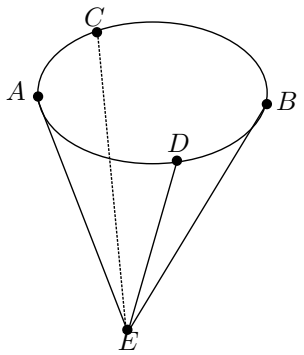
$$A - B - C \text{ in } \mathbb{E} \iff A - B - C \text{ in } \mathbb{M}$$

- **Betweenness among the planes**: We have

$$\begin{aligned}A\text{-}B\text{-}C \text{ in } \mathbb{E} &\iff A\text{-}B\text{-}C \text{ in } \mathbb{M}\\ A\text{-}B\text{-}C \text{ in } \mathbb{E} &\iff A\text{-}B\text{-}C \text{ in } \mathbb{G}\\ A\text{-}B\text{-}C \text{ in } \mathbb{E} &\iff A\text{-}B\text{-}C \text{ in } \mathbb{H}\end{aligned}$$

- **Inside out**: Consider $\mathbb{P} = \{A, B, C, D, E, F\}$, $\mathbb{L}:\ \ell = \{A, B, C, D\}, m = \{A, E\}, n = \{C, E\}, v = \{D, E\}$, and distance

	A	B	C	D	E
A	0	3	1	2	4
B	3	0	2	1	4
C	1	2	0	3	4
D	2	1	3	0	4
E	4	4	4	4	0



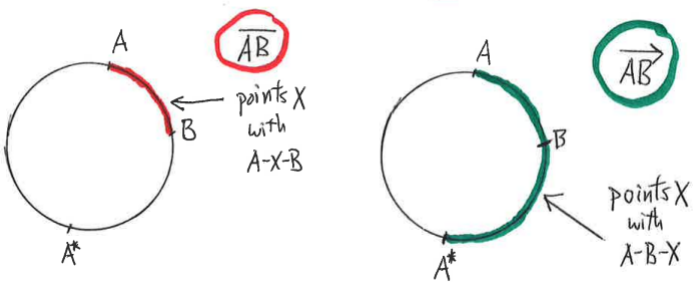
The seven axioms hold, $\mathbb{D} = \{0, 1, 2, 3, 4\}, \omega = 4$, and all betweenness occurs for points on ℓ

$$A - C - B \quad A - D - B \quad C - A - D \quad C - B - D$$

- **Segments and rays**: Let $A \neq B$ be points in $\mathbb{P}$ with $AB < \omega$. Then, there is a unique line through A, B, call it $\overleftrightarrow{AB}$
  - **The segment** $\overline{\{}AB\} = \{A, B\} \cup \{X : A - X - B\}$
  - **The ray** $\overrightarrow{\{}AB\} = \{A, B\} \cup \{X : A - X - B\} \cup \{X : A - B - X\}$

  Note: $\{X : A - X - B\} \cup \{X : A - B - X\} = \varnothing$

  Notation: $\overline{AB}, \overrightarrow{AB}, \overleftrightarrow{AB}$ denote sets of points, with $\{A, B\} \subseteq \overline{AB} \subseteq \overrightarrow{AB} \subseteq \overleftrightarrow{AB}$
- **Carriers**: We call the line $\overleftrightarrow{AB}$ the **carrier** of $\overrightarrow{AB}$
- **Segments and rays on** $\mathbb{S}$



Ray $\overrightarrow{AB}$ goes from A, through B, around to A^*. Since $A - B - A^*$, $\overrightarrow{AB}$ includes A^*

We have

  - **Segment** $\overline{AB} = \{A, B\} \cup \{X : A - X - B\}$ as usual
  - **Ray** $\overrightarrow{AB} = \{A, B\} \cup \{X : A - X - B\} \cup \{X : B - X - A^*\} \cup \{A^*\}$, where A^* is the antipode of A
- **Proving results about general (abstract) planes** $\mathbb{P}$: We only use the undefined terms point, line, distance, the definitions, the assumed axioms, previously proved results, arithmetic of $\mathbb{R}$, and logic.

Sketches from $\mathbb{E}$, while sometimes useful, are not valid for general proofs. General planes include many examples besides $\mathbb{E}$, and Euclidean pictures may not apply to them, and may be misleading.

We assume plane $\mathbb{P}$, in which we have points, lines, and the first seven axioms satisfied.

Recall, for points $A \neq B$, $AB < \omega$,

- **Betweenness**: $A-B-C$ if A, B, C are different, collinear, and $AB+BC = AC$
- **Segment** $\overline{AB} = \{A, B\} \cup \{X : A - X - B\}$
- **Ray** $\overrightarrow{AB} = \{A, B\} \cup \{X : A - X - B\} \cup \{X : A - B - X\}$

- **Proposition 6.3: Segments and lines**:

  Proposition.

  (a) $\overline{AB}$ lies in one line, the line $\overleftrightarrow{AB}$
  (b) $\overline{AB} = \overline{BA}$
  (c) If $x \in \overline{AB}$, with $X \neq B$, then $AX < AB$

  Proof a.) Since $\overline{AB}$ exists, we have $AB < \omega$. Thus, by incidence axioms three and four, there is exactly one line containing points A, and B. Namely, $\overleftrightarrow{AB}$. If X is any other point in $\overline{AB}$, then $A - X - B$ by definition of $\overline{AB}$. Thus, X is collinear with A, B by definition of betweenness, and hence, $x \in \overleftrightarrow{AB}$

  b.) We have

  $$\overline{AB} = \{A, B\} \cup \{X : A - X - B\} \tag{1}$$
  $$\overline{BA} = \{B, A\} \cup \{X : B - X - A\} \tag{2}$$

  But, since ordering in sets doesn't matter, $\{B, A\} = \{A, B\}$, and we have seen previously that $B - X - A = A - X - B$. Thus, (2) is precisely (1). That is, $\{B, A\} \cup \{X : B - X - A\} = \{A, B\} \cup \{X : A - X - B\}$, and therefore $\overline{AB} = \overline{BA}$

  c.) We have

  $$\overline{AB} = \{A, B\} \cup \{X : A - X - B\}$$

  Let $x \in \overline{AB}$, with $X \neq B$. Then, we have $A - X - B$, and $AX + XB = AB$. This implies that AB greater than both AX, XB, which means $AB > AX$.

  Note: Ray $\overrightarrow{AB}$ is also contained in exactly one line, the line $\overleftrightarrow{AB}$

  Also, $\overrightarrow{AB} = \overrightarrow{BA}$ mostly does not hold. We have

  $$\overrightarrow{AB} = \{A, B\} \cup \{X : A - X - B\} \cup \{X : A - B - X\}$$
  $$\overrightarrow{BA} = \{A, B\} \cup \{X : A - X - B\} \cup \{X : B - A - X\}$$

  Since $\{X : A-B-X\} \neq \{X : B-A-X\}$, it is not generally the case that $\overrightarrow{AB} = \overrightarrow{BA}$ (In general). The scenario where $\overrightarrow{AB} = \overrightarrow{BA}$ is when $\overline{AB}, \overline{BA}$ are exactly the line where they are contained. This fact is true in the IO example, since $\overline{AB} = \overline{BA} = \overline{CD} = \overline{DC} = \{A, B, C, D\} = \ell$

- **Proposition 6.4**

Proposition: Let A, B, C, D be collinear points with $0 < AB < \omega$, $0 < CD < \omega$, and $\overline{AB} = \overline{CD}$, then

(a) Either $\{A, B\} = \{C, D\}$ or $\{A, B\} \cap \{C, D\} = \varnothing$

(b) $AB = CD$

Proof (a) Part a says that $\{A, B\}$ and $\{C, D\}$ can have two (all) elements in common, or no elements in common. Thus, we show that it cannot be the case that they have one element in common

Suppose for the sake of contradiction that $\{A, B\}$ and $\{C, D\}$ have exactly one element in common. Assume it is $A = C$. Then, $A \neq B$, $B \neq C$, and $B \neq D$.

By definition of a segment, $D \in \overline{CD} = \{AB\}$, which implies A-D-B since $D \neq A$ and $D \neq B$. Also, $B \in \overline{AB} = \overline{CD}$ implies C-B-D (since $B \neq C$ and $B \neq D$). But, since $A = C$, we have C-B-$D = A$-B-D which cannot happen by the UMT since we know we have A-D-B. Thus, our assumption that $A = C$ must be false. A similar argument for the other equality pairs shows that the two sets must not contain exactly one common element.

Therefore, $\{A, B\} = \{C, D\}$ or $\{A, B\} \cap \{C, D\} = \varnothing$

Proof (b) If $\{A, B\} = \{C, D\}$ then $AB = CD$ by substitution. So, we may assume that $\{A, B\} \cap \{C, D\} = \varnothing$.

We have $C, D \in \overline{CD} = \overline{AB}$, and $C, D \neq A$ or B, which implies A-C-B and A-D-B. Similarly, $A, B \in \overline{AB} = \overline{CD}$ yields C-A-D and C-B-D. Hence, we have $AC + AD = CD$ and $CB + BD = CD$. Adding these two equations and making suitable substitutions yields $2CD = 2AB$, hence, $CD = AB$ ■

4.6 Three axioms for the line

- **Proposition 7.1**

  Proposition. If A-B-C and A-C-D, then A, B, C, D are distinct and collinear

  Proof. Be the definition of betweenness, A, C, D are distinct and collinear, and $AC + CD = AD$. Since $CD > 0$, $AC < AD$. But, since $AD \leqslant \omega$, it must be that $AC < \omega$. Thus, A, C are together in a unique line (the line $\overleftrightarrow{AC}$)

  Also, A, B, C are distinct and collinear. Thus, B, D are both collinear with A and C which implies all four points must be in $\overleftrightarrow{AC}$, and hence they are all collinear.

  The only way two of A, B, C, D could be equal is if $B = D$. But then, substituting B for D in A-C-D, we get A-C-B. This contradicts A-B-C and the UMT. Thus, all four points are different. ■

- **Definition:** Define A-B-C-D to mean the following betweenness relations are all satisfied

  $$A\text{-}B\text{-}C \quad A\text{-}B\text{-}D \quad A\text{-}C\text{-}D \quad B\text{-}C\text{-}D$$

  Also, for collinear points A, B, C, D

  $$A\text{-}B\text{-}C\text{-}D \implies AB + BC + CD = AD$$

- **Proposition 7.2** If A-B-C-D, then A, B, C, D are distinct and collinear, and D-C-B-A

  Proof. A-B-C-D implies A-B-C, A-B-D, A-C-D, B-C-D. Since A-B-C and A-C-D are true, then A, B, C, D are distinct and collinear, if we switch the order on the four betweenness relations (first point and last for each of them), we get precisely

  $$D\text{-}C\text{-}B\text{-}A$$

  ■

- **Betweenness of points axiom (Ax. BP)**: If A, B, C are distinct, collinear points, and if $AB + BC \leqslant \omega$, then there exists a betweenness relation among A, B, C

  What this is really saying is that if **any** of $AB + BC$, $BA + AC$, $AC + CB$ is $\leqslant \omega$, then there is a betweenness relation.

  Note: If Ax.BP is true for a plane $\mathbb{P}$, and if $AB + BC \leqslant \omega$ for distinct collinear A, B, C, then there is a betweenness relation, but not necessarily A-B-C

  When $\omega = \infty$, then for any distinct collinear A, B, C, $AB + BC < \infty = \omega$, so there will be a betweenness relation

- **What would make Ax.BP false?** Three collinear points A, B, C so that at least one of $AB + BC \leqslant \omega, AC + CB \leqslant \omega, BA + AC \leqslant \omega$, and no betweenness relation for A, B, C exists

Note: If there are no lines with three points, then the axiom is vacuously true.

- **Planes with first 8 axioms**: Consider a general plane $\mathbb{P}$ with points, lines, distance, and all 8 axioms true. We can establish some important properties of all these planes
- **Triangle inequality for the line**: If A, B, C are any three distinct, collinear points, then

$$AB + BC \geqslant AC$$

Note: Don't worry about why the word triangle is in the name. Also, the triangle inequality is not necessarily true without Ax.BP

Proof. We examine two cases, either $AB + BC > \omega$, or $AB + BC \leqslant \omega$. In this first case, $AB + BC > \omega$ implies $AB + BC > AC$, since by the definition of ω, $AC \leqslant \omega$.

Next, we consider $AB + BC \leqslant \omega$. In this case, by Ax.BP, there must exist a betweenness relation among A, B, C. One of the following must be satisfied

$$A\text{-}B\text{-}C$$
$$B\text{-}A\text{-}C$$
$$A\text{-}C\text{-}B$$

Assume its A-B-C, then by definition of A-B-C, we have $AB + BC = AC$, which implies $AB + BC \geqslant AC$ is satisfied.

Next, assume its B-A-C, then $BA + AC = BC$. We have

$$\begin{aligned} AC &= BC - AB \\ AC + 2AB &= BC - AB + 2AB \\ AC + 2AB &= BC + AB \\ AC + 2AB &= AB + BC \end{aligned}$$

In this case, since $2AB > 0$ by distance axiom 2, we have $AC + 2AB \geqslant AC$. Thus, $AC + 2AB = AB + BC \geqslant AC$

Lastly, assume the relation we have is A-C-B, then $AC + CB = AB$. We have

$$\begin{aligned} AC + CB &= AB \\ AC &= AB - CB \\ AC + 2CB &= AB - CB + 2CB \\ AC + 2CB &= AB + CB \\ AC + 2CB &= AB + BC \end{aligned}$$

Similar to the previous argument, since $AC + 2CB \geqslant AC$, we have $AC + 2CB = AB + BC \geqslant AC$

Therefore, $AB + BC \geqslant AC$ ■

- **Rule of insertion**:
  - If A-B-C and A-X-B, then A-X-B-C
  - If A-B-C and B-X-C, then A-B-X-C

Proof (a). Since A-B-C, and A-X-B, then we know that A, X, B, C are distinct, collinear.

By the definition of betweenness, we have

$$AB + BC = AC$$
$$AX + XB = AB$$

Thus,

$$AX + XB + BC = AC$$

By the triangle inequality, we have $XB + BC \geqslant XC$. Thus,

$$AC = AX + XB + BC \geqslant AX + XC$$

But the triangle inequality also implies that

$$AX + XC \geqslant AC$$

Thus, since $AX + XC \leqslant AC \leqslant AX + XC$. Thus, it must be that $AC = AX + XC$. Hence, A-X-C. Next, plugging $AC = AX + XC$ into $AC = AX + XB + BC$ yields

$$\begin{aligned} & AX + XC = AX + XB + BC \\ \implies\ & XB + BC = XC \end{aligned}$$

Thus, X-B-C.

Proof (b) If A-B-C and B-X-C, then C-B-A and C-X-B, so by part (a), we have

$$C\text{-}X\text{-}B\text{-}A$$

Which means A-B-X-C ■

- **What does the betweenness of points axiom get us?** The triangle inequality and the insertion theorem
- **Quadrichotomy Axiom for Points (Ax.QP)**: If A, B, C, X are distinct, collinear points, and if A-B-C. Then, at least one of the following must hold

$$X\text{-}A\text{-}B, \quad A\text{-}X\text{-}B, \quad B\text{-}X\text{-}C, \quad \text{or} \quad B\text{-}C\text{-}X$$

Thus, Ax.QP says that whenever A-B-C (say on line ℓ), then any other point X on line ℓ is in either $\overrightarrow{BA}$ or $\overrightarrow{BC}$. That is,

$$\ell = \overrightarrow{BA} \cup \overrightarrow{BC}$$

Ax.QP is true for

- $\mathbb{E}$
- $\mathbb{M}$
- $\mathbb{G}$
- $\mathbb{H}$
- $\mathbb{S}$
- $\mathbb{R}^3$
- $\hat{\mathbb{E}}$ (bumpy plane)

It is also true vacuously for the white stripes model, the TDM, and the Fano plane

Ax.QP is also true for the Inside Out (IO) example. It is vacuously two for the 2-point lines, but we can also check that it is satisfied for $\ell = \{A, B, C, D\}$

Note: If the first 8 axioms are true for a plane $\mathbb{P}$, and if $\omega = \infty$, then the statement of Ax.QP can be proved to hold true in $\mathbb{P}$. A key reason for this is that because $\omega = \infty$, any three collinear points must have a betweenness relation.

- **When is Ax.QP false?**: In a plane with at least four collinear points A, B, C, X with A-B-C, and none of

$$X\text{-}A\text{-}B, \quad A\text{-}X\text{-}B, \quad B\text{-}X\text{-}C, \quad \text{or} \quad B\text{-}C\text{-}X$$

Are true

- **Proposition 7.5**: If $X \neq Y$ are points distinct from A on ray $\overrightarrow{AB}$, then at least one of A-X-Y or A-Y-X or X, Y in $\overline{AB}$ is true.

Proof. If $Y = B$, then either A-X-Y or A-Y-X by definition of $\overrightarrow{AB}$, so we may assume that $Y \neq B$, and similarly that $X \neq B$.

Now either A-X-B or A-B-X, and either A-Y-B or A-B-Y by definition of $\overrightarrow{AB}$, we consider each of the four possibilities.

Suppose that A-B-X and A-Y-B are true. Then, the rule of insertion says that A-Y-B-X, which gets us A-Y-X. Similarly, if A-B-Y and A-X-B are true, then we get A-X-Y by the rule of insertion

Suppose that A-B-X and A-B-Y are true. A-B-X and Ax.QP says that one of

$$Y\text{-}A\text{-}B, \quad A\text{-}Y\text{-}B, \quad B\text{-}Y\text{-}X, \quad \text{or} \quad B\text{-}X\text{-}Y$$

Is true. But, since A-B-Y is true, then we can't have either Y-A-B or A-Y-B because this would contradict the UMT. Thus, we know we have either B-Y-X or B-X-Y. If B-Y-X then by the rule of insertion we get A-Y-X. If B-X-Y then the rule of insertion gets us A-X-Y

Finally, suppose that A-X-B and A-Y-B. The best we can do is assert that X, Y are in $\overline{AB}$ by the definition of a segment.

- **What we have so far**: Nine axioms for a general plane, which are satisfied in the following examples that we have seen
  - $\mathbb{E}$
  - $\mathbb{M}$
  - $\mathbb{S}$
  - $\mathbb{H}$
  - $\mathbb{G}$
  - $\hat{\mathbb{E}}$
  - ws
  - $\mathbb{R}^3$
  - Fano

– TDM
– IO

None of the axioms so far says that rays have to exist on any particular line, or even in the plane overall. We need another axiom to guarantee that rays (and segments) exist on every line.

- **Nontriviality Axiom (Ax.N)**: For any point A on a line ℓ there exists a point B on ℓ with $0 < AB < \omega$

  This axiom is true for the planes in which $\omega = \infty$ ($\mathbb{E}$, $\mathbb{M}$, $\mathbb{H}$, $\mathbb{G}$, $\mathbb{R}^3$, $\hat{\mathbb{E}}$, ws)

  This axiom is also true for $\mathbb{S}$ and Fano, where $\omega < \infty$

  Ax.N is false for TDM, and for the two point lines in IO.

  We now assume the IO axioms for our general plane $\mathbb{P}$

- **Theorem 7.6**: This next theorem is the only one in chapters 6-9 that is about points on more than a single line

  Theorem. For any point A on a line ℓ there exists a point C not on ℓ with $0 < AC < \omega$

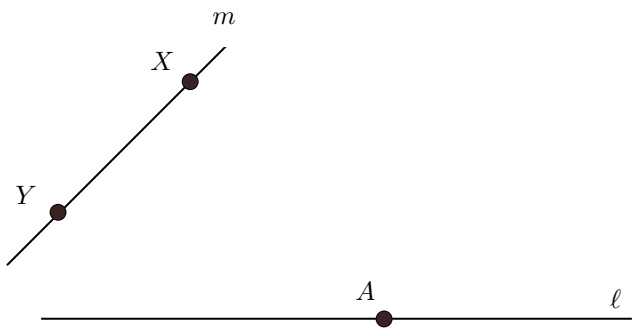
  To prove this, we first show that there is a point C not on ℓ, then we show that this point C satisfies $0 < AC < \omega$

  Proof.

  First, we get a point X not on ℓ. By incidence axiom 1, there is a line $m \neq \ell$, by incidence axiom 2, there is a point X on m. By Ax.N, there is a point y on m with

  $$0 < XY < \omega$$

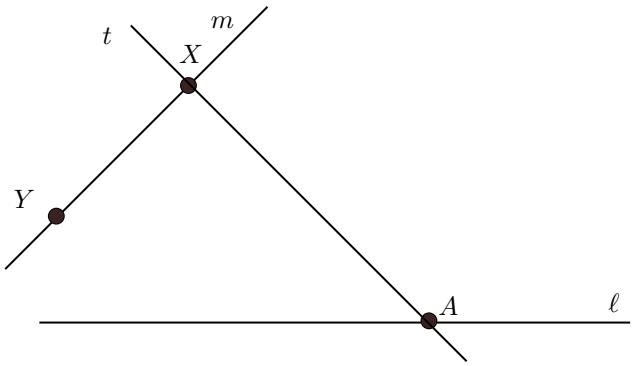
  Thus, $\overleftrightarrow{XY}$ is the **unique** line through X and Y. If X, Y were both on ℓ, then $\ell = \overleftrightarrow{XY} = m$, which contradicts there being a unique line through X, Y. Thus, at least one of X or Y, say X is not on ℓ.



  Second, we need to get a point C not on ℓ, with $0 < AC < \omega$

There is a line t through A and X, with $t \neq \ell$, since X is on t, but not on ℓ. Ax.N says that there is a point C on t with $0 < AC < \omega$. By incidence axiom 4, t is the only line through A and C. If C were on ℓ, then both A and C on ℓ would imply $\ell = t$, which is a contradiction.

Therefore, C is not on ℓ, and $0 < AC < \omega$.



- **Note about Ax.N** This axiom stops us from construction examples of planes in which all points are collinear. (See proof above)
- **Important fact**: Suppose X is a point on a ray $\overrightarrow{AB}$ in a general plane.
  1. If A-X-B then $AX < AB$
  2. If A-B-X then $AX > AB$
  3. IF $X = B$ then $AX = AB$

4.7 Exam 1 Axioms definitions and theorems

- **Euclids fifth postulate**: If a straight line intersects two straight lines such that the interior angles on one side add up to less than two right angles, then the two straight lines, if extended indefinitely, will meet on the side where the angles are less than two right angles.
- **Playfairs postulate**: "Through a given point not on a line, there is exactly one line parallel to the given line"
- $\mathbb{E}$
  - **Points:** (x, y)
  - **Lines:** Each *nonvertical line* ℓ in $\mathbb{E}$ consists of all points (x, y), where $y = mx+b$ for some fixed m and b. Each *vertical line* ℓ consists of all (x, y), where $x = a$ for some fixed a
  - **Distance**: $e(AB) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = |x_1 - x_2| \sqrt{1 + m^2}$
- $\mathbb{M}$
  - **Points:** (x, y)
  - **Lines**: Same as in $\mathbb{E}$
  - **Distance**: $d_{\mathbb{M}}(Ab) = |x_2 - x_1| + |y_2 - y_1| = |x_1 - x_2| (1 + |m|)$
- $\mathbb{S}$:
  - **Points:** (x, y)
  - **Lines:** Great circle through two points
  - **Distance**:
$$d_{\mathbb{S}} = r\theta = r \cos^{-1}\left(\frac{ax + by + cz}{r^2}\right)$$
- $\mathbb{H}$
  - **Points**: (x, y)
  - **Lines**: Chords on unit circle through two given points
  - **Distance**: $d_{\mathbb{H}} = \ln\left(\frac{e(AN)e(BM)}{e(AM)e(BM)}\right)$

    Use the following formulas if M, N have the chance of being misplaced
$$\left|\ln\left(\frac{e(AN)e(BM)}{e(AM)e(BN)}\right)\right| = \left|\ln\left(\frac{e(AM)e(BN)}{e(AN)e(BM)}\right)\right|$$
- $\mathbb{G}$
  - **Points:** (x, y), $x \leqslant 0$,or $x > 1$
  - **Lines:** Lines in $\mathbb{G}$ are defined to be the same as in $\mathbb{E}$, except that for any nonvertical line $y = mx + b$, the part in the missing strip is deleted. So a typical nonvertical line ℓ consists of all (x, y) with $y = mx+b$ $(m, b$ fixed) and with $x \leqslant 0$ or $x > 1$
  - **Distance**:
$$d_{\mathbb{G}}(AB) = \begin{cases} e(AB) & \text{for } A, B \text{ on the same side of the gap} \\ e(AB) - e(CD) & \text{for } A, B \text{ on the opposite sides of the gap} \end{cases}$$

  - $\hat{\mathbb{E}}$
  - **Fano**
  - **White stripes**
  - **Trivial discrete**
  - **Inside out**

- **Interior and Exterior angles**: Interior angles are the angles inside the triangle. Each vertex of the triangle has one interior angle. The sum of the interior angles of a triangle is always 180°

  Exterior angles are the angles formed outside the triangle when one side of the triangle is extended. At each vertex, an exterior angle is supplementary to the interior angle (they add up to 180°

  If an interior angle at a vertex is A, the corresponding exterior angle E is:

  $$E = 180° - A$$

  The sum of the exterior angles of a triangle (one at each vertex) is always 360° , regardless of the shape of the triangle.

- **Remote angles**: Remote angles refer to the interior angles of a triangle that are not adjacent to a given exterior angle

- **More on points**:
  - **Collinear points**: Points that lie on the same straight line.
  - **Noncollinear points**: Points that do not lie on the same straight line.
  - **Coplanar points**: Points that lie on the same plane.
  - **Concurrent Points**: Points where three or more lines intersect.
  - **Equidistant Points**: Points that are all the same distance from a particular point or object.
  - **Lattice Points**: Points with integer coordinates.
  - **Interior points:** Points that lie inside a given shape.
  - **Exterior points:** Points that lie outside a given shape.

- **Congruent triangles**: Congruent triangles are triangles that are exactly the same in shape and size. This means that all corresponding sides and angles of one triangle are equal to those of the other triangle.

- **Vertical (opposite) angles**: Vertical angles (also called opposite angles) are the angles that are formed by two intersecting lines and are opposite to each other

- **Theorem (*Exterior angle inequality*)**: An exterior angle of a triangle is greater than either remote interior angle. That is, if $\triangle ABC$ is a any triangle, and point D is on the extension of segment $\overline{BC}$ through C, then

  $$\angle ACD > \text{ both } \angle A \text{ and } \angle B$$

- **Relationship of angles**:

- **Interior angles**: Interior angles are the angles that are inside the transversal configuration. Angles a, b, c, d are interior
- **Exterior angles**: Exterior angles are the angles that are inside the transversal configuration. Angles e, f, g, h are exterior
- **Consecutive interior angles**: Pairs of interior angles that are on the same side of the transversal. Angles c, d are consecutive interior, and a, b are consecutive interior
- **Consecutive exterior angles**: Pairs of exterior angles that are on the outside of the transversal configuration. Angles e, g are consecutive exterior, angles f, h are consecutive exterior
- **Alternate interior angles**: Pairs of interior angles that are on opposite sides but not complementary, angles b, d and a, c are alternate interior
- **Alternate exterior angles**: Pairs of exterior angles that are on opposite sides but not complementary, angles e, h, and f, g are alternate exterior
- **Vertical angles**: Angles that are opposite each other, formed when two lines intersect. Vertical angles are of equal measure. Pairs d, h - a, g - e, b - and f, c are vertical
- **Supplementary angles**: Angle pairs that sum to 180, pairs a, h - d, g - f, b - and e, c are supplementary
- **Complementary angles**: Angle pairs that sum to 90, none in the transversal configuration

- **Proposition (Equal alternate interior angles)**. Suppose $a + b = 180$, then $b = d$, and $c = a$.
- **Upper bounds**: Suppose S is a set of real numbers, we define $b \in \mathbb{R}$ as an *upper bound* for S if for all $x \in S, x \leqslant b$

  The negation of this definition is, there exists $x \in S$ such that $x \nleq b$, or $x > b$. Thus, to prove some b is not an upper bound for S, we can show that some element of S is greater than b

  There are of course sets that do not have any upper bounds. Consider the set $S = \{n : \ n \in \mathbb{N} \text{ and } n > 0\}$. This set has no upper bound.

  If $S = \varnothing$, then every $b \in \mathbb{R}$ is an upper bound for S. This statement is vacuously true.
- **Least upper bound (supremum)**: $c \in \mathbb{R}$ is a *least upper bound* of a set S of real numbers if
  1. c is an upper bound for S
  2. $c \leqslant b$ for all upper bounds b of S

  Note: The supremum of a set S is denoted $b = \sup(S)$, where b is the supremum of the set
- **Least upper bound property of** $\mathbb{R}$: If S is a nonempty set of real numbers that has an upper bound in $\mathbb{R}$, then S has a least upper bound (l.u.b) in $\mathbb{R}$

This justifies, among other things, that infinite decimals exist as real numbers, since an infinite decimal can be defined as the least upper bound of the set of all its finite truncations. For example, suppose S is the set of all finite decimal expansions of π.

$$S = \{3, 3.1, 3.14, 3.141, 3.1415, ...\}$$

Then, S as an $l.u.b$ π, and $\pi \notin S$

- **Least upper bound proposition**

  Proposition. Let S be a nonempty set of real numbers that has a least upper bound $b \in \mathbb{R}$. Let $t \in \mathbb{R}$ such that $t < b$. Then, there exists some $s \in S$ such that $t < s \leqslant b$.

- **Lower bounds**: Let S be a nonempty set of real numbers. Then $g \in \mathbb{R}$ is a *lower bound* for S if $g \leqslant x$ for all $x \in S$.

- **Greater lower bounds (Infimum)**: $h \in \mathbb{R}$ is a *greatest lower bound*, also called the *infimum*, or *inf* for S if h is a lower bound for S and $h \geqslant g$ for all lower bounds g of S

- **Infimum proposition**

  Proposition. Let S be a nonempty set or real numbers that has a lower bound in $\mathbb{R}$. Then S has a infimum in $\mathbb{R}$

- **Undefined terms**:
  - $\mathbb{P}$**:** Set of elements, called **points.**
  - $\mathbb{L}$**:** Collection of subsets of $\mathbb{P}$, called **lines**
  - A function $d: \mathbb{P} \times \mathbb{P} \to \mathbb{R}$, called a **distance function**

- **Notation, terminology**
  - A line is a set of points
  - If P is on the line m ($P \in m$), we say that "P is on m", or "m goes through P"
  - If two or more points are on the same line, we say they are **collinear**
  - Denote distance $d(P, Q)$, or $d(PQ)$, or just PQ

- **Axiom of distance**: For all points P, Q
  1. $PQ \geqslant 0$
  2. $PQ = 0 \iff P = Q$
  3. $PQ = QP$

- **Axioms of incidence**
  1. There are at least two different lines
  2. Each line contains at least two different points
  3. Each pair of points are together in at least one line
  4. Each pair of points P, Q, with $PQ < \omega$ are together in at most one line

- **Betweenness of points axiom (Ax. BP)**: If A, B, C are distinct, collinear points, and if $AB + BC \leqslant \omega$, then there exists a betweenness relation among A, B, C

  What this is really saying is that if **any** of $AB + BC$, $BA + AC$, $AC + CB$ is $\leqslant \omega$, then there is a betweenness relation.

Note: If Ax.BP is true for a plane $\mathbb{P}$, and if $AB + BC \leqslant \omega$ for distinct collinear A, B, C, then there is a betweenness relation, but not necessarily A-B-C

When $\omega = \infty$, then for any distinct collinear A, B, C, $AB + BC < \infty = \omega$, so there will be a betweenness relation

- **Quadrichotomy Axiom for Points (Ax.QP)**: If A, B, C, X are distinct, collinear points, and if A-B-C. Then, at least one of the following must hold

$$X\text{-}A\text{-}B, \quad A\text{-}X\text{-}B, \quad B\text{-}X\text{-}C, \quad \text{or} \quad B\text{-}C\text{-}X$$

Thus, Ax.QP says that whenever A-B-C (say on line ℓ), then any other point X on line ℓ is in either $\overrightarrow{BA}$ or $\overrightarrow{BC}$. That is,

$$\ell = \overrightarrow{BA} \cup \overrightarrow{BC}$$

- **Nontriviality Axiom (Ax.N)**: For any point A on a line ℓ there exists a point B on ℓ with $0 < AB < \omega$

This axiom is true for the planes in which $\omega = \infty$ ($\mathbb{E}$, $\mathbb{M}$, $\mathbb{H}$, $\mathbb{G}$, $\mathbb{R}^3$, $\hat{\mathbb{E}}$, ws)

This axiom is also true for $\mathbb{S}$ and Fano, where $\omega < \infty$

- **Betweenness**: Let $\mathbb{P}$ be a plane with points, lines, distance, and satisfy the seven axioms (3 distance, 4 incidence). Define

Definition. Point B lies **between** points A and C, denoted $A - B - C$ provide that

1. A, B, and C are different and collinear
2. $AB + BC = AC$

- **Uniqueness Middle Theorem (UMT)**:

Theorem: If $A - B - C$ then $B - A - C$ and $A - C - B$ are false.

- **Segments and rays**: Let $A \neq B$ be points in $\mathbb{P}$ with $AB < \omega$. Then, there is a unique line through A, B, call it $\overleftrightarrow{AB}$
  - **The segment** $\overline{\{}AB\} = \{A, B\} \cup \{X : A - X - B\}$
  - **The ray** $\overrightarrow{\{}AB\} = \{A, B\} \cup \{X : A - X - B\} \cup \{X : A - B - X\}$

Note: $\{X : A - X - B\} \cup \{X : A - B - X\} = \varnothing$

Notation: $\overline{AB}, \overrightarrow{AB}, \overleftrightarrow{AB}$ denote sets of points, with $\{A, B\} \subseteq \overline{AB} \subseteq \overrightarrow{AB} \subseteq \overleftrightarrow{AB}$

- **Segments and rays on** $\mathbb{S}$
  - **Segment** $\overline{AB} = \{A, B\} \cup \{X : A - X - B\}$ as usual
  - **Ray** $\overrightarrow{AB} = \{A, B\} \cup \{X : A - X - B\} \cup \{X : B - X - A^*\} \cup \{A^*\}$, where A^* is the antipode of A

- **Proposition: Segments and lines**:

Proposition.

(a) $\overline{AB}$ lies in one line, the line $\overleftrightarrow{AB}$

(b) $\overline{AB} = \overline{BA}$

(c) If $x \in \overline{AB}$, with $X \neq B$, then $AX < AB$

- **Proposition**

  Proposition: Let A, B, C, D be collinear points with $0 < AB < \omega$, $0 < CD < \omega$, and $\overline{AB} = \overline{CD}$, then

  (a) Either $\{A, B\} = \{C, D\}$ or $\{A, B\} \cap \{C, D\} = \varnothing$

  (b) $AB = CD$

- **Proposition**

  Proposition. If A-B-C and A-C-D, then A, B, C, D are distinct and collinear

  Proof. Be the definition of betweenness, A, C, D are distinct and collinear, and $AC + CD = AD$. Since $CD > 0$, $AC < AD$. But, since $AD \leqslant \omega$, it must be that $AC < \omega$. Thus, A, C are together in a unique line (the line $\overleftrightarrow{AC}$)

  Also, A, B, C are distinct and collinear. Thus, B, D are both collinear with A and C which implies all four points must be in $\overleftrightarrow{AC}$, and hence they are all collinear.

  The only way two of A, B, C, D could be equal is if $B = D$. But then, substituting B for D in A-C-D, we get A-C-B. This contradicts A-B-C and the UMT. Thus, all four points are different. ■

- **Definition:** Define A-B-C-D to mean the following betweenness relations are all satisfied

  $$A\text{-}B\text{-}C \quad A\text{-}B\text{-}D \quad A\text{-}C\text{-}D \quad B\text{-}C\text{-}D$$

  Also, for collinear points A, B, C, D

  $$A\text{-}B\text{-}C\text{-}D \implies AB + BC + CD = AD$$

- **Proposition.** If A-B-C-D, then A, B, C, D are distinct and collinear, and D-C-B-A

- **Triangle inequality for the line**: If A, B, C are any three distinct, collinear points, then

  $$AB + BC \geqslant AC$$

  Note: Don't worry about why the word triangle is in the name. Also, the triangle inequality is not necessarily true without Ax.BP

- **Rule of insertion**:
  - If A-B-C and A-X-B, then A-X-B-C
  - If A-B-C and B-X-C, then A-B-X-C
- **What does the betweenness of points axiom get us?** The triangle inequality and the insertion theorem
- **Proposition 7.5**: If $X \neq Y$ are points distinct from A or ray $\overrightarrow{AB}$, then at least one of A-X-Y or A-Y-X or X, Y in $\overline{AB}$ is true.
- **Theorem 7.6**: This next theorem is the only one in chapters 6-9 that is about points on more than a single line

  Theorem. For any point A on a line ℓ there exists a point C not on ℓ with $0 < AC < \omega$

  To prove this, we first show that there is a point C not on ℓ, then we show that this point C satisfies $0 < AC < \omega$
- **Important fact**: Suppose X is a point on a ray $\overrightarrow{AB}$ in a general plane.
  1. If A-X-B then $AX < AB$
  2. If A-B-X then $AX > AB$
  3. IF $X = B$ then $AX = AB$

4.8 The real ray axiom, Antipodes, and opposite rays

- **Real ray Axiom (Ax.RR)**: For any ray $\overrightarrow{AB}$, and for any real number s with $0 \leqslant s \leqslant \omega$, there is a point X in $\overrightarrow{AB}$ with $AX = s$ This axiom says that every nonnegative real number not exceeding ω produces at least one point on the ray.

  So by $Ax.RR$, the distances AX, for all points X, $\overrightarrow{AB}$ covers all real numbers in $[0, \infty]$ if $\omega < \infty$, and all real numbers in $[0, \infty)$ if $\omega = \infty$

- **Theorem 8.1**: If $\omega = \infty$, then $\mathbb{D} = [0, \infty)$; if $\omega < \infty$, then $\mathbb{D} = [0, \omega]$

  Since there are infinitely many real numbers in the interval $[0, \omega]$, for $\omega < \infty$, as well as in $[0, \infty)$, there must be infinitely many points on ray $\overrightarrow{AB}$

  By Ax.N, any line ℓ in $\mathbb{P}$ contains points $A \neq B$, with $AB < \omega$ contains points $A \neq B$, with $AB < \omega$, hence ℓ contains ray $\overrightarrow{AB}$. Since $\overrightarrow{AB}$ has infinitely many points, so does ℓ

  The points in $\overline{AB}$ are exactly the points X in $\overrightarrow{AB}$ with $AX = s \leqslant AB$. Since $[0, AB]$ contains infinitely many real numbers, there are also infinitely many on $\overline{AB}$. These remark proves the next theorem

- **Theorem 8.2** Each segment, ray, and line has infinitely many points.

  Note that Ax.RR is false for WS and Fano, so this theorem does not hold. Also, Ax.RR is vacuously true for TDM, since there are no rays.

- **Proposition (needs proof) 8.11** Let A, B be any two points on line m, with $0 < AB < \omega$. Then, there exists a point C on m with C-A-B and $CB < \omega$.

- **Theorem 8.3**

  Proposition. If $X \neq Y$ are points different from A on ray $\overrightarrow{AB}$, then one of A-X-Y or A-Y-X is true.

  Proof. Suppose toward a contradiction that the conclusion is false. So not A-X-Y and not A-Y-X

  By prop 7.5, X, Y are in $\overline{AB}$. If $Y = B$, then X in $\overline{AB}$ implies A-X-B, so A-X-Y, which is a contradiction

  Similarly, if $X = B$, then Y in $\overline{AB}$ implies A-Y-B, so A-Y-X, contradiction.

  So neither X nor $Y = B$, hence A-X-B, A-Y-B

  Now, we use Ax.RR just to produce one more point on $\overrightarrow{AB}$. We know $AB < \omega$ by definition of $\overrightarrow{AB}$. So, we pick a point E such that $AE = s$, with $AB < s \leqslant \omega$ (there are infinitely many). Since $AE > AB$, E must be in the ray $\overrightarrow{AB}$ but not in the segment $\overline{AB}$. Thus, A-B-E

  We have $AB + BE = AE$. Also, $BE = EB < AE \leqslant \omega$. So, $EB < \omega$, and $\overrightarrow{EB}$ is defined

We have A-X-B and A-B-E, so by ROI, we have A-X-B-E implies E-B-X

Also, A-Y-B and A-B-E by the ROI says we have A-Y-B-E, which implies E-B-Y

So, points X, Y are in ray $\overrightarrow{EB}$ but not in $\overline{EB}$. Let's now shift our focus to the ray $\overrightarrow{EB}$. We can apply prop 7.5 to points X, Y on ray $\overrightarrow{EB}$. Either E-X-Y or E-Y-X. Equivalently, either Y-X-E or X-Y-E

Recall that we have A-Y-B-E, which gives us A-Y-E, and A-X-B-E gives us A-X-E

If Y-X-E, then A-Y-E and ROI gives us A-Y-X-E, which implies A-Y-X

If X-Y-E, then A-X-E and ROI gives us A-X-Y-E, which implies A-X-Y

So, A-Y-X or A-X-Y holds anyway, contradicting our initial supposition

Therefore, A-X-Y or A-Y-X is true ■

- **Theorem 8.4**

  Proposition. If C is any point on ray $\overrightarrow{AB}$ with $0 < AC < \omega$, then $\overrightarrow{AC} = \overrightarrow{AB}$

  Proof. If $C = B$, then trivially $\overrightarrow{AC} = \overrightarrow{AB}$. So we may assume that $C \neq B$, and we already know that $C \neq A$ since $AC > 0$. Since $0 < AC < \omega$, ray $\overrightarrow{AC}$ is defined.

  Now, C on $\overrightarrow{AB}$ with $C \neq A$ or B implies A-C-B or A-B-C. These betweenness relations imply $B \in \overrightarrow{AC}$ by definition of $\overrightarrow{AC}$

  If $X \neq A$ or C is any other point on $\overrightarrow{AB}$, theorem 8.3 applied to $\overrightarrow{AB}$ implies A-X-C or A-C-X, which implies $x \in \overrightarrow{AC}$. Therefore, $\overrightarrow{AB} \subseteq \overrightarrow{AC}$

  If $X \neq A$ or B is any other point on $\overrightarrow{AC}$, theorem 8.3 applied to $\overrightarrow{AC}$ implies A-X-B or A-B-X, which implies $x \in \overrightarrow{AB}$. Therefore, $\overrightarrow{AC} \subseteq \overrightarrow{AB}$

  Hence, $\overrightarrow{AB} = \overrightarrow{AC}$ and there is nothing sacred about the point B in the ray $\overrightarrow{AB}$ ■

- **Endpoints**:

  Definition. Point A is called an endpoint of ray $\overrightarrow{AB}$

- **Proposition**

  Proposition 8.5: A ray has at most two endpoints

  Proof. Suppose toward a contradiction that ray h has three different endpoints A, C, E. Thus,

  $$\overrightarrow{AB} = \overrightarrow{CD} = \overrightarrow{EF}$$

  With A, C, E distinct, we apply theorem 8.3 three times

1. Points A, C, E on $\overrightarrow{AB}$ implies A-C-E or A-E-C, so not C-A-E by UMT
2. Points A, C, E on $\overrightarrow{CD}$ implies C-A-E or C-E-A, by (1) it cannot be C-A-E, thus it must be C-E-A
3. Points A-C-E on $\overrightarrow{EF}$ implies E-A-C or E-C-A. This contradicts C-E-A by UMT

■

- **More on Ax.RR**: Given a real number s with $0 \leqslant s \leqslant \omega$, there is nothing explicit in $Ax.RR$ about there being only one point X in $\overrightarrow{AB}$ with $AX = s$, we now prove that this is indeed the case.
- **Theorem 8.6 (Unique distances for Rays (UDR))**

  Proposition For any ray $\overrightarrow{AB}$ and any real number s with $0 \leqslant s \leqslant \omega$, there is a **unique** point X on $\overrightarrow{AB}$ with $AX = s$. X is in $\overline{AB}$ if and only if $s \leqslant AB$

  Proof. For $s = 0$, $X = A$ is the unique point in $\overrightarrow{AB}$ with $AX = 0$, by Ax.D2

  If $X \neq Y$ are points distinct from A on $\overrightarrow{AB}$, theorem 8.3 implies either A-X-Y or A-Y-X. So either $AX < AY$ or $AX > AY$. Either way, $AX \neq AY$, so no two different points will be the same distance s from A
- **Planes with the 11 axioms**: All theorems proved thus far hold for a plane that satisfies the 11 axioms. Major examples in which the 11 axioms are true as well as all theorems of propositions stated are $\mathbb{E}, \hat{\mathbb{E}}, \mathbb{M}, \mathbb{G}, \mathbb{S}, \mathbb{R}^3$, and $\mathbb{H}$

  With these axioms and theorems stated thus far, we are soon arriving at a point in which all lines on any plane that satisfies the 11 axioms will either behave like a Euclidean line if $\omega = \infty$, or a spherical line (great circle) if $\omega < \infty$
- **Proposition 8.7**: Let $\overline{AB}$ be a segment and $X, Y \in \overline{AB}$. Then, $XY \leqslant AB$, and if $XY = AB$, then $\{X, Y\} = \{A, B\}$

  Proof. If X or Y equals A, say $X = A$ and $Y \neq B$, then A-Y-B. So, $XY = AY < AB$ by proposition 6.3. Similarly, if X or Y equals B and the other point is not A, then $XY < AB$

  So, we may assume that A-X-B and A-Y-B.

  Since $\overline{AB} \subseteq \overrightarrow{AB}$, we apply theorem 8.3, which yields that either A-X-Y or A-Y-X

  Thus, $XY <$ (AY or AX) $< AB$ ■
- **Proposition 8.8** If $\overline{AB} = \overline{CD}$, then $\{A, B\} = \{C, D\}$

  Proof. If $\overline{AB} = \overline{CD}$, then $AB = CD$ by proposition 6.4.

  Thus, by proposition 8.7, $\{C, D\} = \{A, B\}$ ■
- **More on endpoints**: We can now say that a segment uniquely determines two endpoints
- **Definition (Interior points and length for a segment):** Given a segment $\overline{AB}$, A and B are called its endpoints. All other points of $\overline{AB}$ are called **Interior points** of $\overline{AB}$

Distance AB is called the **length** of $\overline{AB}$

- **Definition**: The interior of $\overline{AB}$, denoted Int$\overline{AB}$ or $\overline{AB}^0$, means the set of all interior points of $\overline{AB}$. That is, Int$\overline{AB} = \overline{AB}^0 = \{X : A\text{-}X\text{-}B\}$
- **Proposition 8.9**: In each segment $\overline{AB}$ there is a unique point M, called the **midpoint** of $\overline{AB}$, with the property that $AM = \frac{1}{2}AB$. Further, $AM = MB$

Proof. Let $s = \frac{1}{2}AB$. By UDR, there is a unique point M in $\overrightarrow{AB}$ with $AM = s = \frac{1}{2}AB$. Since $s < AB$, M is in $\overline{AB}$. Since $AM \neq 0$ and $AM \neq AB$, $M \neq A$ or B. Thus, $A\text{-}M\text{-}B$, and

$$AM + MB = AB \implies \frac{1}{2}AB + MB = AB$$
$$\therefore MB = \frac{1}{2}AB$$

■

- **Theorem 9.1 (Antipode on line theorem)**: Let A be a point on a line m (in a plane with the 11 axioms). Assume that $\omega < \infty$. Then, there exists a unique point A^*_m on m such that $AA^*_m = \omega$. Further, if X is any other point on m, then $A\text{-}X\text{-}A^*_m$

Proof. Let X be any point on m with $0 < AX < \omega$. (Such points exist by Ax. N). So ray $\overrightarrow{AX}$ is defined.

Thm 8.6 (uDR) for $A = \omega$ implies There is a unique point A^*_m on $\overrightarrow{AX}$ with $AA^*_m = \omega$. $AA^*_m > AX$ and the Important Fact $\Rightarrow A - X - A^*_m$.

Now, we show that there is no other point P anywhere on m with $AP = \omega$

Suppose toward a contradiction that P is another point on m with $AP = \omega$. Then, P is not in $\overrightarrow{AX}$.

Ax.QP applied to $A\text{-}X\text{-}A^*_m$ and point P implies one of

$$P\text{-}A\text{-}X, \quad A\text{-}P\text{-}X, \quad X\text{-}P\text{-}A^*_m, \quad X\text{-}A^*_m\text{-}P$$

If $P\text{-}A\text{-}X$, then $PX > PA = \omega$, a contradiction.

If $A\text{-}P\text{-}X$, then $AX > AP = \omega$, another contradiction

If $X\text{-}P\text{-}A^*_m$, then $A\text{-}X\text{-}A^*_m$ and ROI yields $A\text{-}X\text{-}P\text{-}A^*_m$ yields $A\text{-}X\text{-}P$, which implies $P \in \overrightarrow{AX}$, contradiction.

Thus, $X\text{-}A^*_m\text{-}P$. Consequently, $A^*_m P < XP \leqslant \omega$, so $\overline{A^*_m P}$ is defined

UDR says there's a point U with $A^*_m\text{-}U\text{-}P$. So, $UP < A^*_m P < \omega$, which implies $U \neq A$ as $AP = \omega$

Ax.QP applied to A^*_m-U-P and point A yields one of

$$A\text{-}A^*_m\text{-}U, \quad A^*_m\text{-}A\text{-}U, \quad U\text{-}A\text{-}P, \quad U\text{-}P\text{-}A$$

Which implies one of

$$AU > AA^*_m = \omega, \quad A^*_mU > A^*_mA = \omega, \quad UP > AP = \omega, \quad UA > PA = \omega$$

All contradictions. Thus, A^*_m is the only point on m with $AA^*_m = \omega$

So, if X is any point on m other than A or A^*_m, then $0 < AX < \omega$. Hence, ray $\overrightarrow{AX}$ is defined.

Now the first part of this proof applies to $\overrightarrow{AX}$, and so A-X-A^*_m

■

- **Definition**. Assume $\omega < \infty$. Let A be a point on a line m. The unique point A^*_m on m such that $AA^*_m = \omega$ is called the **antipode** of A on m. Thus,

$$\begin{cases} A, A^*_m \text{ are on m, } AA^*_m = \omega \\ \text{and } A\text{-}X\text{-}A^*_m \text{ for all other points } X \text{ on } m \end{cases}$$

- **Theorem 9.2 (Almost-uniqueness for Quadrichotomy)**: Suppose that A, B, C, X are distinct points on a line m, and that $A-B-C$. Then ***exactly one*** of the following holds:

$$X - A - B, \quad A - X - B, \quad B - X - C, \quad B - C - X$$

with the ***only exception*** that both $X - A - B$ and $B - C - X$ are true when $\omega < \infty$ and $X = B^*_m$.

(Note that $B^*_m - A - B$ and $B - C - B^*_m$ ***are both true*** by Thm. 9.1)

Proof. By Axiom QP, at least one of $X - A - B$, $A - X - B$, $B - X - C$, $B - C - X$ holds.

Suppose $A - X - B$**.** Then $A - B - C$ and the Rule of I. $\Rightarrow A - X - B - C$.

So $A-X-B$ and $X-B-C$ are true (by definition of $A-X-B-C$), hence $X-A-B$, $B - X - C$, $B - C - X$ are false by the UMT (Thm. 6.2). Thus, $A - X - B \Rightarrow$ none of the other three relations hold.

Suppose $B - X - C$**.** Then $A - B - C$ and the Rule of I. $\Rightarrow A - B - X - C$.

So $A - B - X$ and $B - X - C$ are true, hence the UMT $\Rightarrow X - A - B$, $A - X - B$, $B - C - X$ are false.

Thus, $B - X - C \Rightarrow$ none of the other three relations hold.

So if more than one of $X - A - B$, $A - X - B$, $B - X - C$, $B - C - X$ holds, they must be exactly $X - A - B$ and $B - C - X$.

Now assume that $X - A - B$ and $B - C - X$ are true.

Suppose (toward a contradiction) that $BX < \omega$.

Then ray $\overrightarrow{BX}$ is defined, and $X - A - B$, $B - C - X \Rightarrow A, C$ are in $\overrightarrow{BX}$.

So Thm. 8.3 $\Rightarrow$ one of $B - A - C$ or $B - C - A$ is true.

This contradicts $A - B - C$ and the UMT (Thm. 6.2).

Therefore, $BX = \omega$, hence $X = B^*_m$.

Corollary 8.5 showed that any ray has at most two endpoints. Prop. 9.3 will show that when $\omega < \infty$, any ray $\overrightarrow{AB}$ with carrier m ($m = \overleftrightarrow{AB}$) has a second endpoint, namely A^*_m. This generalizes what happens on $\mathbb{S}$, where $\overrightarrow{AB} = A^*B$.

- **Proposition 9.3 (needs proof)**: Assume $\omega < \infty$. Let A, B be points on line m with $0 < AB < \omega$. Then
  (a) $\overrightarrow{AB} = \overline{AB} \cup \overline{BA^*_m}$ and $\overline{AB}^\circ \cap \overline{BA^*_m}^\circ = \varnothing$.
  (b) $\overrightarrow{AB} = \overrightarrow{A^*_m B}$, so that if A is an endpoint of a ray with carrier m, then so is A^*_m.
- **Theorem 9.4**. If h is a ray with two endpoints A and P, then $\omega < \infty$ and $P = A^*_m$, where m is the carrier of h ($h \subseteq m$).

  Proof. Suppose (toward a contradiction) that $AP < \omega$.

  Since P is an endpoint of h, and A is on h with $0 < AP < \omega$, Thm. 8.4 $\Rightarrow h = \overrightarrow{PA}$.

  But A is also an endpoint of h, so Thm. 8.4 $\Rightarrow h = \overrightarrow{AP}$.

  Let a be any number with $AP < a \leqslant \omega$.

  Axiom RR or Thm. 8.6 $\Rightarrow$ there is a point X on $\overrightarrow{AP}$ with $AX > a > AP$. So the Important Fact $\Rightarrow A - P - X$.

  Since $h = \overrightarrow{PA}$ and X is on h, one of $X = A$, $X = P$, $P - X - A$, or $P - A - X$ is true, by definition of $\overrightarrow{PA}$. This contradicts $A - P - X$, by the UMT.

  Therefore, $AP = \omega$, which implies that $\omega < \infty$ and $P = A^*_m$.
- **Definition (interior points of a ray)**: Let $h = \overrightarrow{AB}$ be a ray. All points of h that are not endpoints of h are called *interior points* of h.

  The *interior* of h is the set of all interior points of h, and is denoted by h°, $\overline{AB}^\circ$, or Int $\overrightarrow{AB}$.
- **Definition (Opposite rays)**: Two rays with the same endpoint whose union is a line are called **opposite rays**
- **Theorem 9.6 (Opposite ray theorem)**: If B-A-C, then $\overrightarrow{AB}$ and $\overrightarrow{AC}$ are opposite rays

Also, for $m = \overleftrightarrow{AB}$

$$\overrightarrow{AB} \cap \overrightarrow{AC} = \begin{cases} \{A\} & \text{if } \omega = \infty \\ \{A, A^*_m\} & \text{if } \omega < \infty \end{cases}$$

Proof. B-A-C implies both $AB, AC < BC \leqslant \omega$. So, rays $\overrightarrow{AB}, \overrightarrow{AC}$ are defined, and $\overleftrightarrow{AB}$ is the unique line through A, B, and $\overleftrightarrow{AC}$ is the unique line through A, C

B-A-C implies B, A, C collinear which implies $\overleftrightarrow{AB} = \overleftrightarrow{AC}$. Call this line m.

We have $\overrightarrow{AB}, \overrightarrow{AC} \subseteq m$

If $X \neq A, B, C$ is on m, then Ax.QP say one of

$$X\text{-}B\text{-}A \quad B\text{-}X\text{-}A \quad A\text{-}X\text{-}C \quad A\text{-}C\text{-}X$$

Must be satisfied. In other words, X is in $\overrightarrow{AB}$ or $\overrightarrow{AC}$. So, $m \subseteq \overrightarrow{AB} \cup \overrightarrow{AC}$, hence $m = \overrightarrow{AB} \cup \overrightarrow{AC}$

Since $\overrightarrow{AB}$ and $\overrightarrow{AC}$ have the same endpoint, and $\overrightarrow{AB} \cup \overrightarrow{AC} = m$, $\overrightarrow{AB}$ and $\overrightarrow{AC}$ are opposite rays

What about $\overrightarrow{AB} \cap \overrightarrow{AC}$? B-A-C implies not A-B-C or A-C-B, so $B \notin \overrightarrow{AC}$, and $C \notin \overrightarrow{AB}$

So neither B nor C is in $\overrightarrow{AB} \cap \overrightarrow{AC}$

Let X be any point $\neq A, B, C$ in m. Suppose $X \in \overrightarrow{AB} \cap \overrightarrow{AC}$

$$X \in \overrightarrow{AB} \implies X\text{-}B\text{-}A \text{ or } B\text{-}X\text{-}A$$
$$X \in \overrightarrow{AC} \implies A\text{-}X\text{-}C \text{ or } A\text{-}C\text{-}X$$

So two of X-B-A, B-X-A, A-X-C, A-C-X are true

Theorem 9.2 applied to B-A-C and point X implies it must be X-B-A and A-C-X, with $\omega < \infty$ and $X = A^*_m$

Therefore, $\overrightarrow{AB} \cap \overrightarrow{AC} \subseteq \{A, A^*_m\}$

A is on both rays, by definition of a ray, and A-B-A^*_m, A-C-A^*_m with theorem 9.1 implies A^*_m is on $\overrightarrow{AB}$ and $\overrightarrow{AC}$, so $\overrightarrow{AB} \cap \overrightarrow{AC} = \{A, A^*_m\}$ when $\omega < \infty$, and $\{A\}$ when $\omega = \infty$

- **Corollary 9.7**: Each ray has a unique opposite ray.

Proof. Let $\overrightarrow{AB}$ be a ray. Proposition 8.11 says there's a point C on $\overrightarrow{AB}$ with C-A-B. Then, $AC < BC \leqslant \omega$, and the opposite ray theorem implies $\overrightarrow{AC}$ is an opposite ray to $\overrightarrow{AB}$

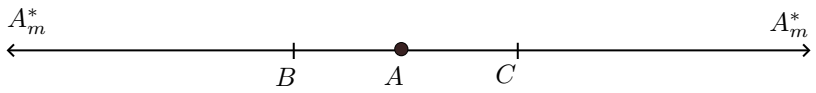
Let h be any ray opposite to $\overrightarrow{AB}$, so that h has endpoint A, and $h \cup \overrightarrow{AB} = \overleftrightarrow{AB}$

C is not in $\overrightarrow{AB}$ (from C-A-B or from $\overrightarrow{AB} \cap \overrightarrow{AC} = \{A, A_m^*\}$), hence $C \in h$

h has endpoint A, C in h with $0 < AC < \omega$ implies $h = \overrightarrow{AC}$ by theorem 8.4

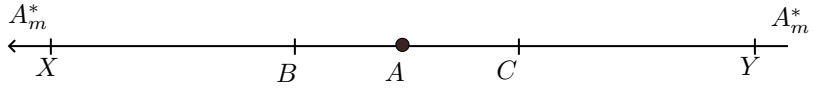
- **Notation:** Denote the ray opposite to ray h by h'. So, $\overrightarrow{AB}'$ means the ray opposite $\overrightarrow{AB}$
- **Corollary 9.8**: Let A, B be points on line m with $0 < AB < \omega < \infty$. Then $\overrightarrow{AB}' = \overrightarrow{AB_m^*}$

  Proof. Theorem 9.1 implies B-A-B_m^*. Then, theorem 9.6 implies $\overrightarrow{AB_m^*} = \overrightarrow{AB}'$
- **Corollary 9.9 (needs proof)**: Let A, B be points on line m with $0 < AB < \omega < \infty$. Then, $m = \overline{AB} \cup \overline{BA_m^*} \cup \overline{A_m^* B_m^*} \cup \overline{B_m^* A}$, with the interiors of these segments being disjoint.
- **Theorem 9.10 (Needs proof)**: Let A, B be points on line m with $0 < AB < \omega < \infty$. Let $C \neq A, B, A_m^*, B_m^*$ be another point on m. Then there is no betweenness relation for A, B, C if and only if $C \in \overline{A_m^* B_m^*}^0$
- **When $\omega < \infty$, any line m is "like a circle"**: Suppose B-A-C on m



$\overrightarrow{AB}, \overrightarrow{AC}$ are opposite rays by theorem 9.6. A_m^* is on both $\overrightarrow{AB} = \overrightarrow{A_m^* B}$ and $\overrightarrow{AC} = \overrightarrow{A_m^* C}$ by prop 9.3

So is it really like a circle? Take point X on $\overrightarrow{AB}$ near A_m^*, and point Y on $\overrightarrow{AC}$ near A_m^*



Say distances XA_m^*, YA_m^* are small enough so that $XA_m^* + YA_m^* \leqslant \omega$. To show that this line should be a circle, we need

$$X\text{-}A_m^*\text{-}Y$$

Because $XA_m^* + YA_m^* \leqslant \omega$, Ax.BP says there is a B.R among X, Y, A_m^*

Suppose A_m^*-X-Y, can we get a contradiction?

$A_m^*X < \omega$, X on $\overrightarrow{AB} = \overrightarrow{A_m^*B}$ implies $\overrightarrow{A_m^*B} = \overrightarrow{A_m^*X}$ (Thm 8.4)

So if A_m^*-X-Y, then Y is in $\overrightarrow{A_m^*X} = \overrightarrow{AB}$. But, $Y \in \overrightarrow{AC}$, and $\overrightarrow{AC}, \overrightarrow{AB}$ have only points A, A_m^* in common. A similar argument reveals why it also cannot be A_m^*-Y-X

Thus, it must be X-A_m^*-Y, and the line is "like a circle"

4.9 Separation

- **Proposition between** Let $\overrightarrow{AB}$ and $\overrightarrow{AC}$ be opposite rays, and points $X \in \text{Int}\overrightarrow{AB}$, $Y \in \text{Int}\overrightarrow{AC}$ with $AX + AY \leqslant \omega$, then X-A-Y

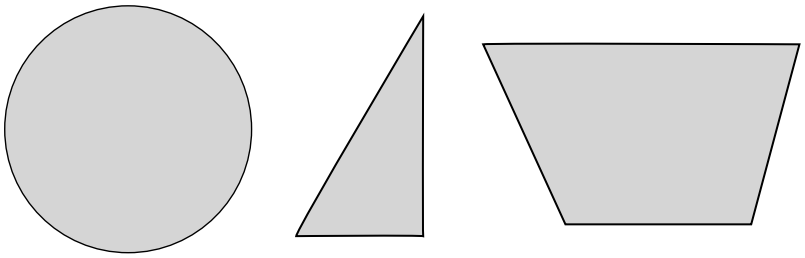
  Proof. X in $\text{Int}\overrightarrow{AB}$ implies $0 < AX < \omega$. So, by theorem 8.4, $\overrightarrow{AB} = \overrightarrow{AX}$. Similarly, $\overrightarrow{AC} = \overrightarrow{AY}$.

  Ax.BP and $AX + AY \leqslant \omega$ implies there is B.R among A, X, Y

  If A-X-Y then the definition of a ray implies $Y \in \overrightarrow{AX} = \overrightarrow{AB}$. But, $Y \in \text{Int}\overrightarrow{AC}$, and $\overrightarrow{AB} \cap \text{Int}\overrightarrow{AC} = \varnothing$, which by the opposite ray theorem (9.6) is a contradiction. So, not A-X-Y, and similarly not A-Y-X. Thus, we have X-A-Y

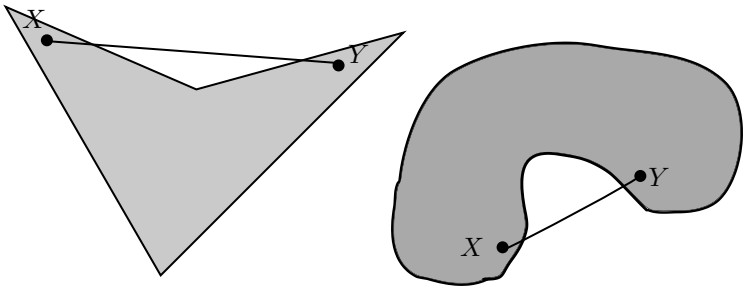
- **Definition**. A subset S of $\mathbb{P}$ is **convex** if for each pair of points $X \neq Y$ in S with $XY < \omega$, $\overline{XY} \subseteq S$ holds.

Examples of convex sets in $\mathbb{E}$



Each of these are convex sets in $\mathbb{E}$, with or without the boundary

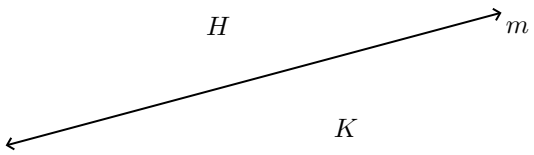
Examples of non-convex sets in $\mathbb{E}$



Each of these are non-convex in $\mathbb{E}$, with or without the boundary

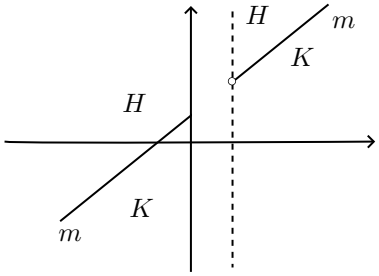
- **Theorem 10.1**: If S_1 and S_2 are convex sets in $\mathbb{P}$, then so is $S_1 \cap S_2$
- **Theorem 10.2**: Segments, rays, and lines are convex.
- **Definition**: A pair of sets H, K in $\mathbb{P}$ is called **opposed around a line** m if
  - $H, K \neq \varnothing$
  - H, K are convex
  - $H \cap K = \varnothing$
  - $H \cup K = \mathbb{P} - m$

In $\mathbb{E}, \hat{\mathbb{E}}, \mathbb{M}$

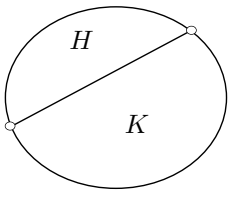


So we can imagine this line m that extends indefinitely, everything above the line is in H, below the line is in K, and the sets H, K are said to be **opposed around the line** m

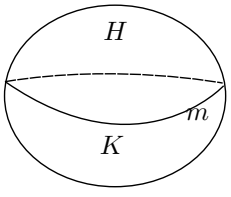
In $\mathbb{G}$



In $\mathbb{H}$



In $\mathbb{S}$



- **H and K in** $\mathbb{R}^3$: Imagine the wall in front of you as part of a plane in 3-dim space that extends to infinity in every direction.

Picture m as vertical line on the wall. Let H be all the points on your side of the wall (plane), together with all points on the wall (plane) to the right of m. In particular, you are in H

Let K be all the points on the other side of the wall (plane), together with all points on the wall (plane) to the left of m

- **Theorem 10.3** Let H, K be sets opposed around a line m in $\mathbb{P}$. Suppose that A, C are points so that $C \in m$, $A \in H$, $AC < \omega$. Then, Int$\overrightarrow{CA} \subseteq H$, and Int$\overrightarrow{CA'} \subseteq K$

Proof. Let $\ell = \overleftrightarrow{AC}$. $\ell \neq m$, since $A \notin m$. ℓ, m meet only at C (or also at $C_m^* = C_\ell^*$ if $\omega < \infty$ and if $C_m^* = C_\ell^*$). So, Int$\overrightarrow{CA} \subseteq H \cup K$

Let $CA = a < \omega$. Pick a number b with $0 < b < \omega - a$. Ax.RR implies there is a point B on $\overrightarrow{CA'}$ with $CB = b$, so $\overrightarrow{CA'} = \overrightarrow{CB}$.

For all points Y with C-Y-B or $Y = B$, $CY \leqslant b < \omega - a$

Let $y = CY$. Then, $y < \omega - a$, which implies $a + y < \omega$, which implies $CA + CY < \omega$

Proposition between implies A-C-Y, so $AY = AC + CY < \omega$. Thus, $\overline{AY}$ is defined, and $C \in \overline{AY}$

If $Y \in H$, then $A \in H$ and H is convex (by the definition of opposed sets). This implies $\overline{AY} \subseteq H$, which implies $C \in H$

But, $C \in m$, and $H \subseteq \mathbb{P} - m$, contradiction. Therefore, $Y \in K$ for all $Y \neq C$ in $\overline{CB}$

Now, let X be any point in Int$\overrightarrow{CA}$, let $x = CX$, so $0 < x < \omega$. Pick a number y with $0 < y < \begin{cases} b \\ w - x \end{cases}$

Ax.RR implies there is a point Y on $\overrightarrow{CA'} = \overrightarrow{CB}$ with $CY = y < b$. So, $Y \in \overline{CB}$, hence $Y \in K$

Also, $XC + CY = x + y < x + \omega - x = \omega$

Prop between implies X-C-Y, so $XY = XC + CY < \omega$. Thus, $\overline{XY}$ is defined, and $C \in \overline{XY}$.

If $X \in K$, then $Y \in K$ and K is convex (defn of opposed sets). This implies $\overline{XY} \subseteq K$, which implies $C \in K$. But, $C \in m$, and $K \subseteq \mathbb{P} - m$ (defn. opposed sets). A contradiction.

Therefore, $X \in H$ **for all** $X \in$ Int$\overrightarrow{CA}$, and Int$\overrightarrow{CA} \subseteq H$

A similar argument, starting with $B \in K$ reveals Int$\overrightarrow{CB} \subseteq K$; that is Int$\overrightarrow{CA'} \subseteq K$

- **Corollary 10.4**: let H, K be sets opposed around a line m, let A, B be points not on m, with A-X-B for some point $X \in m$. Then, A, B lie one in each of H and K, in some order.

Proof. A, B are in $\mathbb{P} - m = H \cup K$. We may assume $A \in H$. Thm 9.6 implies $\overrightarrow{XB} = \overrightarrow{XA'}$. So, Thm 10.3 implies $\text{Int}\overrightarrow{XA} \subseteq H$, $\text{Int}\overrightarrow{XB} \subseteq K$, hence $B \in K$

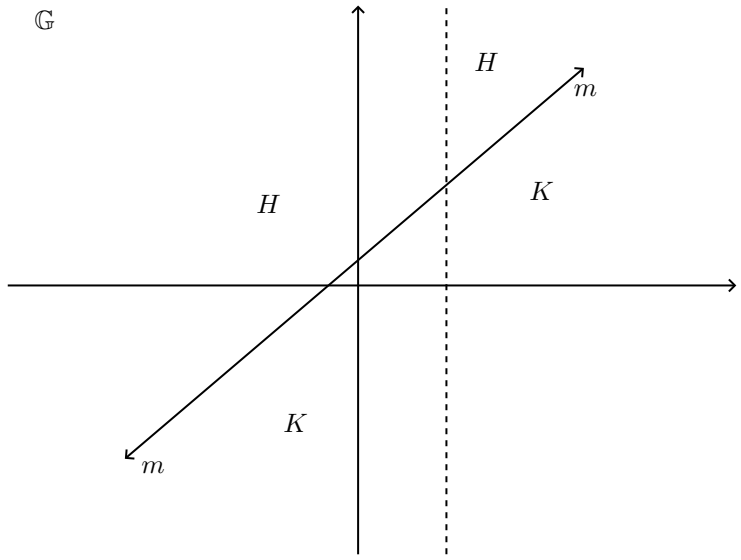
- **Definition**: Let m be a line. Sets H, K are called **opposite halfplanes with edge** m if:

  H, K are opposed around m, and whenever $X \in H, Y \in K$ and $XY < \omega$, then, $\overline{XY} \cap m \neq \varnothing$

  Planes that satisfy this are

  $$\mathbb{E}, \hat{\mathbb{E}}, \mathbb{M}, \mathbb{H}, \mathbb{S}$$

  This definition is not true for $\mathbb{G}$, or $\mathbb{R}^3$

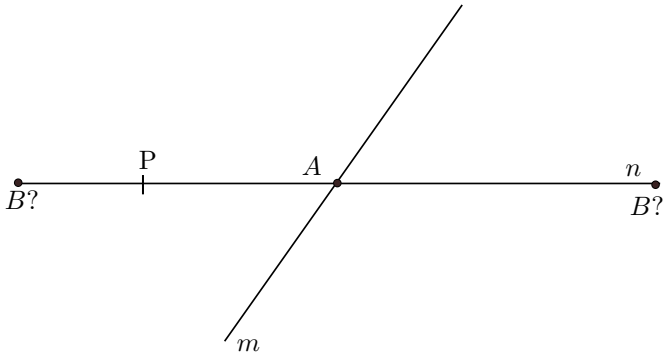


The existence of opposite halfplanes yields a number of important properties. When $\omega < \infty$, they guarantee the uniqueness of an antipode throughout the plane, not just line-by-line

- **Theorem 10.5**: Suppose that m is a line so that there exists a pair H, K of opposite half planes with edge m. Suppose also that $\omega < \infty$ and A is a point on m. If B is any point in $\mathbb{P}$ with $AB = \omega$, then $B \in m$ (so $B = A^*_m$, and there is only one point B in all of $\mathbb{P}$ with $AB = \omega$)

  This is what happens on $\mathbb{S}$, where we see that every line m does have a pair of opposite halfplanes with edge m

  Proof. Suppose toward a contradiction that $B \notin m$. Axioms I3 and N imply there is a line n through A and B, and a point P on n with $0 < AP < \omega$



Antipode on line thm (9.1) implies $B = A^*_m$, and A-P-B, so $BP < \omega$.

$B \notin m$ and Ax.I4 implies $m \cap n = \{A\}$

Prop 9.3 implies $\overrightarrow{AP} = \overrightarrow{A^*_m P} = \overrightarrow{BP}$. Note that $n = \overleftrightarrow{BP}$

Prop 8.11 implies there's a point Q with Q-B-P and $PQ < \omega$.

Thm 9.6 (opp. ray thm) implies $\overrightarrow{BQ} = \overrightarrow{BP'} = \overrightarrow{AP'}$, so $Q \in \text{Int}\overrightarrow{AP'}$. We may assume that $P \in H$, so Thm 10.3 implies $Q \in K$. Then, H, K are opposite halfplanes, and $PQ < \omega$, which implies $\overline{PQ} \cap m = \varnothing$

But, $P, Q \in n$ implies $\overline{PQ} \subseteq n$, which implies $\overline{PQ} \cap m \subseteq n \cap m = \{A\}$. So, $A \in \overline{PQ}$, and Q-B-P, which implies $B \in \overline{PQ}$

Prop 8.7 implies $AB \leqslant PQ$, but $AB = \omega$, and $PQ < \omega$, a contradiction. ■

- **Theorem 10.6**: Suppose that there is a pair H, K of opposite halfplanes with edge m. Let $A \neq B$ be points not on m. Then,

  A, B lie one in each of H, K $\iff$ there is a point X on m such that A-X-B

  Proof. If A-X-B for some point X on m, then coroll. 10.4 (which only required that H, K be opposed around m) implies A, B lie one in each of H, K.

  Suppose that A, B lie one in each of H, K. If $AB < \omega$, then $\overline{AB} \cap m \neq \varnothing$ by the definition of opposite halfplanes, so A-X-B for some X in m

  If $AB = \omega$, then Ax.I3 and the antipode on line thm (9.1) implies B is the antipode P on n, and A-P-B for all other points P on n. If some such point P is also on m, then we are done. Otherwise, since A, B lie one in each of H, K, and P is in H or K. We may assume that A, P are in H and B is not in K.

  Now, A-P-B implies $PQ < \omega$. So the defn of opposite halfplanes implies P-X-B for some point $X \in m$. A-P-B, P-X-B, and the rule of insertion implies

  $$A\text{-}P\text{-}X\text{-}B \implies A\text{-}X\text{-}B$$

- **Corollary 10.7 (Needs proof)**: Suppose that there is a pair H, K of opposite halfplanes with edge a line m. Then, H, K is the only pair of sets opposed around m.

Proof.

Note: It follows from coroll 10.7 that since lines m in $\mathbb{G}$ and in $\mathbb{R}^3$ have sets opposed around them that are not opposite halfplanes, then there are no other pairs of sets that are opposite halfplanes with edge m

- **A difference between $\mathbb{R}^3$ and $\mathbb{G}$**: No line m in $\mathbb{R}^3$ has a pair of opposite halfplanes with edge m. Some lines m in $\mathbb{G}$ have a pair of opposite halfplanes with edge m
- **Separation Axiom Ax.S**: for each line m, there exists a pair of opposite halfplanes with edge m.

  Ax.S is true for $\mathbb{E}, \hat{\mathbb{E}}, \mathbb{M}, \mathbb{H}, \mathbb{S}$. But, is false for $\mathbb{G}, \mathbb{R}^3$
- **Definition**: Let H, K be opposite halfplanes with edge m. Two points in the same halfplane are said to be on the **same side** of m.

  Two points in opposite halfplanes are said to be **opposite sides** of m
- **Theorem 10.8**: Suppose that $\omega < \infty$. For each point A, there is exactly one point A^* in $\mathbb{P}$ with $AA^* = \omega$. Also, every line through A goes through A^* as well.
- **Definition**: A^* is called the **antipode** of A

  Note that the antipode property of the sphere $\mathbb{S}$ is now true for every plane $\mathbb{P}$ with 12 axioms, when $\omega < \infty$

  Whats also true when $\omega < \infty$ is that for all points $A \in \mathbb{P}$ and all points $X \in \mathbb{P}$ with $X \neq A$ or A^*, then

  $$A\text{-}X\text{-}A^*$$

  This is because there is a line m through A and X (Ax.I3). In fact, it is the unique line $\overleftrightarrow{AX}$, since $X \neq A^*$ implies $AX < \omega$

  Thm 10.8 implies A^* is on m, and then the antipode on line theorem (9.1) implies $A\text{-}X\text{-}A^*$
- **Corollary 10.9**: Suppose that $\omega < \infty$. For any line m and point P, there are just two possibilities:

  $$\begin{cases} P, P^* & \text{both on } m \\ P, P^* & \text{on opposite sides of } m \end{cases}$$

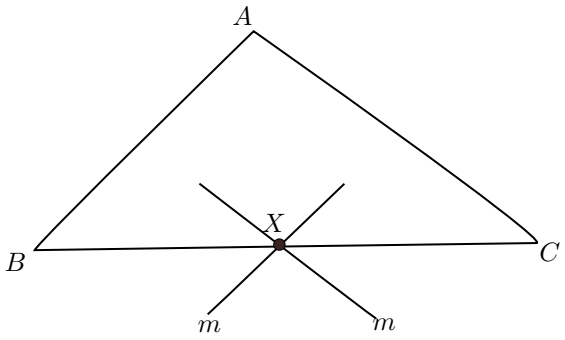
  Proof. If either P or P^* is on m, then so is the other, by Thm 10.8, so we need to only consider when neither P nor P^* is on m. Let X be any point on m. By what we showed above, $P\text{-}X\text{-}P^*$. Then, thm 10.6 implies P, P^* are in opposite halfplanes with edge m
- **Proposition Noncollinear**: If A, B, C are three noncollinear points (not all on the same line), then AB, AC, BC all less than ω.

  Proof. Suppose (toward a contradiction) that $AB = \omega$. Then, $B = A^*$, so $A\text{-}C\text{-}B$. In particular, A, C, B are collinear, a contradiction.

The following theorem was given as an axiom, in place of axioms, by Moritz pasch in 1882. It is equivalent to the separation Axiom: We prove here that it holds as a consequence of Ax.S (along with the other axioms); and also one can assume the statement of the theorem, and prove that the statement of Ax.S must be true.

- **Theorem 10.10 (Pasch's Axioms) (needs proof)**: Let A, B, C be three non-collinear points. Let X be a point with B-X-C, and m a line through X but not through A, B, or C. Then, exactly one of

  1. m contains a point Y with A-Y-C
  2. m contains a point Z with A-Z-B



Note that there are two things to prove.

  – a or b happens, and not both at once, for line m

Hint: Opposite halfplanes.

- **Theorem 10.11**: Assume that $\omega < \infty$. Then, any two distinct lines must have a point (in fact, a pair of antipodes) in common.

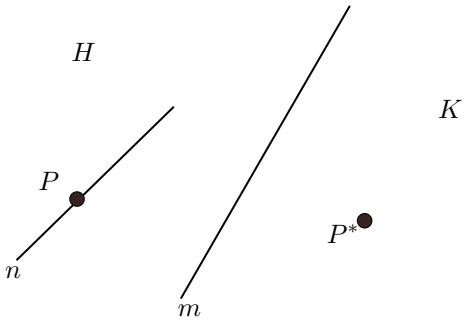
Proof. Suppose toward a contradiction that there are lines m, n with $m \cap n = \varnothing$. Let H, K be the opposite halfplanes with edge m, and let P be a point on n. Since $P \notin m$, we may assume that $P \in H$

Coroll. 10.9 implies $P^* \in K$. $P^* \in n$ by thm 10.8. Let Q be any other point on n. Then, QP, $QP^* < \omega$. So, $\overline{QP}, \overline{QP^*}$ are defined, and are contained in n.

If $Q \in H$, then Ax.S implies $\overline{QP^*} \cap m \neq \varnothing$, which implies $n \cap m \neq \varnothing$

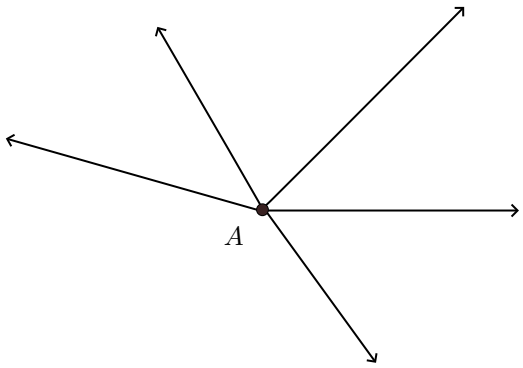
If $Q \in K$, then Ax.S implies $\overline{QP} \cap m \neq \varnothing$, which implies $n \cap m \neq \varnothing$,

Both of these a contradiction.

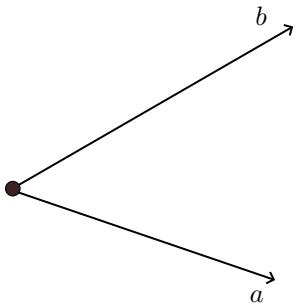


4.10 Pencils and Angles

- **Definition:** ***Coterminal rays***: Rays with the same endpoint



- **Definition:** ***Angle***: $\underline{ab} = a \cup b$, where a, b are coterminal rays

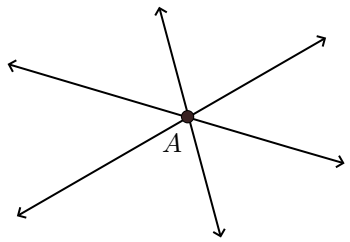


- **duality**: To develop properties of coterminal rays, our steps will be closely analogous to those we took to study collinear points. The analogy is called **duality**

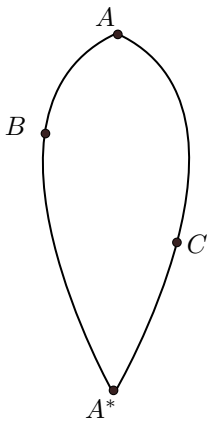
The theory of coterminal rays is dual to the theory of collinear points when $\omega < \infty$

But, the theory of rays will be good whether $\omega < \infty$ or $\omega = \infty$

- **Definition:** ***Pencil of rays at point*** A: The set of all rays with endpoint A: denote by P_A or just P

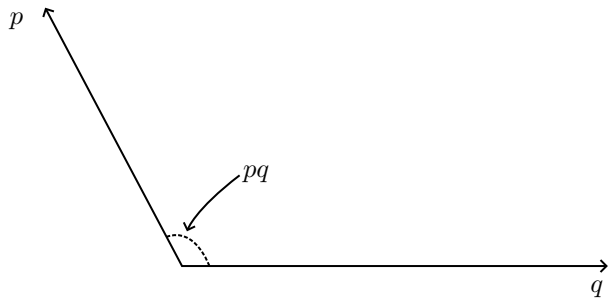


When $\omega < \infty$, each ray $h = \overrightarrow{AB} = \overrightarrow{A^*B}$, so $P_A = P_{A^*}$. h' is the opposite ray to h, as before

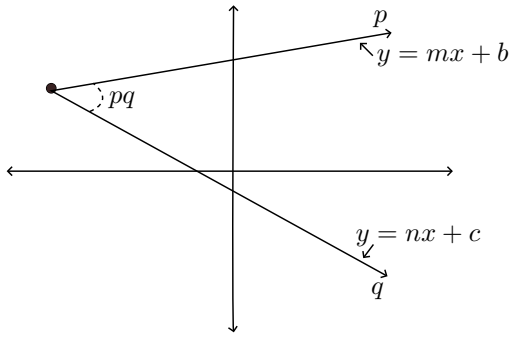


- **Undefined Term** ***Angle distance function, or angle measure***: A function μ from all pairs (p, q) of coterminal rays to $\mathbb{R}$

We abbreviate the angular distance between rays p, q, or the angle measure of the angle pq, $\mu(p, q)$ as pq



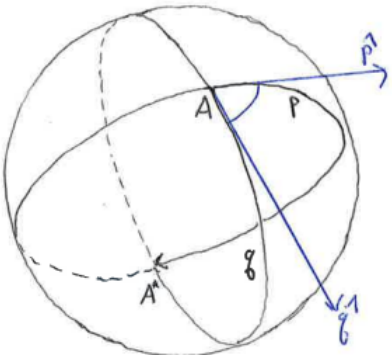
- **Angular distance in each of $\mathbb{E}$, $\hat{\mathbb{E}}$, $\mathbb{M}$, $\mathbb{S}$, $\mathbb{H}$**
  - $\mathbb{E}$, $\hat{\mathbb{E}}$, $\mathbb{M}$: The usual measure in degrees (0 to 180)



$$pq = \cos^{-1}\left(\frac{1+mn}{\sqrt{1+m^2}\sqrt{1+n^2}}\right)$$

From the law of cosines

  - $\mathbb{S}$:



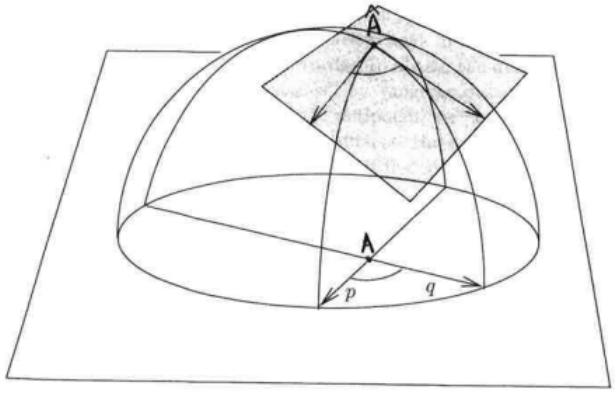
p̂ tangent to p

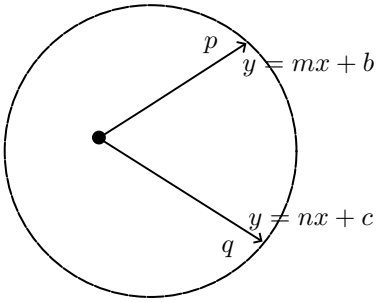
q̂ tangent to q

Rays $\hat{p}, \hat{q}$ in plane tangent to $\mathbb{S}$ at A. pq is defined as $\hat{p}\hat{q}$ in the tangent plane

- **Definition *(Ordinary angle)***: An ordinary angle is any angle that is less than 180 degrees.
- $\mathbb{H}$: Consider $\mathbb{H}$ in a horizontal plane, cover $\mathbb{H}$ with a hemispherical dome. Rays p, q in $\mathbb{H}$, coterminal at point A. Project vertically up to dome: p, q project to circular arcs on the dome, A to $\hat{A}$

  $\mu_{\mathbb{H}}(p, q)$ is the measure of ordinary angle between the two lines in the tangent plane at $\hat{A}$





$$\mu_{\mathbb{H}}(p, q) = \cos^{-1}\left(\frac{1 + mn - bc}{\sqrt{1 + m^2 - b^2}\sqrt{1 + n^2 - c^2}}\right)$$

- **Measure axioms**

  M1 : For all coterminal rays p, q, $0 \leqslant pq \leqslant 180$

  M2 : $pq = 0 \iff p = q$

  M3 : $pq = qp$

  M4 : $pq = 180 \iff q = p'$

Note that $M4$ implies that 180 is the suppremum (and the max) for angular distance. Observe that this fact is analogous to distances (when $\omega < \infty$), and $PQ = \omega \iff Q = P^*$

Thus,

$$180 \leftrightarrow \omega$$
$$p' \leftrightarrow P^*$$

- **Definition *(betweenness for rays)***: Ray b lies **between** rays a and c (a-b-c) provided that
  (a) a, b, c are different, coterminal
  (b) $ab + bc = ac$
- **Theorem 11.1 *(symmetry of betweenness)***: a-b-$c \iff c$-b-a
- **Theorem 11.3 *UMT***: If a-b-c, then b-a-c and a-c-b are false.

Proof. The definition of a-b-c implies a-b-c are different, coterminal, and $ab+bc = ac$.

Observe that each of $ab, bc, ac > 0$ by axioms M1,M2. Thus, $ab + bc = ac$ implies ac is larger than each of ab, bc

Suppose toward a contradiction that b-a-c is also true. Then, bc larger than each of $ba = ab$, and ac. This contradicts ac larger than ab, bc

Therefore, b-a-c is false. Similarly, a-c-b is also false ■

- **Note about Incidence axioms**: There are no duals for the four incidence axioms we studied previously for points

We needed those axioms to begin developing how the undefined terms point and line are related. For instance, to guarantee points A, B are in a **unique** line together, we require $AB < \omega$ (Ax.I4). Point A can be in many different lines

But, rays and pencils are defined concepts, and by the definitions, and by what we have proved about rays, ray $a = \overrightarrow{AB}$ is in **just one pencil** P_A

Coterminal rays $a = \overrightarrow{AB}, b = \overrightarrow{AC}$ are automatically in the unique pencil P_A, no proof needed.

The dual of axiom N is a true statement for rays, but we don't need to assume it; we can prove it.

- **Theorem 11.2 (non-triviality)**: For any ray p there is a coterminal ray q so that $0 < pq < 180$

Proof. Let $p = \overrightarrow{AB}$ be on line ℓ. Thm 7.6 implies there is a point C not on ℓ with $0 \leq AC < \omega$. Then, ray $\overrightarrow{AC}$ is defined, and is not contained in ℓ, since $C \notin \ell$. So, $\overrightarrow{AC} \neq p$, and $\neq p'$.

Let $q = \overrightarrow{AC}$. Then,

$$\begin{aligned} 180 \geqslant & pq \geqslant 0 \quad \text{(Ax.M1)} \\ & pq \neq 0 \quad \text{(Ax.M2, } q \neq p\text{)} \\ & pq \neq 180 \quad \text{(Ax.m4, } q \neq p'\text{)} \end{aligned}$$

Therefore, $0 < pq < 180$

- **Definition** ***(Wedge, fan)***: Let p, q be coterminal rays with $0 < pq < 180$.
  - **Wedge** $\overline{pq} = \{p, q\} \cup \{r : p\text{-}r\text{-}q\}$
  - **Fan** $\overrightarrow{pq} = \{p, q\} \cup \{r : p\text{-}r\text{-}q\} \cup \{r : p\text{-}q\text{-}r\}$

  (The duals of segment and ray)

- **Betweenness of rays axiom (Ax.BR)**: If a, b, c are distinct, coterminal rays, and if $ab + bc \leqslant 180$, then there exists a betweenness relation among a, b, c

  Thus, if no betweenness relation exists, then

$$\begin{aligned} ab + bc &> 180 \\ ac + cb &> 180 \\ ba + ac &> 180 \end{aligned}$$

- **Definition** ***(quad betweenness)***: $a\text{-}b\text{-}c\text{-}d$ means that all four of

$$a\text{-}b\text{-}c \quad a\text{-}b\text{-}d \quad a\text{-}c\text{-}d \quad b\text{-}c\text{-}d$$

  are true

- **Theorem** ***(Triangle inequality for rays)***: If a, b, c are three distinct, coterminal rays, then $ab + bc \geqslant ac$

- **Theorem 11.5** ***(Rule of insertion for rays)***:

  (a) If $a\text{-}b\text{-}c$ and $a\text{-}r\text{-}b$, then $a\text{-}r\text{-}b\text{-}c$

  (b) If $a\text{-}b\text{-}c$ and $b\text{-}r\text{-}c$, then $a\text{-}b\text{-}r\text{-}c$

- **Quadrichotomy of Rays Axiom (Ax.QR)**: If a, b, c, x are distinct, coterminal rays, and if $a\text{-}b\text{-}c$, then at least one of the following must hold

$$x\text{-}a\text{-}b \quad a\text{-}x\text{-}b \quad b\text{-}x\text{-}c \quad b\text{-}c\text{-}x$$

  So, Ax.QR says that whenever $a\text{-}b\text{-}c$ (say in pencil P), then any other ray in P is in either fan $\overrightarrow{ba}$ or fan $\overrightarrow{bc}$ (so $P = \overrightarrow{ba} \cup \overrightarrow{bc}$)

- **Real fan axiom (Ax.RF)**: For any fan $\overrightarrow{ab}$ and for any real number t with $0 \leqslant t \leqslant 180$, there is a ray r in $\overrightarrow{ab}$ with $ar = t$

  Ax.RF says every real number from 0 to 180 produces at least one ray in the fan

  Note: Ax.RF is one version of what is sometimes called the **Protractor Axiom**

- **Notes**: Axioms $M1 - M4$, BR, QR, RF are true for $\mathbb{E}, \hat{\mathbb{E}}, \mathbb{M}, \mathbb{H}, \mathbb{S}$

All the results of chapters 8,9 have dual results for rays in a pencil. We do not state them all. But some that we do not state are needed for the proofs of some that we do.

- **Theorem 11.6 (Unique angular distance for fans)**: For any fan $\overrightarrow{pq}$ and any real number t with $0 \leqslant t \leqslant 180$, there is a unique ray r in $\overrightarrow{pq}$ with $pr = t$. r is in $\overline{pq}$ if and only if $t \leqslant pq$
- **Theorem 11.8**: If ray a lies in pencil P, then a-r-a' for every other ray r in P

  Note: We assumed in Ax.M4 that a' is the unique ray in P with angular distance 1800 from a, so most of the proof of thm 9.1 does not need to be dualized. Alternatively, we could omit the assumption that $pq = 180 \implies q = p'$ and prove that this must be so by writing the dual of the full proof of thm 9.1.
- **Theorem 11.9 (Almost uniqueness of quadrichotomy for rays)**: Suppose that a, b, c, r are distinct rays in a pencil P, and that a-b-c. Then, **exactly** one of

  $$r\text{-}a\text{-}b \quad a\text{-}r\text{-}b \quad b\text{-}r\text{-}c \quad b\text{-}c\text{-}r$$

  With the exception that both r-a-b and b-c-r are true when $r = b'$

  Proof: We proceed by dualizing the proof of theorem 9.2.

  By Axiom.QR, at least one of

  $$r\text{-}a\text{-}b \quad a\text{-}r\text{-}b \quad b\text{-}r\text{-}c \quad b\text{-}c\text{-}r.$$

  Suppose we have a-r-b. Then, a-b-c and the rule of insertion yields a-r-b-c

  So, a-r-b and r-b-c are true. Which, by the UMT guarantees that both b-r-c and b-c-r are false.

  Next, suppose that b-r-c is true. Then, a-b-c and the rule of insertion yields a-b-r-c. So, a-b-r and b-r-c are true, and by the UMT, all three of r-a-b, a-r-b, b-c-r are false. Thus, none of the other three relations hold.

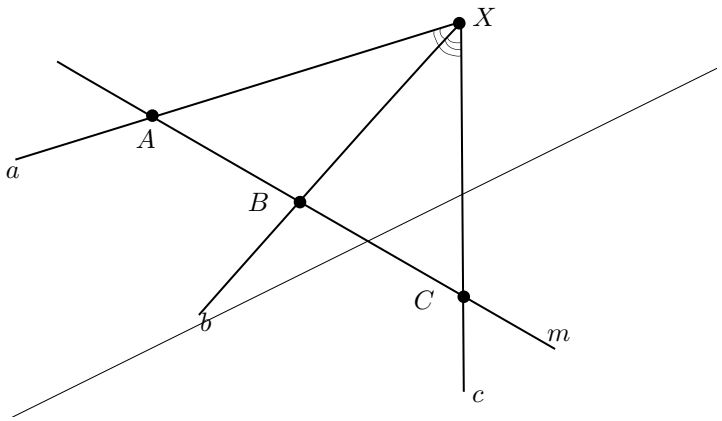
  So, if more than one of r-a-b, a-r-b, b-r-c, b-c-r holds, they must be exactly r-a-b and b-c-r

  Assume that r-a-b and b-c-r are true. Suppose toward a contradiction that $br < 180$. Then, fan $\overrightarrow{br}$ is defined, and r-a-b, b-c-r implies a, c are in $\overrightarrow{br}$. By the dual of theorem 8.3 (stated above), one of

  $$b\text{-}a\text{-}c \quad \text{or} \quad b\text{-}c\text{-}a$$

  is true. But, this contradicts a-b-c by the UMT.

  Therefore, $br = 180$, hence $r = b'$. ■
- **Theorem 11.10 (Opposite fan theorem)**: Let p, q, r be rays in pencil P such that q-p-r. Then, $\overrightarrow{pq} \cup \overrightarrow{pr} = P$, and $\overrightarrow{pq} \cap \overrightarrow{pr} = \{p, p'\}$
- **Corollary 11.11**: If p, q are rays in pencil P with $0 < pq < 180$, then $P = \overrightarrow{pq} \cup \overrightarrow{pq'}$ and $\overrightarrow{pq} \cap \overrightarrow{pq'} = \{p, p'\}$
- **Compatibility Axiom (Ax.C)**: Let A, B, C be points on line m, and X a point not on m. If A-B-C, then $\overrightarrow{XA}$-$\overrightarrow{XB}$-$\overrightarrow{XC}$



Notice $AB + BC = AC \implies ab + bc = ac$

- **Notation and terminology**: Recall that $\not{pq}$ means $p \cup q$, then union of the rays. Measure of $\not{pq}$ means the angular distance pq

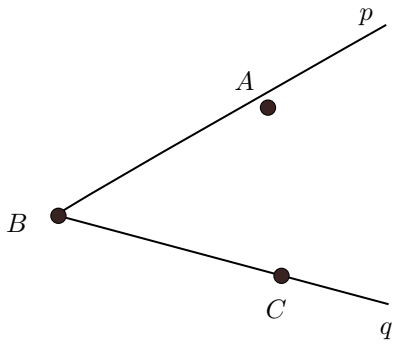
  Suppose $p = \overrightarrow{BA}$, $q = \overrightarrow{BC}$. Then, write

  $$\not{pq} = \underline{\angle ABC} = \underline{\angle CBA}$$

  Or just $\underline{\angle B}$ when clear, and

  $$pq = \angle ABC = \angle CBA$$

  or just $\angle B$.



- **Definition**:

- **Zero angle:** $\underline{pq}$ is a **zero angle** if $pq = 0$ ($\iff p = q$)
- **Straight angle**: If $pq = 180$($\iff p = q'$)
- **Proper angle:** if $0 < pq < 180$
- **acute angle**: if $0 < pq < 90$
- **right angle**: if $pq = 90$
- **obtuse angle**: if $90 < pq < 180$

- **Proposition 11.14**

  (a) If $\omega < \infty$, then $\angle ABC = \angle AB^*C$

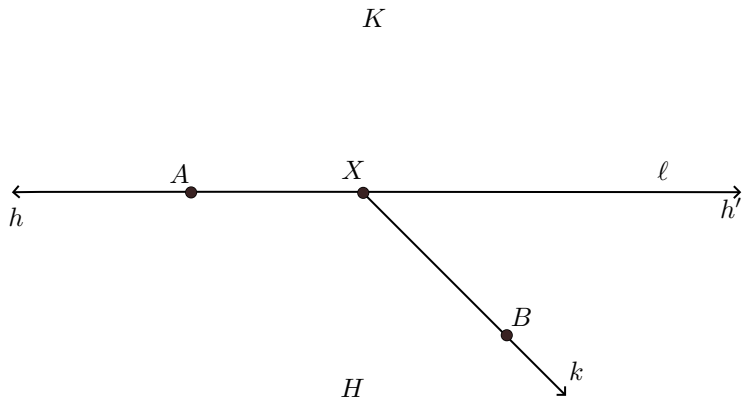
  (b) If $P \in \overrightarrow{BA}^0$ and $Q \in \overrightarrow{BC}^0$, then $\angle ABC = \angle PBQ$

Proof (a) If $\omega < \infty$, then thm 10.8 and prop 9.3 implies $\overrightarrow{BA} = \overrightarrow{B^*A}$, $\overrightarrow{BC} = \overrightarrow{B^*C}$. So, $\underline{\angle ABC} = \overrightarrow{BA} \cup \overrightarrow{BC} = \overrightarrow{B^*A} \cup \overrightarrow{B^*C} = \angle AB^*C$

(b) If $P \in \overrightarrow{BA}^0$ and $Q \in \overrightarrow{BC}^0$, then $0 < BP < \omega$ and $0 < BQ < \omega$ by thm 9.4. Then, $\overrightarrow{BA} = \overrightarrow{BP}$ and $\overrightarrow{BC} = \overrightarrow{BQ}$ by thm 8.4, so $\angle ABC = \overrightarrow{BA} \cup \overrightarrow{BC} = \overrightarrow{BP} \cup \overrightarrow{BQ} = \angle PBQ$

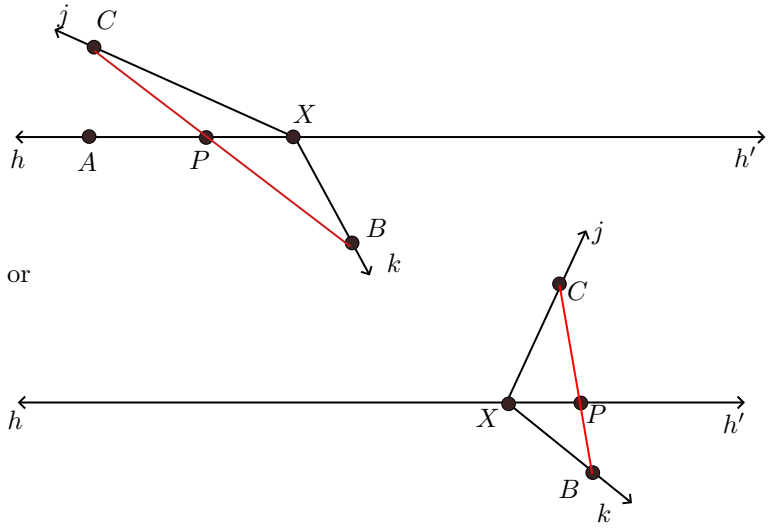
- **Proposition 11.15 (Midpoint)**: If $\underline{pq}$ is a proper angle, then there is exactly one ray b in the wedge $\overline{pq}$ so that $pb = \frac{1}{2}pq$
- **Definition**: The ray b is called the **bisector** of angle $\underline{pq}$
- **Theorem 12.2 (Fan: halfplane)**: Let H, K be opposite halfplanes with edge line ℓ, point $B \in H$. Let X, A be points on ℓ with $0 < AX < \omega$. Let $h = \overrightarrow{XA}$, $k = \overrightarrow{XB}$. Then, H consists of all points on all rays of the fan $\overrightarrow{hk}$, except for the points of ℓ

  That is, $P \in H \iff P \in j^0$, where j^0 is the interior of some ray $j \in \overrightarrow{hk}$, $j \neq h$ or h'



Proof. Since $B \in H$ and $k = \overrightarrow{XB}$, by theorem 10.3 $k^0 \subseteq H$. Suppose that j is a ray in $\overrightarrow{hk}$ with $j \neq k, h, h'$. So, either h-j-k or h-k-j by the definition of Fan.

Suppose toward a contradiction that for some point $C \in j^0$, then $C \in K$. Theorem 8.4 implies $j = \overrightarrow{XC}$, and theorem 10.3 implies $j^0 \in K$



$B \in H$, $C \in K$, so by Theorem 10.6, B-P-C for some $p \in \ell$. If $P = X$, then B-X-C, which implies $\overrightarrow{XC} = \overrightarrow{XB'}$ by the opposite ray theorem (9.6)

If $P = X^*$ then B-X^*-C, which implies $\overrightarrow{X^*C} = \overrightarrow{X^*B'}$ by Theorem 9.6

Then, B, X, C, X^* collinear (theorem 10.8), so proposition 9.3 implies $\overrightarrow{X^*C} = \overrightarrow{XC} = j$, $\overrightarrow{X^*B} = \overrightarrow{XB} = k$

So, if $P = X$ or X^*, then $j = k'$. Then, theorem 11.8 implies k-h-j. But, this contradicts h-j-k or h-k-j by Theorem 11.3 (UMT for rays). Therefore, $P \neq X$ or X^*, so X is not collinear with B, P, C. Also, $P \in \ell$ so either $P \in h^0$ or $P \in \text{Int}h'$. By theorem 8.4, $\overrightarrow{XP} = h$ or h'

By Ax.C and B-P-C, $\overrightarrow{XB}$-$\overrightarrow{XP}$-$\overrightarrow{XC}$, which implies either k-h-j or k-h'-j

Again, $j \in \overrightarrow{hk}$, which implies h-j-k or h-k-j, so by UMT, k-h-j is false. Thus, k-h'-j = j-h'-k

If h-j-k then ROI yields h-j-h'-k, which gives h-h'-j, contradicting h-j-h'.

So, for all rays $j \in \overrightarrow{hk}$ with $j \neq h$ or h', then $j^0 \subseteq H$.

Now, Int$k' \subseteq K$ by thm 10.3, and the proof thus far shows: for all $j \in \overrightarrow{hk'}$ with $j \neq h$ or h', then $j^0 \subseteq K$

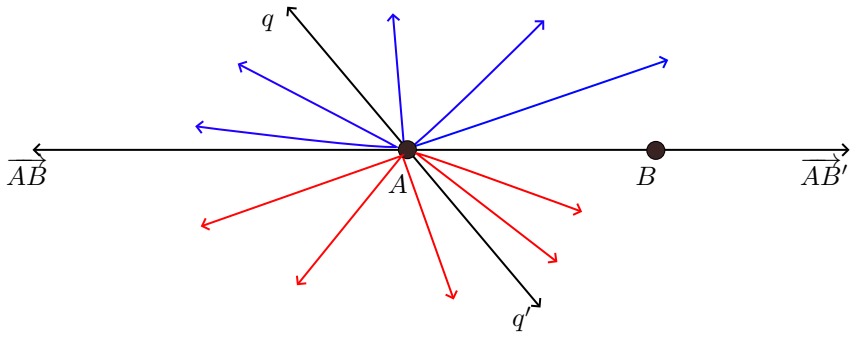
For any point $D \in H$, $\overrightarrow{XD} \in P_X$, but $\overrightarrow{XD} \notin \overrightarrow{hk'}$ since $D \notin K$.

By coroll. 11.11, $\overrightarrow{XD} \in \overrightarrow{hk}$. So, points of H equals points on all $j^0, j \in \overrightarrow{hk}, j \neq h, h'$ ■

- **Corollary 12.3**: Let z by any number with $0 < z < 180$. For any ray $\overrightarrow{AB}$ there are exactly two rays h, k in P_A such that $\overrightarrow{AB}h = z = \overrightarrow{AB}k$. Furthermore, h^0 and k^0 lie in opposite halfplanes with edge $\overleftrightarrow{AB}$

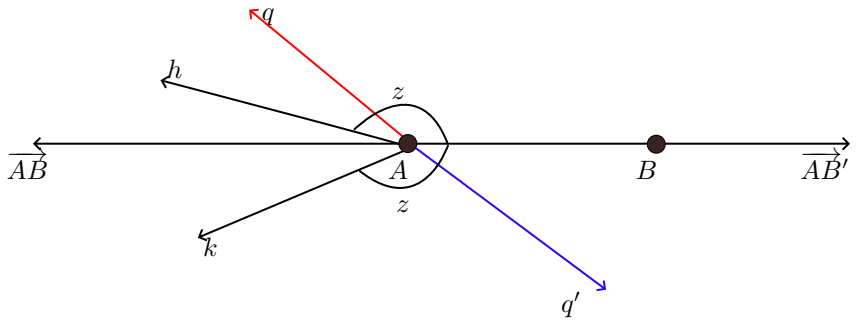
Proof. There is a ray q with $0 < \overrightarrow{AB}q < 180$ (Thm. 11.2). Thus, fan $\overrightarrow{\overrightarrow{AB}q}$ is defined.

Coroll 11.11 implies $P_A = \overrightarrow{\overrightarrow{AB}q} \cup \overrightarrow{\overrightarrow{AB}q'}$, the union of two fans that have only $\overrightarrow{AB}, \overrightarrow{AB'}$ in common.



Thm 11.6 implies there's a unique ray h in $\overrightarrow{\overrightarrow{AB}q}$ with $\overrightarrow{AB}h = z$, and a unique ray k in $\overrightarrow{\overrightarrow{AB}q'}$ with $\overrightarrow{AB}k = z$. So, there are only two such rays in $P_A = \overrightarrow{\overrightarrow{AB}q} \cup \overrightarrow{\overrightarrow{AB}q'}$

Thm 12.2 implies h^0, k^0 lie in opposite halfplanes with edge $\overleftrightarrow{AB}$



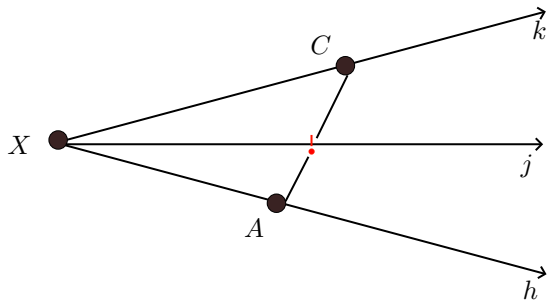
4.11 The Crossbar Theorem

- **Intro:** This gives us a sufficient condition (a guarantee) that a ray will meet a line (in fact, a line segment)

  Axiom C does not do this. It deals with rays that are already assumed to meet a line.

  The context for the crossbar theorem is any general plane with the 20 axioms.

- **Theorem 12.4 (The Crossbar Theorem)**: If $\underline{hk}$ is a proper angle with vertex (common endpoint) X, if $A \in h^0$ (so $h = \overrightarrow{XA}$), $C \in k^0$ (so $k = \overrightarrow{XC}$), and h-j-k, then there is an interior point B of j with A-B-C

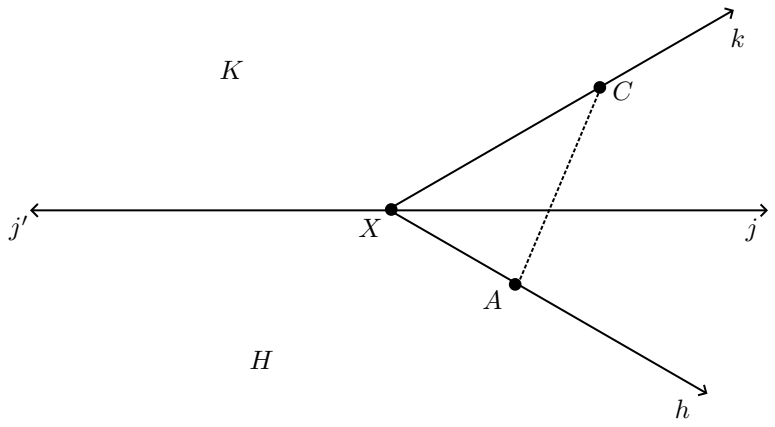


Note: $\underline{hk}$ proper means that $h \neq k$ or k', $k \neq h$ or h'. Hence, $C \notin h \cup h' = \overleftrightarrow{XA}$, $A \notin k \cup k' = \overleftrightarrow{XC}$. Then, X, A, C noncollinear, so $AC < \omega$.

Thus, segment $\overline{AC}$ is defined, and the crossbar theorem says that j^0 meets $\overline{AC}^0$. Note that

$$j^0 \cap \overline{AC}^0 \neq \varnothing.$$

Proof. Assume $h = \overrightarrow{XA}, k = \overrightarrow{XC}$, with $\underline{hk}$ a proper angle. Assume h-j-k



h-j-k implies h, j, k distinct, which implies $0 < hj < hk < jk$

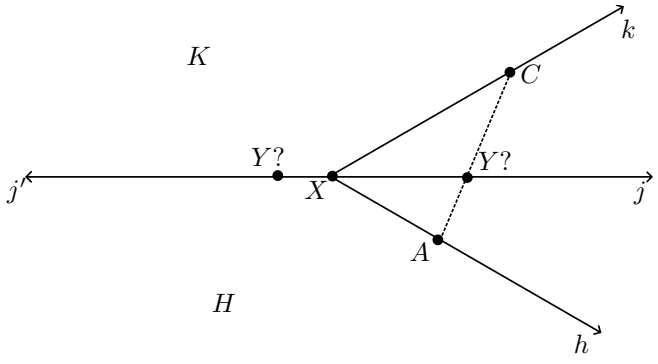
h-j-k implies $hj + jk = hk < 180$, so $0 < hj < 180, 0 < jk < 180$, which implies fans $\overrightarrow{jh}, \overrightarrow{jk}$ defined. h-j-k also implies $P_X = \overrightarrow{jh} \cup \overrightarrow{jk}$, with $\overrightarrow{jh} \cap \overrightarrow{jk} = \{j, j'\}$ (thm 11.10)

Let $m = \text{line } j \cup j'$; we have H, K opposite halfplanes with edge m (Ax.S), we may assume that $A \in H$

Theorem 12.2 implies H is the set of all points of all r^0, for $r \in \overrightarrow{jh}$, $r \neq j$ or j'

$\overrightarrow{jk}$ is the "opposite fan" to $\overrightarrow{jh}$, so theorem 12.2 implies $k^0 \subseteq K$, so $C \in K$

Now, theorem 10.6 implies there is a point Y on m with A-Y-C



If $Y = X$ or X^*, then A-X-C or A-X^*-C, which implies A, X, C collinear. By theorem 10.8, if A, X^*, C collinear, then A, X^*, C, X collinear, which contradicts $\underline{\angle AXC} = \underline{hk}$ proper.

So, Y on m with $Y \neq X, X^*$, which implies $Y \in j^0$ or Intj', which implies $\overrightarrow{XY} = j$ or j'

A, Y, C on line $\overleftrightarrow{AY}$, and X not on $\overleftrightarrow{AY}$, so by Ax.C and A-Y-C, $\overrightarrow{XA}$-$\overrightarrow{XY}$-$\overrightarrow{XC}$, which implies h-$\overrightarrow{XY}$-k

Suppose $\overrightarrow{XY} = j'$, then h-j'-k. By hypothesis, h-j-k. So, k in fans $\overrightarrow{hj'}$ and $\overrightarrow{hj}$. But, coroll 11.11 implies $\overrightarrow{hj'} \cap \overrightarrow{hj} = \{h, h'\}$, and $k \neq h$ or h' ($\underline{hk}$ proper). This is a contradiction, which implies $\overrightarrow{XY} = j$, hence $Y \in j^0$, and A-Y-C ■

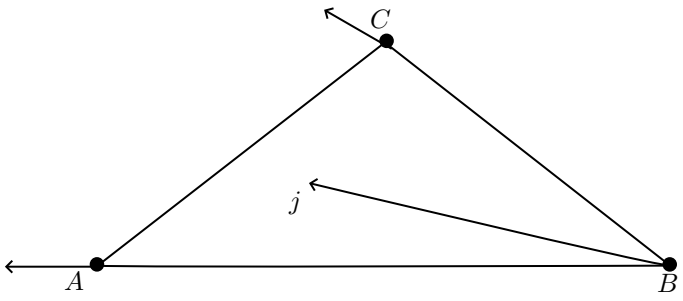
- **Definition**: Let A, B, C be three noncollinear points (So AB, BC, AC all $< \omega$ by prop noncollinear).

  - The **triangle** $\triangle ABC$ is $\overline{AB} \cup \overline{BC} \cup \overline{CA}$
  - The **sides** of $\triangle ABC$ are $\overline{AB}, \overline{BC}, \overline{CA}$
  - The **vertices** of $\triangle ABC$ are A, B, C
  - The **angles** of $\triangle ABC$ are

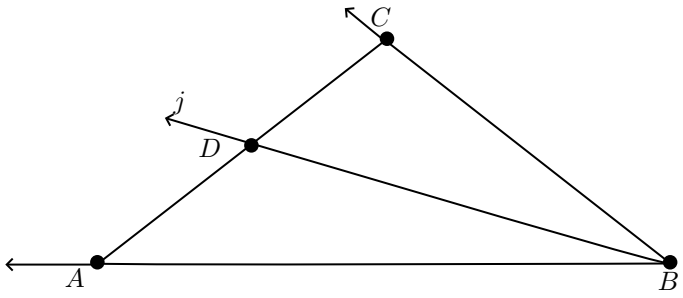
$$\underline{\angle CAB} = \underline{\overrightarrow{AC}\overrightarrow{AB}} = \angle A = \angle A$$
$$\underline{\angle ABC} = \underline{\overrightarrow{BA}\overrightarrow{BC}} = \angle B$$
$$\underline{\angle BCA} = \underline{\overrightarrow{CB}\overrightarrow{CA}} = \angle C.$$

  - $\angle CAB$ are vertex A are called **opposite** $\overline{BC}$, etc...
  - The **angle sum** $\sigma(ABC) = \angle A + \angle B + \angle C$ (This is not necessarily 180)

- **Note:** Say we have $\triangle ABC$ and ray j with $\overrightarrow{BA}$-j-$\overrightarrow{BC}$ $\angle ABC$ is proper, since A, B, C



are noncollinear. So, the crossbar theorem implies j^0 meets $\overline{AC}^0$. Thus, there is a point D on j^0 with A-D-C



Note: Euclid does this a lot, with no explicit justification.

- **Note about the crossbar theorem**: The proof of the crossbar theorem depends on the fan: halfplane theorem (12.2), which in turn depends on the separation axiom (Ax.S)

  In $\mathbb{G}$, where $Ax.S$ is false, the crossbar theorem is also false.

4.12 Duals of results from chapters 8 and 9

4.12.1 Theorems (14)

- **Theorem 8.1D**: The set of angle measures $\mathbb{D} = [0, 180]$
- **Theorem 8.2D**: All wedges, fans, pencils have infinitely many rays
- **Theorem 8.3D**: Let $x \neq y$ be distinct from a on fan $\overrightarrow{ab}$. Then, exactly one of

$$a\text{-}x\text{-}y \quad \text{or} \quad a\text{-}y\text{-}x.$$

- **Theorem 8.4D**: Let $\overrightarrow{ab}$ be a fan. If $c \in \overrightarrow{ab}$, $0 < c < 180$, then $\overrightarrow{ab} = \overrightarrow{ac}$
- **Theorem 8.6D**: Stated in theorem 11.6
- **Theorem 9.1D**: Let ray a be in pencil P, there exists a unique fan $a' \in P$ such that $aa' = 180$. For all other rays $x \in P$, $a\text{-}x\text{-}a'$
- **Theorem 9.2D**: Stated in theorem 11.8
- **Theorem 9.4D**: If $ap = 180$ in some fan h, then $p = a'$.
- **Theorem 9.6D**: Stated in theorem 11.9
- **Theorem 9.7D**: Each fan has a unique opposite fan.
- **Theorem 9.8D**: Let rays $a, b \in P$, if $0 < ab < 180$, then fan $\overrightarrow{ab'} = \overrightarrow{ab'}$
- **Theorem 9.9D**: Let rays $a, b \in P$, if $0 < ab < 180$, then $P = \overline{ab} \cup \overline{ab'} \cup \overline{ba'} \cup \overline{b'a'}$, where the interiors of these wedges are disjoint.
- **Theorem 9.10D**: Let rays $a, b \in P$, if $0 < ab < 180$, and c is some other ray in P, then there exists no betweenness relation among a, b, c if and only if $c \in \overline{a'b'}$

4.12.2 Propositions

- **Proposition 8.11D**: Let $a, b \in P$, $0 < ab < 180$, there exists $c \in P$ such that $c\text{-}a\text{-}b$, $cb < 180$
- **Proposition 8.5D**: A fan has at most two terminal rays
- **Proposition 8.7D**: Let $\overline{ab}$ be a wedge, for all $x, y \in \overline{ab}$, $xy \leqslant ab$, if $xy = ab$, then $\{x, y\} = \{a, b\}$
- **Proposition 8.8D**: If $\overline{ab} = \overline{cd}$, then $\{a, b\} = \{c, d\}$
- **Proposition 8.9D**: Stated in proposition 11.15
- **Proposition 9.3D**: Let $a, b \in P$ such that $0 < ab < 180$. Then,
  - Fan $\overrightarrow{ab} = \overline{ab} \cup \overline{ba'}$, with $\overline{ab} \cap \overline{ba'} = \varnothing$
  - Fan $\overrightarrow{ab} = \overrightarrow{a'b}$

4.13 Side angle side congruence

- **Intro**: We know present the 21 axiom for a general plane, in fact the last axiom that will be covered here. Euclid called it a theorem, but we'll show that it is not a consequence of the first 20 axioms, and so cannot be proved from them.

  Our context is a general plane with the 20 axioms. These include $\mathbb{E}, \hat{\mathbb{E}}, \mathbb{M}, \mathbb{H}, \mathbb{S}$

- **Definition: Congruence**: Two segments $\overline{AB}$ and $\overline{XY}$ are **congruent** ($\cong$) if they have the same length: $\overline{AB} \cong \overline{XY}$ means $AB = XY$

  Two angles $\angle CAB$ and $\angle ZXY$ are congruent if they have the same angle measure

  Two triangles $\triangle ABC$ and $\triangle ZXY$ are congruent under the correspondence $A \leftrightarrow X$, $B \leftrightarrow Y, C \leftrightarrow Z$ (Write as $ABC \leftrightarrow XYZ$) if

  $$\overline{AB} \cong \overline{XY}, \quad \overline{BC} \cong \overline{YZ}, \quad \overline{AC} \cong \overline{XZ}.$$

  and

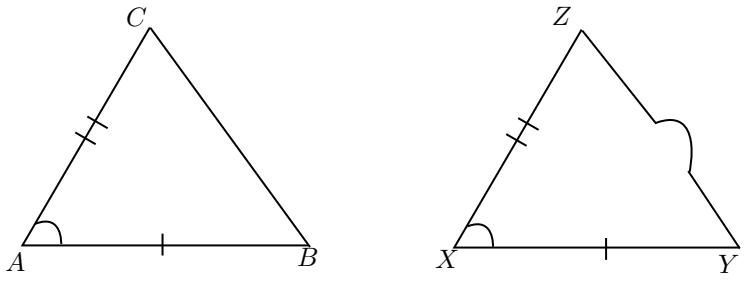
  $$\angle ABC \cong \angle XYZ, \quad \angle CAB \cong \angle ZXY, \quad \angle BCA \cong \angle YZX.$$

  denote this by $\triangle ABC \cong \triangle XYZ$

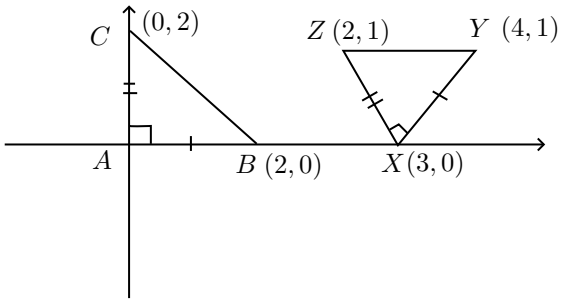
  Note: The correspondence of vertices is an essential part of the congruence of triangles.

- **Side-angle-side axiom (Ax.SAS)**: If under the correspondence $ABC \leftrightarrow XYZ$ between the vertices of $\triangle ABC$ and those of $\triangle XYZ$, two sides of $\triangle ABC$ are congruent to the corresponding two sides of $\triangle XYZ$, and the angle included between these two sides of $\triangle ABC$ is congruent to the corresponding angle of $\triangle XYZ$, then $\triangle ABC \cong \triangle XYZ$

- **The bumpy plane** $\hat{\mathbb{E}}$: Observe that Ax.SAS is false for the bumpy plane



- **The taxicab plane**: Observe that Ax.SAS is false for the taxicab plane



Note: Since the first 20 axioms are true for $\hat{\mathbb{E}}, \mathbb{M}$, but the 21st is false, is is therefore not possible that the 21st is a consequence of the first 20.

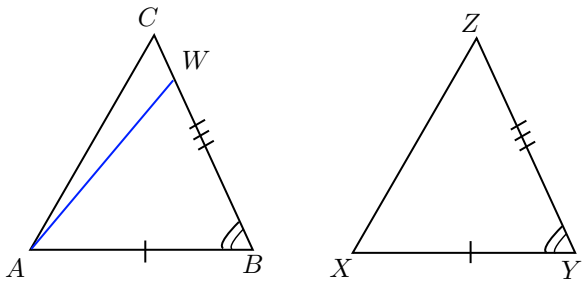
Ax.SAS is true for $\mathbb{E}, \mathbb{S}, \mathbb{H}$

- **Definition: Absolute plane**: An **absolute plane** $\mathbb{P}$ is a set of points $\mathbb{P}$ with lines, distance, and angular distance (all undefined terms), such that all 21 axioms are true. The three planes above are absolute planes
- **Theorem 13.1 (ASA)**: If under the correspondence $ABC \leftrightarrow XYZ$, two angles and the included side of $\triangle ABC$ are congruent, respectively, to the corresponding two angles and included side of $\triangle XYZ$, then $\triangle ABC \cong \triangle XYZ$

Proof. If we can show that $BC = YZ$, then we can apply Ax.SAS: $\overline{BA} \cong \overline{YX}$, $\overline{BC} \cong \overline{YZ}, \angle ABC \cong \angle XYZ$ implies $\triangle ABC \cong \triangle XYZ$

Suppose toward a contradiction that $BC \neq YZ$. Since we are given exactly the same information about the two triangles, we may choose notation so that $BC > YZ$

Thm 8.6 implies there is a point W on $\overrightarrow{BC}$ with $BW = YZ$, then $BW < BC$, which implies B-W-C



$\overrightarrow{BC} = \overrightarrow{BW}$ (Thm 8.4), and Thm 10.3 implies C, W are in the same halfplane with edge $\overleftrightarrow{AB}$. IN particular, A, B, W are noncollinear, and we have $\triangle ABW$ with

$$\overline{BA} \cong \overline{YX},\ \overline{BW} \cong \overline{YZ},\ \angle ABW = \angle ABC \cong \angle XYZ.$$

Note that $\angle ABW = \angle ABC$ by prop 11.14. With these facts, and by Ax.SAS, $\triangle ABW \cong \triangle XYZ$, which implies $\angle BAW \cong \angle YXZ$.

$\angle YXZ = \angle X \cong \angle A = \angle BAC$, so $\angle BAW = \angle BAC$

B-W-C and Ax.C implies $\overrightarrow{AB}$-$\overrightarrow{AW}$-$\overrightarrow{AC}$, which implies $\overrightarrow{AB}\overrightarrow{AW} + \overrightarrow{AW}\overrightarrow{AC} = \overrightarrow{AB}\overrightarrow{AC}$, which implies $\angle BAW + \angle WAC = \angle BAC$

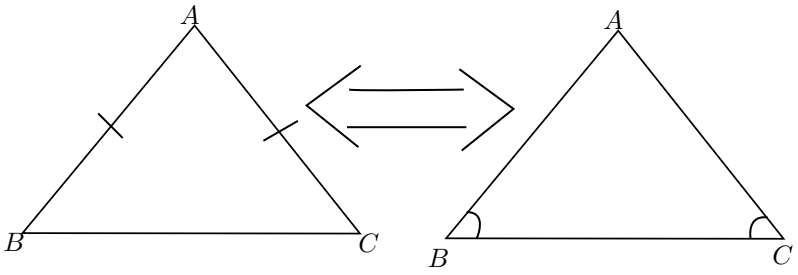
Thm 10.3 implies W, B in a halfplane with edge $\overleftrightarrow{AC}$, which implies W, A, C noncollinear, which implies $\angle WAC$ is proper. Thus, $\angle WAC > 0$, and $\angle BAW < \angle BAC$, which is a contradiction ■

- **Definition: types of triangles**
  - A triangle is **isosceles** if two sides have the same length
  - **Equilateral** if all three sides have the same length
  - **Equiangular** if all three angles have the same measure

  Note: A triangle can be called **scalene** if all all three sides have different lengths and all three angles have different measures

- **Theorem 13.2 (pons asinorum ("Bride of asses"))** In any $\triangle ABC$,

$$AB = AC \iff \angle ACB = \angle ABC$$

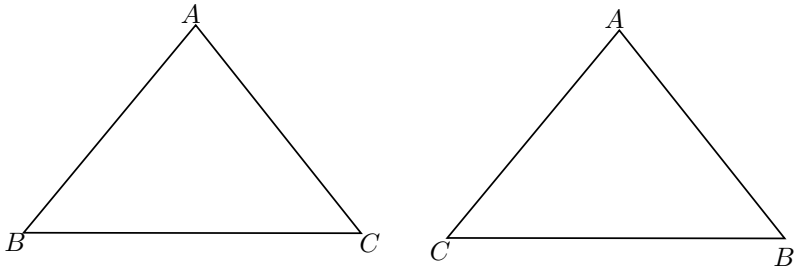


Proof. Consider the correspondence $ABC \leftrightarrow ACB$ between the vertices of $\triangle ABC$ and itself

Suppose that $AB = AC$, then

$$\overline{AB} \cong \overline{AC}, \quad \overline{AC} \cong \overline{AB}, \quad \angle BAC \cong \angle CAB.$$

So, Ax.SAS implies $\triangle ABC \cong \triangle ACB$, which implies $\angle ACB = \angle ABC$



Suppose that $\angle ACB = \angle ABC$. Then, $\angle ACB \cong \angle ABC$, $\angle ABC \cong \angle ACB$, $\overline{CB} = \overline{BC}$

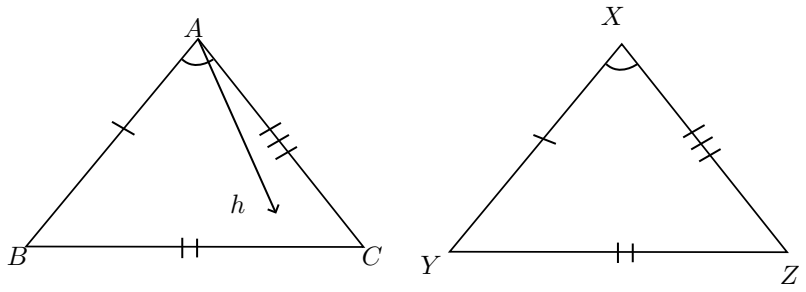
So, Thm 13.1 (ASA) implies $\triangle ABC \cong \triangle ACB$, which implies $AB = AC$ by congruence

- **Corollary 13.3**: A triangle is equilateral if and only if it is equiangular
- **Theorem 13.4 (SSS)**: If in $\triangle ABC$ and $\triangle XYZ$, $\overline{AB} \cong \overline{XY}$, $\overline{BC} \cong \overline{YZ}$ and $\overline{CA} \cong \overline{ZX}$, then

$$\triangle ABC \cong \triangle XYZ.$$

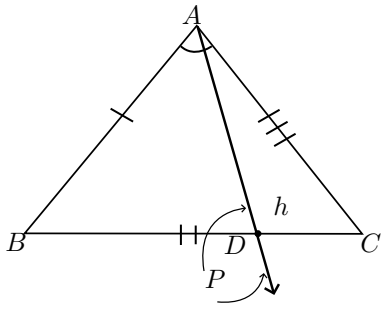
Proof. We'll show that $\angle BAC = \angle YXZ$. Then, $Ax.SAS$ will imply $\triangle ABC \cong \triangle XYZ$.

So, assume for the sake of contradiction that $\angle A \neq \angle X$. We may assume that $\angle A > \angle X$ (otherwise, just switch the notation for vertices from one triangle to the other). So, $\angle BAC > \angle YXZ$, which means that $\overrightarrow{AB}\overrightarrow{AC} > \angle X$. We apply Ax.RF (or thm 11.6) to the fan $\overrightarrow{\overrightarrow{AB}\overrightarrow{AC}}$. Thus, there is a ray h in $\overrightarrow{\overrightarrow{AB}\overrightarrow{AC}}$ with $\overrightarrow{AB}h = \angle X$



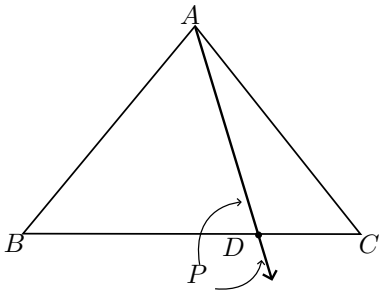
Because $\overrightarrow{AB}h = \angle X < \overrightarrow{AB}\overrightarrow{AC}$, we have $\overrightarrow{AB}$-h-$\overrightarrow{AC}$. $\underline{\overrightarrow{AB}\overrightarrow{AC}}$ is a proper angle, since A, B, C noncollinear. So, we may apply the Crossbar Theorem

Therefore, there is a point $D \in h^0$ with B-D-C. So, $h = \overrightarrow{AD}$ (thm 8.4)



Ax.RR implies there is a point P on h with $AP = XZ$. So, $h = \overrightarrow{AP}$. $\overline{AB} \cong \overline{XY}$, $\overline{AP} \cong \overline{XZ}, \underline{\angle BAP} = \underline{\overrightarrow{AB}h} \cong \underline{\angle YXZ}$

Ax.SAS implies $\triangle ABP \cong \triangle XYZ$, which implies $BP = YZ = BC$ (So $P \neq D$, since $BD < BC$), and $AP = XZ = AC$



Pons Asinorum (13.2) for $\triangle BCP$ implies $\angle BCP = \angle BPC$, and for $\triangle ACP$, $\angle ACP = \angle APC$

B-D-C and Ax.C implies $\overrightarrow{PB}$-$\overrightarrow{PD}$-$\overrightarrow{PC}$, which implies $\angle BPD + \angle DPC = \angle BPC$, implies $\angle BPC > \angle DPC$

Now, we consider separately two cases for P on $\overrightarrow{AD}$

1.) A-P-D. Then, $\overrightarrow{PD} = \overrightarrow{PA'}$ (Thm 9.6), so Thm 11.8 implies $\overrightarrow{PA}$-$\overrightarrow{PC}$-$\overrightarrow{PD}$

Thus, $\angle APC + \angle DPC = \angle APD = 180$ (Ax.M4)

A-P-D and Ax.C implies $\overrightarrow{CA}$-$\overrightarrow{CP}$-$\left(\overrightarrow{CD} = \overrightarrow{CB}\right)$, which implies $\angle ACP + \angle BCP = \angle ACB$.

Since $\angle ACP = \angle APC$ and $\angle BCP = \angle BPC > \angle DPC$, $\angle ACB = \angle ACP + \angle BCP = \angle APC + \angle BPC > \angle APC + \angle DPC = 180$, which contradicts Ax.M1

2.) A-D-P (needs proof)

4.14 Perpendiculars

- **Intro:** We've introduced 21 axioms, and defined an **absolute plane**, which hase
  - **Undefined terms**: Point, line, distance, angle measure
  - **21 Axioms**

  Examples are $\mathbb{E}, \mathbb{H}, \mathbb{S}$

  Next, we introduce and study **perpendicular lines**.

- **Definition: Supplementary angles**: Two angles are **supplementary** if their measures sum to 180.

- **Theorem 14.1 (Supplementary angles theorem)**: If h, j are coterminal rays, then $\underline{hj}$ and $\underline{jh'}$ are supplementary

  Proof. If $j = h$, then $hj = 0$ by Ax.M2, and $jh' = 180$ by Ax.M4. So, $hj + jh' = 180$.

  If $j \neq h$ or h', thm 11.8 implies h-j-h', which means $hj + jh' = hh'$, and $hh' = 180$ by Ax.M4

- **Definition**: Angles $\underline{hk}, \underline{rs}$ are **vertical** if $\{r, s\} = \{h', k'\}$

- **Theorem 14.2 (Vertical angles theorem)**: Vertical angles are congruent

  Proof. Given vertical angles $\underline{hk}, \underline{h'k'}$, Thm 14.1 implies $\underline{hk}$ and $\underline{kh'}$ are supplementary, and $\underline{kh'}$ and $h'k'$ are supplementary, which implies $kh + kh' = 180 = kh' + h'k'$, which implies $hk = h'k'$

  Note: When two lines intersect, four angles are formed, angles $\underline{hk}, \underline{kh'}, \underline{h'k'}$, and $\underline{k'h}$, the measure of any one determines the measure of the others, by thms 14.1, 14.2. In particular, if $hk = 90$, then all four angle measures are 90

- **Definition: Perpendicular**: Two intersecting lines m, n are **perpendicular** (at point of intersection B) if the four angles they determine at B are right angles, we write $m \perp n$ (at B)

  Note: Prop 11.14 implies if $m \perp n$ at B, then $m \perp n$ at B^* (when $\omega < \infty$)

- **Theorem 14.3**: Through any point A on a line m, there is exactly one line n perpendicular to m

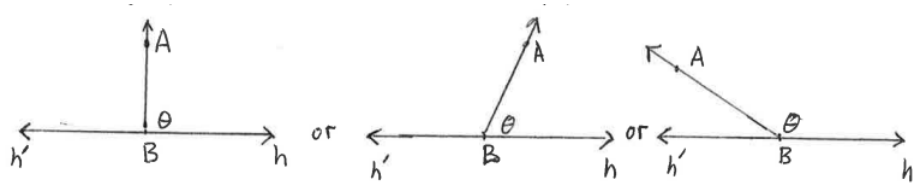
  Proof. Let h be a ray on m with endpoint A (so $m = h \cup h'$)

  Since $0 < 90 < 180$, coroll 12.3 implies there are exactly two rays k, j in P_A with $hk = hj = 90$

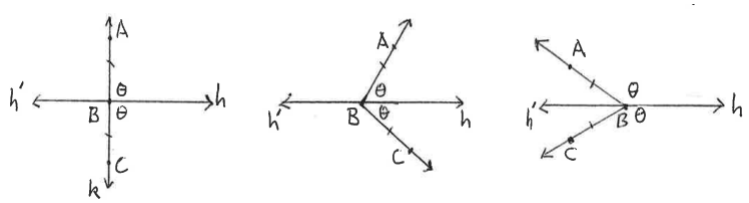
Let $n = k \cup k'$. $hk = 90$ implies $hk' = 90$ by thm 14.1, so $j = k'$. That is, k, k' are the only rays in P_A that form a right angle with h. So, n is the only line through A with $m \perp n$

- **Definition: The perpendicular bisector**: The **perpendicular bisector** of a segment $\overline{AB}$ is the line perpendicular to $\overleftrightarrow{AB}$ at the midpoint M of $\overline{AB}$
- **Theorem 14.9 (needs proof)**: Every point of the perpendicular bisector of a segment is equidistant from the endpoints of the segment: $AX = BX$ for all X on the perpendicular bisector
- **Theorem 14.10 (converse of 14.9)**: Let $m = \overleftrightarrow{AB}$, suppose that line $n \neq m$ meets m at the midpoint M of $\overline{AB}$. Suppose that there is some point X on n, not on m, so that $AX = BX$. Then, $n \perp n$ at M
- **Note about 14.9 and 14.10**: Theorems 14.9 and 14.10 say that the perpendicular bisector of $\overline{AB}$ consists exactly of the points X in $\mathbb{P}$ such that $AX = BX$
- **Theorem 14.4**: Through a point A not on a given line m there is at least one line n perpendicular to m

Proof. Choose any point B on m. Since A is not on m, $AB < \omega$. So, there is a unique line $\overleftrightarrow{AB}$ through A and B, and ray $\overrightarrow{BA}$ is defined. Let h be a ray in m with endpoint B, so that $m = h \cup h'$. Let $\theta = \overrightarrow{BA}h$



A is in a halfplane H with edge m. Coroll. 12.3 implies there is a second ray k with endpoint B so that k^0 is in the opposite halfplane K and $hk = \theta$



Ax.RR implies there's a point C on k with $BC = BA$. Thm 10.6 implies there's a point X on m with A-X-C. Then, $\overrightarrow{XC} = \overrightarrow{XA}'$

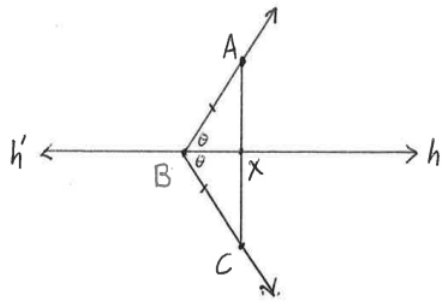
We consider some cases

1.) If $X = B$, then $\overrightarrow{BC} = \overrightarrow{BA}'$, so $\overrightarrow{BA}h + \overrightarrow{BC}h = 180$ (Thm 14.1), this implies $\theta + \theta = 180$, thus $\theta = 90$. So, $\overleftrightarrow{AB} \perp m$. Note that this is the lucky case. Our random choice of B gave us a perpendicular line.

2.) If $X = B^*$, then $\overrightarrow{B^*C} = \overrightarrow{B^*A'}$. Thus, h has endpoint B^* (Prop 9.3). So, Thm 14.1 implies $\overrightarrow{B^*A}h + \overrightarrow{B^*C}h = 180$

In this case, prop 11.14 implies $\overrightarrow{B^*A}h = \overrightarrow{BA}h$, and $\overrightarrow{B^*C}h = \overrightarrow{BC}h$ $\overrightarrow{BA}h + \overrightarrow{BC}h$, and thus $\theta + \theta = 180$, so $\theta = 90$. Thus, we have $\overleftrightarrow{AB} \perp m$

3.) If X is in h^0, then $h = \overrightarrow{BX}$.

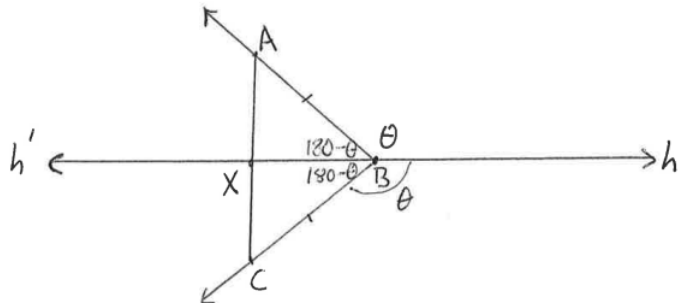


$\theta = \angle ABX = \angle CBX$, so $\underline{\angle ABX} \cong \underline{\angle CBX}$, $\overline{BA} \cong \overline{BC}, \overline{BX} \cong \overline{BX}$, and thus Ax.SAS implies $\triangle ABX \cong \triangle CBX$, which implies $\angle AXB = \angle CXB$ by definition of congruent triangles

From here, thm 14.1 implies $180 = \overrightarrow{XA}\overrightarrow{XB} + \overrightarrow{XC}\overrightarrow{XB} = \angle AXB + \angle CXB$.

Therefore, $\angle AXB = \angle CXB = 90$. So, $\overleftrightarrow{AX} \perp m$ at X, $\overleftrightarrow{AX}$ goes through A.

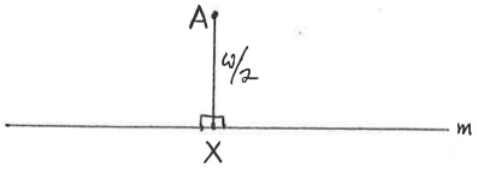
If X is on Inth', thm 14.1 implies $\angle ABX = 180 - \theta = \angle CBX$. So again, $\overline{BA} \cong \overline{BC}$, $\overline{BX} \cong \overline{BX}, \underline{\angle ABX} \cong \underline{\angle CBX}$ and Ax.SAS implies $\triangle ABX \cong \triangle CBX$, which implies $\angle AXB = \angle CXB$.



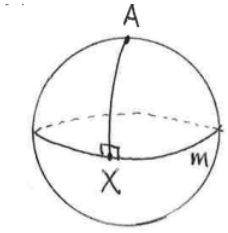
By thm 14.1, $180 = \overrightarrow{XA}\overrightarrow{XB} + \overrightarrow{XC}\overrightarrow{XB} = \angle AXB + \angle CXB$, so $\angle AXB = \angle CXB = 90$, hence $\overleftrightarrow{AX} \perp m$, and $\overleftrightarrow{AX}$ goes through A. ■

- **Definition: Pole**: Point A is a **Pole** of line m if there exists a point X on m such that

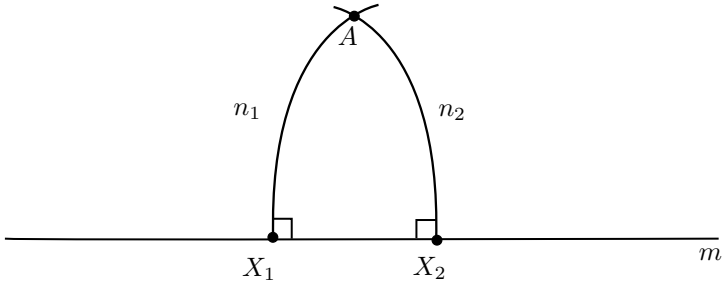
$$\overleftrightarrow{AX} \perp m \text{ and } AX = \frac{\omega}{2}.$$



So, poles will exist only when $\omega < \infty$. Think m = equator, A = north pole.



- **Theorem 14.5**: If there are two different lines through a point A and perpendicular to a line m, then A is a pole of m.

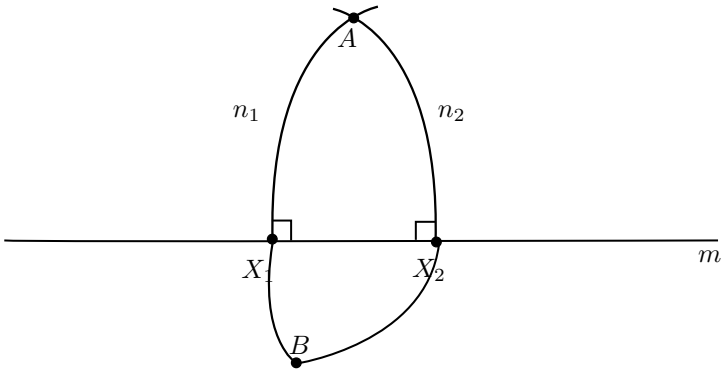


Proof. We'll show that $AX_1 = \frac{\omega}{2}$. Note that $X_1X_2 < \omega$, since $n_1 \neq n_2$

$AX_1 < \omega$ (Since A is not on m), so $\overrightarrow{X_1A}$, and $\overrightarrow{X_1A'}$ exist. Ax.RR implies there is a point B on $\overrightarrow{X_1A'}$ with $X_1B = X_1A$

$\overline{X_1A} \cong \overline{X_1B}, \overline{X_1X_2} \cong X_1X_2, \angle AX_1X_2 = 90 = \angle BX_1X_2$, thus $\underline{\angle AX_1X_2} \cong \underline{\angle BX_1X_2}$. So, by Ax.SAS, $\triangle AX_1X_2 \cong \triangle BX_1X_2$, which implies $\angle BX_2X_1 = \angle AX_2X_1 = 90$

Thus, $\overrightarrow{X_2X_1}\overrightarrow{X_2B} = \overrightarrow{X_2X_1}\overrightarrow{X_2A'}$ with $\overrightarrow{X_2B}^0, \overrightarrow{X_2A'}$ in same halfplane with edge m, which implies $\overrightarrow{X_2B} = \overrightarrow{X_2A'}$ (Coroll. 12.3)



This implies $\overrightarrow{X_2B} \subseteq n_2$, n_1, n_2 meet in A and B

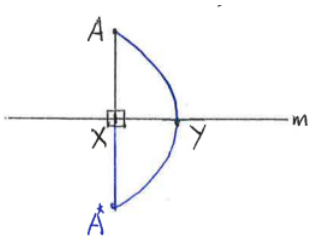
Ax.I4 implies $AB = \omega$. Then, A-X_1-B and $AX_1 = BX_1$, which implies $AX_1 = BX_1 = \frac{\omega}{2}$ ■

- **Theorem 14.6**: If A is a pole of line m, then every line through A is perpendicular to m, and meets m at a point distance $\frac{\omega}{2}$ from A. Also, every line perpendicular to m goes through A

Proof. Let X be a point on m given by the definition of pole: $\overleftrightarrow{AX} \perp m$ (at X), and $AX = \frac{\omega}{2}$

Then, $\overleftrightarrow{AX^*} \perp m$ at X^* by prop 11.14, and $AX^* = \omega - AX = \omega - \frac{\omega}{2} = \frac{\omega}{2}$

Let Y be any point on m, $Y \neq X$ or X^*. Then, X, Y, A are noncollinear, so $\triangle AXY$ is defined.



A^* is on the opposite side of m from A (Coroll 10.9) and A-X-A^* (thm 10.8, 9.1), which implies $\frac{\omega}{2} + XA^* = AX + XA^* = AA^* = \omega$, implies $XA^* = \frac{\omega}{2} = XA$; along with $XY = XY$, $\angle AXY = 90 = \angle A^*XY$

So, Ax.SAS implies $\triangle AXY \cong \triangle A^*XY$, which implies $\angle AYX = \angle A^*YX$

Since $\overleftrightarrow{YA^*} = \overleftrightarrow{YA'}$ (coroll 9.8).

Thm 14.1 implies $\overrightarrow{YA}\overrightarrow{YX}$ and $\overrightarrow{YA^*}\overrightarrow{YX}$ are supplementary, which implies $\angle AYX + \angle A^*YX = 180$

Therefore, $\angle AYX = \angle A^*YX = 90$, hence $\overleftrightarrow{AY} \perp m$ at Y

Also, A-Y-A^* implies $AY + YA^* = AA^* = \omega$. $\triangle AXY \cong \triangle A^*XY$ implies $AY = A^*Y$, which implies $AY = A^*Y = \frac{\omega}{2}$

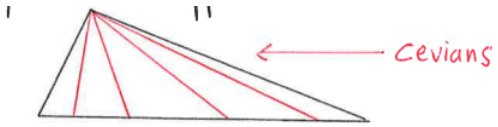
If ℓ is any line perpendicular to m, say at point Z, we just proved that $\overleftrightarrow{AZ} \perp m$. By Thm 14.3, $\overleftrightarrow{AZ} = \ell$, so ℓ goes through A ■

- **Corollary 14.7**: Suppose $\omega < \infty$, each line m has exactly two poles, A and A^*
- **Definition:** ***Right triangle***: A **right triangle** is a triangle with exactly **one** right angle.
- **Definition:** ***Hypotenuse***: In a right triangle, the **hypotenuse** is the side opposite the right angle. The **legs** are the other two sides
- **Definition:** ***Birectangular triangle***: A triangle with exactly **two** right angles is a **birectangular** (e.g $\triangle ABC$ on $\mathbb{S}$ with B, C on equator, $A =$ north pole).
- **Definition:** ***Trirectangular triangle***: A triangle with three right angles is **trirectangular**
- **Definition:** ***small triangle***: A triangle is **small** if all sides have length $< \frac{\omega}{2}$. (So when $\omega = \infty$, every triangle is small).

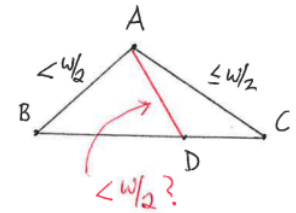
If $\triangle ABC$ has more than one right angle (say $\angle B = \angle C = 90$), then $\overleftrightarrow{AB}, \overleftrightarrow{AC}$ both perpendicular to $\overleftrightarrow{BC}$, so thm 14.5 implies A is a pole for $\overleftrightarrow{BC}$. Then, Thm 14.6 implies $AB = AC = \frac{\omega}{2}$, which implies $\triangle ABC$ is **not** small.

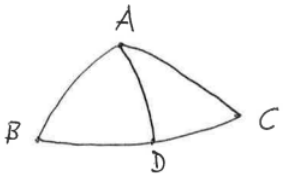
4.15 The Exterior Angle Inequality and the Triangle Inequality

- **Definition: *Cevian*:** A **Cevian** is a segment from a vertex of a triangle to a point on the opposite side.



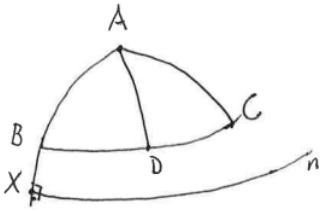
- **Theorem 15.1 (Cevian theorem)**: Suppose $\omega < \infty$, if $AB < \frac{\omega}{2}$, and $AC \leqslant \frac{\omega}{2}$ in $\triangle ABC$, and if B-D-C (so $\overline{AD}$ is a cevian of $\triangle ABC$), then $AD < \frac{\omega}{2}$





Proof. Ax.RR implies there's a point X on $\overrightarrow{AB}$ with $AX = \frac{\omega}{2}$. $AX = \frac{\omega}{2} > AB$ implies A-B-X. So, $\overleftrightarrow{AB} = \overleftrightarrow{AX}, \overrightarrow{AB} = \overrightarrow{AX}$

Thm. 14.3 implies there's a line $n \perp \overleftrightarrow{AB}$ at X

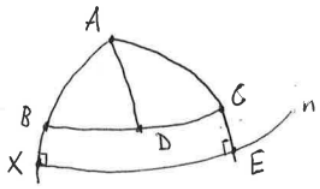


$AX = \frac{\omega}{2}$ and $\overleftrightarrow{AX} \perp n$ implies A is a pole of n. $\overleftrightarrow{AC}$ meets n twice, at a pair of antipodes (Thm. 10.11), one of which will be on $\overrightarrow{AC}$, the other on $\overrightarrow{AC}'$. So, $\overrightarrow{AC}$ meets n at a point E.

Thm 14.6 implies $AE = \frac{\omega}{2}$, so either A-C-E (if $AC < \frac{\omega}{2}$), or $C = E$ (if $AC = \frac{\omega}{2}$).

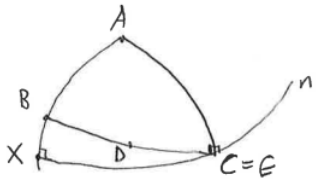
Let H be the halfplane with edge n that contains A. Thm 10.3 and X-B-A implies $B \in H$

If A-C-E



Then Thm 10.3 implies $C \in H$. H convex implies $\overline{BC} \subseteq H$, so B-D-C implies $D \in H$.

If $C = E$,

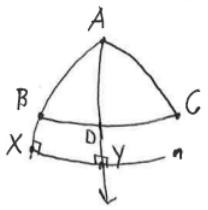


Then B-D-E and Thm 10.3 implies $D \in H$ in this case also

$\overleftrightarrow{AD}$ meets n in a pair of antipodes, so $\overleftrightarrow{AD}$ meets n in a point Y, and $AY = \frac{\omega}{2}$ by Thm 14.6

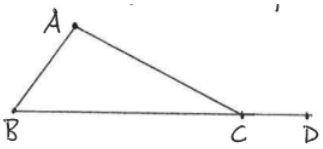
$D \in H$, $Y \in n$ implies $D \neq Y$. If A-Y-D then Thm 10.6 implies A, D are in opposite halfplanes, which is false.

So, $Y \in \overrightarrow{AD}$, which implies we must have A-D-Y. Therefore $AD < AY = \frac{\omega}{2}$

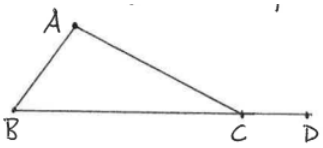


■

- **Definition:** ***exterior and remote interior angles*****:** Given $\triangle ABC$, and D a point with B-C-D, then $\underline{\angle ACD}$ is called an **exterior angle** of $\triangle ABC$, and $\underline{\angle A}, \underline{\angle B}$ are called the **remote interior angles** (relative to $\underline{\angle ACD}$)



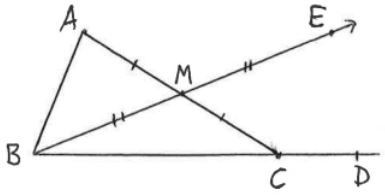
- **Theorem 15.3 (EAI)**: An exterior angle of a small triangle has larger measure than either remote interior angle

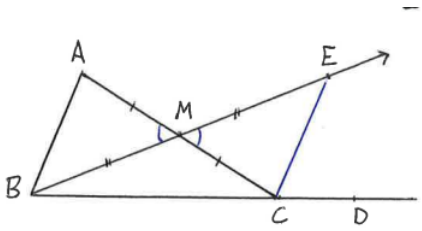


Proof. We aim to show $\angle ACD > \angle A,\ \angle ACD > \angle B$

$\overline{AC}$ has midpoint M, Ceviain Thm and $BA, BC < \frac{\omega}{2}$ (definition of small triangle) implies $BM < \frac{\omega}{2}$

Then, $2BM < \omega$, so Ax.RR implies there is a point E on ray $\overrightarrow{BM}$ with $BE = 2BM > BM$. So, defn of ray implies B-M-E, hence $ME = BM$





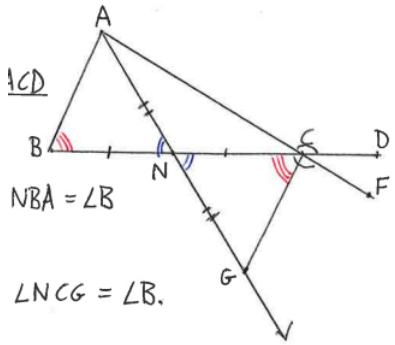
A-M-C, B-M-E and thm 9.6 implies $\overrightarrow{MC} = \overrightarrow{MA}', \overrightarrow{ME} = \overrightarrow{MB}'$, so $\underline{\angle AMB}, \underline{\angle CME}$ are vertical. Then, Thm 14.2 implies $\underline{\angle AMB} \cong \underline{\angle CME}$

So, $\overline{MA} \cong \overline{MC},\ \overline{MB} \cong \overline{ME}$, and AX.SAS implies $\triangle AMB \cong \triangle CME$.

Hence, $\angle MCE = \angle MAB = \angle CAB$. Ax.C and B-M-E implies $\overrightarrow{CB}$-$\overrightarrow{CM}$-$\overrightarrow{CE}$. B-C-D implies $\overrightarrow{CD} = \overrightarrow{CB}'$, which implies $\overrightarrow{CB}$-$\overrightarrow{CE}$-$\overrightarrow{CD}$ (Thm 11.8)

Then, Thm 11.5 (ROI for rays) implies $\overrightarrow{CB}$-$\overrightarrow{CM}$-$\overrightarrow{CE}$-$\overrightarrow{CD}$, which implies $\overrightarrow{CM}$-$\overrightarrow{CE}$-$\overrightarrow{CD}$. So, $\angle ACD = \angle MCD = \angle MCE + \angle ECD > \angle MCE = \angle CAB = \angle A$

From point F with A-C-F, exterior angle $\underline{\angle BCF}$ vertical to $\underline{\angle ACD}$. N = midpoint of $\overline{BC}$, $G \in \overrightarrow{AN}$, $NA = NG$ implies $\triangle BNA \cong \triangle CNG$, which implies $\angle NCG = \angle NBA = \angle B$. Then, $\angle ACD = \angle BCF = \angle NCG + \angle GCF > \angle NCG = \angle B$

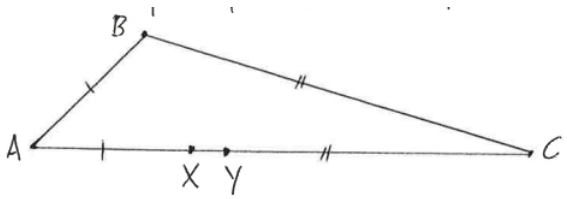


- **Corollary 15.4 (needs proof)**: The nonright angles of a small right triangle are accute
- **Corollary 15.5 (needs proof)**: The base angles of an isosceles triangle whose congruent sides are $< \frac{\omega}{2}$ are acute.
- **Triangle inequality informal proof sketch**: In any $\triangle ABC$ (in any absolute plane),

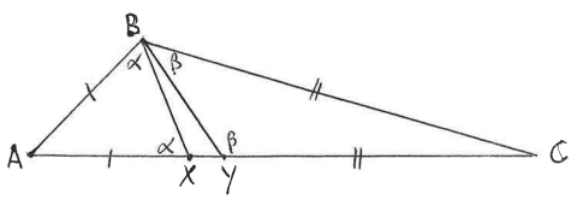
$$AB + BC > AC.$$

Suppose toward a contradiction that in $\triangle ABC$, $AB + BC < AC$ (we'll worry later about contradicting $AB + BC = AC$)

Then, there's a point X in $\overline{AC}$ with $AX = AB$, and a point Y in $\overline{XC}$ with $YC = BC$



Pons asinorum for $\triangle ABX$ implies $\angle ABX = \angle BXA(\alpha)$, and $\triangle BYC$ implies $\angle CBY = \angle CYB(\beta)$



If the EAI applies to $\triangle BXY$, then

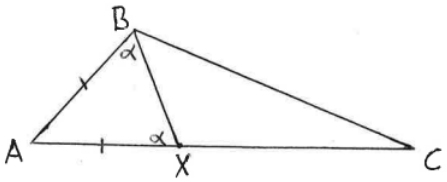
$$\beta = \angle BYC > \angle BXY = 180 - \alpha.$$

Which, implies $\alpha + \beta > 180$. But, at B, $\alpha + \beta < \angle ABC < 180$, a contradiction

- **Proposition 15.6**: If $AB < \frac{\omega}{2}$, and $BC \leqslant \frac{\omega}{2}$ in $\triangle ABC$, then $AB + BC > AC$

Proof. Suppose toward a contradiction that $AB + BC \leqslant AC$. Then, $0 < BC \leqslant AC - AB$. So, $AB < AC$

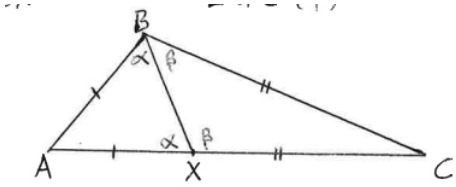
By Ax.RR, there is a point X on $\overrightarrow{AC}$ with $AX = AB$, hence A-X-C



Pons asinorum for $\triangle ABX$ implies $\angle ABX = \angle BXA$ (α).

$AX + BC = AC$ implies $XC = AC - AX = AC - AB \geqslant BC$

Case 1) Suppose that $XC = BC$. Pons asinorum for $\triangle BCX$ implies $\angle XBC = \angle BXC$ (β)

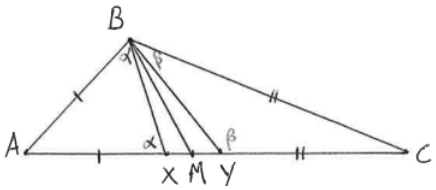


A-X-C and Ax.C implies $\overrightarrow{BA}$-$\overrightarrow{BX}$-$\overrightarrow{BC}$, which implies $\angle ABC = \alpha + \beta = \angle AXB + \angle BXC = 180$ (Thm 14.1)

But, A, B, C not collinear, implies $\underline{\angle ABC}$ is proper, which implies $\angle ABC \neq 180$ by Ax. M4, a contradiction.

Case 2) Suppose that $XC > BC$. Ax.RR for $\overrightarrow{CX}$ implies there is a point Y on $\overrightarrow{CX}$ with $CY = BC$. Then, $CY < CX$ implies C-Y-X by definition of $\overrightarrow{CX}$, which implies X-Y-C.

Let M be the midpoint of $\overline{XY}$. $BA < \frac{\omega}{2}, BC \leqslant \frac{\omega}{2}$ (by hypothesis) so Theorem 15.1 implies $BX, BM, BY < \frac{\omega}{2}$



$XM = MY = \frac{1}{2}XY < \frac{\omega}{2}$, so $\triangle BMX, \triangle BMY$ are small.

Pons asinorum for $\triangle BYC$ implies $\angle CBY = \angle CYB$ (β). EAI for $\triangle BMY$ implies $\beta = \angle BYC > \angle BMY$. EAI for $\triangle BMX \implies \alpha = \angle BXA > \angle BMX$

So, $\alpha + \beta > \angle BMX + \angle BMY = 180$ (Thm 14.1), but $180 > \angle ABC$ (Ax.M1, M4). We have

$$\begin{aligned} 180 &> \angle ABC \\ &= \angle ABX + \angle XBC \\ &= \angle ABX + \angle XBY + \angle YBC \\ &= \alpha + \angle XBY + \beta > \alpha + \beta. \end{aligned}$$

A contradiction, so $AB + BC \leqslant AC$ is false, therefore $AB + BC > AC$

- **Theorem 15.7 (The triangle inequality)**: In any $\triangle ABC$,

$$AB + BC > AC.$$

Proof. If $AB \geqslant \frac{\omega}{2}$ and $BC \geqslant \frac{\omega}{2}$, then $AB + BC \geqslant \omega > AC$. So, we may assume that one of AB, BC, say AB is less than $\frac{\omega}{2}$.

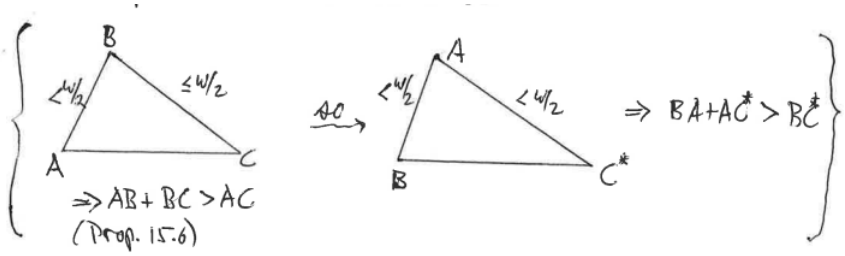
If $BC \leqslant \frac{\omega}{2}$, then $AB < \frac{omega}{2}$ and prop 15.6 implies $AB + BC > AC$. So, we may assume that $BC > \frac{\omega}{2}$and $AB < \frac{\omega}{2}$. In particular, $\omega < \infty$

If $AC \leqslant \frac{\omega}{2}$, then $AB + BC > BC > \frac{omega}{2} \geqslant AC$ and we're done. So, we may also assume that $AC > \frac{\omega}{2}$. We now have $AB < \frac{omega}{2}$, $BC > \frac{omega}{2}, AC > \frac{\omega}{2}$

A, B, C^* are not collinear, since if C^* is on line $\overleftrightarrow{AB}$, then so is C by Thm. 10.8, a contradiction. So, $\triangle ABC^*$ is defined

Now, $BC^* = \omega - BC$ and $AC^* = \omega - AC$. $AC > \frac{\omega}{2}$ implies $AC^* < \omega - \frac{\omega}{2} = \frac{\omega}{2}$

So, in $\triangle ABC^*$, we have $BA < \frac{\omega}{2}$, and $AC^* < \frac{\omega}{2}$. Then, prop 15.6 implies that $BA + AC^* > BC^*$



So, $BA + (\omega - AC) > (\omega - BC)$ implies $AB + BC > AC$ ■

- **Corollary 15.8**: For any points A, B, C, $AB + BC \geqslant AC$.

  Proof. If A, B, C are not collinear, Thm 15.7 implies $AB + BC > AC$. If A, B, C are collinear and distinct, Thm 7.3 implies $AB + BC \geqslant AC$

  If $B = A$ or C, then $AB + BC = AC$. If $A = C$, then $AB + BC \geqslant 0 = AC$. So, the inequality holds in every case.

- **Theorem 16.1 (Comparison theorem)**: If one angle of a triangle is larger than a second, then the side opposite the lager angle is longer than the side opposite the smaller angle; and conversely.

  That is, in $\triangle ABC$,

  $$\angle B > \angle C \iff AC > AB.$$

  Proof. Assume $\angle B > \angle C$. Ax.RF implies there is a ray j in Fan $\overrightarrow{\overrightarrow{BC}\overrightarrow{BA}}$ so that $\overrightarrow{BC}j = \angle C$. Then, $\overrightarrow{BC}j < \angle B = \angle CBA = \overrightarrow{BC}\overrightarrow{BA}$, which implies $\overrightarrow{BC}$-j-$\overrightarrow{BA}$

  $\underline{\overrightarrow{BC}\overrightarrow{BA}}$ is proper, since A, B, C are noncollinear, so the crossbar theorem may be applied.

  j^0 meets $\overline{AC}^0$ at a point D, so A-D-C and $j = \overrightarrow{BD}$

  Now, $\angle CBD = \overrightarrow{BC}j = \angle ACB = \angle DCB$ (prop 11.14), so pons asinorum for $\triangle DBC$ implies $CD = BD$. Triangle inequality for $\triangle ADB$ implies

  $$\begin{aligned} AB &< AD + BD \\ &= AD + CD = AC. \end{aligned}$$

  Therefore, $AC > AB$

  If $\angle C > \angle B$, then the same argument, reversing the notation $B \leftrightarrow C$ implies $AB > AC$.

  If $\angle B = \angle C$, then pons asinorum implies $AC = AB$. Thus,

  $$\angle B > \angle C \iff AC > AB.$$

- **Corollary 16.2 (Needs proof)**: The hypotenuse of a small right triangle is its longest side

- **Theorem 16.3**: Suppose that in $\triangle ABC$, $\angle C = 90$ and $AC < \frac{\omega}{2}$. Then, $\underline{\angle B}$ is acute and $AB > AC$.

  Proof. If we can show that $AB > AC$, then the Comparison Theorem will imply that $\angle C > \angle B$, so $90 = \angle C > \angle B$, hence $\underline{\angle B}$ will be acute. Thus, we aim to show that $AB > AC$

By hypothesis, $AC < \frac{\omega}{2}$. So, if $AB \geqslant \frac{\omega}{2}$, then $AB > AC$ and we're done. So, we may assume that $AB < \frac{\omega}{2}$

Let M be the midpoint of $\overline{BC}$. Then, $BM = CM = \frac{1}{2}BC < \frac{\omega}{2}$. $AM < \frac{\omega}{2}$ by the Cevian Theorem (15.1), so $\triangle ABM$ and $\triangle ACM$ are both small.

In particular, $\triangle ACM$ is a small right triangle. Coroll. 16.2 implies $AM > AC$ and $90 = \angle ACM > \angle AMC$

EAI (15.3) for $\triangle AMB$ implies $\angle AMC > \angle ABM$, so

$$\begin{aligned} & \angle ACM > \angle ABM \\ \implies & \angle ACB > \angle ABC. \end{aligned}$$

So, the Comparison theorem 16.1 implies $AB > AC$

- **Definition:** for any line m and point A, the **distance between A and m**, denoted $d(A, m)$, is the minimum distance AX for all points X on m.

  Note: If A is on m, then $d(A, m) = AA = 0$

- **Theorem 16.8**: Let m be a line, $C \in m$, $A \notin m$, $\overleftrightarrow{AC} \perp m$
  (a) If $AC < \frac{\omega}{2}$ then $d(A, m) = AC$; and $AC < AX$, all $X \neq C$ on m
  (b) If $AC = \frac{\omega}{2}$ (so $\omega < \infty$), then $d(A, m) = \frac{\omega}{2} = AX$, all $X \in m$
  (c) If $AC > \frac{\omega}{2}$ (so $\omega < \infty$), then $d(A, m) = \omega - AC = AC^*$; and $AC^* < AX$, all $X \neq C^*$ on m

  Proof.

  (a) Suppose $\overleftrightarrow{AC} \perp m$ and $AC < \frac{\omega}{2}$. If $X \in m$ with $X \neq C$ or C^*, then A, C, X are not collinear, so $\triangle ACX$ exists. Then, Thm. 16.3 implies $AC < AX$. If $X = C^*$, then $AX = AC^* = \omega - AC > \frac{\omega}{2}$, since $AC < \frac{\omega}{2}$ so again $AC < AX$

  (b) Suppose $\overleftrightarrow{AC} \perp m$ and $AC = \frac{\omega}{2}$. Then A is a pole of m, so Thm 14.6 implies $AX = \frac{\omega}{2}$ for all X on m

  (c) Suppose $\overleftrightarrow{AC} \perp m$ and $AC > \frac{\omega}{2}$. Then, $AC^* = \omega - AC < \frac{\omega}{2}$. C^* is on m (Thm 10.8), and $\overleftrightarrow{AC^*} \perp m$ at C^* (prop 11.14), so part (a) implies $AC^* < AX$ for all $X \neq C^*$ on m

4.16 Extra

- **Definition:** ***parallel lines***: Two lines $m \neq n$ are called **parallel** if $m \cap n = \varnothing$. If so, we write $m \parallel n$

  Suppose a point P is not on a line m. There are exactly three mutually exclusive possibilities

  (i) There is no line through P parallel to m

  (ii) There is exactly one line through P parallel to m

  (iii) There are at least two lines through P parallel to m

- **Definition** An absolute plane in which (i) holds for every line m and point P not on m (ie no parallel lines) is called **spherical**

  An absolute plane in which (ii) holds for every line m and point P not on m is called **Euclidean**

  An absolute plane in which (iii) holds for every line m and point P not on m is called **Hyperbolic**

  These properties do not mix, only one must hold per absolute plane.

- **Theorem**: In any absolute plane $\mathbb{P}$, exactly one of the following must occur.

  (i) $\mathbb{P}$ is Spherical; $\sigma(ABC) = \angle A + \angle B + \angle C > 180$ for all $\triangle ABC$, and $\omega < \infty$

  (ii) $\mathbb{P}$ is Euclidean; $\sigma(ABC) = \angle A + \angle B + \angle C = 180$ for all $\triangle ABC$, and $\omega = \infty$

  (iii) $\mathbb{P}$ is Hyperbolic; $\sigma(ABC) = \angle A + \angle B + \angle C < 180$ for all $\triangle ABC$, and $\omega = \infty$

  $\mathbb{S}$, $\mathbb{E}$, and $\mathbb{H}$ are essentially the only examples, up to isometry.

- **Theorem (*AAA*)**: Suppose that $\triangle ABC$ and $\triangle XYZ$ are triangles in a non-Euclidean absolute plane with $\angle A = \angle X$, $\angle B = \angle Y$, and $\angle C = \angle Z$. Then, $\triangle ABC \cong \triangle XYZ$

Set-theoretic asides

5.1 Sets and structure

- **Binary relation**: Consider a set X, a binary relation $\leqslant_X$ on X is defined to be any subset of $X \times X$, so

$$\leqslant_X \subseteq X \times X.$$

  Thus, $\leqslant_X$ is a set of ordered pairs (x, y), with $x, y \in X$. If $(x, y) \in \leqslant_X$ for $x, y \in X$, then

$$x \leqslant_X y.$$

  So, $(x, y) \in \leqslant_X \iff x \leqslant_X y$

- **Ordered sets**: An ordered set is a set equipped with a relation that allows you to compare elements in a consistent way.

  An ordered set is a pair

$$(X, \leqslant_X),$$

  where X is a set and $\leqslant_X$ is a binary relation on X satisfying certain axioms. The exact axioms depend on the type of order. This notation is shorthand for "the ordered set X with order $\leqslant_X$"

  When we write

$$(X, \leqslant_X),$$

  we are saying For any two elements $x, y \in X$, the relation $\leqslant_X$ specifies whether x is below y, equal to y, or incomparable with y. Without the relation, X is just an unordered set.

  so, $(X, \leqslant_X)$ means we are taking the unordered set X, and giving it the structure $\leqslant_X$ so that we can order the elements in X. If $(x, y) \in \leqslant_X$, then

$$x \leqslant_X y,$$

  and we can say that x is below y in the order of X. If $(x, y) \notin \leqslant_X$, these elements are still in X, but we cannot say anything about their ordering relative to each other in X.

  So, $(X, \leqslant_X)$ is not a subset of X, it is a set equipped with structure, a pair than consists of a set, and a binary relation that uses that set for elements.

- **Partial order**: A relation $\leqslant_X$ on X is a **partial order** if, for all $x, y, z \in X$

  1. **Reflexive**: $x \leqslant_X x$
  2. **Antisymmetric**: If $x \leqslant_X y$ and $y \leqslant_X x$, then $x = y$
  3. **Transitive**: If $x \leqslant_X y$, and $y \leqslant_X z$, then $x \leqslant_X z$

  A set equipped with a partial order is called a partially ordered set, or poset.

- **Total (linear) orders**: A partial order $\leqslant_X$ is a **total order** if for all $x, y \in X$,

$$x \leqslant_X y \quad \text{or } y \leqslant_X x.$$

- **Strict orders**: Sometimes the strict relation $<_X$ is used instead of $\leqslant_X$. A strict order satisfies

  1. **Irreflexivity**: $x \not<_X x$
  2. **Transitivity**: $x <_X y$ and $y <_X z$ implies $x <_X z$

  Strict and non-strict orders are interdefinable:

  $$x <_X y \iff x \leqslant_X y \text{ and } x \neq y.$$

- **Number sets**: When we use the number sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \bar{\mathbb{Q}}, \mathbb{R}$, or $\mathbb{C}$ we are talking about that set equipped with the usual numerical order $\leqslant$. This is the standard total (linear) order.

- **Natural numbers partially ordered by divisibility**: Consider

  $$(\mathbb{N}, |).$$

  So,

  $$| \subseteq \mathbb{N} \times \mathbb{N}.$$

  If $(x, y) \in |$, for $x, y \in \mathbb{N}$, then

  $$x \mid y,$$

  or x "divides" y, so there exists $k \in \mathbb{N}$ such that

  $$y = kx.$$

  $x \mid x$ if there exits $k \in \mathbb{N}$ such that $x = kx$. Fortunately, $1 \in \mathbb{N}$, so $x = 1x = x$. Thus, divisibility on $\mathbb{N}$ is reflexive.

  Suppose that $x \mid y$, and $y \mid x$, then there exists $k, \ell \in \mathbb{N}$ such that

  $$y = kx, \quad \text{and } x = \ell y.$$

  So

  $$y = k\left(\ell y\right),$$

  which implies $k = \frac{1}{\ell}$, and $\ell = \frac{1}{k}$. But, $k, \ell \in \mathbb{N}$, so $k = \ell = 1$. Thus,

  $$x = 1y = y.$$

  So, divisibility on $\mathbb{N}$ is antisymmetric. Next, suppose $x \mid y$, and $y \mid z$. So, there exists $k, \ell \in \mathbb{N}$ such that

  $$y = kx, \quad \text{and } z = \ell y.$$

  We can say that $x \mid z$ if there exists $s \in \mathbb{N}$ such that $z = sx$. Since $y = kx$ and $z = \ell y$,

  $$z = \ell(kx).$$

  By commutativity of multiplication,

  $$z = (\ell k)x.$$

  And, since $\ell k \in \mathbb{N}$, $x \mid Z$. So, divisibility on $\mathbb{N}$ is transitive.

Therefore, $(\mathbb{N}, |)$ is partially ordered. But, consider $x, y \in \mathbb{N}$, where $x = 2\ell$, and $y = 2k + 1$ for $k, \ell \in \mathbb{N}$. Then, $x \mid y$ implies that

$$2\ell \mid 2k + 1.$$

Assume for the sake of contradiction that $2\ell \mid 2k + 1$. Then, there exist $m \in \mathbb{N}$ such that

$$2k + 1 = m(2\ell) = 2(m\ell).$$

But, notice that the $2k + 1$ is odd, while $2(m\ell)$ is even. Since an odd number cannot equal an even number, a contradiction. So, we found two numbers, $x, y \in \mathbb{N}$ that is not a member of the binary relation $|$. So, divisibility on $\mathbb{N}$ is not a total ordering.

- **Disjoint unions**: Let $\{A_i\}_{i \in I}$ be a family of sets, the notation

$$\bigsqcup A_i$$

means the union of the sets of A_i, with the additional information that they are considered pairwise disjoint. Formally, even if the sets overlap, the disjoint union treats them as distinct by tagging elements with their index.

The disjoint union is defined as

$$\bigsqcup_{i \in I} A_i := \bigcup_{i \in I} (A_i \times \{i\}).$$

So, an element is not just a set $a \in A_i$, but a pair (a, i). Formally, even if the sets overlap, the disjoint union treats them as distinct by tagging elements with their index.

For example, suppose $A_1 = \{1, 2\}$, and $A_2 = \{2, 3\}$. The ordinary union is

$$\bigcup_{i=1}^{2} A_i = A_1 \cup A_2 = \{1, 2, 3, \},$$

whereas the disjoint union is

$$\bigsqcup_{i=1}^{2} = A_1 \sqcup A_2 = \{(1, 1), (2, 1), (2, 2), (3, 2)\}.$$

5.2 Spaces

- **The product space**: The product space is the natural mathematical construction that allows us to treat pairs of elements as single objects, while retaining the structure of each component space.

  Let A and B be sets. The **product space** (or Cartesian product) is

  $$A \times B = \{(a, b) : \ a \in A, \ b \in B\}.$$

- **Product spaces and vector spaces**: If V and W are vector spaces over the same field $\mathbb{F}$, then

  $$V \times W$$

  is again a vector space, with operations defined componentwise, for $(v_1, w_1), (v_2, w_2) \in V \times W$,

  $$(v_1, w_1) + (v_2, w_2) = (v_1 + v_2, w_1 + w_2) \in V \times W.$$

  For $(v, w) \in V \times W$, and $\alpha \in \mathbb{F}$,

  $$\alpha(v, w) = (\alpha v, \alpha w) \in V \times W.$$

  Also, its dimension satisfies

  $$\dim (V \times W) = \dim (V) + \dim (W).$$

  For example,

  $$\mathbb{R}^2 \times \mathbb{R} \cong \mathbb{R}^3.$$

  Thus, there exists a linear structure preserving bijection

  $$\Phi : \ \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3, \quad \Phi((x, y), z) = (x, y, z).$$

  With inverse,

  $$\Phi^{-1}(x, y, z) = ((x, y), z).$$

- **Ambient spaces**: The ambient space is the larger space in which the objects under study naturally live. It provides the surrounding structure—coordinates, topology, metric, algebraic operations—needed to define and analyze those objects.

  An ambient space is not the object of primary interest; it is the context that makes the object meaningful. The object is typically a subset, subspace, or embedded structure inside the ambient space.

  Formally, if $X \subseteq A$, then A is the ambient space for X.

  A circle defined by $x^2 + y^2 = 1$ has ambient space $\mathbb{R}^2$. The circle itself is one-dimensional, but distances, angles, and curvature are defined using the surrounding plane.

  For a function $f : \ \mathbb{R}^n \to \mathbb{R}$, the graph

  $$\mathcal{G}(f) = \{(x, f(x)) : \ x \in \mathbb{R}^n\}$$

  has ambient space $\mathbb{R}^{n+1}$. Although the graph is n-dimensional, it lives inside a higher-dimensional space

A plane through the origin in $\mathbb{R}^3$ is a 2-dimensional subspace. Its ambient space is $\mathbb{R}^3$, which determines dot products, orthogonality, and projections.

- **Higher dimensional ambient spaces**: If an object is n-dimensional, then any space that can contain an embedding of it may serve as an ambient space.

  Formally, if

  $$X \hookrightarrow \mathbb{R}^{n+k},$$

  then $\mathbb{R}^{n+k}$ is a valid ambient space for X, for any $k \geqslant 1$. There is no rule forcing $k = 1$. However, the ambient space $\mathbb{R}^{n+1}$ is common because it is often the minimal convenient choice. A graph of a function $f : \mathbb{R}^n \to \mathbb{R}$ naturally lives in $\mathbb{R}^{n+1}$.

- **Codimension**: The number

  $$k = (\text{dimension of ambient space}) - (\text{dimension of object})$$

  is called the **codimension**. There is no upper bound on codimension in principle.

- **Morphism spaces**: Given two sets (or spaces) A and B, there are two fundamentally different kinds of objects you can consider:

  - **Points** in A or B
  - **Maps** from A to B

  A morphism space is the collection of all maps of a specified type between two objects.

  For two objects A and B,

  $$\mathrm{Hom}(A, B)$$

  denotes the set (or space) of all structure-preserving maps from A to B.

  Note: Hom means homomorphism, because a homomorphism is a map that preserves the relevant algebraic structure.

5.3 Functions

- **Image under a function**: Suppose A and B are sets, and

  $$f: A \to B$$

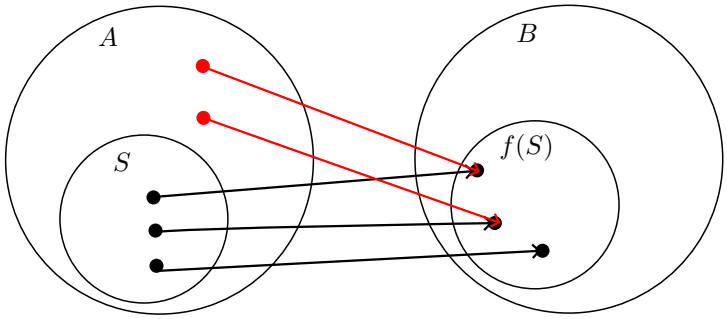
  is a function. For an element $a \in A$, the **image** of a under f is the element

  $$f(a) \in B.$$

  For a subset $S \subseteq A$, the **image** of S under f is

  $$f(S) := \{f(a) \in B : a \in S\}.$$

  So, the set of all outputs obtained by applying f to elements of S.



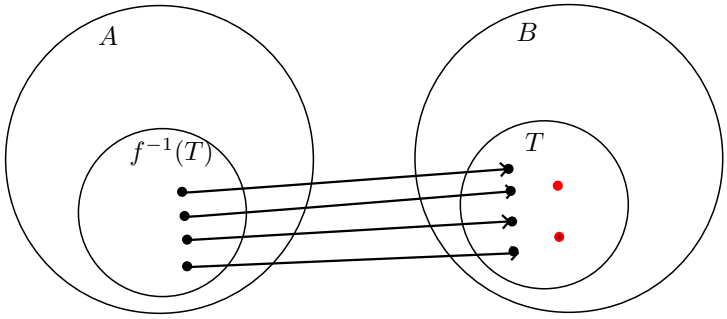
  Note: The red inputs may exist if f is not injective.

- **Preimage under a function**: For a subset $T \subseteq B$, the **preimage** of T under f is

  $$f^{-1}(T) := \{a \in A : f(a) \in T\}.$$

  For an element $b \in B$, the preimage is the set

  $$f^{-1}(\{b\}) := \{a \in A : f(a) = b\}.$$



Note: The red outputs may exist if f is not surjective. If f is not surjective, then it could be the case that some elements of T are not hit by f. In this case, not all elements of T have a preimage in A.

- **The image of the preimage, encoding surjectivity**: Suppose A and B are sets, and

$$f: A \to B$$

is a function from A to B. Let $T \subseteq B$. Then, the preimage of T under f is the set

$$f^{-1}(T) := \{a \in A: f(a) \in T\}.$$

Then, the image of the preimage of T, $f^{-1}(T)$, is the set

$$f(f^{-1}(T)) = \{f(a): a \in f^{-1}(T)\} = \{f(a): f(a) \in T\} \subseteq T.$$

If f is not surjective, and some elements of T do not have a preimage in A, then $f(f^{-1}(T)) \subset T$. Otherwise, f is surjective and $f(f^{-1}(T)) = T$.

So, $f(f^{-1}(T))$ removes the elements of T that are not hit by f. Thus, we can say that f is surjective if and only if for any subset $T \subseteq B$,

$$f(f^{-1}(T)) = T.$$

Otherwise, if $f(f^{-1}(T)) \subset T$, then f is not surjective. So, we say that $f(f^{-1}(T))$ encodes surjectivity.

- **The preimage of the image, encoding injectivity**: Suppose A and B are sets, and

$$f: A \to B$$

is a function from A to B. Let $S \subseteq A$. Then, the image of S under f is the set

$$f(S) = \{f(a) \in B: a \in S\}.$$

So, the preimage of $f(S)$ is the set

$$f^{-1}(f(S)) = \{a \in A: f(a) \in f(S)\} = \{a \in A: \exists s \in S \text{ with } f(a) = f(s)\}.$$

Now, notice that if f is injective, then all elements in $f(S)$ must have their preimage in S. So, if we take the preimage of the entirety of $f(S)$, then we get back S. Therefore, if f is injective,

$$f^{-1}(f(S)) = S.$$

However, suppose f is not injective. Consider some element $s \in S$ with image $f(s) \in f(S)$. It could be the case that there exists some $a \in A$, $a \notin S$ with $f(a) = f(s)$. Notice that this element $a \in A \setminus S$ then belongs to the preimage of $f(S)$, $f^{-1}(f(S))$.

Hence, if f is not injective, then

$$S \subseteq f^{-1}(f(S)).$$

So, $f^{-1}(f(S))$ encodes injectivity. $f^{-1}(f(S))$ is the set S, along with any additional points in A that have images in $f(S)$. These additional points only appear if f is not injective.

Therefore, for any subset $S \subseteq A$, f is injective if and only if

$$f^{-1}(f(S)) = S.$$

If $f^{-1}(f(S)) \supset S$ for any subset S of A, then f is not injective.

- **Properties of image and preimage**: For sets A, B, and a function $f : A \to B$, the following properties hold for any subset $S \subseteq A$ and $T \subseteq B$:

  1. $f(f^{-1}(T)) \subseteq T$
  2. $f^{-1}(f(S)) \supseteq S$

  Also, preimages preserve set operations,

  $$f^{-1}(T_1 \cup T_2) = f^{-1}(T_1) \cup f^{-1}(T_2),$$
  $$f^{-1}(T_1 \cap T_2) = f^{-1}(T_1) \cap f^{-1}(T_2).$$

- **Endofunction**: If X is a set, a function

  $$f : X \to X$$

  is called an **endofunction on** X

  Note: If f is a bijection, then we can say f is a **bijection on** X, whereas if the codomain is some other set Y, we would say that f is a bijection from X to Y.

- **Bijective endofunction**: If X is a set, a function

  $$f : X \to X$$

  that is bijective is called a permutation of X. The set of all permutations of X forms the symmetric group on X, denoted $\mathrm{Sym}(X)$. If $X = \{1, \ldots, n\}$, then this group is S_n.

  If f is a bijection on X, then f is a member of $\mathrm{Sym}(X)$.

- **Bijection**: A function

  $$f : X \to Y$$

  that is both surjective and injective (bijective) is called a **bijection** from X to Y.

- **Surjection**: A function

  $$f : X \to Y$$

  that is surjective is called a **surjection** from X to Y.

- **Injection**: A function

  $$f : X \to Y$$

  that is injective is called an **injection** from X to Y.

- **Structure preserving**: "Structure-preserving" means that the map respects the operations and axioms that define the object you are working with. The phrase is intentionally generic; its precise meaning depends on what structure the set carries.

A structure on a set is:

- A collection of operations (e.g., addition, scalar multiplication),
- Possibly distinguished elements (e.g., zero),
- And axioms relating them.

For a vector space V, this structure is

- Vector addition
- Scalar multiplication by elements of a field $\mathbb{F}$
- The vector space axioms.

A map preserves structure if doing the operation before or after applying the map gives the same result

$$\text{Apply operation} \leftrightarrow \text{apply map}$$

The map "commutes" with the operations.

- **Morphisms**: A morphism is a structure-preserving map between objects of the same type.

  A morphism

  $$f : X \to Y$$

  must preserve the structure of X as expressed inside Y. So, f is a morphism if it preserves the structure **carried by** X in a way that is compatible with the **structure on** Y.

  Performing an operation in X, then applying f, gives the same result as first applying f and then performing the corresponding operation in Y.

  The structure being preserved is the type of structure common to both X and Y.

  A morphism is defined relative to a specified structure. If a map preserves some operations but not all of the operations that define a structure, then the map is **not** a morphism of that structure. However, it may still be a morphism of a reduced structure obtained by forgetting some operations.

- **Homomorphism**: A homomorphism is a structure-preserving map between algebraic objects of the same type.

- **Linear maps over vector spaces**: Let

  $$T : V \to W$$

  be a linear map, where V and W are vector spaces over a field $\mathbb{F}$. The structure of a vector space consists of vector addition and scalar multiplication. A map preserves structure if applying a vector space operation before the map yields the same result as applying the map after the operation.

Let $v_1, v_2 \in V$. Since

$$T(v_1 + v_2) = T(v_1) + T(v_2),$$

the map T preserves vector addition. Moreover, for $\alpha \in \mathbb{F}$ and $v \in V$,

$$T(\alpha v) = \alpha T(v),$$

so T also preserves scalar multiplication. Therefore, linear maps are morphisms of vector spaces.

- **Endomorphism**: If X is a set equipped with some structure, for example a vector space, then a map

  $$f : X \to X$$

  that preserves the structure of X is called an endomorphism. A linear operator $T : V \to V$ is an endomorphism of X

- **Isomorphism**: If X and Y sets with some structure, and

  $$f : X \to Y$$

  is a bijective morphism whose inverse is also a morphism, then f is called an **isomorphism**, or **isomorphic**.

  This means that X and Y have the same structure

  Note: If two spaces V, and W are isomorphic, then we can use the symbol $\cong$,

  $$V \cong W$$

  to signify that there exists an isomorphism between the two spaces.

- **Automorphism**: If

  $$f : X \to X$$

  is a bijective endomorphism, the f is an **automorphism**. For example, an invertible linear operator.

- **Summary, self maps (endofunctions)**: If X is a set, and

  $$f : X \to X$$

  is an **endofunction** of X, then f is an **endomorphism** of X if f preserves the structure of X, and an **automorphism** of X if f is bijective and structure preserving (endomorphism).

  So,

  $$\text{automorphisms } \subset \text{ endomorphisms } \subset \text{ endofunctions.}$$

- **Non-self maps**: Suppose X and Y are sets, and

  $$f : X \to Y$$

  is a function from X to Y. If f preserves

- **Homogeneous function**: A homogeneous function is a function whose value scales in a predictable way when its input is scaled.

Let V and W be vector spaces over a field $\mathbb{F}$, a function

$$f : V \to W$$

is called **homogeneous of degree** k, or **degree-k homogeneous** if, for all $v \in V$, and $\alpha \in \mathbb{F}$,

$$f(\alpha v) = \alpha^k f(v).$$

For example,

$$f(\alpha v) = \alpha f(v)$$

is a function homogeneous of degree 1. This is often called positively homogeneous or 1-homogeneous. If

$$f(\alpha v) = f(v),$$

this function is degree-0 homogeneous. The function is invariant under scaling.

Every linear map $T : V \to W$ is degree-1 homogeneous. The determinant is degree-n homogeneous, for $A \in \mathbb{R}^{n \times n}$, since

$$\det(\alpha A) = \alpha^n \det(A).$$

- **Monotone (order-preserving) functions**: Let $(X, \leqslant_X)$ and $(Y, \leqslant_Y)$ be ordered sets. A function

$$f : X \to Y$$

is order-preserving (or monotone increasing) if

$$x_1 \leqslant_X x_2 \implies f(x_1) \leqslant_Y f(x_2).$$

Monotone increasing functions respect the ordering. Larger inputs do not map to smaller outputs.

For example, suppose $(X, \leqslant)$, $(Y, \leqslant)$, and

$$f : X \to Y \quad f(x) = ax$$

for $a \neq 0$. Let $x_1, x_2 \in X$, with $x_1 \leqslant x_2$. Then, $f(x_1) = ax_1$, and $f(x_2) = ax_2$. So,

$$ax_1 \leqslant ax_2 \iff x_1 \leqslant x_2.$$

So, f is monotone increasing.

- **Antitone (Order reversing) functions**: A function

$$f : X \to X$$

is order-reversing (antitone) if

$$x_1 \leqslant_X x_2 \implies f(x_1) \geqslant_Y f(x_2).$$

The function flips the order, larger inputs map to smaller outputs. For example, $f(x) = -x$ on $\mathbb{R}$.

- **Idempotent**: A function

$$f : X \to X$$

is idempotent if

$$f \circ f = f \quad \text{i.e.} \quad f(f(x)) = f(x) \text{ for all } x \in X.$$

Applying the function **more than once has no additional effect**. For example, projection onto a subspace $P^2 = P$. Idempotent maps often encode **projection or stabilization**.

- **Identity map**: Let M be a map,

$$M : X \to X.$$

The **identity map** id for the underlying set (X) is the map

$$\text{id } X \to X, \quad \text{id}(x) = x$$

for all $x \in X$.

- **Zero map**: Let X be a set and Y be a set equipped with a distinguished element $0 \in Y$. The zero map

$$Z : X \to Y$$

is defined by

$$Z(x) = 0 \quad \text{for all } x \in X.$$

- **Involution**: A function

$$f : X \to X$$

is an **involution** if

$$f \circ f = \text{id}_X.$$

The function is its own inverse. Matrix transposition is an involution. Recall the transpose map

$$T : M_{n\times n}(\mathbb{F}) \to M_{n\times n}(\mathbb{F}), \quad T(A) = A^T.$$

Consider $T(T(A))$,

$$T(T(A)) = T(A^T) = (A^T)^T = A.$$

Thus, $T \circ T = \text{id}$, so T is an involution.

- **Nilpotent**: A function

$$f : X \to X$$

is nilpotent if there exists $k \geqslant 1$ such that

$$f^k = 0,$$

where 0 is the zero map. Repeated application of the map **kills everything**.

- **A function and its graph**: A function is a triple

$$f:\ A \to B$$

consisting of

  - A domain A
  - A codomain B
  - A rule assigning to each $a \in A$ a **single element** $f(a) \in B$

A function is not a set of points in space, it lives in a morphism space, not in the ambient space.

For example,

$$f(x, y) = x^2 + y^2$$

is a rule from $\mathbb{R}^2$ to $\mathbb{R}$. The **graph** of f is a **subset** of the product space $A \times B$

$$\text{Graph}(f) = \{(a, b) \in A \times B :\ b = f(a)\}.$$

For $f:\ \mathbb{R}^2 \to \mathbb{R}$,

$$\text{Graph}(f) = \{(x, y, z) \in \mathbb{R}^3 :\ z = f(x, y)\}.$$

This is the geometric object. A function by itself has no geometry until you embed it into a set of points. The graph is such an embedding, the embedding into the product space.

So, a function lives in a space of maps, the graph lives in the product space.

5.4 Sums and products

- **Bijective reindexing**: Let A be a finite set, and

  $$f : A \to \mathbb{R}$$

  be any function. If

  $$\phi : A \to A$$

  is a bijection, then

  $$\sum_{a \in A} f(a) = \sum_{a \in A} f(\phi(a)).$$

  Since a bijective endofunction over A permutes the elements of A, and the outputs of f are commutative, the sums are equal.

Numerical Linear Algebra

6.1 Introduction

- **Matrix Notation**: For a matrix $A \in \mathbb{R}^{m \times n}$, we say

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

  with $a_{ij} \in \mathbb{R}$.

- **Vector notation**: For a vector $x \in \mathbb{R}^n$ (or $\mathbb{R}^{n \times 1}$), we have

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

  for $x_i \in \mathbb{R}$.

- **Submatrix notation (rows)**:

$$A(i,:) \in \mathbb{R}^{1 \times n} \iff A(i,:) = \begin{bmatrix} a_{i1} & a_{i2} & \cdots & a_{in} \end{bmatrix}.$$

- **Submatrix notation (columns)**:

$$A(:,j) \in \mathbb{R}^{m \times 1} \iff A(:,j) = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}.$$

- **Sparse Matrix**: A sparse matrix or sparse array is a matrix in which most of the elements are zero. There is no strict definition regarding the proportion of zero-value elements for a matrix to qualify as sparse but a common criterion is that the number of non-zero elements is roughly equal to the number of rows or columns.

- **Dense Matrix**: if most of the elements are non-zero, the matrix is considered dense

- **Sparsity**: The number of zero-valued elements divided by the total number of elements is sometimes referred to as the sparsity of the matrix.

- **Band Matrix**: a band matrix or banded matrix is a sparse matrix whose non-zero entries are confined to a diagonal band, comprising the main diagonal and zero or more diagonals on either side.

$$A(i_1:i_2,:) \in \mathbb{R}^{(i_2-i_1+1)\times n} \iff A(i_1:i_2,:) = \begin{bmatrix} a_{i_1 1} & a_{i_1 2} & \cdots & a_{i_1 n} \\ a_{i_1+1,1} & a_{i_1+1,2} & \cdots & a_{i_1+1,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_2 1} & a_{i_2 2} & \cdots & a_{i_2 n} \end{bmatrix}.$$

$$A(:,j_1:j_2) \in \mathbb{R}^{m\times(j_2-j_1+1)} \iff A(:,j_1:j_2) = \begin{bmatrix} a_{1j_1} & a_{1,j_1+1} & \cdots & a_{1j_2} \\ a_{2j_1} & a_{2,j_1+1} & \cdots & a_{2j_2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{mj_1} & a_{m,j_1+1} & \cdots & a_{mj_2} \end{bmatrix}.$$

Where

$A(i_1:i_2,:)$: all rows between i_1 and i_2, across all columns,

$A(:,j_1:j_2)$: all columns between j_1 and j_2, across all rows.

- **Transposition**: $\mathbb{R}^{m\times n} \to \mathbb{R}^{n\times m}$

$$C = A^\top \iff c_{ij} = a_{ji}.$$

- **Transpose map**: The map

$$T:\ M_{n\times n}(\mathbb{F}) \to M_{n\times n}(\mathbb{F}), \quad T(A) = A^T.$$

- **Addition** $(\mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n})$

$$C = A + B \implies c_{ij} = a_{ij} + b_{ij}.$$

- **Scalar-matrix Multiplication:** $(\mathbb{R} \times \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n})$

$$C = \alpha A \implies c_{ij} = \alpha a_{ij}.$$

- **Matrix-matrix Multiplication:** $(\mathbb{R}^{m\times p} \times \mathbb{R}^{p\times n} \to \mathbb{R}^{m\times n})$

$$C = AB \implies c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}.$$

- **Matrix-vector Multiplication:** $(\mathbb{R}^{m\times n} \times \mathbb{R}^{n} \to \mathbb{R}^{m})$

$$y = Ax \implies y_i = \sum_{j=1}^{n} a_{ij} x_j.$$

- **Inner product (or dot product):** $(\mathbb{R}^{n} \times \mathbb{R}^{n} \to \mathbb{R})$

$$c = x^T y \implies c = \sum_{i=1}^{n} x_i y_i.$$

- **Outer product:** $(\mathbb{R}^{m} \times \mathbb{R}^{n} \to \mathbb{R}^{m\times n})$

$$C = xy^T \implies c_{ij} = x_i y_j.$$

- **Flops**: A flop is a floating-point operation between numbers stored in a floating-point format on a computer.

If x and y are numbers stored in a floating point format, then the following operations are each one flop

$$x+y \quad x-y \quad xy \quad x/y.$$

- **Empty sum**: In standard mathematical convention, if the lower bound exceeds the upper bound, the sum is defined to be zero

$$\sum_{i=k}^{j} f(k) = 0 \quad \text{if } i > j$$

- **Conformable matrix**: Simply a matrix that has the right dimensions to partcipate
- **Relationship between the sign of a 2-degree polynomial and its discriminant**: Recall that for a degree two polynomial, $p(x) \in P_2$, $p(x) = ax^2 + bx + c$, the discriminant is given by

$$D = b^2 - 4ac.$$

  - $p(x) \geqslant 0$ for all x if and only if $a > 0$ $D \leqslant 0$
  - $p(x) \leqslant 0$ for all x if and only if $a < 0$ $D \leqslant 0$

- **Useful fact I**: Let $a, b \in \mathbb{R} \setminus \{0\}$, if $a \geqslant b$, then

$$\frac{1}{a} \leqslant \frac{1}{b}.$$

- **Useful fact II**: Let $a, b, c, d \in \mathbb{R}$, if $0 \leqslant a \leqslant b$, and $0 < d \leqslant c$, then

$$\frac{a}{c} \leqslant \frac{b}{d}.$$

  Proof. Assume $a, b, c, d \in \mathbb{R}$, with $0 \leqslant a \leqslant b$, and $0 < d \leqslant c$. We have

$$a \leqslant b \implies \frac{a}{c} \leqslant \frac{a}{c}.$$

  Since $c \geqslant d$, $\frac{1}{c} \leqslant \frac{1}{d}$, which implies that

$$\frac{b}{c} \leqslant \frac{b}{d}.$$

  Combining these two facts gives

$$\frac{a}{c} \leqslant \frac{b}{c} \leqslant \frac{b}{d}.$$

  Therefore, $\frac{a}{c} \leqslant \frac{b}{d}$ ■

- **Transposition of the product of n matrices**: Let $A_1, A_2, ..., A_{n-2}, A_{n-1}, A_n \in \mathbb{R}^{n \times n}$, then

$$(A_1 A_2 \cdots A_{n-2} A_{n-1} A_n)^T = A_n^T A_{n-1}^T A_{n-2}^T \cdots A_2^T A_1^T.$$

  Proof.

$$\begin{aligned}
(A_1 A_2 \cdots A_{n-2} A_{n-1} A_n)^T &= A_n^T (A_1 A_2 \cdots A_{n-2} A_{n-1})^T \\
&= A_n^T A_{n-1}^T (A_1 A_2 \cdots A_{n-2} A_{n-1})^T \\
&= A_n^T A_{n-1}^T A_{n-2}^T (A_1 A_2 \cdots)^T \\
&= \dots \\
&= A_n^T A_{n-1}^T A_{n-2}^T \cdots A_2^T A_1^T.
\end{aligned}$$

  ■

- **What does it mean to multiply by** $\cos(\theta)$ **or** $\sin(\theta)$: From Euclidean geometry, if a vector

$$v = \begin{pmatrix} x \\ y \end{pmatrix}$$

is measured from the origin $(0, 0)$, then

$$\cos(\theta) = \frac{x}{\|v\|_2},$$
$$\sin(\theta) = \frac{y}{\|v\|_2}.$$

Thus, $\|v\| \cos(\theta) = x$ gives the **component of** v **in the direction of** $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ (the x-axis), and $\|v\| \sin(\theta) = y$ gives the **component of** v **in the direction of** $e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ (the y-axis).

- **The dot product**: For $v, u \in \mathbb{R}^n$, the dot product $v^T u$ measures how much of u lies in the direction of v. Recall the projection formula,

$$\text{proj}_v(u) = \left(\frac{v^T u}{v^T v}\right) v = \left(\frac{v^T u}{\|v\|_2^2}\right) v = \frac{v^T u}{\|v\|_2} \cdot \frac{v}{\|v\|_2}.$$

The scalar $(v^T u)/\|v\|_2 \in \mathbb{R}$ measures the length of u's projection onto v, and the vector $v/\|v\|_2$ is a vector with length one that points in the direction of v.

Thus, $v^T u$ measures how much of u aligns with v, scaled by the length of v ($\|v\|_2$).

We also have that

$$v^T u = \|v\| \, \|u\| \cos(\theta),$$

where θ is the angle between u and v. Notice that above we have that

$$\frac{v^T u}{\|v\|_2}$$

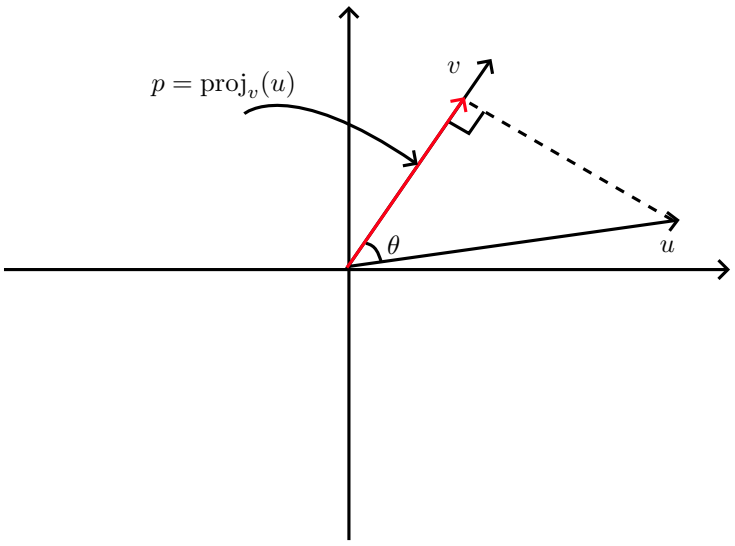
is the length of the projection of u onto v. So,

$$\frac{v^T u}{\|v\|_2} = \|u\|_2 \cos(\theta).$$

This tells us that the length of the projection of u onto v is the length of $\|u\|$ multiplied by the cosine of the angle between u and v.

As you can see we dropped a perpendicular from u down to v. Using the definition of cosine in right triangles,

$$\cos(\theta) = \frac{\text{projection length}}{\|u\|_2}$$
$$\implies \quad \text{projection length} \; = \|u\|_2 \cos(\theta).$$



Using the fact that $v^T u = u^T v = \|v\| \|u\| \cos(\theta)$, we see that the dot product depends on three things

1. The length of v
2. The length of u
3. The cosine of the angle between them.

Let $v \neq 0$ be fixed, and let $u = t\hat{u}$, where $\hat{u}$ is in the direction of u with $\|\hat{u}\| = 1$. Let ϕ be the fixed angle between v and $\hat{u}$. So,

$$v^T u = \|v\| \, t \cos(\phi).$$

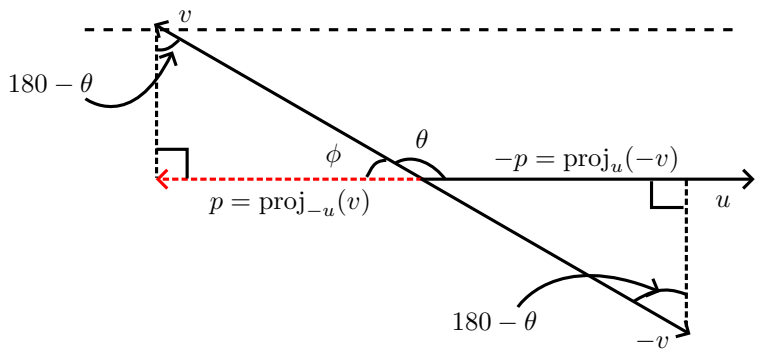
1. As $t \to \infty$, $\|u\|_2 \to \infty$, and $v^T u \to +\infty$ if $\cos(\phi) > 0$, $\to -\infty$ if $\cos(\phi) < 0$, and stays zero if $\cos(\phi) = 0$
2. As $t \to -\infty$, $\|u\|_2 \to \infty$, and $v^T u \to -\infty$ if $\cos(\phi) > 0$, $\to +\infty$ if $\cos(\phi) < 0$, and stays zero if $\cos(\phi) = 0$
3. As $\|u\|_2 = |t| \to 0$, $v^T u \to 0$

Now, fix the length of u, but let the direction vary,

1. If $\theta \in \{0, 2\pi\}$, $v^T u = \|v\|_2 \|u\|_2$
2. As θ goes from zero to π, $v^T u$ goes from $+\|v\|_2 \|u\|_2$ to $-\|v\|_2 \|u\|_2$, crossing 0 at $\theta = \frac{\pi}{2}$
3. As θ goes from π to 2π, $v^T u$ goes from $-\|v\|_2 \|u\|_2$ to $+\|v\|_2 \|u\|_2$, crossing 0 at $\theta = \frac{3\pi}{2}$

Note: If $v^T u < 0$, then $\cos(\theta) < 0$, meaning the angle between v and u is obtuse; greater than 90°

That tells you that the projection of one vector onto the other points opposite to the other's direction. In other words, moving along u decreases your progress in the direction of v.



First, we notice that $\phi = 180 - \theta$, so

$$\cos(180-\theta) = \cos(\phi) = \cos(180)\cos(\theta) + \sin(180)\sin(\theta) = -\cos(\theta).$$

We will look at $\text{proj}_u(v)$, $\text{proj}_{-u}(v)$, and $\text{proj}_u(-v)$. First,

$$\text{proj}_u(v) = \frac{v^T u}{\|u\|_2}\frac{u}{\|u\|_2} = \|v\|_2 \cos(\theta)\frac{u}{\|u\|_2}.$$

Next,

$$\begin{aligned}\text{proj}_{-u}(v) &= \frac{v^T(-u)}{\|-u\|_2}\frac{-u}{\|-u\|_2} = \frac{v^T u}{\|u\|_2}\frac{u}{\|u\|_2} = \|v\|_2\cos(\phi)\frac{-u}{\|u\|_2} \\ &= -\|v\|_2\cos(\theta)\frac{-u}{\|u\|_2} = \text{proj}_u(v).\end{aligned}$$

Last, we have that

$$\begin{aligned}\text{proj}_u(-v) &= \frac{-v^T u}{\|u\|_2}\frac{u}{\|u\|_2} = \|-v\|_2\cos(\phi)\frac{u}{\|u\|_2} \\ &= -\|v\|_2\cos(\theta)\frac{u}{\|u\|_2} = -\text{proj}_u(v).\end{aligned}$$

So,

$$\text{proj}_u(v) = \text{proj}_{-u}(v) = -\text{proj}_{-u}(v).$$

Suppose that

$$v = \begin{pmatrix} x \\ y \end{pmatrix}, \quad e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \ (x\text{-axis}).$$

Then

$$v^T e_1 = e_1^T v = x.$$

Thus, $v^T e_1$ gives the **component of v in the direction of** e_1, i.e., the amount of v that lies along the x-axis. Similarly, if

$$e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \ (y\text{-axis}),$$

then

$$v^T e_2 = e_2^T v = y.$$

Hence, $v^T e_2$ gives the **component of v in the direction of** e_2, i.e., the amount of v that lies along the y-axis.

- **Dot product and projection of vectors on the same line**: Suppose $a, x \in \mathbb{R}^n$ are on the same line. Then, $x = ka$, and

$$\frac{x}{\|x\|_2} = \frac{ka}{\|ka\|_2} = \frac{k}{|k|}\frac{a}{\|a\|_2} = \text{sgn}(k)\frac{a}{\|a\|_2}.$$

If x, a point in opposite directions, then $\text{sgn}(k) = -1$, and

$$a^T x = \|a\|_2 \|x\|_2 (-1) = -\|a\|_2 \|x\|_2, \quad \frac{x}{\|x\|_2} = -\frac{a}{\|a\|_2}.$$

So, the projection is

$$\text{proj}_a(x) = \frac{a^T x}{\|a\|_2}\frac{a}{\|a\|_2} = \frac{-\|a\|_2 \|x\|_2}{\|a\|_2}\left(-\frac{x}{\|x\|_2}\right) = x.$$

If x, a point in the same direction, $\text{sgn}(k) = 1$, and

$$a^T x = \|a\|_2 \|x\|_2 (1) = \|a\|_2 \|x\|_2, \quad \frac{x}{\|x\|_2} = \frac{a}{\|a\|_2}.$$

The projection in this case is

$$\text{proj}_a(x) = \frac{a^T x}{\|a\|_2}\frac{a}{\|a\|_2} = \frac{\|a\|_2 \|x\|_2}{\|a\|_2}\frac{x}{\|x\|_2} = x.$$

Hence, the projection is the same in both cases. In the raw dot product, notice that

$$a^T x = \|a\|_2 \|x\|_2 \cos(\theta) = \pm \|a\|_2 \|x\|_2.$$

So, if a and x lie on the same line, then the projection of x onto a has length $\|x\|_2$. If they are in opposite directions, $a^T x < 0$, and if they point in the same direction, $a^T x > 0$.

- **Describing dot product in words**: Suppose that $a, x \in \mathbb{R}^n$. The quantity $a^T x$ measures how much of x lies in the direction of a. If $a^T x = b$, then the projection of x onto a has length $\dfrac{b}{\|a\|}$.
  - If x is *orthogonal* to a, then $a^T x = 0$, so the projection of x onto a is the zero vector.
  - If $a^T x > 0$, then x has a component in the same direction as a.
  - If $a^T x < 0$, then x has a component in the direction opposite to a.

6.2 Gaussian Elimination and its variants (1)

6.2.1 Matrix Multiplication

- **Matrix Multiplication**: In general, if A is a real matrix with m rows and n columns, and x is a real vector with n entries, then

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n} \quad \text{and} \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n.$$

  If $b = Ax$, then $b \in \mathbb{R}^m$ and

$$b_i = \sum_{j=1}^{n} a_{ij}x_j = a_{i1}x_1 + \cdots + a_{in}x_n, \quad i = 1, \ldots, m.$$

  Thus, b_i is the **inner-product** between the i-row of A,

$$A(i,:) = [a_{i1} \ \cdots \ a_{in}], \quad (i = 1, \ldots, m)$$

  and the vector x.

  Also,

$$b = A(:,1)x_1 + \cdots + A(:,n)x_n,$$

  so b is a **linear combination** of the columns of A, i.e.,

$$A(:,j) = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}, \quad j = 1, \ldots, n.$$

- **Matrix-Matrix Multiplication**: Let $A \in \mathbb{R}^{m\times n}$ and $X \in \mathbb{R}^{n\times p}$.

  If $B = AX$ then $B \in \mathbb{R}^{m\times p}$ and

$$b_{ij} = \sum_{k=1}^{n} a_{ik}x_{kj}, \quad i = 1, \ldots, m, \ \ j = 1, \ldots, p.$$

  That is, b_{ij} is the inner-product between row i of A and column j of X.

  Also, each column of B is a linear combination of the columns of A.

  Total flops required for matrix multiplication is

$$\sum_{i=1}^{m}\sum_{j=1}^{p}\sum_{k=1}^{n} 2 = 2mnp.$$

  If $A, X \in \mathbb{R}^{n\times n}$, then computing $B = AX$ requires $2n^3 = O(n^3)$ flops.

We can see this by describing the algorithm for Matrix-Matrix multiplication

```
for i = 1:m
    for j = 1:n
        for k = 1:p
            C[i,j] += A[i,k]B[k,j]
        end
    end
end
```

The multiplication $A[i,j]B[k,j]$ is one flop, followed by the addition. Therefore, two flops per iteration of the innermost loop.

- **Block Matrices**: Partition $A \in \mathbb{R}^{m\times n}$ and $X \in \mathbb{R}^{n\times p}$ into blocks:

$$A = \begin{array}{c} \\ m_1 \\ m_2 \end{array} \begin{array}{c} \begin{array}{cc} n_1 & n_2 \end{array} \\ \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \end{array}, \qquad X = \begin{array}{c} \\ n_1 \\ n_2 \end{array} \begin{array}{c} \begin{array}{cc} p_1 & p_2 \end{array} \\ \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \end{array}$$

where $n = n_1 + n_2$, $m = m_1 + m_2$, and $p = p_1 + p_2$.

If $B = AX$, and

$$B = \begin{array}{c} \\ m_1 \\ m_2 \end{array} \begin{array}{c} \begin{array}{cc} p_1 & p_2 \end{array} \\ \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \end{array},$$

then

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = B = AX = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11}X_{11} + A_{12}X_{21} & A_{11}X_{12} + A_{12}X_{22} \\ A_{21}X_{11} + A_{22}X_{21} & A_{21}X_{12} + A_{22}X_{22} \end{bmatrix}$$

That is,

$$B_{ij} = \sum_{k=1}^{2} A_{ik}X_{kj}, \qquad i,j = 1,2.$$

- **Transpose of block matrices**: Let $A \in \mathbb{R}^{n\times n}$, with

$$A = \begin{bmatrix} A_{11} & a_{12} \\ A_{21} & a_{22} \end{bmatrix}.$$

Then,

$$A^\top = \begin{bmatrix} A_{11}^\top & A_{21}^\top \\ A_{12}^\top & A_{22}^\top \end{bmatrix}$$

- **Transpose of a block vector**: Similarly, if $x \in \mathbb{R}^n$ is decomposed into blocks

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

with $x_1 \in \mathbb{R}^{n_1},\ x_2 \in \mathbb{R}^{n_2},\ n = n_1 + n_2$, then

$$x^\top = \begin{pmatrix} x_1^\top & x_2^\top \end{pmatrix}$$

6.2.2 Systems of Linear Equations

- **Systems of linear equations**: Let $A \in \mathbb{R}^{n \times n},\ b \in \mathbb{R}^m$, our goal is to find $x \in \mathbb{R}^m$ such that $Ax = b$
- **Singularity**: A **singular matrix** is a square matrix that does not have an inverse.

  A **nonsingular** matrix is a square matrix that does have an inverse.

  The following are equivalent, if any one holds, they all hold

  - $Ax = b$ has a unique solution
  - $\det(A) \neq 0$
  - A^{-1} exists
  - There is no nonzero vector $y \in \mathbb{R}^m$ such that $Ay = 0$
  - The columns of A are linearly independent
  - The rows of A are linearly independent
  - Given any vector b, there is exactly one vector x such that $Ax = b$

  If any one of the following are true, they all are true, and A is nonsingular
- **Solution to** $Ax = b$: If A is nonsingular, then A^{-1} exists, and

  $$x = A^{-1}b.$$

  Which is the unique solution to $Ax = b$

  Note: Practically, it is not wise to compute A^{-1}, as this can be expensive.

6.2.3 Triangular systems

- **Upper triangular**: A square matrix $A = U$ of the form

$$A = U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

  is called **upper triangular**.

  Formally, a matrix A is upper triangular if $a_{ij} = 0$ whenever $i > j$

- **Lower triangular**: A square matrix $A = L$

$$A = L = \begin{bmatrix} \ell_{11} & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nn} \end{bmatrix}$$

  is called **lower triangular**

  Formally, a matrix A is lower triangular if $a_{ij} = 0$ whenever $i < j$

- **Theorem 1.3.1**: Let A be a triangular matrix. Then, A is nonsingular if and only if $g_{ii} \neq 0$ for $i = 1, 2, ..., n$

- **Solutions to triangular systems**: Consider the system

$$\begin{bmatrix} \ell_{11} & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

  So,

$$\begin{aligned} \ell_{11}x_1 &= b_1 \\ \ell_{21}x_1 + \ell_{22}x_2 &= b_2 \\ &\vdots \\ \ell_{n1}x_1 + \ell_{n2}x_2 + \cdots + \ell_{nn}x_n &= b_n. \end{aligned}$$

  Then,

$$x_1 = \frac{b_1}{\ell_{11}}$$

  and,

$$\begin{aligned} \ell_{22}x_2 &= b_2 - \ell_{21}x_1 \\ \implies x_2 &= \frac{b_2 - \ell_{21}x_1}{\ell_{22}}. \end{aligned}$$

  In general, we have

$$x_i = \frac{b_i - \sum_{j=1}^{i-1} \ell_{ij}x_j}{\ell_{ii}}$$

  for $i = 1, 2, ..., n$. This method is called **Forward Substitution**.

A similar process is used on upper triangular matrices and is called **Backward Substitution**.

- **Counting flops for the forward substitution method**: We have

```
for i = 1:n
    for j=1:i-1
        b[i] = b[i] - ell[i,j]b[j]
    end
    b[i] = b[i] / ell[i,i]
end
```

Thus, the count of flops is

$$\begin{aligned} n + \sum_{i=1}^{n} 2(i-1) &= n + 2\sum_{i=1}^{n}(i-1) = n + 2\left(\sum_{i=1}^{n} i - \sum_{i=1}^{n} 1\right) \\ &= n + 2\left(\sum_{i=1}^{n} i - n\right) = n + 2\left(\frac{n(n+1)}{2} - n\right) \\ &= n + n^2 - n = n^2 \end{aligned}$$

So, forward substitution is $\mathcal{O}(n^2)$

- **Column oriented forward substitution**: Suppose we have $Lx = b$ when L is lower triangular, we split the matrix into the following blocks

$$\begin{bmatrix} \ell_{11} & 0 \\ \hat{\ell} & \hat{L} \end{bmatrix} \begin{bmatrix} x_1 \\ \hat{x} \end{bmatrix} = \begin{bmatrix} b_1 \\ \hat{b} \end{bmatrix}.$$

With $\hat{\ell} \in \mathbb{R}^{n-1}$, $\hat{L} \in \mathbb{R}^{n-1 \times n-1}$, $\hat{x} \in \mathbb{R}^{n-1}$, $\ell_{11}, x_1, b_1 \in \mathbb{R}$. Note that $\hat{L}$ is also lower triangular.

We have

$$\begin{aligned} \ell_{11}x_1 = b_1 &\implies x_1 = \frac{b_1}{\ell_{11}} \\ \hat{\ell}x_1 + \hat{L}\hat{x} = \hat{b} &\implies \hat{L}\hat{x} = \hat{b} - \hat{\ell}x_1 \end{aligned}$$

Thus, we reduced the dimension by one. We repeat this process for the remaining x_i. The process is

1. Compute $x_1 = \frac{b_1}{\ell_{11}}$
2. Compute $\hat{b} - \hat{\ell}x_1 = \tilde{b} \in \mathbb{R}^{n-1}$
3. Find $\hat{L}x = \tilde{b}$

- **Counting flops for column oriented forward substitution**: Let f_n be the flop count, we have

$$f_n = 1 + 2(n-1) + f_{n-1}.$$

With

$$\begin{aligned} f_{n-1} &= 1 + 2(n-2) + f_{n-2}, \\ f_{n-2} &= 1 + 2(n-3) + f_{n-3}, \\ &\vdots \end{aligned}$$

Until

$$f_{n-(n-1)} = f_1 = 1 + 2((n-(n-1))-1) + f_0 = 1 + 2(0) + f_0 = 1$$

with $f_0 = 0$

So,

$$\begin{aligned} f_n &= 1 + 2(n-1) + 1 + 2(n-2) + ... + 1 + 2((n-(n-1))-1) \\ &= \sum_{i=1}^{n} 1 + 2(n-1) = n + 2n^2 - 2\sum_{i=1}^{n} i \\ &= ... = n^2. \end{aligned}$$

Thus, column oriented forward substitution is also $\mathcal{O}(n^2)$

- **Solutions of an upper triangular system (Backward substitution)**: Let $A \in \mathbb{R}^{n\times n},\ x \in \mathbb{R}^n,\ b \in \mathbb{R}^n$, with

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Then,

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

implies

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{nn}x_n &= b_n. \end{aligned}$$

So,

$$\begin{aligned} x_1 &= \frac{b_1 - (a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n)}{a_{11}}, \\ x_2 &= \frac{b_2 - (a_{23}x_3 + a_{24}x_4 + \cdots + a_{2n}x_n)}{a_{22}}, \\ x_n &= \frac{b_n}{a_{nn}}. \end{aligned}$$

In general, we have that

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}}, \quad i = n, n-1, ..., 1$$

- **Column-oriented backward substitution**: Let $U \in \mathbb{R}^{n\times n}$ be upper triangular, $x \in \mathbb{R}^n$, and $b \in \mathbb{R}^n$ which gives the system

$$\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Split the system into the following block decomposition

$$\begin{bmatrix} \hat{U} & u \\ 0^\top & u_{nn} \end{bmatrix} \begin{bmatrix} \hat{x} \\ x_n \end{bmatrix} = \begin{bmatrix} \hat{b} \\ b_n \end{bmatrix}$$

Then,

$$\hat{U}\hat{x} + ux_n = \hat{b} \implies \hat{U}\hat{x} = \hat{b} - ux_n = \tilde{b},$$
$$u_{nn}x_n = b_n \implies x_n = \frac{b_n}{u_{nn}}$$

Thus, the column-oriented backward substitution algorithm is defined by the following steps

1. Compute $x_n = \frac{b_n}{u_{nn}}$
2. Compute $\tilde{b} = \hat{b} - ux_n$
3. Run the algorithm on $\hat{U}, \tilde{b}$. That is, $\text{Alg}(\hat{U}, \tilde{b})$

The non-recursive pseudocode in the spirit of 1.3.5 and 1.3.13 is

```
for i = n, ..., 1
if U[i, i] = 0, set error flag, exit

b[i] = b[i]/U[i, i]

for j = i − 1, ..., 1
b[j] = b[j] − U[j, i] · b[i]
end
end
```

- **Counting flops for the above backward substitution algorithm**: The general expression for x_i can be expressed as

```
for i = n:-1:1
    for j = i+1:n
        B[i] = B[i] - A[i,j] * B[j]
    end
    B[i] = B[i]/A[i,i]
end
```

So, the flop count is

$$\begin{aligned} f(n) &= n + \sum_{i=1}^{n} 2(n - (i+1) + 1) \\ &= n + \sum_{i=1}^{n} 2(n - i) \\ &= n + 2\sum_{i=1}^{n} n - 2\sum_{i=1}^{n} i \\ &= n + 2n^2 - 2 \cdot \frac{n(n+1)}{2} \\ &= n + 2n^2 - (n^2 + n) \\ &= n + 2n^2 - n^2 - n \\ &= n^2 \end{aligned}$$

So, the backward substitution algorithm is $\mathcal{O}(n^2)$

- **Triangular matrices after multiplication**: triangular matrices stay triangular after multiplication, provided you multiply matrices of the same triangular type:
  - Upper triangular $\times$ upper triangular = upper triangular
  - Lower triangular $\times$ lower triangular = lower triangular
- **Triangular matrices after transpose**
  - The transpose of an upper triangular matrix is a lower triangular matrix
  - The transpose of an lower triangular matrix is a upper triangular matrix
- **Triangular matrices after inversion**: The inverse of a lower triangular matrix is lower triangular, and the inverse of an upper triangular matrix is upper triangular
- **Diagonal matrices**: Are both upper triangular and lower triangular (at the same time).

6.2.4 Positive Definite Systems

- **Positive definite matrix**: A matrix A is **positive definite** provided that the following two conditions are satisfied

  1. A is symmetric. That is, $A = A^\top$
  2. $x^\top Ax > 0$ for all $x \neq 0$

- **Positive Definiteness Characterizations**: Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then the following are equivalent:

  1. A is positive definite.
  2. All eigenvalues of A are positive.
  3. All *leading principal minors* of A are positive (Sylvester's criterion).

  Moreover:

  - (1) $\implies$ $a_{ii} > 0$ for all $i = 1, 2, \ldots, n$.
  - (1) $\implies$ every principal submatrix of A is positive definite.

- **Properties of positive definite (p.d) matrices**:

  1. If A is p.d then A is *nonsingular*

     Note: Since A is nonsingular there is no $y \in \mathbb{R}^n$, $y \neq 0$ such that $Ay = 0$
  2. If $A = M^\top M$ for some M nonsingular than A is p.d
  3. A is p.d if and only if all eigenvalues of A are positive

     Recall that λ is an eigenvalue of A if there exists $x_\lambda \neq 0$ such that $Ax_\lambda = \lambda x_\lambda$
  4. If A is p.d then all principal submatrices are p.d
  5. A is p.d if and only if all leading principal minors are positive
  6. If A is p.d than $\det(A) > 0$
  7. A is p.d if and only if there exists a unique upper triangular matrix R such that $A = R^\top R$ (Cholesky factorization described below)

  Note: Property two is a key property.

  Proof of (1): Assume A is a p.d matrix. Further assume (for the sake of contradiction) that A is singular. Then, there exits $y \in \mathbb{R}^n$, $y \neq 0$ with $Ay = 0$

  Since $Ay = 0$, $y \neq 0$, then $y^\top Ay = 0$ when $y \neq 0$, so A is not p.d.

  Therefore if A is p.d, then A is nonsingular ■

  Proof of (2): Assume that $A = M^\top M$ for some M nonsingular. Then,

  $$A^\top = (M^\top M)^\top = M^\top (M^\top)^\top = M^\top M = A$$

  So A is symmetric ($A = A^\top$)

Next, let $x \neq 0$

$$\begin{aligned} x^\top Ax &= x^\top (M^\top M)x \\ &= (x^\top M^\top)(Mx) \\ &= (Mx)^\top (Mx) \end{aligned}$$

let $y = Mx$, then

$$\begin{aligned} y^\top y &= \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix} \begin{pmatrix} y_1 & & \\ y_2 & & \\ \vdots & & y_n \end{pmatrix} \\ &= y_1^2 + y_2^2 + ... + y_n^2 > 0 \end{aligned}$$

So, A is p.d ■

Note that since $x \neq 0$, $y \neq 0$.

Recall that $y^\top y = \|y\|^2$

Proof of (3): Assume that A is p.d. Let λ be an eigenvalue for A. Then,

$$\begin{aligned} Ax_\lambda &= \lambda x_\lambda \\ \implies x_\lambda^\top Ax_\lambda &= x_\lambda^\top \lambda x \\ &= \lambda x_\lambda^\top x_\lambda \end{aligned}$$

First, observe that $x^\top Ax > 0$ since A p.d. Thus,

$$\lambda x_\lambda^\top x > 0$$

and since $x_\lambda^\top x = \|x_\lambda\|^2$, we have

$$\lambda \|x_\lambda\|^2 > 0$$

since $\|x_\lambda\|^2 > 0$, and $\lambda \|x_\lambda\| > 0$, it follows that $\lambda > 0$

- **Principal submatrices**: A **principal submatrix** of a square matrix $A \in \mathbb{R}^{n \times n}$ is obtained by selecting a subset of indices $I \subseteq \{1, \ldots, n\}$, and then keeping only the rows and the columns of A with those same indices.

  Examples

  $$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

  - Choosing $I = \{1, 2\}$ gives the principal submatrix

    $$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

  - Choosing $I = \{2, 3\}$ gives

    $$\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}.$$

  - Choosing $I = \{1, 3\}$ gives

    $$\begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix}$$

  - Choosing $I = \varnothing$ gives the empty matrix, usually denoted $0_{M_{0,0}}$. This empty matrix is in fact a principal submatrix of A

Note that if I is the subset of indices, then the submatrix is denoted $A[I]$, or $A[I, I]$

If $A \in \mathbb{R}^{n \times n}$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

Then the principal submatrices are

$$A_1 = \begin{bmatrix} a_{11} \end{bmatrix}$$
$$A_2 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$
$$A_3 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

- **Leading principal minors**: Let $A \in \mathbb{R}^{n \times n}$, with

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

  - Take $I_1 = \{1\}$

$$A_{I_1} = \begin{bmatrix} a_{11} \end{bmatrix}$$

  - Take $I_2 = \{1, 2\}$

$$A_{I_2} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

  - Take $I_k = \{1, 2, ..., k\}$ for $k < n$

$$A_{I_k} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix}$$

  - Take $I_n = \{1, 2, ..., n\}$ (The whole matrix)

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

These matrices are a special chain of principal submatrices called *leading principal submatrices*

This family of principal submatrices are the ones most often used in certain matrix theory results.

- **Determinant of the empty matrix**: We define
$$\det(\varnothing) = 1$$
- **Principal minors**: A principal minor is simply the determinant of a principal submatrix.
- **Cholesky decomposition and the Cholesky Factor**: Let $A \in \mathbb{R}^{n\times n}$ be p.d, then $A = R^\top R$ where R is upper triangular with $r_{ii} > 0$. The matrix R is called the **Cholesky factor**.

  If $A = R^\top R$, then $Ax = b$ can be written as
$$R^\top Rx = b$$
  where
$$\begin{cases} Rx & = y \quad \text{(Lower triangular)} \\ R^\top y & = b \quad \text{(Upper triangular)} \end{cases}$$
  and since these new systems are triangular, they can be solved quickly with forward or backward substitution.
- **Inner product formulas to compute R (Cholesky factor)**: We have the formulas
$$r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2} \quad i = 1, 2, ..., n$$
$$r_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}}{r_{ii}} \quad j = i+1, ..., n$$
- **Recursive column oriented method to find the Cholesky factor R (Outer product method)**: Let $A \in \mathbb{R}^{n\times n}$. Assume that A is positive definite, so $A = A^\top$, and $A = R^\top R$ for a unique upper triangular matrix R. We have,
$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} r_{11} & 0 & \cdots & 0 \\ r_{12} & r_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \cdots & r_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}$$
  We then perform a matrix decomposition
$$\begin{bmatrix} a_{11} & a^\top \\ a & \hat{A} \end{bmatrix} = \begin{bmatrix} r_{11} & 0^\top \\ r & \hat{R}^\top \end{bmatrix} \begin{bmatrix} r_{11} & r^\top \\ 0 & \hat{R} \end{bmatrix}.$$
  Where $\hat{A} = \hat{A}^\top \in \mathbb{R}^{n-1\times n-1}$, $a \in \mathbb{R}^{n-1}$, $\hat{R}^\top \in \mathbb{R}^{n-1\times n-1}$ lower triangular, and $\hat{R} \in \mathbb{R}^{n-1\times n-1}$ upper triangular. Further,
$$\hat{A} = \begin{bmatrix} a_{22} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n2} & \cdots & a_{nn} \end{bmatrix}, \; a = \begin{pmatrix} a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{pmatrix},$$
$$\hat{R} = \begin{bmatrix} r_{22} & \cdots & r_{2n} \\ 0 & \ddots & \vdots \\ 0 & \cdots & r_{nn} \end{bmatrix}, \; r = \begin{pmatrix} r_{12} \\ r_{13} \\ \vdots \\ r_{1n} \end{pmatrix}.$$

So, given this decomposition, we see that

$$a_{11} = r_{11}^2 \implies r_{11} = \sqrt{a_{11}}.$$

Recall that in the definition of the Cholesky factor R, $r_{ii} > 0$ for $i = 1, 2, ..., n$

Continuing the matrix multiplication, we have that

$$a = r_{11} r \implies r = \frac{a}{r_{11}}$$

and,

$$\hat{A} = rr^\top + \hat{R}\hat{R}^\top \implies \hat{R}^\top \hat{R} = \hat{A} - rr^\top = \tilde{A}$$

Thus, the recursive column oriented algorithm to compute the Cholesky factor R is given by the following steps

1. $r_{11} = \sqrt{a_{11}}$
2. $r = \frac{a}{r_{11}}$
3. $\tilde{A} = \hat{A} - rr^\top$
4. $\text{Alg}(\tilde{A}) = \hat{R}$

- **Counting flops for the recursive algorithm above**:

  1. (Step 1): One flop
  2. (Step 2): $n - 1$ flops
  3. (Step 3): $(n-1)^2$ flops. Notice that $r \in \mathbb{R}^{n-1}$, $r^\top \in \mathbb{R}^{n-1}$, and the outer product

     $$rr^\top \in \mathbb{R}^{n-1 \times n-1}$$

     and requires $(n-1)^2$ flops.

     Since $\hat{A} \in \mathbb{R}^{n-1 \times n-1}$, $\hat{A} - rr^\top$ requires an addition $(n-1)^2$ flops. So, step 3 requires $2(n-1)^2$ flops

Thus,

$$\begin{aligned} f_n &= 1 + (n-1) + 2(n-1)^2 + f_{n-1} \\ &= n + 2(n-1)^2 + f_{n-1}. \end{aligned}$$

Where

$$\begin{aligned} f_{n-1} &= 1 + (n-2) + 2(n-2)^2 + f_{n-2} = n - 1 + 2(n-2)^2 + f_{n-2} \\ &\vdots \\ f_{n-(n-1)} &= f_1 = 1 + 0 + 0 + f_0. \end{aligned}$$

Note that $f_0 = 0$. So, $f_1 = 1$. In total, we have

$$\begin{aligned} f_n &= n + 2(n-1)^2 + n - 1 + 2(n-2)^2 + n - 2 + 2(n-3)^2 + ... + 1 \\ &= \sum_{k=1}^{n} k + 2(k-1)^2 \\ &= \sum_{k=1}^{n} k + 2\sum_{k=1}^{n} k^2 - 2k + 1 \\ &= \sum_{k=1}^{n} k + 2\sum_{k=1}^{n} k^2 - 4\sum_{k=1}^{n} k + 2\sum_{k=1}^{n} 1 \\ &= 2\sum_{k=1}^{n} k^2 - 3\sum_{k=1}^{n} k + 2k \\ &= 2\left(\frac{n(n+1)(2n+1)}{6}\right) - 3\left(\frac{n(n+1)}{2}\right) + 2k = \mathcal{O}(n^3) \end{aligned}$$

So, the recursive algorithm is $\mathcal{O}(n^3)$

- **Consequence of proof n.1.2:** If A has any $a_{ii} \leqslant 0$, A is not positive definite.
- **Bordered form of Choleskys method**: Suppose $A \in \mathbb{R}^{n\times n}$ is positive definite. Then, A admits a decomposition $A = R^\top R$, for a unique upper triangular matrix R called the Cholesky factor, with $r_{ii} > 0$ for $i = 1, 2, ..., n$. So,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} r_{11} & 0 & \cdots & 0 \\ r_{12} & r_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \cdots & r_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}$$

We then perform a matrix decomposition

$$\begin{bmatrix} \hat{A} & a \\ a^\top & a_{nn} \end{bmatrix} = \begin{bmatrix} \hat{R}^\top & 0 \\ r^\top & r_{nn} \end{bmatrix} \begin{bmatrix} \hat{R} & r \\ 0 & r_{nn} \end{bmatrix}.$$

So,

$$\begin{aligned} \hat{A} &= \hat{R}^\top \hat{R}, \\ a &= \hat{R}^\top r, \\ a_{nn} &= r^\top r + r_{nn}^2 \implies r_{nn} = \sqrt{a_{nn} - r^\top r}. \end{aligned}$$

So, the steps for the algorithm are

1. Recurse $\hat{A}$ until $A \in \mathbb{R}^{1\times 1}$
2. Solve the lower triangular system $\hat{R}^\top r = a$ by forward substitution
3. Compute $r_{nn} = \sqrt{a_{nn} - r^\top r}$
4. Return the step two on the previous call

The above algorithm is postorder recursion and requires $\mathcal{O}(n^3)$ flops.

6.2.5 Banded Matrices

- **Cholesky factor R in a diagonal matrix**: If $A = D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$ then

$$R = \begin{bmatrix} \sqrt{a_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{a_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{a_{nn}} \end{bmatrix}$$

- **Banded matrix**: A banded matrix is a sparse matrix whose nonzero entries are confined to a diagonal band, consisting of the main diagonal and a fixed number of diagonals on either side of it.

  Let $A \in \mathbb{R}^{m \times n}$. Then A is called a **banded matrix** if there exist nonnegative integers p, q (called the *lower* and *upper bandwidths*) such that

$$a_{ij} = 0 \quad \text{whenever } i - j > p \text{ or } j - i > q.$$

  – The *lower bandwidth* p is the number of subdiagonals (below the main diagonal) that may contain nonzero entries.

  – The *upper bandwidth* q is the number of superdiagonals (above the main diagonal) that may contain nonzero entries.

  The *total bandwidth* is sometimes defined as $p + q + 1$, counting the main diagonal as well.

  A tridiagonal matrix has lower bandwidth $p = 1$ and upper bandwidth $q = 1$:

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix}.$$

  The total bandwidth here is $1 + 1 + 1 = 3$.

- **Column envelope**: The column envelope of A is the set of indices (i, j) in the upper triangular part of A (including the main diagonal). Define

$$\text{colenv}\{A\} = \{(i, j)\} : \ i \leqslant j \text{ and } a_{kj} \neq 0 \text{ for } k \leqslant i\}$$

- **Theorem**: Let A be p.d, if R is the Cholesky factor of A, then

$$\text{colenv}\{R\} = \text{colenv}\{A\}$$

6.2.6 Gaussian Elimination and LU Decompositions

- **Intro to LU decomposition**: Consider a matrix $A \in \mathbb{R}^{n\times n}$. If we can factor A as $A = LU$, for L lower triangular, U upper triangular, then the system $Ax = b$, for vectors $x, b \in \mathbb{R}^n$ turns into

$$LUx = b.$$

We can then split this system as follows

$$\begin{cases} Ly &= b \\ Ux &= y \end{cases}$$

First, we solve $Ly = b$ with forward substitution to find y. We can then solve $Ux = y$ with backward substitution to find the target x.

Recall that the forward and backward substitution methods for solving linear systems requires $\mathcal{O}(n^2)$ flops.

- **Elementary operations on systems that do not change the solution set**. We have the operations

1. Interchange rows.
2. Multiply an equation by a nonzero constant.
3. Add a multiple of one equation to another equation.

We show that these elementary operations (E.O) leave the solution set unchanged. Let the original system be S and the modified system be S'. To show that the solution set is unchanged is to show that if a vector x is a solution to S then it is also a solution to S', and vice versa.

Consider the system S, $Ax = b$ for $A \in \mathbb{R}^{n\times n}$, $x, b \in \mathbb{R}^n$. So, the system is

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n &= b_n. \end{aligned}$$

Proof (2). Multiply an arbitrary equation by a nonzero scalar k, suppose we choose the second equation. S' is then

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= b_1, \\ k(a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n) &= k(b_2), \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n &= b_n. \end{aligned}$$

Let $(c_1, c_2, ..., c_n)$ be a solution to the original system S. That is, it satisfies all equations. Let's look at the second equation

$$a_{21}c_1 + a_{22}c_2 + ... + a_{2n}c_n = b_2$$

If we multiply by k, we get

$$k(a_{21}c_1 + a_{22}c_2 + ... + a_{2n}c_n) = k(b_2).$$

Which means $(c_1, c_2, ..., c_n)$ also satisfies the second equation in S'. Since all other equations were left unchanged, $(c_1, c_2, ..., c_n)$ satisfies those equations as well. So, the solution set is the same for both systems.

Note: If $k = 0$, the second equation would collapse to $0 = 0$, which would enlarge the solution set. In this case, a constraint would be removed from the system S'. The second equation is now tautological, it imposes no restriction. The solution set would be

$$\{x \in \mathbb{R}^n : \ Ax = b \text{ for all rows except row two}\}$$

Every solution of S is also a solution of S', but the converse need not hold: S' could have solutions that don't satisfy the second original equation.

So S' has at least as many solutions, and possibly more. If the second equation was independent of the others, then yes - you've enlarged the solution set.

Proof (3). Add an arbitrary equation to a different equation. Suppose we add the first equation to the second equation. Note that we leave the first unchanged. S' is then

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= b_1, \\ (a_{21} + a_{11})x_1 + (a_{22} + a_{12})x_2 + ... + (a_{2n} + a_{1n})x_n) &= b_2 + b_1, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n &= b_n. \end{aligned}$$

Let $c = (c_1, c_2, ..., c_n)$ be a solution to S, so c is a solution to the first, second, and the remaining equations. That is,

$$\begin{aligned} a_{11}c_1 + a_{12}c_2 + ... + a_{1n}c_n &= b_1, \\ a_{21}c_1 + a_{22}c_2 + ... + a_{2n}c_n &= b_2, \\ &\vdots \\ a_{n1}c_1 + a_{n2}c_2 + ... + a_{nn}c_n &= b_n. \end{aligned}$$

Add the first equation to the second, we get

$$\begin{aligned} a_{11}c_1 + a_{12}c_2 + ... + a_{1n}c_n &= b_1, \\ (a_{21} + a_{11})c_1 + (a_{22} + a_{12})c_2 + ... + (a_{2n} + a_{1n})c_n) &= b_2 + b_1, \\ &\vdots \\ a_{n1}c_1 + a_{n2}c_2 + ... + a_{nn}c_n &= b_n \end{aligned}$$

which is precisely S', so c satisfies S'. Next, let $c = (c_1, c_2, ..., c_n)$ be a solution to S'. So,

$$\begin{aligned} a_{11}c_1 + a_{12}c_2 + ... + a_{1n}c_n &= b_1, \\ (a_{21} + a_{11})c_1 + (a_{22} + a_{12})c_2 + ... + (a_{2n} + a_{1n})c_n) &= b_2 + b_1, \\ &\vdots \\ a_{n1}c_1 + a_{n2}c_2 + ... + a_{nn}c_n &= b_n \end{aligned}$$

Subtract the first equation from the second, and we get back S. So, the solution set remains unchanged.

- **Elimination matrix** E: An elimination matrix is just a special matrix that performs a single step of Gaussian elimination when you multiply it by another matrix.

Suppose you want to eliminate the entry in row i, column j of A. In elimination, you would replace row i by

$$\text{row}_i - m \cdot \text{row}_j,$$

where $m = \frac{a_{ij}}{a_{jj}}$.

The **elimination matrix** E is the identity matrix, except in position (i, j), where it has $-m$. So,

$$E = I - me_i e_j^\top = I - mE_{ij},$$

where e_i and e_j are standard basis vectors (all zeros except a one at the i^{th} or j^{th} position), and $E_{ij} = e_i e_j^T$

Multiplying E by A from the left actually performs that row operation:

$$EA = A \quad \text{with entry } (i, j) \text{ zeroed out.}$$

Let

$$A = \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix}.$$

We want to eliminate the entry in the bottom-left, the multiplier is

$$m = \frac{4}{2} = 2.$$

The elimination matrix is

$$E = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}.$$

Now check

$$EA = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix},$$

- **Type three elementary operations using left matrix multiplication**: Suppose we have $A \in \mathbb{R}^{n \times n}$, and we want to add a multiple of the j^{th} row to the i^{th} row, call $\tilde{A}$ the matrix obtained after this elementary operation

$$\begin{aligned} &\text{If } k \neq i, \ \text{row}_k(\tilde{A}) = \ \text{row}_k(A), \\ &\text{If } k = i, \ \text{row}_k(\tilde{A}) = \ \text{row}_k(A) + m \cdot \text{row}_j(A). \end{aligned}$$

We assert that $\tilde{A} = MA$, where $M \in \mathbb{R}^{n \times n}$ is the identity matrix, except for m at position i, j. If $i > j$, M is lower triangular, and if $i < j$, M is upper triangular. If we use this fact during Gaussian elimination, M will always be lower triangular, since we only care about zeroing out the lower triangular part of A.

Let E_{ij} be the zero matrix except for a one at e_{ij}. Thus,

$$M = I + mE_{ij}.$$

Observe that $E_{ij} = e_i e_j^T$, so

$$M = I + me_i e_j^T.$$

From this fact, we have

$$MA = (I + me_i e_j^T)A = A + me_i(e_j^T A).$$

Recall that $e_j^T A$ is the j^{th} row of A, so

$$MA = A + me_i \cdot \text{row}_j(A).$$

Further observe that $e_i \cdot \text{row}_j(A)$ is a matrix of size $n \times n$, where the i^{th} row is $\text{row}_j(A)$, and all other rows are zero.

So, we see that

$$\text{If } k \neq i,\ \text{row}_k(E_{ij}A) = 0,\ \text{ so row}_k(MA) = \text{ row}_k(A),$$
$$\text{If } k = i,\ \text{row}_k(E_{ij}A) = \text{row}_j(A),\ \text{ so row}_k(MA) = \text{row}_i(A) + m \cdot \text{row}_j(A).$$

Thus, $\tilde{A} = MA$ ■

- **Type two elementary operations using left matrix multiplication**: Suppose $\tilde{A}$ is obtained from A by multiplying the ith row by the nonzero constant c. We wish to find a matrix M such that $MA = \tilde{A}$

Suppose we have $A \in \mathbb{R}^{2\times 2}$, where

$$A = \begin{bmatrix} \alpha & \beta \\ \gamma & \varphi \end{bmatrix}.$$

Now, suppose we want to scale the second row by c, where $c \in \mathbb{R}$, then

$$\begin{aligned} \tilde{A} = \begin{bmatrix} \alpha & \beta \\ c\gamma & c\varphi \end{bmatrix} &= \begin{bmatrix} \alpha & \beta \\ \gamma & \varphi \end{bmatrix} + \begin{bmatrix} \alpha & \beta \\ (c-1)\gamma & (c-1)\varphi \end{bmatrix} \\ &= A + (c-1)\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} A \\ &= A + (c-1)E_{22}A. \end{aligned}$$

So, suppose we wish to scale the i^{th} row of A by a constant c, then

$$\tilde{A} = A + (c-1)E_{ii}A = (I + (c-1)E_{ii})A = MA.$$

Thus, $M = I + (c-1)E_{ii}$.

It seems that the inverse operation is multiplying row i by $\frac{1}{c}$. Thus, we propose that the inverse of M is

$$M^{-1} = I + \left(\frac{1}{c} - 1\right) E_{ii}.$$

We have

$$\begin{aligned} MM^{-1} &= (I + (c-1)E_{ii})\left(I + \left(\frac{1}{c} - 1\right) E_{ii}\right) \\ &= I + \left(\frac{1}{c} - 1\right) E_{ii} + (c-1)E_{ii} + (c-1)\left(\frac{1}{c} - 1\right) E_{ii}^2. \end{aligned}$$

But,

$$\begin{aligned} E_{ii}^2 = (e_i e_i^T)(e_i e_i^T) &= e_i(e_i^T e_i)e_i^T \\ &= (e_i^T e_i)e_i e_i^T = (e_i^T e_i)E_{ii} \\ &= \|e_i\|^2 E_{ii} = E_{ii}. \end{aligned}$$

So, $E_{ii}^2 = E_{ii}$, and

$$\begin{aligned} MM^{-1} &= II + \left(\frac{1}{c} - 1\right) E_{ii} + (c-1)E_{ii} + (c-1)\left(\frac{1}{c} - 1\right) E_{ii}^2 \\ &= I + \left(\frac{1}{c} - 1 + c - 1\right) E_{ii} + (c-1)\left(\frac{1}{c} - 1\right) E_{ii} \\ &= I + \left(\frac{1}{c} - 1 + c - 1 + (c-1)\left(\frac{1}{c} - 1\right)\right) E_{ii} \\ &= I + \left(\frac{1}{c} - 1 + c - 1 + 1 - c - \frac{1}{c} + 1\right) E_{ii} \\ &= I + 0E_{ii} = I. \end{aligned}$$

Thus, $M^{-1} = I + \left(\frac{1}{c} - 1\right) E_{ii}$

Observe that the determinant of M is

$$\det(M) = \det(I + (c-1)E_{ii}) = \prod_{k=1}^{n} m_{kk}.$$

But, notice that $m_{kk} = 1$, except at $k = i$, where we have $m_{ii} = 1 + (c-1) = c$. Thus, $\det(M) = c$, and

$$\det(\tilde{A}) = \det(MA) = \det(M)\det(A) = c\det(A).$$

Since $c \neq 0$, $\det(\tilde{A}) = 0 \iff \det(A) = 0$. Hence, $\tilde{A}$ is nonsingular if and only if A is. ■

- **LU Factorization without E.O (1)**: We perform Gaussian Elimination on the augmented system $[A|b]$ to yield a new system $[U|y]$.

  We move down the main diagonal selecting a_{ii} as the **pivot element**, and row i as the **pivot row**. We do this for $i = 1, 2, ..., n$. For each pivot element, we get the elements $a_{ki} = 0$ for $k = i+1, i+2, ..., n$ we can accomplish this without interchanging rows so long as the pivot elements are nonzero.

  We perform elementary operations of the form

  $$-m_{ki}r_i + r_k \to r'_k \quad \text{for } k = i+1, i+2, ..., n$$

  where r_i is the i^{th} row, r_k is the k^{th} row, and m_{ki} is the **multiplier** $m_{ki} = \frac{a_{ki}}{a_{ii}}$

  Upon completion of the Gaussian elimination, the collected multipliers together with ones in the main diagonal and zeros in the entries above the main diagonal form the matrix L.

  Additional information: We perform Gaussian elimination on the augmented matrix $[A|b] \to [U|y]$. Note that after we achieve U, we have the system $Ux = y$ with the same solution set as $Ax = b$.

  Notice that each step has its own elimination matrix $E_1, E_2, ...$ If we apply them in sequence,

  $$E_k \cdots E_2 E_1 A = U$$

  Where U is the final upper triangular matrix. It follows that

  $$A = (E_k \cdots E_2 E_1)^{-1} U$$

  Define

  $$(E_k \cdots E_2 E_1)^{-1} = L$$

  It's lower triangular because each elimination matrix is lower triangular, and the inverse of a lower triangular matrix is also lower triangular.

For example, consider the system

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 7 \\ 2 & 2 & -1 & 3 \\ 4 & -1 & 6 & 20 \end{array}\right]$$

We start with $a_{11} = 2$ as the pivot element, and row one as the pivot row. We perform elementary operations of the form above to get $a_{21} = a_{31} = 0$. To get $a_{21} = 0$, we have the operation

$$-1r_1 + r_2 \to r_2', \quad m_{21} = 1$$

To get $a_{31} = 0$, we perform the operation

$$-2r_1 + r_3 \to r_3', \quad m_{31} = 2$$

After the two operations, we have

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 7 \\ 0 & 1 & -2 & -4 \\ 0 & -3 & 4 & 6 \end{array}\right]$$

We move to the next pivot element $a_{22} = 1$. To get $a_{32} = 0$, we perform the operation

$$-(-3)r_2 + r_3 \to r_3', \quad m_{32} = -3$$

So after the last operation we have

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 7 \\ 0 & 1 & -2 & -4 \\ 0 & 0 & -2 & -6 \end{array}\right]$$

- **Theorem**: Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. Then, we can solve the system $Ax = b$, $b \in \mathbb{R}^n$ using Gaussian Elimination without row interchanges if and only if all landing principal sub-matrices of A are nonsingular.
- **Theorem**: Let $A \in \mathbb{R}^{n\times n}$. Then A admits an LU factorization

  $$A = LU,$$

  where $L \in \mathbb{R}^{n\times n}$ is unit lower triangular and $U \in \mathbb{R}^{n\times n}$ is upper triangular, **without row interchanges**, if and only if all leading principal submatrices of A are nonsingular.
- **Row oriented algorithm to find LU factorization**: Let $A \in \mathbb{R}^{n\times n}$. If A can be factored into products LU, for $U \in \mathbb{R}^{n\times n}$ upper triangular, $L \in \mathbb{R}^{n\times n}$ unit lower triangular, then

  $$A = LU$$

  implies

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{1n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}.$$

  Let's first examine the formula for u_{ij} by solving for each row in U

1. **Row 1**: for $j = 1, \ldots, n$
$$u_{1j} = a_{1j}.$$
2. **Row 2**: for $j = 2, \ldots, n$
$$u_{2j} = a_{2j} - \ell_{21}\, u_{1j}.$$
3. **Row 3**: for $j = 3, \ldots, n$
$$u_{3j} = a_{3j} - \ell_{31}\, u_{1j} - \ell_{32}\, u_{2j}.$$
4. **Row** i: for $j = i, \ldots, n$
$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} \ell_{ik}\, u_{kj}.$$
5. **Row** n: (just the diagonal entry)
$$u_{nn} = a_{nn} - \sum_{k=1}^{n-1} \ell_{nk}\, u_{kn}.$$

Next, we look at the formula for ℓ_{ij} by solving for each column in L

1. **Column 1**: for $i = 2, \ldots, n$
$$\ell_{i1} \;=\; \frac{a_{i1}}{u_{11}}.$$
2. **Column 2**: for $i = 3, \ldots, n$
$$\ell_{i2} \;=\; \frac{a_{i2} - \ell_{i1} u_{12}}{u_{22}}.$$
3. **Column 3**: for $i = 4, \ldots, n$
$$\ell_{i3} \;=\; \frac{a_{i3} - \ell_{i1} u_{13} - \ell_{i2} u_{23}}{u_{33}}.$$
4. **Column** j: for $i = j + 1, \ldots, n$
$$\ell_{ij} \;=\; \frac{a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} u_{kj}}{u_{jj}}.$$
5. **Column** n: (no entries below the diagonal to compute if $j = n$)
$$\text{Only } \ell_{nn} = 1 \text{ (by unit lower convention).}$$

So, we see that

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} \ell_{ik} u_{kj} \quad j = i, i+1, ..., n \tag{1}$$

$$\ell_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} u_{kj}}{u_{jj}} \quad i = j+1, j+2, ..., n \tag{2}$$

To use these formulas to find each u_{ij} we first need to plug $i = 1$ into (1), then after we get the first row of U, we can plug in $j = 1$ into (2) to get the first column of L, and so on.

- **Column oriented recursive algorithm to find the LU factorization**: Assume $A \in \mathbb{R}^{n\times n}$ admits an LU factorization for $L \in \mathbb{R}^{n\times n}$ unit lower triangular, U upper triangular. Then,

$$A = LU$$

implies

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}.$$

Decompose $A = LU$ into the blocks

$$\begin{bmatrix} a_{11} & b^\top \\ a & \hat{A} \end{bmatrix} = \begin{bmatrix} 1 & 0^\top \\ \ell & \hat{L} \end{bmatrix} \begin{bmatrix} u_{11} & u^\top \\ 0 & \hat{U} \end{bmatrix}.$$

We see that

$$\begin{aligned} a_{11} &= u_{11}, \\ b^\top &= u^\top, \\ \ell u_{11} = a &\implies \ell = \frac{a}{u_{11}}, \\ \hat{A} = \ell u^\top + \hat{L}\hat{U} &\implies \hat{L}\hat{U} = \hat{A} - \ell u^\top. \end{aligned}$$

Define $\tilde{A} = \hat{A} - \ell u^\top$. The recursive algorithm is then defined by the following steps.

1. $u_{11} = a_{11}$ (zero flops)
2. $u^\top = b^\top$ (zero flops)
3. $\ell = \frac{a}{u_{11}}$ ($n-1$ flops)
4. $\tilde{A} = \hat{L}\hat{U} = \hat{A} - \ell u^\top$ ($2(n-1)^2$ flops)
5. Alg($\tilde{A}$)

Let $f(n)$ be the flop count for the above algorithm. We have

$$\begin{aligned} f(n) &= 0 + 0 + (n-1) + 2(n-1)^2 + f_{n-1} = (n-1) + 2(n-1)^2 + f_{n-1} \\ f(n-1) &= ((n-1)-1) + 2((n-1)-1)^2 + f_{n-2} \\ &\vdots \\ f(n-(n-2)) = f_2 &= ((n-(n-2))-1) + 2((n-(n-2))-1)^2 + f_{n-(n-1)} \\ f_{n-(n-1)} &= f_1 = 0 \end{aligned}$$

So, the total number of flops is given by the sum

$$\sum_{k=2}^{n} (k-1) + 2(k-1)^2.$$

Let $i = k - 1$. When $k = 2$, $i = 1$. When $k = n$, $i = n - 1$. So, the sum becomes

$$\sum_{i=1}^{n-1} 2i^2 + i$$

Remark. We have the summation rules

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \quad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Plug in $n-1$ for each,

$$\sum_{i=1}^{n} i = \frac{(n-1)n}{2}, \quad \sum_{i=1}^{n} i^2 = \frac{(n-1)n(2n-1)}{6}$$

So,

$$\begin{aligned}\sum_{i=1}^{n-1} 2i^2 + i &= 2\left(\frac{(n-1)n(2n-1)}{6}\right) + \frac{(n-1)n}{2} \\ &= \frac{2}{3}n^2 - \frac{1}{2}n^2 - \frac{1}{6}n = \frac{2}{3}n^3 + \mathcal{O}(n^2)\end{aligned}$$

Therefore, the number of flops required for the recursive outer product method to find the LU factorization is $\frac{2}{3}n^3 + \mathcal{O}(n^2)$

- **Bordered form LU decomposition algorithm**: Assume $A \in \mathbb{R}^{n\times n}$ admits an LU factorization for $L \in \mathbb{R}^{n\times n}$ unit lower triangular, U upper triangular. Then,

$$A = LU$$

implies

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}.$$

Decompose $A = LU$ into the blocks

$$\begin{bmatrix} \hat{A} & b \\ a^\top & a_{nn} \end{bmatrix} = \begin{bmatrix} \hat{L} & 0 \\ \ell^\top & 1 \end{bmatrix} \begin{bmatrix} \hat{U} & u \\ 0^\top & u_{nn} \end{bmatrix}$$

Where $\hat{A}, \hat{L}, \hat{U}$ are the $(n-1)^{\text{th}}$ leading principal submatrices of A, L, U.

Then,

$$\begin{aligned} \hat{A} &= \hat{L}\hat{U} \\ \hat{L}U &= b \\ \ell^\top \hat{U} = a^\top &\implies \hat{U}^\top \ell = a \\ u_{nn} &= a_{nn} - \ell^\top u^\top \end{aligned}$$

- **Intro to row interchanges (pivoting)**: Without pivoting, Gaussian elimination can behave as if the problem were ill-conditioned even when it is not, because tiny pivots can amplify rounding errors. This is why we use partial pivoting.

Consider the system

$$\begin{bmatrix} 0.0003 & 1.566 \\ 0.3454 & -2.436 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.569 \\ 1.018 \end{bmatrix}.$$

Solving the system by Gaussian Elimination without pivoting, we get $m_{21} = \frac{0.3454}{0.0003} = 1151.3333$. After Gaussian Elimination, we get

$$x_1 = 3.333, \quad x_2 = 1.001.$$

We note that the exact solution to the system is $x_1 = 10, x_2 = 1$. So what happened? We see that x_2 is far from the true solution.

If we instead swap the rows to use the second row as the pivot row, we get

$$m_{21} = \frac{0.0003}{0.3454} = 0.0008686.$$

Then, $x_1 = 10.01,\ x_2 = 1$.

We we round off a number α, we get $\alpha \to \bar{\alpha}$, where $\alpha = \bar{\alpha} + \epsilon$, for some small ϵ. If this number is then multiplied by a scalar m, we get

$$m\alpha = m\bar{\alpha} + m\epsilon.$$

So, the error grows as m grows. Our goal is to select the pivot such that m is minimized.

If we select the largest element in the k^{th} column (at step k), then we can guarantee $m \leqslant 1$.

- **Partial pivoting**: At iteration k of Gaussian Elimination, we swap row k with some row below so that the new a_{kk} has the largest absolute value compared to all entries below in column k.
- **Permutation matrix**: A permutation matrix is a special kind of square matrix that represents a permutation of elements. Formally: It is obtained from the identity matrix by rearranging its rows (or equivalently, its columns).

  Each row and each column has exactly one entry equal to 1, and all other entries are 0.

  Multiplying a vector (or another matrix) by a permutation matrix reorders its entries.

  Suppose P is formed by taking I and interchanging rows one and two. Then,

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

  Then,

$$P \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} b \\ a \\ c \end{pmatrix}.$$

  So, it swaps the first two entries.

- **Permutation matrix left multiplication vs right multiplication**: If P is a permutation matrix, and $A \in \mathbb{R}^{n \times n}$, then AP permutes the columns of A. Recall that $\text{col}_j(AP) = A\text{col}_j(P)$, so

$$\text{col}_j(AP) = A\text{col}_j(P).$$

  But, $\text{col}_j(P) = e_k$, where e_k is the k^{th} standard basis vector in $\mathbb{R}^n$, so

$$\text{col}_j(AP) = Ae_k.$$

  Notice that $Ae_k = \text{col}_k(A)$. Suppose column one is swapped with column two in P, then $\text{col}_1(P) = e_2$, so $\text{col}_1(AP) = Ae_2 = \text{col}_2(A)$.

Similarly, PA permutes the rows of A. Recall that $\text{row}_i(AB) = \text{row}_i(A)B$, so

$$\text{row}_i(PA) = \text{row}_i(P)A = e_k^T A = \text{row}_k(A).$$

Suppose that row one is swapped with row two in P, then $\text{row}_1(P) = e_2^T$, and $\text{row}_1(PA) = e_2^T A = \text{row}_2(A)$

- **Gaussian elimination with partial pivoting**: Eliminates entries below the pivots to produce an upper triangular system.

  At each step, you swap the current row with one below it to bring the largest (by absolute value) element in the pivot column into the pivot position. This improves numerical stability.

- LU **factorization with partial pivoting**: We we partial pivot rows in Gaussian Elimination (while building L, U), we make the same swap in P. If L is not being stored in A during the process, then we also need to swap the rows of L, but keep the main diagonal the same. The main diagonal of L is always remains ones, and the upper triangular part is always zero.

  At the end, we get

  $$PA = LU.$$

  Then, we also see that

  $$A = P^{-1}LU = P^T LU.$$

  We can use this fact to solve systems $Ax = b$. We have

  $$\begin{aligned} & Ax = b \\ \implies & PAx = Pb \\ \implies & LUx = Pb. \end{aligned}$$

  Just like in standard LU decomposition, we split the system in two triangular systems that can both be solved by substitution in n^2 flops. We have

  $$\begin{cases} Ly &= Pb \quad \text{(Lower triangular)} \\ Ux &= y \quad \text{(Upper triangular)} \end{cases}.$$

6.3 Outer Products, inner products, and transposition tricks

- **Build an $m \times n$ matrix from n vectors in $\mathbb{R}^m$**: Suppose we have vectors $x_1, x_2, x_3, ..., x_n \in \mathbb{R}^m$, and we wish to construct the matrix formed by combining each vector x_k for $k = 1, 2, 3, ..., n$. Algebraically, we have

$$\begin{aligned} X &= \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix} = x_1 e_1^\top + x_2 e_2^\top + x_3 e_3^\top + ... + x_n e_n^\top \\ &= \sum_{k=1}^{n} x_k e_k^\top \end{aligned}$$

  where e_ℓ for $\ell = 1, 2, 3, ..., n$ are the standard basis vectors in $\mathbb{R}^n$.

  For example, if $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$, and $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathbb{R}^2$ then

$$\begin{aligned} X = xe_1^\top + ye_2^\top &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \begin{pmatrix} 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} x_1 & 0 \\ x_2 & 0 \end{pmatrix} + \begin{pmatrix} 0 & y_1 \\ 0 & y_2 \end{pmatrix} = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix} \end{aligned}$$

- **Construct a $m \times n$ matrix with a single element in some position**: Suppose we want an $m \times n$ matrix with k in position a_{ij} We take the outer product

$$A = me_i e_j^\top$$

  Where e_i is the i^{th} standard basis vector in $\mathbb{R}^m$, and e_j is the j^{th} standard basis vector in $\mathbb{R}^n$.

- **Matrix multiplication in terms of column**: For $A, B \in \mathbb{R}^{n \times n}$, and $C = AB \in \mathbb{R}^{n \times n}$, we have that

$$\text{col}_j(C) = A\text{col}_j(B).$$

  Thus, $C = AB$, where $A, B \in \mathbb{R}^{n \times n}$ requires solving n linear systems.

- **Matrix multiplication in terms of rows**: Let $A, B \in \mathbb{R}^{n \times n}$, and $C = AB$. Since $C = AB$, $C^T = B^T A^T$, and using the fact above, we see

$$\text{col}_j(C^T) = B^T \text{col}_j(A^T).$$

  But, $\text{col}_j(C^T) = (\text{ row}_j(C))^T$, and $\text{col}_j(A^T) = (\text{ row}_j(A))^T$. So,

$$\begin{aligned} \text{col}_j(C^T) &= B^T \text{col}_j(A^T) \\ \implies (\text{row}_j(C))^T &= B^T (\text{row}_j(A))^T . \end{aligned}$$

  Taking the transposition of both sides gives

$$\text{row}_j(C) = \text{ row}_j(A)B.$$

  Therefore, changing the name of index, we get the result

$$\text{row}_i(C) = \text{ row}_i(A)B.$$

- **Getting the rows and columns of a matrix algebraically**: Let $A \in \mathbb{R}^{n \times n}$, let e_i be the i^{th} standard basis vector in $\mathbb{R}^n$. That is, a vector of size n with a one in the i^{th} position, and zeros everywhere else. Then,

$$\text{col}_j(A) = Ae_j.$$

  This comes directly from how matrix–vector multiplication works in linear algebra. Multiplying a matrix A by a standard basis vector e_j picks out the j-th column of A, because the 1 in position j selects that column while all other zeros eliminate the rest.

If we take the transpose of both sides,

$$(\text{col}_j(A))^T = e_j^T A^T,$$

which implies that

$$\left(\text{col}_j(A^T)\right)^T = e_j^T A, .$$

But, we know that $\text{row}_j(A) = (\ \text{col}_j(A^T))^T$, and $\text{col}_j(A) = (\text{row}_j(A^T))^T$. Thus,

$$\begin{aligned} \left(\text{col}_j(A^T)\right)^T &= e_j^T A \\ \implies \text{row}_j(A) &= e_j^T A. \end{aligned}$$

- **Retrieving an element of a matrix algebraically**: Let $A \in \mathbb{R}^{n \times n}$, and $e_i \in \mathbb{R}^n$ be the i^{th} standard basis vector, then

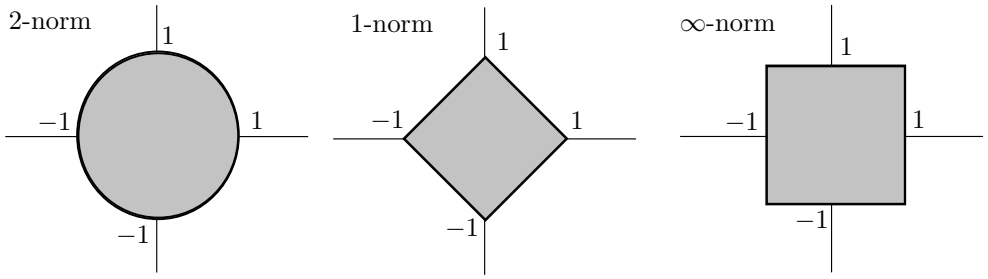
$$e_i^T A e_j = a_{ij},$$

since $Ae_j = \text{col}_j(A)$, and $e_i^T \text{col}_j(A) = 1(\text{col}_j)_i = a_{ij}$

6.4 Sensitivity of linear systems (2)

6.4.1 Vector and matrix norms

- **Norm:** A norm is an operation $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+ : x \to \|x\| \geqslant 0$ that satisfies
  1. $\|x\| = 0 \iff x = 0$
  2. $\|\alpha x\| = |\alpha| \, \|x\|$
  3. $\|x + y\| \leqslant \|x\| + \|y\|$ (triangle inequality)
- **Euclidean norm (2-norm)**: The standard Euclidean distance. For $x \in \mathbb{R}^n$,
$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}.$$
- **Manhattan norm (1-norm)**: Denoted L^1, and also called **Taxicab norm**. For $x \in \mathbb{R}^n$,
$$\|x\|_1 = |x_1| + |x_2| + ... + |x_n| \, .$$
- L**-Infinity (max) norm (∞-norm)**: Denoted L^∞. for $x \in \mathbb{R}^n$,
$$\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i| = \max\{|x_1| \, , |x_2| \, , ..., |x_n|\}.$$
- **Unit balls of norms in $\mathbb{R}^2$**:



  - **2-norm**: $B_{\|x\|_2}(0,1) = \{x \in \mathbb{R}^n : \; x_1^2 + x_2^2 \leqslant 1\}$
  - **1-norm**: $B_{\|x\|_1}(0,1) = \{x \in \mathbb{R}^n : \; |x_1| + |x_2| \leqslant 1\}$
  - **∞-norm**: $B_{\|x\|_\infty}(0,1) = \{x \in \mathbb{R}^n : \; \max_{1 \leqslant i \leqslant 2} |x_i| \leqslant 1\}$
- p**-norm**: In $\mathbb{R}^n$, A more general norm is
$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} = (|x_1^p| + |x_2|^p + ... + |x_n|^p)^{\frac{1}{p}}$$
for $1 \leqslant p < \infty$. The general p-norm satisfies all three properties of a norm only when $p \geqslant 1$. For smaller p, the triangle inequality does not hold.
- **Entrywise (Bad) Matrix-norms**: Consider the isomorphism $\phi : \; \mathbb{R}^{n \times n} \to \mathbb{R}^{n \cdot n}$. For example,
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

This way, we can use our vector norms defined above on matrices. The norms we have seen so far would be

$$\|A\|_p = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^p \right)^{\frac{1}{p}},$$
$$\|A\|_1 = \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|,$$
$$\|A\|_2 = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}}.$$

Note: The matrix 2-norm $\|A\|_2$ is also called the *Frobenius* norm, denoted $\|A\|_F$.

We see that in the Frobenius norm, $\|I\|_F = \sqrt{n} \neq 1$. In general, we would like our matrix norms to have $\|I\| = 1$ for all dimensions, and to not grow as the dimension increases.

These entrywise norms treat the matrix as a big vector and ignore its action on other vectors.

- **Properties of matrix norms**: Matrix norms satisfy the three required properties of norms.
  1. $\|A\| = 0 \iff A = 0$
  2. $\|\alpha A\| = |\alpha| \|A\|$
  3. $\|A + B\| \leqslant \|A\| + \|B\|$ (Triangle inequality)
- **Induced (operator) matrix norms**: For all $A \in \mathbb{R}^{n \times m}$, we define

$$\|A\|_p := \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|y\|_p = 1} \|Ay\|_p .$$

- **Properties of induced matrix norms**
  - **Sub-multiplicativity**: $\|AB\|_p \leqslant \|A\|_p \|B\|_p$
  - **Consistency**: $\|Ax\|_p \leqslant \|A\|_p \|x\|_p$
  - **Normalization**: $\|I\|_p = 1$

  These are what entrywise ("flattened") norms lack.
- **Induced matrix norms special cases**:

p	**Name**	**Explicit formula**
1	Maximum column sum	$\|A\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{m} \|a_{ij}\|$
2	Spectral norm	$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$
∞	Maximum row sum	$\|A\|_\infty = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{n} \|a_{ij}\|$

- **Singular values**: For $A \in \mathbb{R}^{m\times n}$, its **singular values** are the numbers

$$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0,$$

defined as the square roots of the eigenvalues of A^TA

$$\sigma_i(A) = \sqrt{\lambda_i(A^TA)}.$$

If A has rank r, then there are r positive singular values. The rest are zero.

Singular values measure how much A stretches vectors

$$\sigma_1 = \max_{\|x\|_2=1} \|Ax\|_2$$
$$\sigma_r = \min_{\|x\|_2=1} \|Ax\|_2 .$$

  - σ_1 is the maximum expansion factor of A
  - σ_r is the minimum expansion factor

- **Singular values of A^{-1}**: The singular values of A^{-1} are the reciprocals of the singular values of A.

$$\sigma_1(A^{-1}) = \frac{1}{\sigma_1(A)}, \sigma_2(A^{-1}) = \frac{1}{\sigma_2(A)}, \cdots, \sigma_n(A^{-1}) = \frac{1}{\sigma_n(A)}.$$

- **Eigenvalues of A^TA**: Consider the singular values of A,

$$\sigma_i = \sqrt{\lambda_i(A^TA)}.$$

Thus,

$$\lambda_i(A^TA) = \sigma_i(A)^2.$$

- **Singular values of A^TA**:

$$\sigma_i(A^TA) = \sqrt{\lambda_i((A^TA)^T(A^TA))} = \sqrt{\lambda_i(A^TA)^2}.$$

But, to get the eigenvalues for the square of a matrix you square the eigenvalues for the matrix. So, $\lambda_i(A^TA)^2 = (\lambda_i(A^TA))^2$. Thus,

$$\sigma_i(A^TA) = \sqrt{(\lambda_i(A^TA))^2} = \lambda_i(A^TA) = \sigma_i(A)^2.$$

So, the singular values for A^TA are the squares of the singular values of A. Thus, the set of eigenvalues for A^TA is the same as the set of singular values for A^TA

- **Spectral norm**: We have

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \sigma_1 = \sqrt{\lambda_{\max}(A^TA)}.$$

Note: Since $\frac{1}{\sigma_n}$ is the largest singular value for A^{-1},

$$\left\|A^{-1}\right\|_2 = \frac{1}{\sigma_n} = \frac{1}{\sqrt{\lambda_{\min}}(A^TA)}.$$

- **Derived property of matrix norm**: For $A \in \mathbb{R}^{n\times n}$, we have

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} \geqslant \frac{\|Ax\|}{\|x\|}$$
$$\implies \|Ax\| \leqslant \|A\| \|x\| .$$

- **Cauchy Schwarz inequality for 2-norm (vector norm)**: states

$$\left|x^Ty\right| \leqslant \|x\|_2 \|y\|_2 .$$

Proof. Let $t \in \mathbb{R}$. We know that $0 \leqslant \|x + ty\|_2^2$. Recall that $x^T x = \|x\|_2^2 = \|x\|_2 \|x\|_2$.

We have

$$\begin{aligned} 0 \leqslant (x+ty)^T(x+ty) &= x^T x + x^T ty + ty^T x + t^2 y^t y \\ &= \|x\|_2^2 + 2t(x^T y) + t^2 \|y\|_2^2 . \end{aligned}$$

Observe that this is a 2-degree polynomial in t, call it $p_2(t)$.

$$p_2(t) = \|y\|_2^2 t^2 + 2(x^t y)t + \|x\|_2^2 \geqslant 0.$$

Since $p_2(t)$ is greater than or equal to zero, we know that the discriminant is less than or equal to zero. That is, $p_2(t) \geqslant 0$ implies $D \leqslant 0$, where $D = (2(x^T y))^2 - 4(\|y\|_2^2)(\|x\|_2^2)$. Thus,

$$\begin{aligned} (2(x^T y))^2 - 4(\|y\|_2^2)(\|x\|_2^2) &\leqslant 0 \\ \implies 4(x^T y)^2 - 4(\|y\|_2^2)(\|x\|_2^2) &\leqslant 0 \\ \implies (x^T y)^2 - (\|y\|_2^2)(\|x\|_2^2) &\leqslant 0 \\ \implies (x^T y)^2 &\leqslant \|y\|_2^2 \|x\|_2^2 \\ \implies \left|x^T y\right| &\leqslant \|x\|_2 \|y\|_2 . \end{aligned}$$

This property of vector norms also holds for norms induced by an inner product. In an inner product space with norm $\|\cdot\|_P$ is induced by an inner product $\langle x, y\rangle$ if $\langle x, x\rangle = \|x\|_P^2$

6.4.2 Condition number

- **Numerical error when solving systems, residual vector**: Suppose we want to solve a linear system $Ax = b$. In practice, due to floating-point roundoff and the large number of flops required for a big system, the computed solution $\bar{x}$ will generally not satisfy $Ax = b$ exactly.

  We define the **residual** as

  $$r = b - A\bar{x},$$

  which measures how far $\bar{x}$ is from being an exact solution. If $\bar{x}$ is a good approximation, then $r \approx 0$.

- **Iterative approach to improve** $\hat{x}$: Suppose for a system $Ax = b$ numerical techniques yields an approximation $\hat{x}_1$. Then, the residual vector $\hat{r}_1 = b - A\hat{x}_1$, which implies that $b = \hat{r}_1 + A\hat{x}_1$. If $\hat{x}_2$ is a different approximation, where $\hat{x}_2 = \hat{x}_1 + \delta\hat{x}_1$, then

  $$\begin{aligned} & A\hat{x}_2 = b = \hat{r}_1 + A\hat{x}_1 \\ \implies & A(\hat{x}_1 + \delta\hat{x}_1) = \hat{r}_1 + A\hat{x}_1 \\ \implies & A\hat{x}_1 + A\delta\hat{x}_1 = \hat{r}_1 + A\hat{x}_1 \\ \implies & A\delta\hat{x}_1 = \hat{r}_1. \end{aligned}$$

  So, we solve the system for $\delta\hat{x}_1$. Then, since $\hat{x}_2 = \hat{x}_1 + \delta\hat{x}_1$, we see that we need to update the first solution by adding the computed $\delta\hat{x}_1$.

  In general, if $\hat{x}_i$ is the i^{th} numerical solution to $Ax = b$, and $\hat{r}_i$ is the residual vector to the i^{th} solution, then

  $$\hat{x}_{i+1} = \hat{x}_i + A^{-1}\hat{r}_i.$$

  In practice, we don't compute $A^{-1}\hat{r}_i$, as we know that this is an expensive task. Instead, we solve the system correction system

  $$A\delta\hat{x}_i = \hat{r}_i.$$

  In exact arithmetic,

  $$\begin{aligned} & A\delta\hat{x} = b - A\hat{x} \\ \implies & \delta\hat{x} = A^{-1}\left(b - A\hat{x}\right) \\ & \quad = A^{-1}b - A^{-1}A\hat{x} \\ & \quad = x - \hat{x}. \end{aligned}$$

  So,

  $$\hat{x}_{\text{new}} = \hat{x} + \delta\hat{x} = \hat{x} + x - \hat{x} = x.$$

  Thus, in exact arithmetic, we converge to the true solution in one step.

  In practice, computations are done in floating-point arithmetic, so both the residual and the correction are computed approximately. Let

  $$A\,\delta\hat{x} = r + \delta r,$$

where δr represents rounding or truncation errors. When we update

$$\hat{x}_{\text{new}} = \hat{x} + \delta\hat{x},$$

we hope that the new residual

$$r_{\text{new}} = b - A\hat{x}_{\text{new}}$$

is smaller than the previous residual. Each iteration ideally improves the approximation because

$$\delta\hat{x} \approx A^{-1}(b - A\hat{x}) = x - \hat{x}.$$

When floating-point errors are small enough relative to the conditioning of A, this correction moves $\hat{x}$ closer to x. But, if A is ill-conditioned, the corrections may no longer reduce the error — in fact, they can make it worse.

- **Intro to measuring solutions**: Consider a problem (P), where

$$(P): \ Ax = b.$$

Numerical techniques yields a solution $\hat{x}$, which may or may not be the true solution to (P). Let x be the true solution to the system. So, x solves $Ax = b$.

We want to measure the distance between the numerical solution $\hat{x}$ and the true solution x, we hope that the numerical solution $\hat{x}$ is close to x. If the distance is small, then $\hat{x}$ is a good solution.

- **Relative error**: The relative error in $\hat{x}$ is given by

$$\frac{\|\hat{x} - x\|}{\|x\|} = \frac{\|\delta x\|}{\|x\|}$$

where $\hat{x} = x + \delta x$, which implies $x = \hat{x} - \delta x$.

- **Perturbation**: If numerical methods to solve a linear system $Ax = b$ yields $\hat{x}$, then $\hat{x}$ solves $\hat{A}\hat{x} = \hat{b}$. Note that it is possible for $\hat{A} = A$ or $\hat{b} = b$. If both $\hat{A} = A$ and $\hat{b} = b$, then $\hat{x} = x$.

$\hat{A}$ and $\hat{b}$ are called perturbed if they are modified versions of the original. If $\hat{A}$ is a perturbed matrix A, and $\hat{b}$ is a perturbed vector b, then

$$\hat{A} = A + \delta A,$$
$$\hat{b} = b + \delta b.$$

- **Perturbing** b: We can perturb b, but not A such that $\hat{x}$ solves $A\hat{x} = \hat{b}$.

Recall that the residual vector is $\hat{r} = b - A\hat{x}$. If $\hat{b} = A\hat{x}$, then

$$\hat{b} = A\hat{x} = A\hat{x} - b + b = b - b + A\hat{x} = b - (b - A\hat{x}) = b - \hat{r}.$$

Note that in this case, $\hat{A} = A$. We can quantify the change in b by observing that since $\hat{b} = b + \delta b$, and $\hat{b} = b - \hat{r}$, we have

$$b - \hat{r} = b + \delta b$$
$$\implies -\hat{r} = \delta b.$$

- **Condition number**: We wish to find an upper bound for the relative error in x, $\frac{\|\delta x\|}{\|x\|}$. We have the two systems,

$$Ax = b, \quad A\hat{x} = \hat{b}.$$

Thus, we have

$$Ax = b\,, \tag{1}$$
$$A(x + \delta x) = b + \delta b\,. \tag{2}$$

Looking at (1), we see

$$Ax = b \implies \|b\| = \|Ax\|\,.$$

But, by the Cauchy Schwarz inequality, $\|b\| \leqslant \|A\|\,\|x\|$. So,

$$\|b\| \leqslant \|A\|\,\|x\|\,. \tag{1}$$

Looking at (2), we see

$$\begin{aligned} & A(x + \delta x) = b + \delta b \\ \implies\ & Ax + A\delta x = b + \delta b \\ \implies\ & A\delta x = \delta b \\ \implies\ & \delta x = A^{-1}\delta b \\ \implies\ & \|\delta x\| = \left\|A^{-1}\delta b\right\| \\ \implies\ & \|\delta x\| \leqslant \left\|A^{-1}\right\|\,\|\delta b\|\,. \end{aligned} \tag{2}$$

Notice that we can setup (1) so that dividing (2) by (1) gives the relative error in x on the left, and relative error of b on the right. So,

$$\frac{1}{\|x\|} \leqslant \frac{\|A\|}{\|b\|}.$$

Now, we divide (2) by (1), we have

$$\frac{\|\delta x\|}{\|x\|} \leqslant \left\|A^{-1}\right\|\,\|A\|\,\frac{\|\delta b\|}{\|b\|}.$$

We now have the relative error in the numerical solution bounded above by the relative error of b times some constant $\left\|A^{-1}\right\|\,\|A\|$, we call this constant the condition number $\kappa(A)$. That is,

$$\kappa(A) = \left\|A^{-1}\right\|\,\|A\|\,.$$

The condition number of a matrix A measures how sensitive the solution of a linear system $Ax = b$ is to small changes in b (or in A).

We see that as $\kappa(A) \to \infty$, the relative error in x grows without bound.

- **Condition number ($\kappa_2(A)$) in terms of singular values**: Recall that

$$\|A\|_2 = \sigma_1 = \sigma_{\max},$$
$$\left\|A^{-1}\right\|_2 = \frac{1}{\sigma_n} = \frac{1}{\sigma_{\min}}.$$

So,

$$\kappa_2(A) = \left\|A^{-1}\right\|_2\,\|A\|_2 = \sigma_{\max} \cdot \frac{1}{\sigma_{\min}} = \frac{\sigma_{\max}}{\sigma_{\min}}.$$

- **Properties of the condition number**: Let A be a matrix, and $\kappa(A)$ be the condition number that measures the system $Ax = b$. The following two properties hold

  1. $\kappa(A) \geqslant 1$
  2. $\kappa(I) = 1$
  3. $\kappa(A) = \kappa(A^{-1})$

  Proof (1): $\kappa(A) = \left\|A^{-1}\right\| \|A\|$. By Cauchy Schwarz,

$$\begin{aligned} \left\|A^{-1}A\right\| &\leqslant \left\|A^{-1}\right\| \|A\| \\ \implies \|I\| &\leqslant \left\|A^{-1}\right\| \|A\| \\ \implies 1 &\leqslant \left\|A^{-1}\right\| \|A\| = \kappa(A). \end{aligned}$$

  ■

  Proof (2):

$$\kappa(I) = \left\|I^{-1}\right\| \|I\| = \|I\| \|I\| = 1 \cdot 1 = 1.$$

  ■

  Proof (3):

$$\kappa(A^{-1}) = \left\|\left(A^{-1}\right)^{-1}\right\| \left\|A^{-1}\right\| = \left\|A^{-1}\right\| \|A\| = \kappa(A).$$

  ■

- **Theorem *(Relative Error Bound I)***: Let A be nonsingular, $b \neq 0$, and $Ax = b$. If $A(x + \delta x) = b + \delta b$, then

$$\frac{\|\delta x\|}{\|x\|} \leqslant \kappa(A) \frac{\|\delta b\|}{\|b\|}$$

  where $\kappa(A) = \left\|A^{-1}\right\| \|A\|$.

  Proof. Suppose $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$, $b \neq 0$, $Ax = b$, and $A(x + \delta x) = b + \delta b$. Then,

$$\begin{aligned} & A(x + \delta x) = b + \delta b \\ \implies & Ax + A\delta x = b + \delta b \\ \implies & A\delta x = \delta b \\ \implies & \delta x = A^{-1}\delta b \\ \implies & \|\delta x\| = \left\|A^{-1}\delta b\right\| \leqslant \left\|A^{-1}\right\| \|\delta b\|. \end{aligned} \tag{1}$$

  But, since $Ax = b$,

$$\begin{aligned} & Ax = b \\ \implies & \|Ax\| = \|b\| \leqslant \|A\| \|x\| \\ \implies & \|x\| \geqslant \frac{\|b\|}{\|A\|}. \end{aligned} \tag{2}$$

  Dividing (1) by (2) gives

$$\frac{\|\delta x\|}{\|x\|} \leqslant \left\|A^{-1}\right\| \|A\| \frac{\|\delta b\|}{\|b\|} = \kappa(A) \frac{\|\delta b\|}{\|b\|}.$$

Note: Recall that if $a, b, c, d \in \mathbb{R}$, $0 \leqslant a \leqslant b$, and $0 < d \leqslant c$, then

$$\frac{a}{c} \leqslant \frac{b}{d}.$$

Consequences.

1. If $\kappa(A)$ small, the relative error in x is small.
2. If $\kappa(A)$ is large, then it is possible to have the relative error in b small, but the relative error in x large.

- **Well-conditioned and ill-conditioned in terms of** $\kappa(A)$: If $\kappa(A)$ is large, then (P) is ill-conditioned. If $\kappa(A)$ is small (close to one), then (P) is well-conditioned.
- **Pre-conditioned system (preconditioner)**: If the system

$$Ax = b.$$

has a large condition number $\kappa(A)$, the system is ill-conditioned, meaning small perturbations in b cause large changes in x.

To improve this, we introduce a preconditioner B (sometimes written M^{-1}) such that $B \approx A^{-1}$ but is much easier to compute or apply.

We then multiply the equation by B on the left

$$BAx = Bb.$$

Define

$$\tilde{A} = BA, \quad \tilde{b} = Bb.$$

This gives the **preconditioned system**

$$\tilde{A}x = \tilde{b}.$$

We want to choose B so that

1. $\kappa(\tilde{A}) = \kappa(BA) \ll \kappa(A)$, i.e., the new system is better conditioned,
2. Solving $Bz = y$ is cheap

We don't usually form BA explicitly, it usually destroys the structure and sparsity that make the original system efficient to handle.

If A and B are large sparse matrices (which they almost always are in practice), explicitly multiplying them gives a dense matrix BA

That means:

1. much higher memory usage,
2. much slower matrix–vector products,
3. and loss of efficiency in iterative methods.

Forming BA explicitly also introduces round-off errors, especially when B approximates A^{-1} You end up multiplying two ill-conditioned matrices, potentially worsening accuracy before solving anything.

If:

- A is **moderate in size** (not huge, maybe $n \lesssim 10^3$),
- A and B are **dense** anyway (so sparsity isn't being destroyed), and
- you can **compute or approximate** B reliably and cheaply,

then **explicitly forming** BA may sometimes be beneficial.

For example:

- in small to medium dense problems (common in computational linear algebra, not large PDEs),
- or when using **direct solvers** (like LU or QR), where forming BA once is acceptable,
- or when B comes from a **stabilizing transformation** (e.g., scaling rows or columns).

In such cases, if BA has a much smaller condition number than A,

$$\kappa(BA) \ll \kappa(A),$$

then solving

$$BAx = Bb$$

can indeed give a **more accurate and stable solution** than directly solving

$$Ax = b.$$

Usually, in **large or iterative** problems:

- Forming BA destroys **structure** (sparsity, bandedness, symmetry).
- You have to store an entire new matrix (cost $O(n^2)$ memory).
- The cost of computing BA can exceed the cost of solving $Ax = b$ itself.
- Rounding errors during multiplication can **degrade** the benefits of preconditioning.

In these cases, applying B as an **operator** (by solving $Bz = y$ inside each iteration) gives the same effect with **less cost and better numerical behavior**.

If we let $B = A^{-1}$, then the system becomes

$$Ix = A^{-1}b$$

where $\tilde{A} = A^{-1}A = I$, $\tilde{b} = A^{-1}b = x$, and $\kappa(\tilde{A}) = \kappa(I) = 1$. If $\tilde{A} = I$, then

1. **The condition number is ideal:**

   $$\kappa(\tilde{A}) = \kappa(I) = 1.$$

   This is the *best possible conditioning*—there is no amplification of errors at all.
2. **The system becomes trivial:**

   $$Ix = x = \tilde{b}.$$

   You have effectively solved the problem in one step.

3. **Interpretation:**
   - $B = A^{-1}$ is the *perfect preconditioner*.
   - In practice, we cannot use it, because computing A^{-1} explicitly is just as expensive (and less stable) than solving $Ax = b$ directly.
4. **Practical takeaway:** Preconditioning aims to **approximate this ideal case**:

$$B \approx A^{-1},$$

   such that

$$BA \approx I, \qquad \text{and hence} \qquad \kappa(BA) \approx 1,$$

   but without the full cost of inverting A.

When we say

$$\kappa(A) \gg 1,$$

we mean the **mathematical problem** $Ax = b$ is *ill-conditioned*—small perturbations in b cause large changes in x.

If we form

$$\tilde{A} = A^{-1}A = I,$$

then the **preconditioned system** has

$$\kappa(\tilde{A}) = 1,$$

so that transformed system is *perfectly conditioned*. However, to achieve this we would have to **compute** A^{-1} **numerically**, and that step is where the *instability* arises.

- **Numerical stability vs conditioning**
  - **Conditioning**
    - A property of the **mathematical problem** itself.
    - Measures the **sensitivity** of the true solution to small input changes.
    - Example: if $Ax = b$, then

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A)\frac{\|\delta b\|}{\|b\|}.$$

      A large $\kappa(A)$ indicates an **ill-conditioned problem**.
  - **Numerical Stability**
    - A property of the **algorithm** used to solve the problem.
    - Measures how much **round-off and truncation errors** the algorithm introduces or amplifies.
    - A **stable algorithm** gives the exact solution to a *nearby problem*:

$$(A + \delta A)\hat{x} = b + \delta b, \quad \|\delta A\|,\ \|\delta b\| \text{ small.}$$

Note: If an algorithm is **numerically unstable**, then the perturbations δA and/or δb required to explain its result might be **large**:

$$\|(A + \delta A) - A\| \text{ is not small.}$$

Hence, the computed $\hat{x}$ is the exact solution to a *far-away problem*:

$$(A + \delta A)\hat{x} = b + \delta b,$$

where $\|\delta A\|$ or $\|\delta b\|$ are no longer small compared to $\|A\|$ or $\|b\|$.

This means the algorithm's result may not correspond meaningfully to the original system at all.

- **Perturbing A but not b**: Suppose A nonsingular, and $Ax = b$ yields a numerical solution $\hat{x}$. Then, $\hat{x}$ solves $\hat{A}\hat{x} = b$, where

$$\hat{A} = A + \delta A.$$

But, this is only when $\hat{A}$ nonsingular. Let's suppose for a moment that $\hat{A}$ is singular. Then, there exists a $y \neq 0$ such that

$$(A + \delta A)y = 0.$$

This implies

$$\begin{aligned}
Ay + \delta Ay &= 0 \\
\implies Ay &= -\delta Ay \\
\implies y &= -A^{-1}\delta Ay \\
\implies \|y\| &= \left\|-A^{-1}\delta Ay\right\| \\
&= |-1| \left\|A^{-1}\delta Ay\right\| \\
&= \left\|A^{-1}\delta Ay\right\| \\
&\leqslant \left\|A^{-1}\right\| \|\delta Ay\| \\
&\leqslant \left\|A^{-1}\right\| \|\delta A\| \|y\| \\
\implies 1 &\leqslant \left\|A^{-1}\right\| \|\delta A\| \\
\implies \|A\| &\leqslant \|A\| \left\|A^{-1}\right\| \|\delta A\| \\
\implies \|A\| &\leqslant \kappa(A) \|\delta A\| \\
\implies \frac{\|\delta A\|}{\|A\|} &\geqslant \frac{1}{\kappa(A)}.
\end{aligned}$$

So, we have the following theorem

- **Theorem (*Singularity of perturbed* A)**: If

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)}$$

then $A + \delta A$ is nonsingular.

  - A large $\kappa(A)$ means A is **ill-conditioned** — even small perturbations can make it singular.
  - A small $\kappa(A)$ (close to 1) means A is **well-conditioned** — it can tolerate relatively large perturbations without becoming singular.

If $\kappa(A) = 1$, then

$$\frac{\|\delta A\|}{\|A\|} < 1,$$

meaning the perturbation can be as large as the matrix itself before singularity is possible.

This inequality defines a **ball in matrix space** with center A and radius $\|A\| / \kappa(A)$, consistency of nonsingular matrices. The inequality tells us that all matrices within that ball are guaranteed to be nonsingular. All matrices outside that ball (and on the boundary) are singular.

This ball is called a **matrix norm ball** (or **ball in matrix space**), and is the set

$$\mathcal{B}(A, r) = \{A + \delta A : \|\delta A\| < r\},$$

where $r = \frac{\|A\|}{\kappa(A)}$. We can call this specific ball the **Ball of guaranteed nonsingularity**, or the **Neighborhood of nonsingularity**.

Proof. See argument above, the converse reveals the theorem.

- **Theorem *(Relative error bound II)***: Let A be nonsingular, $b \neq 0$, and $Ax = b$. If $(A + \delta A)(x + \delta x) = b$, and

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)},$$

then

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{\kappa(A)\frac{\|\delta A\|}{\|A\|}}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}}.$$

Proof. Suppose for a moment that A nonsingular, $b \neq 0$, $Ax = b$, and $(A + \delta A)(x + \delta x) = b$. Then, it's easy to see that

$$\begin{aligned}
&(A + \delta A)(x + \delta x) = b \\
\implies\ & Ax + A\delta x + \delta Ax + \delta A\delta x = b \\
\implies\ & A\delta x + \delta Ax + \delta A\delta x = 0 \\
\implies\ & A\delta x = -\left(\delta Ax + \delta A\delta x\right) \\
\implies\ & A\delta x = -\delta A(x + \delta x) \\
\implies\ & \delta x = -A^{-1}\delta A(x + \delta x) \\
\implies\ & \|\delta x\| = \left\|-A^{-1}\delta A(x + \delta x)\right\| \\
& \qquad = \left\|A^{-1}\delta A(x + \delta x)\right\| \\
\implies\ & \|\delta x\| \leqslant \left\|A^{-1}\right\| \|\delta A\| \|x + \delta x\| \quad \text{(Cauchy-Schwarz)} \\
& \qquad \leqslant \left\|A^{-1}\right\| \|\delta A\| \left(\|x\| + \|\delta x\|\right) \quad \text{(Triangle inequality)} \\
& \qquad \leqslant \left\|A^{-1}\right\| \|\delta A\| \|x\| + \left\|A^{-1}\right\| \|\delta A\| \|\delta x\| \\
\implies\ & \|\delta x\| - \left\|A^{-1}\right\| \|\delta A\| \|\delta x\| \leqslant \left\|A^{-1}\right\| \|\delta A\| \|x\| \\
\implies\ & \|\delta x\| \left(1 - \left\|A^{-1}\right\| \|\delta A\|\right) \leqslant \left\|A^{-1}\right\| \|\delta A\| \|x\| \\
\implies\ & \frac{\|A\|}{\|A\|} \|\delta x\| \left(1 - \left\|A^{-1}\right\| \|\delta A\|\right) \leqslant \frac{\|A\|}{\|A\|} \left\|A^{-1}\right\| \|\delta A\| \|x\| \\
\implies\ & \|\delta x\| \left(1 - \frac{\|A\|}{\|A\|} \left\|A^{-1}\right\| \|\delta A\|\right) \leqslant \|A\| \left\|A^{-1}\right\| \frac{\|\delta A\|}{\|A\|} \|x\| \\
\implies\ & \|\delta x\| \left(1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}\right) \leqslant \kappa(A)\frac{\|\delta A\|}{\|A\|} \|x\| \\
& \therefore \frac{\|\delta x\|}{\|x\|} \leqslant \frac{\kappa(A)\frac{\|\delta A\|}{\|A\|}}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}}
\end{aligned}$$

as desired. ■

- **Theorem *(Relative error bound III)***: Let A be nonsingular, $b \neq 0$, and $Ax = b$. If $(A + \delta A)(x + \delta x) = b + \delta b$, and

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)},$$

then

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{\kappa(A)\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\right)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}.$$

Remark (a)**.** Consider the quantity $q = \frac{a}{b}$, $b \neq 0$. If $c \leqslant b$, then

$$q = \frac{a}{b} \leqslant \frac{a}{c}.$$

Proof. Suppose for a moment that A nonsingular, $b \neq 0$, $Ax = b$, and $(A + \delta A)(x + \delta x) = b + \delta b$. Then, it's immediately obvious that

$$\begin{aligned}
(A+\delta A)(x+\delta x) &= b+\delta b \\
\implies Ax + A\delta x + \delta A x + \delta A \delta x &= b+\delta b \\
\implies A\delta x + \delta A x + \delta A \delta x &= \delta b \\
\implies A\delta x &= \delta b - (\delta A x + \delta A \delta x) \\
\implies A\delta x &= \delta b - \delta A\,(x+\delta x) \\
\implies \delta x &= A^{-1}\delta b - A^{-1}\delta A\,(x+\delta x) \\
\implies \|\delta x\| &= \left\|A^{-1}\delta b - A^{-1}\delta A\,(x+\delta x)\right\| \\
&\leqslant \left\|A^{-1}\delta b\right\| + \left\|-A^{-1}\delta A\,(x+\delta x)\right\| \\
&= \left\|A^{-1}\delta b\right\| + \left\|A^{-1}\delta A\,(x+\delta x)\right\| \\
&\leqslant \left\|A^{-1}\right\|\|\delta b\| + \left\|A^{-1}\right\|\|\delta A\|\,(\|x+\delta x\|) \\
&\leqslant \left\|A^{-1}\right\|\|\delta b\| + \left\|A^{-1}\right\|\|\delta A\|\,(\|x\|+\|\delta x\|) \\
&= \left\|A^{-1}\right\|\|\delta b\| + \left\|A^{-1}\right\|\|\delta A\|\|x\| + \left\|A^{-1}\right\|\|\delta A\|\|\delta x\| \\
\implies \|\delta x\| - \left\|A^{-1}\right\|\|\delta A\|\|\delta x\| &\leqslant \left\|A^{-1}\right\|\|\delta b\| + \left\|A^{-1}\right\|\|\delta A\|\|x\| \\
\implies \|\delta x\|\left(1-\left\|A^{-1}\right\|\|\delta A\|\right) &\leqslant \left\|A^{-1}\right\|\|\delta b\| + \left\|A^{-1}\right\|\|\delta A\|\|x\| \\
\implies \frac{\|A\|}{\|A\|}\|\delta x\|\left(1-\left\|A^{-1}\right\|\|\delta A\|\right) &\leqslant \frac{\|A\|}{\|A\|}\left(\left\|A^{-1}\right\|\|\delta b\| + \left\|A^{-1}\right\|\|\delta A\|\|x\|\right) \\
\implies \|\delta x\|\left(1-\|A\|\left\|A^{-1}\right\|\frac{\|\delta A\|}{\|A\|}\right) &\leqslant \|A\|\left\|A^{-1}\right\|\frac{\|\delta b\|}{\|A\|} + \|A\|\left\|A^{-1}\right\|\frac{\|\delta A\|}{\|A\|}\|x\| \\
\implies \|\delta x\|\left(1-\kappa(A)\frac{\|\delta A\|}{\|A\|}\right) &\leqslant \kappa(A)\frac{\|\delta b\|}{\|A\|} + \kappa(A)\frac{\|\delta A\|}{\|A\|}\|x\| \\
\implies \frac{\|\delta x\|}{\|x\|}\left(1-\kappa(A)\frac{\|\delta A\|}{\|A\|}\right) &\leqslant \kappa(A)\frac{\|\delta b\|}{\|A\|\,\|x\|} + \kappa(A)\frac{\|\delta A\|}{\|A\|}.
\end{aligned}$$

From here, we use

$$\begin{aligned}
Ax &= b \\
\implies \|Ax\| &= \|b\| \\
\implies \|Ax\| &\leqslant \|A\|\,\|x\| \\
\implies \|b\| &\leqslant \|A\|\,\|x\|
\end{aligned}$$

and remark *(a)* to see that

$$\kappa(A)\frac{\|\delta b\|}{\|A\|\,\|x\|} + \kappa(A)\frac{\|\delta A\|}{\|A\|} \leqslant \kappa(A)\frac{\|\delta b\|}{\|b\|} + \kappa(A)\frac{\|\delta A\|}{\|A\|}.$$

Thus, it follows that

$$\begin{aligned}
\frac{\|\delta x\|}{\|x\|}\left(1-\kappa(A)\frac{\|\delta A\|}{\|A\|}\right) &\leqslant \kappa(A)\frac{\|\delta b\|}{\|A\|\,\|x\|}+\kappa(A)\frac{\|\delta A\|}{\|A\|}\\
&\leqslant \kappa(A)\frac{\|\delta b\|}{\|b\|}+\kappa(A)\frac{\|\delta A\|}{\|A\|}\\
&= \kappa(A)\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right)\\
\therefore \frac{\|\delta x\|}{\|x\|} &\leqslant \frac{\kappa(A)\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\right)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}.
\end{aligned}$$

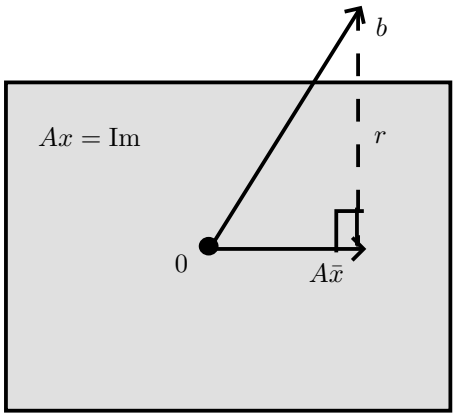
As desired. ■

6.5 The least squares problem and orthogonal matrices

6.5.1 The discrete least squares problem and orthogonal matrices

- **Solving systems with no solution**: Suppose that $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, where $m > n$. When we go to solve $Ax = b$, it could be that there is no such $x \in \mathbb{R}^n$ such that $Ax = b$ for a given b. In this case, we know that the linear map $L : \mathbb{R}^n \to \mathbb{R}^n$ is not surjective.
- **Over-determined systems**: If a system $Ax = b$ has more equations than unknowns ($m > n$), we call the system **over-determined**
- **under-determined systems**: If a system $Ax = b$ has less equations than unknowns ($m < n$), we call the system **under-determined**
- **Determined system**: If a system $Ax = b$ has the same number of equations as unknowns ($m = n$), we call the system **determined**.
- **Well-determined and degenerate**: If a system $Ax = b$ has a unique solution for a given b, then the system is said to be **well-determined**. If the system has no solution or infinitely many, the system is **degenerate**.
- **Geometric interpretation**: Suppose that $A \in \mathbb{R}^{3\times 2}$, so $L : \mathbb{R}^2 \to \mathbb{R}^3$. We know that L cannot be surjective, so the image of L is some proper subset of the codomain $\mathbb{R}^3$. Suppose that $b \notin \text{Im}(L)$

  If we try to solve this system for x, we will find that there is no solution. What can we do? We can project b down onto the image, now we have a vector $\bar{b} \in \text{Im}(L)$, and we can solve the system.



  Notice that $r = b = A\bar{x}$.

- **The discrete least squares problem**: Let $A \in \mathbb{R}^{m\times n}$, where $m \geqslant n$, and $b \in \mathbb{R}^m$. The discrete least squares problem is finding

$$\min_{x\in\mathbb{R}^n} \|r\|_2^2,$$

  where $r = b - Ax$. If $b \notin \text{Im}(L) = \text{col}(A)$, then no solution to $Ax = b$ exists, and we can instead solve the discrete least squares problem to get the best approximation.

Recall that since the closest point in a subspace to a vector is its orthogonal projection, minimizing $\|r\|_2^2$ is equivalent to finding the projection of b onto the column space of A.

- **Data fitting problem**: Suppose we have a function $y(t) = 1+e^t+3e^{-t}$ that generates m points

$$y(t_1),\ y(t_2),\ \ldots,\ y(t_m) = (t_1, y_1),\ (t_2, y_2),\ \ldots,\ (t_m, y_m).$$

Now, suppose we introduce some noise to our points,

$$\tilde{y_i} = y(t_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

Given just the noisey data, can we recover the function? If we suspect or somehow find out that the function has the form

$$x_1(1) + x_2(e^t) + x_3(e^{-t}),$$

so a linear combination of $\{1, e^t, e^{-t}\}$, then $\tilde{y}(t) = x_1(1) + x_2(e^t) + x_3(e^{-t})$. Using the noisey points, we have

$$\begin{pmatrix} 1 & e^{t_i} & e^{-t_i} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \tilde{y}_i,$$

which gives the system of linear equations

$$\begin{pmatrix} 1 & e^{t_1} & e^{-t_1} & \\ 1 & e^{t_2} & e^{-t_2} & \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{t_m} & e^{-t_m} & \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_m \end{pmatrix}.$$

Thus, an over-determined system.

- **Into to QR factorization**: Given a matrix $A \in \mathbb{R}^{m\times n}$, we want to factor A into QR, so that $A = QR$, where $Q \in \mathbb{R}^{m\times m}$ is orthogonal ($QQ^T = I$), and $R \in \mathbb{R}^{m\times n}$ is upper triangular.

$R \in \mathbb{R}^{m\times n}$ is upper triangular if we can split R as follows,

$$R = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix},$$

where $\hat{R} \in \mathbb{R}^{n\times n}$, and is upper triangular. Observe that the bottom half has size $m - n \times n$, and is zero.

- **Orthogonal matrices**: A matrix $Q \in \mathbb{R}^{n\times n}$ is orthogonal if $QQ^T = Q^TQ = I$, so $Q^T = Q^{-1}$.

An orthogonal matrix is a matrix whose columns form a set of orthonormal vectors, each has length one and is perpendicular to the others.

Suppose $\{q_1, q_2, \ldots, q_n\}$ form a set of orthonormal vectors. Let $Q \in \mathbb{R}^{n\times n}$ be the matrix whose columns are the vectors q_i, then

$$Q = \begin{bmatrix} | & | & & | \\ q_1 & q_2 & \vdots & q_n \\ | & | & & | \end{bmatrix}, \; Q^T = \begin{bmatrix} - & q_1^T & - \\ - & q_2^T & - \\ & \cdots & \\ - & q_n^T & - \end{bmatrix}.$$

So, we see that in Q^TQ,

$$\begin{cases} q_iq_j = 0 & \text{if } i \neq j \\ q_iq_j = 1 & \text{if } i = j \end{cases}.$$

Thus,

$$Q^TQ = \begin{bmatrix} q_1^Tq_1 & q_1^Tq_2 & \cdots & q_1^Tq_n \\ q_2^Tq_1 & q_2^Tq_2 & \cdots & q_2^Tq_n \\ \vdots & \vdots & \ddots & \ddots \\ q_n^Tq_1 & q_n^Tq_2 & \cdots & q_n^Tq_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I.$$

Thus, $Q^TQ = I$, so $Q^{-1} = Q^T$. Also,

$$\det(Q^TQ) = \det(I) = 1 \implies \det(Q^T)\det(Q) = 1 \implies \det(Q)^2 = 1$$
$$\therefore \det(Q) = \pm 1.$$

Note: For square matrices, the terms "orthogonal matrix" and "orthonormal matrix" are used interchangeably.

- **Definition of orthogonal matrices**: $Q \in \mathbb{R}^{n\times n}$ is orthogonal if the columns of Q satisfy
  1. $\|q_i\| = 1$ for $i = 1, 2, ..., n$
  2. $\langle q_i, q_j\rangle = 0$ if $i \neq j$
- **Properties of orthogonal matrices**
  1. $Q^TQ = QQ^T = I$
  2. $Q^{-1} = Q^T$
  3. $\det(Q) = \pm 1$
- **Theorem**: If $Q \in \mathbb{R}^{n\times n}$ is orthogonal, then
  1. $\langle Qx, Qy\rangle = \langle x, y\rangle$
  2. $\|Qx\|_2 = \|x\|_2$

Proof. (1). Let $x, y \in \mathbb{R}^n$,

$$\langle Qx, Qy\rangle = (Qx)^T(Qy) = x^TQ^TQy = x^TIy = x^Ty = \langle x, y\rangle.$$

(2).

$$\|Qx\|_2^2 = \langle Qx, Qx\rangle^2 = \langle x, x\rangle = \|x\|_2^2.$$

Thus, $\|Qx\|_2 = \|x\|_2$ ■

Note: Because of the second property ($\|Qx\|_2 = \|x\|_2$), orthogonal matrices preserve length. Thus, they represent a **rotation** or a **reflection**.

- **Solving the discrete least squares problem (LSP) with the QR factorization**: The LSP has us finding

$$\min_{x \in \mathbb{R}^n} \|b - Ax\|_2^2 .$$

If we find the QR decomposition $A = QR$, then

$$\|b - Ax\|_2^2 = \|b - QRx\|_2^2 = \left\|QQ^T b - QRx\right\|_2^2 = \left\|Q(Q^T b - Rx)\right\|_2^2 .$$

Recall that

$$A \in \mathbb{R}^{m \times n},\ Q \in \mathbb{R}^{m \times m},\ R \in \mathbb{R}^{m \times n},\ x \in \mathbb{R}^n,\ b \in \mathbb{R}^m,$$

so $Q^T b \in \mathbb{R}^m$, $Rx \in \mathbb{R}^m$, and $Q^T b - Rx \in \mathbb{R}^m$. Thus,

$$\left\|Q(Q^T b - Rx)\right\|_2^2 = \left\|Q^T b - Rx\right\|_2^2 .$$

Since

$$R = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix}, \quad \hat{R} \in \mathbb{R}^{n \times n},$$

we have that

$$Rx = \begin{bmatrix} \hat{R}x \\ 0 \end{bmatrix}, \quad \hat{R}x \in \mathbb{R}^n.$$

Let $c = Q^T b \in \mathbb{R}^m$. Divide c into blocks

$$c = \begin{bmatrix} \hat{c} \\ \bar{c} \end{bmatrix}, \quad \hat{c} \in \mathbb{R}^n,\ \bar{c} \in \mathbb{R}^{m-n}.$$

So,

$$\left\|Q^T b - Rx\right\|_2^2 = \left\| \begin{bmatrix} \hat{c} - \hat{R}x \\ \bar{c} \end{bmatrix} \right\|_2^2 .$$

Remark. Let $x \in \mathbb{R}^{n+m}$, where

$$x = \begin{pmatrix} \hat{x} \\ \bar{x} \end{pmatrix}, \quad \hat{x} \in \mathbb{R}^n,\ \bar{x} \in \mathbb{R}^m.$$

Then,

$$\|x\|_2^2 = \|\hat{x}\|_2^2 + \|\bar{x}\|_2^2 .$$

Proof.

$$\|x\|_2^2 = \sum_{i=1}^{m+n} |x_i|^2 = \sum_{i=1}^{n} |x_i|^2 + \sum_{i=n+1}^{n+m} |x_i|^2 = \|\hat{x}\|_2^2 + \|\bar{x}\|_2^2 .$$

■

Thus,

$$\min_{x\in\mathbb{R}^n} \|b - Ax\|_2^2 = \min_{x\in\mathbb{R}^n} \left|\left|\begin{bmatrix}\hat{c} - \hat{R}x \\ \bar{c}\end{bmatrix}\right|\right|_2^2 = \min_{x\in\mathbb{R}^n} \left\|\hat{c} - \hat{R}x\right\|_2^2 + \|\bar{c}\|_2^2 .$$

Notice that $\|\bar{c}\|_2^2$ does not depend on x, so we don't need to consider it when finding the minimizer. Second, notice that $\left\|\hat{c} - \hat{R}x\right\|_2^2$ is minimized when its equal to zero, and since $\left\|\hat{c} - \hat{R}x\right\|_2^2 = 0 \iff \hat{c} - \hat{R}x = 0$, we find the minimizer by solving for x when $\hat{c} - \hat{R}x = 0$, which gives the system

$$\hat{R}x = \hat{c}.$$

Therefore,

$$\min_{x\in\mathbb{R}^n} \|b - Ax\|_2^2 \iff \hat{R}x = \hat{c}.$$

Note: We have

$$\min_{x\in\mathbb{R}^n} \|b - Ax\|_2^2 = \min_{x\in\mathbb{R}^n} \left\|\hat{c} - \hat{R}x\right\|_2^2 + \|\bar{c}\|_2^2 .$$

But, the minimum is achieved when $\hat{c} - \hat{R}x = 0$, so $\left\|\hat{c} - \hat{R}x\right\|_2^2 = 0$, and thus

$$\min_{x\in\mathbb{R}^n} \|b - Ax\| = \min_{x\in\mathbb{R}^n} \left\|\hat{c} - \hat{R}x\right\|_2^2 + \|\bar{c}\|_2^2 = 0 + \|\bar{c}\|_2^2 .$$

So, x is found by solving the system $\hat{R}x = \hat{c}$, and the square of the norm of the residual is precisely $\|\bar{c}\|_2^2$. So, $\bar{c}$ represents the residual

- **The residual in the discrete LSP**: Let $A \in \mathbb{R}^{m\times n}$, $m > n$, and $b \in \mathbb{R}^m$. If x^* solves the discrete least squares problem,

$$x^* = \arg\min_{x\in\mathbb{R}^n} \|b - Ax\|_2^2 ,$$

then the residual vector is given by $r = b - Ax^*$. Applying Q^T gives

$$\begin{aligned} Q^T r = Q^T b - Q^T A x^* = Q^T b - Q^T Q R x^* = Q^T b - R x^* = c - R x^* \\ = \begin{bmatrix}\hat{c}\\ \bar{c}\end{bmatrix} - \begin{bmatrix}\hat{R}\\ 0\end{bmatrix} x^* = \begin{bmatrix}\hat{c} - \hat{R}x^*\\ \bar{c}\end{bmatrix} . \end{aligned}$$

But, x^* solves $\hat{R}x = \hat{c}$, so $\hat{c} - \hat{R}x^* = 0$, so

$$Q^T r = \begin{bmatrix}\hat{c} - \hat{R}x^*\\ \bar{c}\end{bmatrix} = \begin{bmatrix}0\\ \bar{c}\end{bmatrix} .$$

Applying Q gives

$$r = Q \begin{bmatrix}0\\ \bar{c}\end{bmatrix}, \quad Q \in \mathbb{R}^{m\times m},\ 0 \in \mathbb{R}^n,\ \bar{c} \in \mathbb{R}^{m-n}.$$

If we partition Q as

$$Q = \begin{bmatrix}Q_1 & Q_2\end{bmatrix}, \quad Q_1 \in \mathbb{R}^{m\times n},\ Q_2 \in \mathbb{R}^{m\times(m-n)},$$

then

$$r = \begin{bmatrix}Q_1 & Q_2\end{bmatrix} \begin{bmatrix}0\\ \bar{c}\end{bmatrix} = Q_2 \bar{c}.$$

Therefore, the norm of the residual is

$$\|r\|_2 = \|Q_2 \bar{c}\|_2 = \|\bar{c}\|_2$$

as expected.

If $Ax = b$ has a unique solution, then $Ax = b$, which implies

$$\begin{aligned} QRx = b &\implies Rx = Q^T b = c \\ &\implies \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} x = \begin{bmatrix} \hat{c} \\ \bar{c} \end{bmatrix} \\ &\implies \bar{c} = 0. \end{aligned}$$

Thus, $\|r\| = \|\bar{c}\| = \|0\| = 0$, as expected.

- **The discrete LSP algorithm**: Let $A \in \mathbb{R}^{m \times n}$, for $m > n$. Define

$$(P): \min_{x \in \mathbb{R}^n} \|x - Ax\|_2^2 .$$

To solve discrete least squares, we take the following steps

1. $c = Q^T b = \begin{bmatrix} \hat{c} \\ \bar{c} \end{bmatrix}$, $c \in \mathbb{R}^m$, $\hat{c} \in \mathbb{R}^n$, $\bar{c} \in \mathbb{R}^{m-n}$
2. Solve $(\hat{P})$: $\hat{R}x = \hat{c}$, solution of $(\hat{P})$ is the solution of (P).
3. $\min_{x \in \mathbb{R}^n} \|b - Ax\|_2^2 = \|\bar{c}\|_2^2$

Note: For now we want to assume that $\text{rank}(A) = n$, so all columns are linearly independent. In this case, $\hat{r}_{ii} \neq 0$.

- QR **factorization of a square matrix**: Let $A \in \mathbb{R}^{n \times n}$, then $Q \in \mathbb{R}^{n \times n}$, $R = \hat{R} \in \mathbb{R}^{n \times n}$, $Q^T b = c = \hat{c}$, and

$$Ax = b \implies QRx = b \implies Rx = Q^T b \implies \hat{R}x = Q^T b = \hat{c}.$$

So, since $\hat{R}$ is upper triangular, we can solve the system using backward substitution.

- **Givens rotations**: A Givens rotation is an orthogonal transformation that acts only in a 2-dimensional coordinate plane — say the plane spanned by the i-th and j-th coordinate axes. It's the identity matrix except for a 2×2 rotation block:

$$G(i,j,\theta) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c & & s & & \\ & & & 1 & & & \\ & & -s & & c & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}, \quad c = \cos(\theta),\ s = \sin(\theta).$$

Everywhere else its the identity matrix I. Only entries $(i,i),(i,j),(j,i),(j,j)$ are modified.

Let $x = \begin{pmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_j & \cdots & x_n \end{pmatrix}^T \in \mathbb{R}^n$, applying $G(i,j,\theta)$ gives

$$G(i,j,\theta)x = \begin{pmatrix} x_1 \\ \vdots \\ cx_i + sx_j \\ \vdots \\ -sx_i + cx_j \\ \vdots \end{pmatrix} \in \mathbb{R}^n.$$

Note that $G(i,j,\theta)$ is orthogonal, so $\|G(i,j,\theta)x\|_2 = \|x\|_2$. Also, if we let $\sqrt{x_i^2 + x_j^2} = y_i$. That is, we rotate such that a vector $\begin{pmatrix} x_1 & \cdots & x_i & \cdots & x_j & \cdots & x_m \end{pmatrix}^T \mapsto \begin{pmatrix} * & \cdots & y_i & \cdots & 0 & \cdots & * \end{pmatrix}^T$. In this situation,

$$c = \cos(\theta) = \frac{x_i}{\|x\|} = \frac{x_i}{y_i},$$
$$s = \sin(\theta) = \frac{x_j}{\|x\|} = \frac{x_j}{y_i}.$$

So,

$$cx_i + sx_j = cy_i c + sy_i s = y_i c^2 + y_i s^2 = y_i(c^2 + s^2) = y_i,$$
$$-sx_i + cx_j = -sy_i c + cy_i s = -y_i sc + y_i sc = 0.$$

Thus,

$$G(i,j,\theta)x = \begin{pmatrix} x_1 \\ \vdots \\ cx_i + sx_j \\ \vdots \\ -sx_i + cx_j \\ \vdots \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ y_i \\ \vdots \\ 0 \\ \vdots \end{pmatrix} \in \mathbb{R}^n.$$

Important: In computing y_i, we do not use the entire vector x, we only use the i, j entries. So, if

$$\begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_j \\ \vdots \\ x_m \end{pmatrix} \mapsto \begin{pmatrix} * \\ \vdots \\ y_i \\ \vdots \\ 0 \\ \vdots \\ * \end{pmatrix}$$

using a Givens rotation matrix Q, then

$$y_i = \left\| \begin{pmatrix} x_i \\ x_j \end{pmatrix} \right\|_2 = \sqrt{x_i^2 + x_j^2}.$$

- **Finding Q and R, 2×2 example**: Suppose in $\mathbb{R}^{2\times 2}$,

$$A = \begin{bmatrix} x_1 & x_3 \\ x_2 & x_4 \end{bmatrix}.$$

$A = QR$, so $Q^T A = R$. Recall that R is upper triangular, and Q is orthogonal. Since Q is orthogonal, it must represent a rotation (or reflection). Let's first focus on the first column of A, which yields the first column of R, we have

$$Q^T A = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} x_1 & * \\ x_2 & * \end{bmatrix} = \begin{bmatrix} y_1 & * \\ 0 & * \end{bmatrix} = R.$$

Thus,

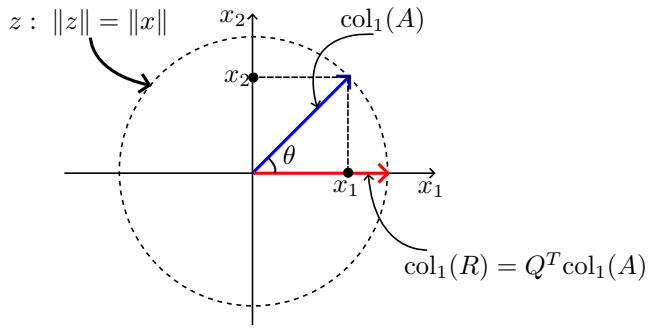
$$Q^T \text{col}_1(A) = \text{col}_1(R) \implies \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ 0 \end{pmatrix}.$$

So,

$$\alpha x_1 + \beta x_2 = y_1 \tag{1}$$
$$\gamma x_1 + \delta x_2 = 0. \tag{2}$$

Applying Q^T to $\text{col}_1(A)$ preserves length, so $\left\|Q^T \text{col}_1(A)\right\| = \|\text{col}_1(A)\| = \|\text{col}_1(R)\|$. Thus,

$$y_1 = \sqrt{x_1^2 + x_2^2} = \|x\|_2 .$$



Notice we have that

$$\sin(\theta) = \frac{x_2}{\|x\|} = \frac{x_2}{y_1},$$
$$\cos(\theta) = \frac{x_1}{\|x\|} = \frac{x_1}{y_1}.$$

From (1), we have

$$\begin{aligned} \alpha\frac{x_1}{y_1} + \beta\frac{x_2}{y_1} = 1 &\implies \alpha\cos(\theta) + \beta\sin(\theta) = 1 \\ &\implies \alpha = \cos(\theta),\ \beta = \sin(\theta). \end{aligned}$$

From (2),

$$\begin{aligned} \gamma x_1 + \delta x_2 = 0 &\implies \gamma y_1 \cos(\theta) + \delta y_1 \sin(\theta) = 0 \\ &\implies \gamma\cos(\theta) + \delta\sin(\theta) = 0 \\ &\implies \gamma = -\sin(\theta),\ \delta = \cos(\theta). \end{aligned}$$

So,

$$Q^T = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} = \begin{pmatrix} \frac{x_1}{y_1} & \frac{x_2}{y_1} \\ -\frac{x_2}{y_1} & \frac{x_1}{y_1} \end{pmatrix}.$$

Now that we have Q^T, we can get the second column of R with $\text{col}_2(R) = Q^T\text{col}_2(A)$.

So, in the 2×2 case, we got Q^T from the first column of A, since we know where the first column should map to under Q^Ts transformation. Then, after we found Q^T, finding the second column of R is simple.

Example: Let $A \in \mathbb{R}^{2\times 2}$,

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}.$$

We have

$$\begin{aligned} y_1 &= \sqrt{1^2 + 1^2} = \sqrt{2}, \\ \cos(\theta) &= \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}, \\ \sin(\theta) &= \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}. \end{aligned}$$

Thus,

$$Q^T = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

So, $\text{col}_1(R) = \begin{pmatrix} y_1 & 0 \end{pmatrix}^T = \begin{pmatrix} \sqrt{2} & 0 \end{pmatrix}^T$, and

$$\text{col}_2(R) = Q^T \text{col}_1(A) = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} \frac{5\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} = \sqrt{2} \begin{pmatrix} \frac{5}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Now we have Q and R,

$$Q^T = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \; R = \sqrt{2} \begin{pmatrix} 1 & \frac{5}{2} \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Suppose that $b = \begin{pmatrix} -1 & -2 \end{pmatrix}^T \in \mathbb{R}^2$. Then,

$$Ax = b \iff QRx = b \iff Rx = Q^T b.$$

So, we can solve the system for x by solving the upper triangular system $Rx = Q^T b$

- **General process for finding Q and R**: We use the fact that for $x \in \mathbb{R}^m$,

$$\begin{pmatrix} \vdots \\ x_i \\ \vdots \\ x_j \\ \vdots \end{pmatrix} \mapsto \begin{pmatrix} \vdots \\ y_i \\ \vdots \\ 0 \\ \vdots \end{pmatrix}.$$

Under the transformation

$$\begin{aligned} (Q_\ell^T)_{ii} &= \cos(\theta), \; (Q_\ell^T)_{ij} = \sin(\theta) \\ (Q_\ell^T)_{ji} &= -\sin(\theta), \; (Q_\ell^T)_{jj} = \cos(\theta), \end{aligned}$$

and with

$$\cos(\theta) = \frac{x_i}{y_i}, \; \sin(\theta) = \frac{x_j}{y_i}, \; y_i = \sqrt{x_i^2 + x_j^2}.$$

For $Q_k \in \mathbb{R}^{m\times m} = I$, except for at the positions above.

In the 2×2 example we only needed to apply one rotation to build Q^T, but for bigger matrices we will need to apply many rotations under we are able to fully build the matrix $\hat{R} \in \mathbb{R}^{n \times n}$. At the end, we will have

$$Q^T A = Q_k^T Q_{k-1}^T \cdots Q_2^T Q_1^T A = R.$$

So,

$$Q^T = Q_k^T Q_{k-1}^T \cdots Q_2^T Q_1^T.$$

We build these matrices in the same way as Gaussian elimination, top to bottom, left to right, choosing two indices at a time, until we have built R.

For example, suppose $A \in \mathbb{R}^{3 \times 2}$,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{pmatrix}.$$

So,

$$Q^T \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

We work top to bottom left to right. Looking at the structure of R in this example, our first transformation will map

$$Q_1^T \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{41} \end{pmatrix} \mapsto \begin{pmatrix} y_1 \\ 0 \\ a_{31} \\ a_{41} \end{pmatrix}.$$

Where

$$Q^T = \begin{pmatrix} \cos(\theta) & \sin(\theta) & 0 & 0 \\ -\sin(\theta) & \cos(\theta) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Notice connection between the targets in A, and the matrix Q, which uses the fact described above. After we have Q_ℓ^T, we can use the following. Eventually, we will have performed enough transformations, and $R_k = R$.

Let $R_1 = A$, then $Q_\ell^T R_\ell = R_{\ell+1}$. So, after the first step,

$$R_2 = Q_1^T R_1 = \begin{pmatrix} y_1 & * \\ 0 & * \\ a_{31} & * \\ a_{41} & * \end{pmatrix},$$

with

$$\begin{pmatrix} * \\ * \\ * \\ * \end{pmatrix} = Q_1^T \operatorname{col}_2(R_1).$$

In the second step, we target y_1, and a_{31}, we want

$$Q_2^T \begin{pmatrix} y_1 \\ 0 \\ a_{31} \\ a_{41} \end{pmatrix} \mapsto \begin{pmatrix} y_2 \\ 0 \\ 0 \\ a_{41} \end{pmatrix},$$

so

$$Q_2^T = \begin{pmatrix} \cos(\theta) & 0 & \sin(\theta) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Thus,

$$R_3 = Q_2^T R_2 = Q_2^T (Q_1^T A),$$

with

$$R_3 = \begin{pmatrix} y_2 & * \\ 0 & * \\ 0 & * \\ a_{41} & * \end{pmatrix}.$$

Notice how we are slowly building the structure of R, similar to how we build the structure of U using Gaussian elimination.

To finish off the first column of R, we will have

$$R_4 = Q_3^T R_3 = Q_3^T (Q_2^T (Q_1^T A)),$$

with

$$R_4 = \begin{pmatrix} y_3 & * \\ 0 & * \\ 0 & * \\ 0 & * \end{pmatrix}.$$

Now, we proceed to the second column, we need the structure to be

$$\begin{pmatrix} r_{12} \\ r_{22} \\ 0 \\ 0 \end{pmatrix}.$$

So, we start by selecting the second and third element of the second column of R_4, so that we can get the third element zero.

Note that as we are working on the second column, the nonzero elements in the columns before will change, but the zero elements are safe, so the structure that we have already built will be preserved.

- **Product of orthogonal matrices is orthogonal**: Let Q_1 and Q_2 be orthogonal, then $Q = Q_1 Q_2$ is orthogonal.

Proof. Q_1, Q_2 are orthogonal, so $Q_1^T Q_1 = I$, and $Q_2^T Q_2 = I$. $Q = Q_1 Q_2$, so

$$Q^T Q = (Q_1 Q_2)^T (Q_1 Q_2) = Q_2^T Q_1^T Q_1 Q_2 = Q_2^T I Q_2 = Q_2^T Q_2 = I.$$

Thus, Q is orthogonal. ■

- **Householder reflection**: A Householder reflection is a linear transformation

$$H = I - 2\frac{vv^T}{v^T v}.$$

Where $v \neq 0$ is a chosen vector. It geometrically represents a reflection across the hyperplane orthogonal to v, this means

1. $Hv = -v$
2. For any vector z orthogonal to v, $Hz = z$

So H keeps the subspace orthogonal to v fixed and reverses the component of a vector along v.

- **Direction vector**: A direction vector describes how you can move within a geometric object without leaving it.
  - In a line, a direction vector tells you which way the line goes.
  - In a plane, direction vectors describe all the ways you can move along the plane.
  - In a curve or surface, a direction vector describes a tangent direction — an infinitesimal direction of movement within the object.

Let $S \subseteq \mathbb{R}^n$. A **direction vector** of S is any vector d such that for some point $x_0 \in S$, we have

$$x_0 + td \in S \quad \text{for all } t \in \mathbb{R}.$$

So, if you start at x_0 and move in the direction d by any scalar multiple t, you stay inside S.

If S is a linear subspace (passes through origin), then every vector $d \in S$ is a direction vector, since

$$0 + td = td \in S.$$

- **Linear subspace**: A linear subspace (often called simply a subspace) $S \subseteq V$ is a subset of a vector space V (the ambient space) that is itself a vector space under the same operations (vector addition and scalar multiplication).

S is a subspace iff it satisfies:

$$\forall u, v \in S, ; \forall \alpha, \beta \in \mathbb{R}, \ \alpha u + \beta v \in S.$$

This property is called the closure under linear combinations. Notice that if $\alpha, \beta = 0$, then $0 \in S$ is a requirement. Thus, there are three key properties for a subset $S \subseteq V$ to be a linear subspace

1. $0 \in S$
2. $u, v \in S \implies u + v \in S$
3. $u \in S, \ \alpha \in \mathbb{R} \implies \alpha u \in S$

If we wanted to be more general, we could swap $\mathbb{R}$ with any field F. I.e the scalars must live in some field F.

- **Affine subspaces**: An affine set is a translation of a linear subspace. A set $A \subseteq V$ is an affine subspace of the ambient set V if it can be written as

$$A = S + v = \{s + v : \ s \in S\},$$

where S is a linear subspace of the ambient space V ($S \subseteq V$), and v is a fixed member of the ambient space V

  - S is called the direction subspace or associated subspace of A
  - v is any base point through which the affine subspace passes.

  Note: So, $a \in A$ implies that $a = s + v = td + v$, for $t \in \mathbb{R}, d \in \mathbb{S}$. Since S is linear $td \in S$. So, every vector $s \in S$ is a direction vector for A.

- **Orientation**: The direction it "points" in — how it's aligned relative to coordinate axes, changes if you rotate or reflect it

  So "orientation" answers questions like:

  - Is this plane horizontal or vertical?
  - Does this line go northeast–southwest or northwest–southeast?
  - Is the coordinate basis right-handed or left-handed?

  A linear subspace $S \subseteq \mathbb{R}^n$ always passes through the origin. Its orientation is determined entirely by the directions it contains — that is, by its basis vectors.

  The orientation of a subspace is defined by the set of directions it spans, or equivalently, by any orthonormal basis for it. Two subspaces have the same orientation if they are related by a rotation, not a reflection.

  An affine subspace (or affine space) is a translation of a linear subspace:

$$S = S_0 + x_0.$$

  S_0 gives the orientation (the directions it extends in), so the affine subspace has the same orientation as its linear part S_0, translating doesn't affect orientation.

- **Intro to Hyperplanes**: In $\mathbb{R}^n$, a hyperplane is an $(n-1)$-dimensional affine subspace defined by a linear equation of the form:

$$H = \{x \in \mathbb{R}^n : \ a^T x = b\},$$

  where

  - $a \in \mathbb{R}^n$ is a nonzero vector (called the **normal vector** to the hyperplane),
  - $b \in \mathbb{R}$ is a scalar constant.

  The vector a determines the orientation of the hyperplane (it's perpendicular to it), the constant b determines its position relative to the origin.

  If $b = 0$, the hyperplane passes through the origin and is called a linear subspace, call this hyperplane space H_0

$$H_0 = \{s \in \mathbb{R}^n : \ a^T s = 0\}.$$

  In this case, all $s \in H_0$ are orthogonal to a.

If $b \neq 0$, the set

$$H = \{x \in \mathbb{R}^n : \ a^T x = b\}.$$

is no longer a linear subspace, it does not pass through the origin. Instead, its a translated copy of the original hyperplane H_0. That is,

$$H = H_0 + x_0 = \{s + x_0 : \ x_0 \in H, \ s \in H_0, \ a^T s = 0\}.$$

So, H is an **affine** hyperplane instead of a linear one. To prove this fact holds, we can show that $H \subseteq H_0 + x_0$, and $H_0 + x_0 \subseteq H$.

First, consider $x = s + x_0$, for $s \in H_0$, and $x_0 \in H$. We have

$$a^T x = a^T(s + x_0) = a^T s + a^T x_0 = 0 + b = b.$$

So, $x \in H$, and $H_0 + x_0 \subseteq H$.

Next, let $x, x_0 \in H$, we aim to show that x can be written as $s + x_0$, for $s \in H_0$, and $x_0 \in H$. At this point we don't know where s lives. If $x = s + x_0$, then $s = x - x_0$, and

$$a^T s = a^T(x - x_0) = a^T x - a^T x_0 = b - b = 0.$$

So, $s \in H_0$, and x can indeed be written as $s + x_0$. This implies that $H \subseteq H_0 + x_0$

Since $H_0 + x_0 \subseteq H$, and $H \subseteq H_0 + x_0$, it must be that $H = H_0 + x_0$ ■

You may have noticed that we said the hyperplane is orthogonal to a. But, in H, $a^T x = b$ implies that vectors in H are not orthogonal to a. Instead, the **directions** lying within the hyperplane are orthogonal to a.

Let $x_0, x_1 \in H$. Define the difference d as

$$d = x_0 - x_1.$$

Notice that since x_0 and x_1 are in H, we have $a^T x_0 = a^T x_1 = b$. Thus,

$$a^T d = a^T(x_0 - x_1) = a^T x_0 - a^T x_1 = b - b = 0.$$

For example, in 3D, if $a = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T$, the plane $a^T x = b$ is

$$0x + 0y + 1z = b \implies z = b.$$

Every point on that plane has $z = b$, and these points are not orthogonal to a. But, any direction vector lying along the plane, for example $(1, 0, 0)$ or $(0, 1, 0)$ *is* orthogonal to a.

So when we say "the hyperplane orthogonal to a", we mean that the plane's orientation is determined by a, not that its points themselves are orthogonal to a.

Some hyperplane examples are...

Space	**Equation**	Description
$\mathbb{R}^2$	$a_1x_1 + a_2x_2 = b$	A **line** (1D hyperplane)
$\mathbb{R}^3$	$a_1x_1 + a_2x_2 + a_3x_3 = b$	A **plane** (2D hyperplane)
$\mathbb{R}^n$	$a_1x_1 + a_2x_2 + a_3x_3 + = b$	An $(n-1)$-dimensional **flat surface**

Why does the hyperplane with orientation perpendicular to $a \in \mathbb{R}^2$, with position b relative to the origin form a 1D line? Let $a \in \mathbb{R}^2$, $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$. Then, $x \in \mathbb{R}^2$ is on the hyperplane if

$$a^T x = b,$$

which implies

$$a_1x_1 + a_2x_2 = b.$$

We know that $a^T x = b$ says that all vectors x in the hyperplane when projected onto a have length $\frac{b}{\|a\|}$.

So, if we take the component of each x in the direction of a, it must equal the same value $(b/\|a\|)$.

If ℓ is a line not orthogonal to a, and y is any vector on that line, then $a^T y$ does not stay fixed as we move along the line.

However, if ℓ is a line orthogonal to a, then any vector x on that line is orthogonal to a, and so the dot product is constant (0).

Let n be a vector orthogonal to a. Then,

$$a^T n = 0.$$

Let x_0 be a vector that satisfies $a^T x_0 = b$, then all vectors x that satisfy $x = x_0 + tn$, for $t \in \mathbb{R}$ satisfy $a^T x = b$, since

$$a^T x = a^T(x_0 + tn) = a^T x_0 + a^T tn = a^T x_0 + ta^T n = b + t \cdot 0 = b.$$

Notice that our vector generator $x = x_0 + tn$ generates all vectors $x \in \mathbb{R}^n$ such that $a^T x = b$, and $x_0 + tn$ is precisely the vector equation for a line. Thus, the hyperplane orthogonal to a (the hyperplane with normal a), for $a \in \mathbb{R}^2$ is a 1D line.

- **Hyperplanes**: Let $\mathbb{R}^n$ be the ambient space. Let a be a member in the ambient space. The space orthogonal to a is a hyperplane. That is,

$$H = \{x \in \mathbb{R}^n : \ a^T x = b\}.$$

Observe that the vectors x in the hyperplane are not orthogonal to a if $b \neq 0$, since their dot product is nonzero. Instead, the direction vectors in H are orthogonal to a. This is why we say the hyperplanes orientation is orthogonal to a.

If $b = 0$, the hyperplane is a linear subspace of $\mathbb{R}^n$. If $b \neq 0$, the hyperplane is an affine subspace of $\mathbb{R}^n$.

We know that $H = H_0 + v$, for v in the ambient space. So, $H = \{s + v : \ s \in \mathbb{H}_0\}$. Thus,

$$x \in H \implies x = s + v = td + v,$$

for $t \in \mathbb{R}$, $d \in H_0$. So, any member of H_0 is a direction vector for H. Let $x \in H$ be written as $x = x_0 + s$, for $x_0 \in H$, and $s \in H_0$. We have

$$a^T x = a^T(x_0 + s) = a^T x_0 + a^T s = b + 0 = b,$$

and

$$a^T s = 0.$$

Thus, the direction of H is orthogonal to a, but the members are not.

- **Intro to Householder reflections for QR factorizations**: Using givens rotations, we could only zero out one entry per Q_k, so building one column of R took many rotations. Suppose instead we wanted to build each column of R with just one transformation. This is not possible with givens rotations, but we can use Householder reflections.

  We could perform the transformation

$$Qx = Q\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_j \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ -\tau_i \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} = y,$$

  with

$$Q = I - \gamma u u^T, \ \tau = \operatorname{sgn}(x_i) \|x\|_2, \ u = \frac{x - y}{\tau + x_i} = \begin{pmatrix} 1 \\ x_2/(\tau + x_1) \\ x_3/(\tau + x_1) \\ \vdots \\ x_n/(\tau + x_1) \end{pmatrix}, \ \|u\|_2 = 1,$$

$$\gamma = \frac{\tau + x_i}{\tau},$$

  and

$$\operatorname{sgn}(x_i) = \begin{cases} 1 & \text{if } x_i > 0 \\ -1 & \text{if } x_i < 0 \\ 0 & \text{if } x_i = 0 \end{cases}.$$

  Q is called the **Householder matrix**.

- **Properties of the Householder matrix**
  1. **Symmetry**: $Q = Q^T$
  2. **Orthogonality**: $Q^T Q = Q Q^T = I$

Proof (1).

$$Q^T = (I - 2uu^T)^T = I - 2(uu^T)^T = I - 2(u^T)^T u^T = I - 2uu^T = Q.$$

(2).

$$\begin{aligned} Q^TQ = QQ &= (I - 2uu^T)(I - 2uu^T) = I - 2uu^T - 2uu^T + 4uu^Tuu^T \\ &= I - 4uu^T + 4u(u^Tu)u^T = I - 4uu^T + 4(u^Tu)uu^T \\ &= I - 4uu^T + 4\|u\|_2 uu^T = I - 4uu^T + 4uu^T = I. \end{aligned}$$

- QR **with Householder reflectors**: Let $A \in \mathbb{R}^{m\times n}$, where

$$A = \begin{bmatrix} | & | & \cdots & | \\ x^1 & x^2 & \cdots & x^n \\ | & | & \cdots & | \end{bmatrix},$$

where x^j denotes the j^{th} column of A. For each column j of A, we have the following steps

1. $\tau_j = \operatorname{sgn}(x_1)\left\|x^j\right\|_2$
2. $\gamma_j = \frac{\tau_j + x_1}{\tau_j}$
3. $u_j = \begin{pmatrix} 1 \\ x_2/(\tau_j + x_1) \\ \vdots \\ x_m/(\tau_j + x_1) \end{pmatrix}$
4. $Q_j = I - \gamma_j u_j u_j^T$

Then, just like with givens rotations, $Q = Q_k \cdots Q_j \cdots Q_2 Q_1$, and $R = QA = Q_k \cdots Q_j \cdots Q_2 Q_1 A$.

- **Householder reflector decompositions**: We don't transform an entire column. Instead, we act on a block of each column, where the first entry is the first nonzero entry starting from the bottom to the top for each column of R. For example, if

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{pmatrix}.$$

Then R looks like

$$R = \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

So, for the first column of A, we do need to act on the entire column. But, for the second column, we only need from the second element onward.

Because the vectors that we need Q_j to act on will be of size one less than the previous, the portion of Q_j that is not I will also decrease in size. Note that Q_j will always be of size $m \times m$. We will have

$$Q_1 = I_m - \gamma_1 u_1 u_1^T,\ Q_2 = I_{m-1} - \gamma_2 u_2 u_2^T, \ldots.$$

Consider Q_2, the first row and the first column will be I, and the remaining $(m - 1 \times m - 1)$ block will be $I_{m-1} - \gamma_2 u_2 u_2^T$. Because of this fact, the first row and first column of A will be unchanged.

If $A \in \mathbb{R}^{m\times n}$, then

$$A_1 = Q_1 A = \left(\begin{array}{c|c} -\tau_1 & a_1^T \\ \hline 0 & \tilde{A}_1 \end{array}\right),$$

where

$$Q_1 = I_m - \gamma_1 u_1 u_1^T.$$

Then,

$$A_2 = Q_2 A_1 = Q_2 Q_1 A = \left(\begin{array}{c|cc} -\tau_1 & a_1^T & \\ \hline & -\tau_2 & a_2^T \\ 0 & \hline & 0 \quad \tilde{A}_2 \end{array}\right),$$

with

$$Q_2 = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & I_{m-1} - \gamma_2 u_2 u_2^T \end{array}\right).$$

Then,

$$A_3 = Q_3 A_2 = Q_3 Q_2 Q_1 A = \left(\begin{array}{c|cc} -\tau_1 & a_1^T & \\ \hline & -\tau_2 & a_2^T \\ 0 & \hline & -\tau_3 \quad a_3^T \\ & 0 & \hline & 0 \quad \tilde{A}_3 \end{array}\right),$$

with

$$Q_3 = \left(\begin{array}{c|cc} 1 & 0 & \\ \hline & 1 & 0 \\ 0 & \hline & 1 \quad I_{m-2} - \gamma_3 u_3 u_3^T \end{array}\right).$$

We proceed in this fashion until $A_n = R$. Note that we use the blocks of Q that contain the reflector to find the blocks of A that are changed in each step. Call this block $\bar{Q}$.

- **Flop counts for QR with Householder reflectors**: Consider the steps of Householder
  1. $\tau_j = \operatorname{sgn}(x_1) \left\|x^j\right\|_2$ ($\mathcal{O}(m)$ flops)
  2. $\gamma_j = \frac{\tau_j + x_1}{\tau_j}$ ($\mathcal{O}(1)$ flops)

3. $u_j = \begin{pmatrix} 1 \\ x_2/(\tau_j + x_1) \\ \vdots \\ x_m/(\tau_j + x_1) \end{pmatrix}$ ($\mathcal{O}(m)$ flops)

4. $Q_j = I - \gamma_j u_j u_j^T$ ($\mathcal{O}(m^2)$ flops)

So, each Q_j requires roughly $\mathcal{O}(m^2)$ flops. But, we compute these steps for each column of A. If A has n columns, then we require roughly $\mathcal{O}(nm^2)$ flops. If A is square, then we need $\mathcal{O}(m^3)$ flops for the QR factorization of A with Householder reflectors.

To be precise, QR factorization with Householder reflectors requires $\frac{4}{3}m^3$ flops. Not as efficient as LU or Cholesky, but more efficient that computing the inverse.

- **Reducing flops in QR factorization with Householder reflectors**: Consider a system $Ax = b$, with the QR factorization,

$$Ax = b \implies QRx = b \implies Rx = Qb.$$

But, $R = QA$, so

$$Rx = Qb \implies QA = Qb \implies Q_k \cdots Q_2 Q_1 A = Q_k \cdots Q_2 Q_1 b.$$

However, we don't really matrix multiply Q_j by A, what we do is multiply columns of A by Q_j. Consider a vector x multiplied by a Householder reflector Q...

$$Qx = (I - \gamma u u_T)x = x - \gamma u u^T x.$$

The expensive part is computing the outer product uu^T. Instead, lets use the associative law of vectors to instead do

$$Qx = x - \gamma u(u^T x).$$

Notice that $u^T x$ is an inner product, just a real number. Thus, we can move it. We have

$$Qx = x - \gamma(u^T x)u.$$

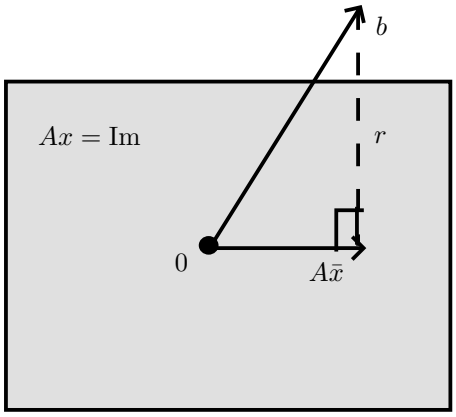
This simple fact reduces the flops for this step by a factor of m, i.e $\mathcal{O}(m)$ flops instead of $\mathcal{O}(m^2)$.

We now don't even need to form Q at all, using the fact that since

$$Rx = Qb \implies QA = Qb,$$

all we need is vector multiplication by the form of Q, the form that we have above $(x - \gamma(u^T x)u)$

- **Normal equations**: Consider again the geometric interpretation of a vector b outside of the column space of A.



Which can happen if our system is over-determined. Since there is clearly no solution to the system for the given b, we instead solve the least squares problem

$$(p): \min_{x \in \mathbb{R}^n} \|b - Ax\|_2^2 .$$

Thus, we aim to find some vector $\bar{x}$ such that the residual $b - A\bar{x}$ is minimized. We know that this vector is found by projecting b onto the column space Ax, and thus the residual $r = b - A\bar{x}$ is orthogonal to $A\bar{x}$, where $A\bar{x}$ is the projection of b onto the column space.

So, we have

$$A\bar{x} \perp r,$$

which implies that

$$\begin{aligned} (A\bar{x})^T r &= 0 \\ \implies (A\bar{x})^T (b - A\bar{x}) &= 0 \\ \implies \bar{x}^T A^T b - \bar{x}^T A^T A\bar{x} &= 0 \\ \implies \bar{x}^T (A^T b - A^T A\bar{x}) &= 0. \end{aligned}$$

But, notice that $\bar{x} \neq 0$ since if it were, then $A\bar{x} = b = 0$. Since the zero vector is in the column space, b is non-zero, and so is $\bar{x}$. Thus,

$$\begin{aligned} A^T b - A^T A\bar{x} &= 0 \\ \implies A^T A\bar{x} &= A^T b. \end{aligned}$$

Therefore, $\bar{x}$ is found by solving the system $A^T Ax = A^T b$, and

$$\bar{x} = \arg\min_{x \in \mathbb{R}^n} \|b - Ax\|_2^2 .$$

Although this method is far easier than QR factorization methods, one can prove that

$$\kappa(A^T A) = \kappa^2(A)$$

So, if A is moderately ill-conditioned, then $A^T A$ can be severely ill-conditioned.

Note: The "normal equations" are the equations

$$A^T Ax = A^T b.$$

They are "normal" because they enforce that the residual is **normal** to the column space.

- **Gram Matrix**: The matrix $A^T A$ is referred to as the *Gram matrix* of the columns of A. Any matrix of inner products is called a *Gram* matrix.

The matrix A^TA is also sometimes referred to as the normal matrix, in the context of the normal equations.

- **Property of the normal matrix**: Consider the normal matrix A^TA, we can verify that this matrix is symmetric,

$$(A^TA)^T = A^T(A^T)^T = A^TA.$$

Thus, it is symmetric. Next, for any nonzero vector x, we have

$$x^TA^TAx = (Ax)^T(Ax) = \|Ax\|_2^2 .$$

Since $\|Ax\|_2 \geqslant 0$ for all x, $A^TA \succeq 0$.

But, notice that if A has full column rank ($\text{rank}(A) = n$), then Ax is injective and the kernel space is trivial

$$Ax = 0 \iff x = 0.$$

Thus for $A \in \mathbb{R}^{m\times n}$,

$$A^TA \succ 0 \iff \text{rank}(A) = n.$$

Therefore, if A has full column rank, then A^TA is positive definite, and we can use Cholesky decomposition for the normal equations.

6.5.2 Singular value decomposition (SVD)

- **Singular values**: For a matrix $A \in \mathbb{R}^{m \times n}$, its singular values are

$$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0,$$

where r is the rank of A. Each singular value is defined by

$$\sigma_i = \sqrt{\lambda_i(A^T A)},$$

where λ_i is the i^{th} eigenvalue for $A^T A$.

- **Rank of $A^T A$**: Let $A \in \mathbb{R}^{m \times n}$. Recall by the rank-nullity theorem,

$$\text{rank}(A) = n - \dim(\ker(A)).$$

The key insight is that $A^T A$ and A share the same null space. Suppose $x \in \text{null}(A^T A)$, then

$$A^T Ax = 0 \iff x^T A^T Ax = 0 \iff (Ax)^T (Ax) = 0 \iff \|Ax\|^2 = 0 \iff Ax = 0.$$

Thus, $x \in \text{null}(A)$. So, since $\ker(A^T A) = \ker(A)$,

$$\text{rank}(A^T A) = n - \dim(\ker(A^T A)) = n - \dim(\ker(A)) = \text{rank}(A).$$

- **SVD**: The Singular Value Decomposition (SVD) is a fundamental factorization in linear algebra that applies to any real or complex matrix, square or rectangular. For a matrix

$$A \in \mathbb{R}^{m \times n},$$

The SVD expresses A as

$$A = U \Sigma V^T$$

where

  - $U \in \mathbb{R}^{m \times m}$ An orthogonal matrix ($U^T U = I$). Its columns are called **left singular vectors**.
  - $\sum \in \mathbb{R}^{m \times n}$ A diagonal (rectangular) matrix whose entries are the singular values

$$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant 0$$

  - $V \in \mathbb{R}^{n \times m}$ An orthogonal matrix. Its columns are the **right singular vectors**

6.6 The Determinant

- **Bilinear function**: Let $f: V \times W \to \mathbb{F}$, where V and W are vector spaces over the field $\mathbb{F}$. f is called bilinear if it is linear in each argument separately. So,

$$f(\alpha v + \beta u, w) = \alpha f(v, w) + \beta f(u, w),$$
$$f(v, \alpha w + \beta u) = \alpha f(v, w) + \beta f(v, u).$$

  Thus,

$$\begin{aligned} f(\alpha v_1 + \beta v_2, \gamma w_1 + \lambda w_2) &= f(\alpha v_1, \gamma w_1 + \lambda w_2) + f(\beta v_2, \gamma w_1 + \lambda w_2) \\ &= f(\alpha v_1, \gamma w_1) + f(\alpha v_1, \lambda w_2) + f(\beta v_2, \gamma w_1) + f(\beta v_2, \lambda w_2) \\ &= \alpha\gamma f(v_1, w_1) + \alpha\lambda f(v_1, w_2) + \beta\gamma f(v_2, w_1) + \beta\lambda f(v_2, w_2). \end{aligned}$$

  Note: As an example, the inner product $\langle v, w \rangle = v^T w$ is a bilinear function.

  The row-column matrix multiplication $(v^T A)w$ is also bilinear in v and w.

- **Multilinear function**: For a single-variable function

$$f : \mathbb{R}^n \to \mathbb{R}$$

  (that takes one vector as input), we say that f is **linear** if for all scalars $c_1, c_2 \in \mathbb{R}$ and vectors $u, w \in \mathbb{R}^n$,

$$f(c_1 u + c_2 w) = c_1 f(u) + c_2 f(w).$$

  Where

  - **Additivity**: $f(u + w) = f(u) + f(w)$
  - **Homogeneity**: $f(cu) = cf(u)$

  This is just the usual definition of linearity that you already know from linear algebra.

  Now suppose f takes several vectors as arguments:

$$f(v_1, v_2, \ldots, v_n).$$

  We cannot talk about linearity in all of them at once (since we could mix them in complicated ways), so we instead require linearity in each one separately, while holding the others fixed.

  Pick one argument, say the i^{th} one, and replace it with a linear combination $c_1 u + c_2 w$. The function is **multilinear** if this linearity property holds in that position:

$$f(v_1, \ldots, c_1 u + c_2 w, \ldots, v_n) = c_1 f(v_1, \ldots, u, \ldots, v_n) + c_2 f(v_1, \ldots, w, \ldots, v_n).$$

- **Properties of multilinear functions**: Suppose $f: V^n \to \mathbb{F}$ is multilinear, then the following properties can be observed

  1. **Zero vector kills the value**: If any argument is the zero vector, then $f(\ldots, 0, \ldots) = 0$. This follows immediately from homogeneity.
  2. **Expansion property (multilinear expansion)**: If each argument is written as a sum, the function expands as a sum over all combinations. For example, if

$$v_i = u_i + w_i \quad \text{for all } i,$$

     then

$$f(v_1, \ldots, v_n) = \sum f(\text{all choices of } u_i \text{ or } w_i).$$

     This is exactly the algebraic mechanism behind the Leibniz determinant formula.

3. **Behavior under scaling all arguments**: If all arguments are scaled by the same scalar α, then

$$f(\alpha v_1, \ldots, \alpha v_n) = \alpha^n f(v_1, \ldots, v_n).$$

This follows from applying homogeneity for all n arguments.

- **Expansion property (multilinear expansion)**: We saw in the bilinear case that

$$\begin{aligned} f(v_1 + v_2, w_1 + w_2) &= f(v_1, w_1 + w_2) + f(v_2, w_1 + w_2) \\ &= f(v_1, w_1) + f(v_1, w_2) + f(v_2, w_1) + f(v_2, w_2) \\ &= f(v_1, w_1) + f(v_1, w_2) + f(v_2, w_1) + f(v_2, w_2). \end{aligned}$$

This is precisely the expansion property of multilinear functions. In this case, the bilinear function $f(v_1, v_2)$ replaced arguments so that each argument was a linear combination. That is,

$$\begin{aligned} v &= v_1 + v_2, \\ w &= w_1 + w_2. \end{aligned}$$

So, by the expansion property,

$$f(v, w) = f(v_1 + v_2, w_1 + w_2) = f\left(\sum_{i=1}^{2} v_i, \sum_{j=1}^{2} w_j\right) = \sum_{i=1}^{2}\sum_{j=1}^{2} f(v_i, w_j).$$

In fact, the decomposition does not need to be limited to two terms, if f has two arguments v_1, v_2, then we can express each argument as a finite sum of vectors,

$$\begin{aligned} v_1 &= v_1^{(1)} + v_1^{(2)} + \cdots + v_1^{(m_1)}, \\ v_2 &= v_2^{(1)} + v_2^{(2)} + \cdots + v_2^{(m_2)}. \end{aligned}$$

Then,

$$f(v_1, v_2) = f\left(\sum_{i=1}^{m_1} v_1^{(i)}, \sum_{j=1}^{m_2} v_2^{(j)}\right) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} f(v_1^{(i)}, v_2^{(j)}).$$

Now, consider a general multilinear function $f : V^n \to \mathbb{F}$, where f has n vector arguments $v_1, \ldots, v_n$. Replace each v_i by a sum,

$$v_i = v_i^{(1)} + v_i^{(2)} + \cdots + v_i^{(m_i)}.$$

Then,

$$\begin{aligned} f(v_1, \ldots, v_n) &= f\left(\sum_{k_1} v_1^{(k_1)}, \ldots, \sum_{k_i} v_i^{(k_i)}, \ldots, \sum_{k_n} v_n^{(k_n)}\right) \\ &= \sum_{k_1=1}^{m_1} \cdots \sum_{k_i=1}^{m_i} \cdots \sum_{k_n=1}^{m_n} f(v_1^{(k_1)}, \ldots, v_i^{(k_i)}, \ldots, v_n^{(k_n)}). \end{aligned}$$

- **Ordinary multiplication is multilinear**: Consider the map

$$m : \mathbb{R}^n \to \mathbb{R} \quad m(x_1, \ldots, x_n) \mapsto x_1 \cdots x_n.$$

Note that in this situation, $\mathbb{R}^n$ is taken to mean $\mathbb{R} \times \cdots \times \mathbb{R}$, an n-tuple of scalars in $\mathbb{R}$, not the vector space $\mathbb{R}^n$.

This map is multilinear, so

$$\begin{aligned} m(x_1,\ldots,\alpha x_i,\ldots,x_n) &= \alpha m(x_1,\ldots,x_i,\ldots,x_n),\\ m(x_1,\ldots,x_i^{(1)}+x_i^{(2)},\ldots,x_n) &= m(x_1,\ldots,x_i^{(1)},\ldots,x_n)\\ &\quad + m(x_1,\ldots,x_i^{(2)},\ldots,x_n). \end{aligned}$$

Which means,

$$\begin{aligned} x_1x_2\cdots\alpha x_i\cdots x_n &= \alpha(x_1x_2\cdots x_i\cdots x_n),\\ x_1\cdots(x_i^{(1)}+x_i^{(2)})\cdots x_n &= (x_1\cdots x_i^{(1)}\cdots x_n)+(x_1\cdots x_i^{(2)}\cdots x_n). \end{aligned}$$

For example, if $n=3$, the map becomes

$$m(x_1,x_2,x_3)=x_1x_2x_3=\prod_{i=1}^{3}x_i.$$

So,

$$\begin{aligned} m(x_1,\alpha x_2,x_3)=x_1\alpha x_2x_3=\alpha(x_1x_2x_3)&=\alpha\prod_{i=1}^{3}x_i\\ m(x_1,x_2^{(1)}+x_2^{(2)},x_3)=x_1(x_2^{(1)}+x_2^{(2)})x_3&=x_1x_2^{(1)}x_3+x_1x_2^{(2)}x_3\\ &=\sum_{k=1}^{2}\prod_{i=1}^{3}x_i^{(k)}, \end{aligned}$$

where $x_1=x_1^{(1)}=x_1^{(2)}$, and $x_3=x_3^{(1)}=x_3^{(2)}$, since these terms are not split into a sum.

Now consider $x_i=x_i^{(1)}+\cdots+x_i^{(m_i)}$, by the expansion property of multilinearity,

$$\begin{aligned} m(x_1,\ldots,x_n)&=m\left(\sum_{k_1=1}^{m_1}x_1^{(k_1)},\ldots,\sum_{k_n=1}^{m_n}x_n^{(k_n)}\right)=\sum_{k_1=1}^{m_1}\cdots\sum_{k_n=1}^{m_n}m(x_1^{(k_1)},\ldots,x_n^{(k_n)})\\ &=\sum_{k_1,\ldots,k_n}m(x_1^{(k_1)},\ldots,x_n^{(k_n)})=\sum_{k_1,\ldots,k_n}x_1^{(k_1)}\cdots x_n^{(k_n)}. \end{aligned}$$

Consider again the $n=3$ case, let $x_i=x_i^{(1)}+\cdots+x_i^{(m_i)}$ for $i=1,2,3$. So,

$$\begin{aligned} m(x_1,x_2,x_3)&=m\left(\sum_{k_1=1}^{m_1}x_1^{(k_1)},\sum_{k_2=1}^{m_2}x_2^{(k_2)},\sum_{k_3=1}^{m_3}x_3^{(k_3)}\right)\\ &=\sum_{k_1=1}^{m_1}\sum_{k_2=1}^{m_2}\sum_{k_3=1}^{m_3}m(x_1^{(k_1)},x_2^{(k_2)},x_3^{(k_3)})=\sum_{k_1,k_2,k_3}m(x_1^{(k_1)},x_2^{(k_2)},x_3^{(k_3)}). \end{aligned}$$

So,

$$\begin{aligned} x_1x_2x_3&=(x_1^{(1)}+\cdots+x_1^{(m_1)})(x_2^{(1)}+\cdots+x_2^{(m_2)})(x_3^{(1)}+\cdots+x_3^{(m_3)})\\ &=\prod_{i=1}^{3}\sum_{k_i=1}^{m_i}x_i^{(k_i)}=\sum_{k_1=1}^{m_1}\sum_{k_2=1}^{m_2}\sum_{k_3=1}^{m_3}x_1^{(k_1)}x_2^{(k_2)}x_3^{(k_3)}\\ &=\sum_{k_1=1}^{m_1}\sum_{k_2=1}^{m_2}\sum_{k_3=1}^{m_3}\prod_{i=1}^{3}x_i^{(k_i)}=\sum_{k_1,k_2,k_3}\prod_{i=1}^{3}x_i^{(k_i)}. \end{aligned}$$

So, if $m_1 = m_2 = m_3 = 2$, then

$$\begin{aligned} x_1x_2x_3 &= (x_1^{(1)} + x_1^{(2)})(x_2^{(1)} + x_2^{(2)})(x_3^{(1)} + x_3^{(2)}) \\ &= x_1^{(1)}x_2^{(1)}x_3^{(1)} + x_1^{(1)}x_2^{(1)}x_3^{(2)} + x_1^{(1)}x_2^{(2)}x_3^{(1)} + x_1^{(1)}x_2^{(2)}x_3^{(2)} \\ &+ x_1^{(2)}x_2^{(1)}x_3^{(1)} + x_1^{(2)}x_2^{(1)}x_3^{(2)} + x_1^{(2)}x_2^{(2)}x_3^{(1)} + x_1^{(2)}x_2^{(2)}x_3^{(2)}. \end{aligned}$$

Now, consider the map

$$m(x, y) = xy.$$

But, let's let $y = x$, and $x = (z_1 + z_2)$, so

$$m(x, y) = xx = (z_1 + z_2)(z_1 + z_2) = \prod_{i=1}^{2} x = \prod_{i=1}^{2}(z_1 + z_2) = \prod_{i=1}^{2}\sum_{k=1}^{2} z_k.$$

By the expansion property,

$$\prod_{i=1}^{2}\sum_{k=1}^{2} z_k = \sum_{k_1=1}^{2}\sum_{k_2=1}^{2} z_{k_1}z_{k_2} = \sum_{k_1=1}^{2}\sum_{k_2=1}^{2}\prod_{i=1}^{2} z_{k_i}.$$

Which if we expand is

$$z_1z_1 + z_1z_2 + z_2z_1 + z_2z_2 = z_1^2 + 2z_1z_2 + z_2^2 = (z_1 + z_2)^2.$$

More generally,

$$(z_1 + z_2)^n = \prod_{i=1}^{n}\sum_{k=1}^{2} z_k = \sum_{k_1=1}^{2}\cdots\sum_{k_n=1}^{2}\prod_{i=1}^{n} z_{k_i}.$$

Even more generally,

$$(z_1 + z_2 + \cdots + z_m)^n = \prod_{i=1}^{n}\sum_{k=1}^{m} z_k = \sum_{k_1=1}^{m}\cdots\sum_{k_n=1}^{m}\prod_{i=1}^{n} z_{k_i} = \sum_{k_1,k_2,\ldots,k_n}\prod_{i=1}^{n} z_{k_i}.$$

Notice that this describes the distributive property.

Note that $\sum_{k_1,\ldots k_n}$ is shorthand for $\sum_{k_1=1}^{m_1}\cdots\sum_{k_n=1}^{m_n}$. Or, we can use the set-theoretic notation

$$\sum_{(k_1,\ldots,k_n)\in\{1,\ldots,m\}^n}$$

to describe the shorthand. Note that $\{1, \ldots, m\}^n$ means the Cartesian product $\{1, \ldots, m\} \times \cdots \times \{1, \ldots, m\}$, where a member of this set is an n-tuple $(k_1, \ldots, k_n)$, where $k_i \in \{1, \ldots, m\}$.

- **Alternating function**: Let V be a vector space, and let

$$f : V^n \to \mathbb{F}$$

be a function of n vector arguments. The function f is called alternating if

$$f(v_1, \ldots, v_i, \ldots, v_j, \ldots, v_n) = -f(v_1, \ldots, v_j, \ldots, v_i, \ldots, v_n).$$

That is, swapping any two inputs flips the sign of the output.

- **Multilinear alternating functions**: Consider a multilinear alternating function $f(v_1, \ldots, v_i, \ldots, v_n)$, if two arguments are equal, say $v_j = v_i$, then

$$\begin{aligned} f(v_1, \ldots, v_i, \ldots, v_j, \ldots, v_n) &= f(v_1, \ldots, v_j, \ldots, v_i, \ldots, v_n) \\ &= f(v_1, \ldots, v_i, \ldots, v_i, \ldots, v_n). \end{aligned}$$

Thus, $f = -f$ implies $f = 0$. So, when two arguments are equal, the function vanishes.

Note: This is often taken as the definition of an alternating multilinear function ($f(v_1, \ldots, v_i, \ldots, v_i, \ldots, v_n) = 0$).

Suppose one argument is a linear combination of the others, let

$$v_k = \alpha_1 v_1 + \cdots + \alpha_{k-1} v_{k-1}.$$

Then,

$$\begin{aligned} f(v_1, \ldots, v_k, \ldots, v_n) &= f(v_1, \ldots, \alpha_1 v_1 + \cdots + \alpha_{k-1} v_{k-1}, \ldots, v_n) \\ &= \sum_{i<k} f(v_1, \ldots, v_i, \ldots, v_i, \ldots, v_n) = 0. \end{aligned}$$

Each term contains two equal arguments, so by alternation each term is zero. Hence the entire sum is zero.

Lastly, if $\sigma \in S_n$ is a permutation of the inputs, then

$$f(v_{\sigma(1)}, \ldots, v_{\sigma(n)}) = \operatorname{sgn}(\sigma) f(v_1, \ldots, v_n).$$

Since the $\operatorname{sgn}(\sigma) = (-1)^k$, where k is the number of swaps in the permutation.

- **Determinant as a function of the rows (or column)**: If you view the determinant as a function of the rows (or columns) of a matrix,

$$\det : \ (\mathbb{R}^n)^n \to \mathbb{R}, \quad (r_1, r_2, ..., r_n) \mapsto \det(A).$$

The determinant det is **linear in each row**, scaling a row by c multiplies the determinant by c. Replacing a row by a sum of two rows adds the corresponding determinants.

$$\det(r_1, \ldots, cr_i, \ldots, r_n) = c \det(r_1, \ldots, r_i, \ldots, r_n), .$$

$$\begin{aligned} \det(A) &= \det(r_1, \ldots, r_i + r_k, \ldots, r_k, \ldots, r_n) \\ &= \det(r_1, \ldots, r_i, \ldots, r_k, \ldots, r_n) + \det(r_1, \ldots, k_i, \ldots, r_k, \ldots, r_n)..\end{aligned}$$

It is also **alternating**: if two rows are equal, the determinant is 0.

Because the determinant is multilinear, if you scale every row of A by α, you multiply the determinant by α once for each row:

$$\det(\alpha A) = \alpha^n \det(A).$$

Thus, the determinant function det is multilinear and alternating.

- **Symmetries of an object**: A symmetry of an object means a transformation that rearranges its components while preserving their structure.

For a set with no additional structure, the only structure-preserving operations are permutations (bijections from the set to itself). Thus, the **symmetries** of a set are the permutations of the set.

- **The symmetric group S_n and permutations**: Consider the set $\{1, 2, ..., n\}$, the symmetries of this set are its permutations. The collection of all such symmetries form the symmetric group S_n. For example, consider the set

$$\{1, 2, 3\}.$$

The symmetries are $(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,2,1),(3,1,2)$, there are exactly 3! symmetries. The collection of these symmetries form the set S_3,

$$S_3 = \{(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,2,1),(3,1,2)\}.$$

Every permutation in S_n is a bijection σ,

$$\sigma : \{1,2,...,n\} \to \{1,2,...,n\},$$

where $\sigma(i)$ tells you which number i is mapped to under the permutation

For example, the permutation $(1,3,2)$ in S_3. For this permutation,

$$\sigma(1) = 1, \quad \sigma(2) = 3, \quad \sigma(3) = 2.$$

The **identity** permutation is

$$\sigma(i) = i.$$

- **Notation for permutations**: We can write a specific permutation as

  $$\sigma = (\sigma(1), \sigma(2), ..., \sigma(n)).$$

  We may also name the permutations with subscripts,

  $$\sigma_1,\ \sigma_2,\ \cdots,\ \sigma_{n!}.$$

  In this case, we denote a specific permutation $\sigma_k \in S_n$ as

  $$\sigma_k = (\sigma_k(1), \sigma_k(2), ..., \sigma_k(n)).$$

  Then, we can write S_n as

  $$S_n = \{\sigma_1, \sigma_2, ..., \sigma_{n!}\}.$$

- **Transpositions and inversions**: A **Transposition** A transposition is a permutation that exchanges two elements and leaves the others fixed. A transposition is written as $(i\ j)$, which means swap i and j. Consider $\{1,2,3\}$, the transposition (1 3) means swap 1 and 3, so the permutation is

  $$(3,2,1).$$

  Any permutation can be written as a product of transpositions. For example,

  $$(3,1,2) = (1\ 3)(1\ 2).$$

  First we swap 1 and 3, then we swap 1 and 2.

  Note: For a given permutation, a decomposition into a product of transpositions is not unique, but the parity of the number of transpositions is unique. You can always insert extra "canceling" swaps or factor swaps differently.

For a permutation written as a list $(\sigma(1), \sigma(2), ..., \sigma(n))$, an **inversion** is a pair (i, j) with $i < j$ but $\sigma(i) > \sigma(j)$

- **Sign of a permutation**: The sign (also called the parity or signature) of a permutation tells you whether the permutation is built from an even or odd number of swaps.

$$\operatorname{sgn}(\sigma) = \begin{cases} +1 & \text{if the permutation is even} \\ -1 & \text{if the permutation is odd} \end{cases}.$$

Given any decomposition of a permutation σ into transpositions,

$$\sigma = \tau_1 \tau_2 \cdots \tau_k,$$

where each τ_i is a transposition, the sign of σ is

$$\operatorname{sgn}(\sigma) = (-1)^k.$$

Similarly, the parity of a permutation can be determined by the number of inversions,

$$\operatorname{sgn}(\sigma) = \begin{cases} +1 & \text{if the number of inversions is even} \\ -1 & \text{if the number of inversions is odd} \end{cases}.$$

Note: S_n contains $n!$ permutations, if $n \geqslant 2$, then there are always $n!/2$ even permutations, and $n!/2$ odd permutations.

- **Composition of permutations**: A composition of permutations means performing one permutation after another. It is the way we "multiply" permutations, and it is the group operation in S_n

If σ and τ are permutations of $\{1, 2, \ldots, n\}$, then their composition is the permutation

$$(\tau \circ \sigma)(i) = \tau(\sigma(i)).$$

For example, if $\sigma = (3, 2, 1)$, and $\tau = (2, 3, 1)$, then the composition $(\tau \circ \sigma)(i) = \tau(\sigma(i))$ is the permutation

$$\begin{aligned}((\tau \circ \sigma)(1), (\tau \circ \sigma)(2), (\tau \circ \sigma)(3)) = (\tau(\sigma(1), \tau(\sigma(2), \tau(\sigma(3))))) &= (\tau(3), \tau(2), \tau(1)) \\ &= (1, 3, 2).\end{aligned}$$

We can write the composition $(\tau \circ \sigma)(i) = \tau(\sigma(i))$ simply as $\tau\sigma$. Both refer to τ composed with σ.

- **Sign of a composition of permutations**: Let $\sigma, \tau \in S_n$. Suppose σ can be written as a product of k transpositions and τ as a product of ℓ transpositions, where k and ℓ are taken modulo 2 (Since transpositions are not unique, but the parity in the number of transpositions is). Then

$$\operatorname{sgn}(\sigma) = (-1)^k, \quad \operatorname{sgn}(\tau) = (-1)^\ell.$$

So,

$$\begin{aligned}\sigma &= \pi_1 \pi_2 \cdots \pi_k, \\ \tau &= \hat{\pi}_1 \hat{\pi}_2 \cdots \hat{\pi}_\ell.\end{aligned}$$

Then, the composition

$$\sigma \circ \tau = \sigma\tau$$

can be constructed by combining the transpositions,

$$\sigma \circ \tau = \sigma\tau = \pi_1\pi_2 \cdots \pi_k\hat{\pi}_1\hat{\pi}_2 \cdots \hat{\pi}_\ell.$$

Thus,

$$\text{sgn}(\sigma \circ \tau) = (-1)^{k+\ell} = (-1)^k(-1)^\ell = \text{sgn}(\sigma)\text{sgn}(\tau).$$

- **The inverse of a permutation**: Consider a permutation σ, we want a permutation τ such that

$$\tau(\sigma(i)) = \sigma(\tau(i)) = i,$$

where i is the identity permutation $i(i) = i$. If $\sigma(i) = j$, then we require $\tau(\sigma(i)) = i$, so $\tau(i) = j$, where $\sigma(j) = i$. For example, if $\sigma = (3, 1, 2)$, then

$$\begin{aligned}\sigma(1) &= 3\\ \sigma(2) &= 1\\ \sigma(3) &= 2.\end{aligned}$$

Thus, $\tau = (2, 3, 1)$, since $\sigma(2) = 1,\ \sigma(3) = 2$, and $\sigma(1) = 3$. Observe that

$$\begin{aligned}(\tau \circ \sigma)(i) &= \tau(\sigma(i)) = (\tau(\sigma(1)), \tau(\sigma(2)), \tau(\sigma(3))) = (\tau(3), \tau(1), \tau(2)) = (1, 2, 3),\\ (\sigma \circ \tau)(i) &= \sigma(\tau(i)) = (\sigma(\tau(1)), \sigma(\tau(2)), \sigma(\tau(3))) = (\sigma(2), \sigma(3), \sigma(1)) = (1, 2, 3).\end{aligned}$$

Thus, $\tau = \sigma^{-1}$

- **Sign of the inverse permutation**: Let $\sigma \in S_n$, and $\sigma^{-1} \in S_n$ be the inverse of σ. Then,

$$\sigma\sigma^{-1} = i,$$

Where i is the identity permutation, $i(i) = i$. Notice that the sign of the identity permutation is $+1$. So,

$$\sigma\sigma^{-1} = i \implies \text{sgn}(\sigma\sigma^{-1}) = \text{sgn}(i) = 1 \implies \text{sgn}(\sigma^{-1}) = \frac{1}{\text{sgn}(\sigma)}.$$

Notice that $\text{sgn}(\sigma) \in \{\pm 1\}$. If $\text{sgn}(\sigma) = 1$, then

$$\text{sgn}(\sigma^{-1}) = \frac{1}{1} = 1 = \text{sgn}(\sigma).$$

If $\text{sgn}(\sigma) = -1$, then

$$\text{sgn}(\sigma^{-1}) = \frac{1}{-1} = -1 = \text{sgn}(\sigma).$$

In either case, $\text{sgn}(\sigma^{-1}) = \text{sgn}(\sigma)$.

- **Inversion map**: Let S_n be the set of all permutations over $\{1, \ldots, n\}$, define the map

$$\Phi :\ S_n \to S_n, \quad \Phi(\sigma) = \sigma^{-1}.$$

Suppose that $\Phi(\sigma_1) = \Phi(\sigma_2)$, then

$$\sigma_1^{-1} = \sigma_2^{-1}.$$

If we take the inverse of both sides,

$$\left(\sigma_1^{-1}\right)^{-1} = \left(\sigma_2^{-1}\right)^{-1} \implies \sigma_1 = \sigma_2.$$

Thus,

$$\Phi(\sigma_1) = \Phi(\sigma_2) \implies \sigma_1 = \sigma_2,$$

so Φ is injective. Next, let $\tau \in S_n$, we require $\sigma \in S_n$ such that $\Phi(\sigma) = \tau$. Choose $\sigma = \tau^{-1}$, then

$$\Phi(\sigma) = \sigma^{-1} = \left(\tau^{-1}\right)^{-1} = \tau.$$

So, every $\tau \in S_n$ is hit by its inverse, and Φ is surjective.

Since Φ is both injective and surjective, Φ is a bijection of S_n.

- **Parity decomposition of S_n**: Consider the set S_n, every even $\sigma \in S_n$ pairs with a permutation of opposite parity. Thus, for any $n \geqslant 2$, there are exactly $n!/2$ even permutations, and $n!/2$ odd permutations.

  Pick any odd permutation τ. The simplest choice is a transposition (a swap), since transpositions are always odd permutations. Let $\tau = (1\ 2)$. Now, define a mapping

  $$f:\ S_n \to S_n, \quad f(\sigma) = \tau\sigma.$$

  This mapping has two crucial properties

  1. Bijective
  2. It flips parity

  $$\text{sgn}(f(\sigma)) = -\text{sgn}(\sigma).$$

  Because composition with a fixed permutation always has an inverse, $f(\sigma)$ is a bijection. Given $f(\sigma) = \tau\sigma$, we can recover σ by multiplying by τ^{-1}

  $$f(\sigma) = \tau\sigma \implies \sigma = \tau^{-1} f(\sigma).$$

  Thus, f is bijective. Furthermore, notice that

  $$\text{sgn}(\tau\sigma) = \text{sgn}(tau)\text{sgn}(\sigma) = (-1)\text{sgn}(\sigma).$$

  Recall that τ is odd, so $\text{sgn}(\tau) = -1$. So, we see that $f(\sigma)$ flips the parity. Thus, $f(\sigma)$ is odd when σ is even, and even when σ is odd.

  Because f is a bijection from S_n to itself, maps odd permutations to even permutations, and even permutations to odd permutations, it must be that the number of even permutations in S_n matches the number of odd permutations in S_n and that number is precisely $n!/2$

  As an example, consider S_3. There are exactly $3! = 6$ permutations, with $6/2 = 3$ of them being odd, and $6/2 = 3$ of them being even. If we generate all the odd permutations by making one swap, then we can generate the remaining three even permutations by defining the bijection $f(\sigma) = \tau\sigma$, where $\tau = (1\ 2)$ is an odd transposition. Let σ_o be the set of odd permutations,

  $$\sigma_o = \{(2,1,3),\ (3,2,1),\ (1,3,2)\}.$$

Then,

$$f(2,1,3) = (1\ 2)(2,1,3) = (1,2,3)$$
$$f(3,2,1) = (1\ 2)(3,2,1) = (2,3,1)$$
$$f(1,3,2) = (1\ 2)(1,3,2) = (3,1,2).$$

So,

$$S_n = \{(2,1,3),\ (3,2,1),\ (1,3,2),\ (1,2,3),\ (2,3,1),\ (3,1,2)\}.$$

- S_n **on the dimensions of a matrix**: Consider $A \in \mathbb{R}^{n\times n}$, $\sigma \in S_n$ tells us how to choose an element from each row. For example, in S_3 if $\sigma_1 = (\sigma_1(1), \sigma_1(2), \sigma_1(3)) = (3,2,1)$,

$$\sigma_1(1) = 3 \quad \text{(choose the element in row 1 column 3)}$$
$$\sigma_1(2) = 2 \quad \text{(choose the element in row 1 column 2)}$$
$$\sigma_1(3) = 1 \quad \text{(choose the element in row 1 column 1).}$$

  Furthermore, $\sigma_1 = (3,2,1) = (1\ 3)$, so $\text{sgn}(\sigma_1) = -1$.

  $\sigma(i) = j$ says from the i^{th} row, choose the element in the j^{th} column. The collection of all permutations S_n tells us all the ways of choosing an element from each row.

  Because permutations are bijections, they guarantee

  - each row gets one column,
  - each column is used exactly once.

- **Geometry of linear transformations and the determinant**: Consider a 1×1 matrix, which is just a number $A = (a)$. It acts on one dimensional vectors (numbers) by multiplication,

$$Ax = ax.$$

  It is a linear map from $\mathbb{R}$ to $\mathbb{R}$. In 1-dimensional space, the only shape is a line. The unit segment is $[0,1]$, with length 1.

  A vector $x \in \mathbb{R}$ is a segment $[0,x]$ with length $|x|$. A transformation Ax on a vector $x \in \mathbb{R}$, where $A = (a)$ scales x by a, so

$$[0,x] \mapsto [0,ax].$$

  Thus, the length is scaled by a. $|x| \mapsto |ax| = |a| \cdot |x|$. Thus, the transformation scales lengths by $|a|$.

  The determinant of this transformation $\det(A)$ is a. So,

$$\text{length after transformation } = |\det(A)| \cdot \text{length before transformation.}$$

  If

  - $a > 0$, the map stretches/compresses without reversing orientation.
  - $a < 0$, the map reflects the line (flips orientation)

  This parallels the 2D and 3D cases

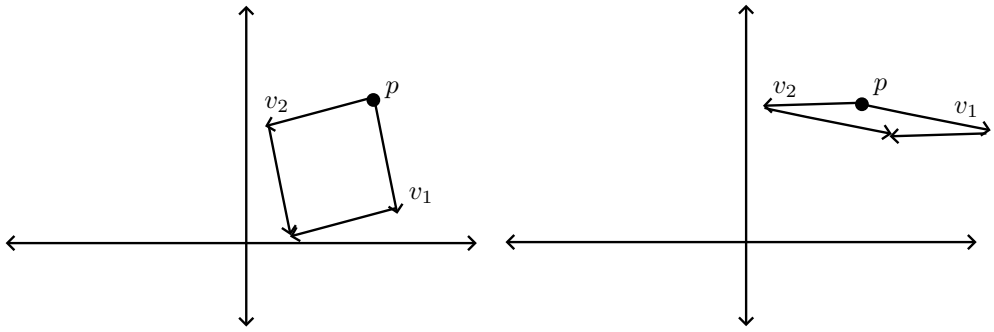
  - **det** $> \mathbf{0}$: Orientation preserved.
  - **det** $< \mathbf{0}$: Orientation reversed.

  In the 2×2 case, $A \in \mathbb{R}^{2\times 2}$ scales area.

Any parallelogram in the plane is determined by two vectors.

$$v_1,\ v_2 \in \mathbb{R}^2.$$

Take any point P in the plane, one side extends from P in the direction of v_1, the adjacent side extends from P in the direction of v_2, and opposite sides run parallel. Any other geometric shape can be approximated by many tiny parallelograms, which is why understanding just parallelograms is enough to understand how a transformation scales area.



$A \in \mathbb{R}^{2\times 2}$ sends each vector to a new vector

$$Av_1, \quad Av_2.$$

So the original parallelogram becomes a new parallelogram with sides

- $v_1' = Av_1$
- $v_2' = Av_2$

So the matrix transforms the shape by transforming the vectors that define it.

For two vectors $v_1, v_2 \in \mathbb{R}^2$, the area of the parallelogram P, $\mathcal{A}(P)$, spanned by v_1 and v_2 is given by

$$\mathcal{A}(P) = \left|\det \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix}\right|.$$

If v_1', v_2' are the vectors transformed by A, where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},\ v_1' = \begin{pmatrix} ax_1 + by_1 \\ cx_1 + dy_1 \end{pmatrix},\ v_2' = \begin{pmatrix} ax_2 + by_2 \\ cx_2 + dy_2 \end{pmatrix},$$

then the area of the parallelogram P' after As effect is

$$\begin{aligned}
\mathcal{A}(P') &= \left|\det \begin{pmatrix} ax_1 + by_1 & ax_2 + by_2 \\ cx_1 + dy_1 & cx_2 + dy_2 \end{pmatrix}\right| = \left|\det\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix}\right)\right| \\
&= \left|\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \det \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix}\right| = \left|\det \begin{pmatrix} a & b \\ c & d \end{pmatrix}\right| \left|\det \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix}\right| \\
&= |\det(A)|\, \mathcal{A}(P).
\end{aligned}$$

Thus, in 2-space, the determinant of A tells us how much the area of the parallelogram P spanned by two vectors v_1 and v_2 is scaled when v_1 and v_2 are transformed by A.

Similarly, for three vectors $v_1, v_2, v_3 \in \mathbb{R}^3$ that define a parallelepiped P, with

$$v_1 = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix}, v_2 = \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix}, v_3 = \begin{pmatrix} x_3 \\ y_3 \\ z_3 \end{pmatrix},$$

the volume of the P is given by

$$\mathcal{V}(P) = \det \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}.$$

Let $A \in \mathbb{R}^{3\times 3}$ be a linear transformation,

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}.$$

Then, A scales P by acting on the vectors v_1, v_2 and v_3. Call the scaled parallelepiped P' with defining sides

$$v_1' = Av_1, \quad v_2' = Av_2, \quad v_3' = Av_3.$$

The new volume is given by

$$\mathcal{V}(P') = \left|\det\left(\begin{bmatrix} Av_1 & Av_2 & Av_3 \end{bmatrix}\right)\right| = \left|\det\left(\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}\begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}\right)\right|$$
$$= \left|\det(A)\det\begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}\right| = |\det(A)|\,\mathcal{V}(P).$$

Thus, in 3-space, the determinant of A tells us how much the volume of the parallelepiped spanned by v_1, v_2 and v_3 is scaled when v_1, v_2, and v_3 are transformed by A. Just like in 1-space and 2-space.

The pattern we have developed in dimensions 1, 2, and 3 generalizes cleanly and conceptually to n-**dimensional space**. The key point is that the determinant always measures how a linear transformation scales n-**dimensional volume**.

Let $A \in \mathbb{R}^{n\times n}$ be a linear transformation, and let

$$v_1, v_2, \ldots, v_n \in \mathbb{R}^n$$

be vectors that define an n-dimensional parallelepiped P. Define the matrix V whose columns define the edges of P,

$$V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

The transformation A acts by sending each edge to

$$v_i' = Av_i.$$

The transformed parallelepiped P' is defined by the columns of

$$AV = \begin{bmatrix} Av_1 & Av_2 & \cdots & Av_n \end{bmatrix}.$$

In $\mathbb{R}^n$, the volume of the n-dimensional parallelepiped spanned by $v_1, v_2, \cdots, v_n$ is

$$\mathcal{V}(P) = |\det(V)|.$$

Which,

– reduces to length in 1D,
– area in 2D,
– volume in 3D,
– and gives the correct n-dimensional volume in general.

The volume of the transformed parallelepiped is

$$\mathcal{V}(P') = |\det(AV)|\,.$$

Using the multiplicativity of determinants,

$$\det(AV) = \det(A)\det(V).$$

Taking absolute values,

$$\mathcal{V}(P') = |\det(A)\det(V)| = |\det(A)|\,\mathcal{V}(P).$$

Thus,

$$\text{For } A \in \mathbb{R}^{n\times n}, \quad \mathcal{V}(Av_1, \ldots, Av_n) = |\det(A)|\,\mathcal{V}(v_1, \ldots, v_n).$$

That is, The determinant of A is the factor by which A scales n-dimensional volume.

– If $\det(A) > 0$, orientation is preserved.
– If $\det(A) < 0$, orientation is reversed.
– If $\det(A) = 0$, all n-dimensional volumes collapse to zero (the image lies in a lower-dimensional subspace)

The determinant is the unique multilinear, alternating function that measures how a linear transformation scales oriented n-dimensional volume.

- **Properties of volume**: Let $v_1, v_2, \ldots, v_k \in \mathbb{R}^n$ be the edge vectors of a k-dimensional parallelepiped embedded in the ambient space $\mathbb{R}^n$. Note that $k \leqslant n$.

  The volume of the k-dimensional shape is given by

$$\mathcal{V}(v_1, \ldots, v_k) = \sqrt{\det\left(\text{Gram}(v_1, \ldots, v_k)\right)},$$

  where

$$G = \text{Gram}(v_1, \ldots, v_k) = (v_i^T v_j)_{i,j=1}^k = \begin{pmatrix} v_1^T v_1 & \cdots & v_1^T v_k \\ \vdots & \ddots & \vdots \\ v_k^T v_1 & \cdots & v_k^T v_k \end{pmatrix} \in \mathbb{R}^{k\times k}.$$

  Suppose we scale one of the edge vectors $v_i \to \lambda v_i = v_i'$, for $\lambda \in \mathbb{R}$. Notice what happens to $v_i'^T v_j$

$$\begin{aligned} v_i'^T v_i' &= \|v_i'\|_2^2 = \|\lambda v_i\|_2^2 = \|\lambda v_i\|_2 \|\lambda v_i\|_2 = \lambda^2 \|v_i\|_2^2\,, \\ v_i'^T v_j &= \lambda v_i^T v_j. \end{aligned}$$

  So,

$$G' = \text{Gram}(v_1, \ldots, v_i', \ldots, v_k) = \begin{pmatrix} v_1^T v_1 & \cdots & \lambda v_1^T v_i & \cdots & v_1^T v_k \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \lambda v_i^T v_1 & \cdots & \lambda^2 v_i^T v_i & \cdots & \lambda v_i^T v_k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_k^T v_1 & \cdots & \lambda v_k^T v_i & \cdots & v_k^T v_k \end{pmatrix}.$$

Notice that one row and one column are scaled by λ, so

$$\det(G') = \lambda^2 \det(G).$$

Which implies

$$\mathcal{V}' = \mathcal{V}(v_1, \ldots, v_i', \ldots, v_k) = \sqrt{\lambda^2 \det(G)} = |\lambda| \, \mathcal{V}(v_1, \ldots, v_k).$$

When an edge vector is replaced by a linear combination, $v_i = u_1 + u_2$,

$$\mathcal{V}(v_1, \ldots, v_i, \ldots, v_k) = \mathcal{V}(v_1, \ldots, u_1, \ldots, v_k) + \mathcal{V}(v_1, \ldots, u_2, \ldots, v_k).$$

Lastly, volume is alternating, swapping two edge vectors swaps the sign of the oriented volume.

- **The goal of the determinant**: The determinant of an $n \times n$ matrix is a single number that measures How the linear transformation encoded by the matrix scales oriented volume
  - A 1×1 matrix scales length.
  - A 2×2 matrix scales area.
  - A 3×3 matrix scales volume.

  Thus, we need a function that given the edge vectors of a n-dimensional parallelepiped, outputs the volume. And, given a linear transformation, outputs the scaling factor of the transformed parallelepiped when the linear transformation acts on the edge vectors of the solid.

- **The Leibniz formula for the determinant**: Let $A = (a_{ij})$ be an $n \times n$ matrix. We want a function $\det(A)$ with certain properties. Specifically, we want
  1. Linear in each row
  2. Alternating (swapping rows changes the sign)
  3. Normalized so $\det(I) = 1$

  The core idea is to choose on entry per row and per column. A determinant term must use
  - One entry from row 1,
  - One entry from row 2,
  - $\cdots$
  - One entry from row n.

  But, you cannot reuse a column, the order of column choices must matter. Thus, the correct way to index the choices is

  $$\text{choose column } \sigma(i) \text{ for row } i,$$

  where σ is permutation. Because permutations are bijections, they guarantee
  - Each row gets one column,
  - Each column is used exactly once.

  There are $n!$ such choices, which is precisely the size of S_n.

For a given permutation σ, consider the product

$$\prod_{i=1}^{n} a_{i,\sigma(i)}.$$

This is the product of the entries chosen by the pattern σ. For example, in S_3, if $\sigma_1 = (3, 2, 1)$, then

$$\prod_{i=1}^{3} a_{i,\sigma_1(i)} = a_{13}a_{22}a_{31}.$$

Each permutation $\sigma_k \in S_n$ gives one such product. There are exactly $n!$ products.

Now, to get the determinant, we need to sum these products, but also include the sign of the permutation. The determinant of a square matrix $A \in \mathbb{R}^{n\times n}$ is given by

$$\det(A) = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}.$$

If we did not include the sign of the permutation, this would force

- matrices with two equal rows to have nonzero determinant (incorrect),
- row-swapping to leave determinant unchanged (incorrect),
- multiplicativity to fail.

So we need something that makes the sum change sign whenever two rows are swapped.

Swapping two rows reverses the order of multiplication in each term. If the original product was

$$a_{i,\sigma(i)}a_{j,\sigma(j)},$$

then after swapping rows i and j, the product becomes

$$a_{j,\sigma(j)}a_{i,\sigma(i)}.$$

This corresponds to replacing each permutation σ by a new permutation

$$(i\ j)\sigma.$$

Where $(i\ j)$ is a transposition. Since a transposition flips the parity of a permutation,

$$\operatorname{sgn}((i\ j)\sigma) = -\operatorname{sgn}(\sigma).$$

This is exactly the behavior needed for the determinant.

- **Determinant identity**: Let $v_1, v_2, \ldots, v_n \in \mathbb{R}^n$. Define $A = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \in \mathbb{R}^{n\times n}$. Recall the Gram matrix

$$G = (v_i^T v_j)_{i,j=1}^{n} = A^T A.$$

So,

$$\det(G) = \det(A^T A) = \det(A^T)\det(A) = (\det(A))^2.$$

Thus,

$$(\det(v_1, \ldots, v_n))^2 = \det \begin{pmatrix} v_1^T v_1 & v_1^T v_2 & \cdots & v_1^T v_n \\ v_2^T v_1 & v_2^T v_2 & \cdots & v_2^T v_n \\ \vdots & \vdots & \ddots & \vdots \\ v_n^T v_1 & v_n^T v_2 & \cdots & v_n^T v_n \end{pmatrix}.$$

- **Determinant of the transpose**: Suppose $A \in \mathbb{R}^{n\times n}$, then $A^T \in \mathbb{R}^{n\times n}$, and

$$\det(A^T) = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} (A^T)_{i,\sigma(i)} = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(i),i}.$$

- **Efficient determinants**: Consider the Leibniz determinant formula. For $A \in \mathbb{R}^{n\times n}$

$$\det(A) = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}.$$

There are exactly

$$nPn = \frac{n!}{(n-n)!} = n!$$

permutations in S_n, and for each permutation we take the product of n terms in A. Thus, using this formula requires

$$n \cdot n!$$

flops. Instead, we can compute the LU decomposition $PA = LU$, which requires roughly $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ flops. Then,

$$PA = LU \implies A = P^T LU.$$

Where P is a permutation matrix, L is unit lower triangular, and U is upper triangular. So,

$$\det(A) = \det(P^T LU) = \det(P^T)\det(L)\det(U).$$

Notice that since P is a permutation matrix, which is just I after k row swaps, $\det(P^T) = \det(P) = \pm 1$. Also, L is unit lower triangular, so

$$\det(L) = \prod_{i=1}^{n} \ell_{ii} = 1,$$

since $\ell_{ii} = 1$ for $i = 1, \ldots, n$. Thus,

$$\det(A) = \operatorname{sgn}(P)\det(U) = \operatorname{sgn}(P)\prod_{i=1}^{n} u_{ii}.$$

Note that if P is obtained from the identity by k row swaps, then

$$\operatorname{sgn}(P) = (-1)^k.$$

Thus, computing the determinant in this way requires only

$$\frac{2}{3}n^3 + n + \mathcal{O}(n^2) \approx \frac{2}{3}n^3 = \mathcal{O}(n^3)$$

flops. If LU is already known, the determinant can be found in only n flops.

6.6.1 Determinant proofs

- **Multiplicative Property of Determinants**: Suppose $A, B \in \mathbb{R}^{n\times n}$, then $\det(AB) = \det(A)\det(B)$

  Proof. Assume $A, B \in \mathbb{R}^{n\times n}$, then

  $$\det(AB) = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} (AB)_{i,\sigma(i)}.$$

  But, by matrix multiplication,

  $$(AB)_{i,\sigma(i)} = \sum_{k=1}^{n} a_{ik} b_{k,\sigma(i)}.$$

  So,

  $$\det(AB) = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} (AB)_{i,\sigma(i)} = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} \sum_{k=1}^{n} a_{ik} b_{k,\sigma(i)}.$$

  Since ordinary multiplication is multilinear, we can use the expansion property. Thus,

  $$\prod_{i=1}^{n} \sum_{k=1}^{n} = \sum_{k_1=1}^{n} \cdots \sum_{k_n=1}^{n} \prod_{i=1}^{n} a_{ik_i} b_{k_i,\sigma(i)} = \sum_{k_1,\ldots,k_n} \prod_{i=1}^{n} a_{ik_i} b_{k_i,\sigma(i)},$$

  where $(k_1, \ldots, k_n) \in \{1, \ldots, n\}^n$. Thus,

  $$\sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} \sum_{k=1}^{n} a_{ik} b_{k,\sigma(i)} = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \sum_{k_1,\ldots,k_n} \prod_{i=1}^{n} a_{ik_i} b_{k_i,\sigma(i)}.$$

  Notice that we can split the product, $\prod_{i=1}^{n} a_{ik_i} b_{k_i,\sigma(i)} = \prod_{i=1}^{n} a_{ik_i} \prod_{i=1} b_{k_i,\sigma(i)}$. So,

  $$\sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \sum_{k_1,\ldots,k_n} \prod_{i=1}^{n} a_{ik_i} b_{k_i,\sigma(i)} = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \sum_{k_1,\ldots,k_n} \prod_{i=1}^{n} a_{ik_i} \prod_{i=1}^{n} b_{k_i,\sigma(i)}.$$

  Now, since we are summing a finite number of terms, we can interchange the sums

  $$\sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \sum_{k_1,\ldots,k_n} \prod_{i=1}^{n} a_{ik_i} \prod_{i=1}^{n} b_{k_i,\sigma(i)} = \sum_{k_1,\ldots,k_n} \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{ik_i} \prod_{i=1}^{n} b_{k_i,\sigma(i)}.$$

  But, notice that $\prod_{i=1}^{n} a_{ik_i}$ does not depend on σ. So, we can move it outside the sum

  $$\sum_{k_1,\ldots,k_n} \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{ik_i} \prod_{i=1}^{n} b_{k_i,\sigma(i)} = \sum_{k_1,\ldots,k_n} \left(\prod_{i=1}^{n} a_{ik_i} \right) \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} b_{k_i,\sigma(i)}.$$

  Now the structure is becoming clear. Let's consider an iteration of the outer sum. So, fix an index tuple $(k_1, \ldots, k_n)$, which is a permutation of the indices in the long form

  $$\sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \cdots \sum_{k_n=1}^{n}.$$

  Now, this index tuple is fixed for all iterations of the inner sum, so this index tuple is used for all $\sigma \in S_n$. If two indices coincide, say $k_p = k_q$ with $p \neq q$, then the inner sum

  $$\sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} b_{k_i,\sigma(i)}$$

  vanishes. This expression is precisely the Leibniz formula for a determinant whose matrix has two identical rows (rows k_p and k_q). Determinants are alternating, so this term is zero.

Thus, only index tuples $(k_1, \ldots, k_n)$ with all entries distinct contribute. If all k_i are distinct and each lies in $\{1, \ldots, n\}$, then

$$(k_1, \ldots, k_n)$$

is a permutation of $(1, \ldots, n)$. So, there exists a unique permutation $\tau \in S_n$ such that

$$k_i = \tau(i).$$

Thus the sum over index tuples collapses to a sum over permutations

$$\sum_{k_1,\ldots,k_n} \left(\prod_{i=1}^{n} a_{ik_i} \right) \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} b_{k_i,\sigma(i)} = \sum_{\tau \in S_n} \left(\prod_{i=1}^{n} a_{i,\tau(i)} \right) \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} b_{\tau(i),\sigma(i)}.$$

Now, notice that $\tau \in S_n$. Thus, τ is a bijection of $\{1, \ldots, n\}$, as i runs over $1, \ldots, n$, so does $\tau(i)$. Let

$$\sigma' = \sigma\tau^{-1}, \quad \sigma = \sigma'\tau.$$

So, $\sigma(i) = \sigma'(\tau(i))$. Thus,

$$\prod_{i=1}^{n} b_{\tau(i),\sigma(i)} = \prod_{i=1}^{n} b_{\tau(i),\sigma'(\tau(i))}.$$

Since τ is a bijection of $\{1, \ldots, n\}$, we can set $\tau(i) = j$, where j runs over $\{1, \ldots, n\}$. Now, we have

$$\prod_{i=1}^{n} b_{\tau(i),\sigma'(\tau(i))} = \prod_{j=1}^{n} b_{j,\sigma'(j)}.$$

Since $\sigma = \sigma'\tau$,

$$\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma'\tau) = \operatorname{sgn}(\sigma')\operatorname{sgn}(\tau).$$

So, we have

$$\sum_{\tau \in S_n} \left(\prod_{i=1}^{n} a_{i,\tau(i)} \right) \sum_{\sigma' \in S_n} \operatorname{sgn}(\sigma')\operatorname{sgn}(\tau) \prod_{j=1}^{n} b_{j,\sigma'(j)}.$$

Notice by the start of the inner sum τ is a constant factor. So we can move it outside,

$$\sum_{\tau \in S_n} \left(\prod_{i=1}^{n} a_{i,\tau(i)} \right) \operatorname{sgn}(\tau) \sum_{\sigma' \in S_n} \operatorname{sgn}(\sigma') \prod_{j=1}^{n} b_{j,\sigma'(j)}.$$

Notice, we now have $\det(B)$,

$$\sum_{\tau \in S_n} \left(\prod_{i=1}^{n} a_{i,\tau(i)} \right) \operatorname{sgn}(\tau) \sum_{\sigma' \in S_n} \operatorname{sgn}(\sigma') \prod_{j=1}^{n} b_{j,\sigma'(j)} = \sum_{\tau \in S_n} \left(\prod_{i=1}^{n} a_{i,\tau(i)} \right) \operatorname{sgn}(\tau) \det(B).$$

Lastly, notice that we can move the $\operatorname{sgn}(\tau)$ so that it is the first term instead of the last,

$$\sum_{\tau \in S_n} \left(\prod_{i=1}^{n} a_{i,\tau(i)} \right) \operatorname{sgn}(\tau) \det(B) = \sum_{\tau \in S_n} \operatorname{sgn}(\tau) \left(\prod_{i=1}^{n} a_{i,\tau(i)} \right) \det(B).$$

Which is precisely $\det(A)\det(B)$. Thus,

$$\sum_{\tau \in S_n} \operatorname{sgn}(\tau) \left(\prod_{i=1}^{n} a_{i,\tau(i)} \right) \det(B) = \det(A)\det(B).$$

And we conclude that $\det(AB) = \det(A)\det(B)$. ■

- **Transpose invariance**: Suppose $A \in \mathbb{R}^{n\times n}$, then

$$\det(A^T) = \det(A).$$

Proof. Assume $A \in \mathbb{R}^{n\times n}$, then $A^T \in \mathbb{R}^{n\times n}$, and

$$\det(A^T) = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(i),i}.$$

Let $j = \sigma(i)$, then $i = \sigma^{-1}(j)$. Recall that $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma^{-1})$, so

$$\sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(i),i} = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma^{-1}) \prod_{j=1}^{n} a_{j,\sigma^{-1}(j)}.$$

Let Φ be the inversion map

$$\Phi :\ S_n \to S_n, \quad \Phi(\sigma) = \sigma^{-1}.$$

Since Φ is a bijection over S_n,

$$\sum_{\sigma\in S_n} f(\sigma) = \sum_{\sigma\in S_n} f(\sigma^{-1}).$$

Let $\tau = \sigma^{-1}$, then

$$\sum_{\sigma\in S_n} \operatorname{sgn}(\sigma^{-1}) \prod_{j=1}^{n} a_{j,\sigma^{-1}(j)} = \sum_{\tau\in S_n} \operatorname{sgn}(\tau) \prod_{j=1}^{n} a_{j,\tau(j)} = \det(A).$$

- **Determinant of the inverse**: Suppose $A \in \mathbb{R}^{n\times n}$ is invertible, then

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

Proof. Assume $A \in \mathbb{R}^{n\times n}$ admits an inverse $A^{-1} \in \mathbb{R}^{n\times n}$, where

$$AA^{-1} = A^{-1}A = I.$$

So,

$$\det(AA^{-1}) = \det(I) = 1 \implies \det(A)\det(A^{-1}) = 1$$
$$\therefore \det(A^{-1}) = \frac{1}{\det(A)}.$$

As desired. ■

Corollary. From above, we see

$$\det(A) = \frac{1}{\det(A^{-1})}.$$

- **Determinant of triangular matrices**:

6.7 Chapter 1: Gaussian Elimination and its variants

6.7.1 Definitions

- **Matrix multiplication**: If A is an $n \times m$ matrix, and X is $m \times p$, we can form the product $B = AX$, which is $n \times p$. The (i, j) entry of B is

$$b_{ij} = \sum_{k=1}^{m} a_{ik} x_{kj}.$$

- **Triangular matrix**: A matrix $G = (g_{ij})$ is *lower triangular* if $g_{ij} = 0$ whenever $i < j$. Thus a lower-triangular matrix has the form

$$G = \begin{bmatrix} g_{11} & 0 & 0 & \cdots & 0 \\ g_{21} & g_{22} & 0 & \cdots & 0 \\ g_{31} & g_{32} & g_{33} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ g_{n1} & g_{n2} & g_{n3} & \cdots & g_{nn} \end{bmatrix}.$$

  Similarly, an *upper triangular* matrix is one for which $g_{ij} = 0$ whenever $i > j$. A *triangular* matrix is one that is either upper or lower triangular.

- **Positive definite matrix**: A square matrix A is positive definite provided it satisfies

  1. $A = A^\top$
  2. $x^\top Ax > 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$

- **Column envelope**: The envelope of a sparse matrix is the set of positions around the diagonal that "must be included" when storing or working with the matrix in a compressed way.

  It essentially captures the profile of where the nonzeros start (or stop) in each column.

  Let $A = (a_{ij})$. Define for each column j

$$q(j) = \min\{i : \ a_{ij} \neq 0\}$$

  I.e the first nonzero entry in column j. Then, the column envelope of column j is

$$\{(i, j) : \ q(j) \leqslant i \leqslant j\}$$

  So, for column j, you start at the first nonzero entry $q(j)$ and include all positions down to the diagonal ($i = j$), whether or not some of them are explicitly zero.

  The column envelope of the matrix A is the union over all columns:

$$\text{colenv}\{A\} = \bigcup_{j=1}^{n} \{(i, j) : \ q(j) \leqslant i \leqslant j\}$$

  with

$$q(j) = \min\{i : \ a_{ij} \neq 0\}$$

- **Row envelope**: For each row i find the first nonzero entry

$$p(i) = \min\{j : \ a_{ij} \neq 0\}.$$

  Then, the row envelope of row i is

$$\{(i,j) : \ p(i) \leqslant j \leqslant i\}$$

  So, the row envelope is in the lower triangular part ($j \leqslant i$)

  The row envelope of A is then

$$\text{rowenv}\{A\} = \bigcup_{i=1}^{m} \{(i,j) : \ p(i) \leqslant j \leqslant i\}$$

- **Envelope**: The envelope is the union of both envelopes. That is,

$$\begin{aligned} \text{env}\{A\} =& \text{rowenv}\{A\} = \bigcup_{i=1}^{m} \{(i,j) : \ p(i) \leqslant j \leqslant i\} \\ \cup& \text{colenv}\{A\} = \bigcup_{j=1}^{n} \{(i,j) : \ q(j) \leqslant i \leqslant j\} \end{aligned}$$

- **Elementary operations on systems**:

  1. Interchange rows.
  2. Multiply an equation by a nonzero constant.
  3. Add a multiple of one equation to another equation.

- **Transpose of block matrices**: Let $A \in \mathbb{R}^{n \times n}$, with

$$A = \begin{bmatrix} A_{11} & a_{12} \\ A_{21} & a_{22} \end{bmatrix}.$$

  Then,

$$A^\top = \begin{bmatrix} A_{11}^\top & A_{21}^\top \\ A_{12}^\top & A_{22}^\top \end{bmatrix}$$

- **Transpose of a block vector**: Similarly, if $x \in \mathbb{R}^n$ is decomposed into blocks

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

  with $x_1 \in \mathbb{R}^{n_1}, \ x_2 \in \mathbb{R}^{n_2}, \ n = n_1 + n_2$, then

$$x^\top = \begin{pmatrix} x_1^\top & x_2^\top \end{pmatrix}$$

- **Nonsingular matrix**: A nonsingular matrix is a matrix that has an inverse
- **Singular matrix**: A singular matrix is a matrix that does not have an inverse
- **Positive definite matrix**: A matrix A is **positive definite** provided that the following two conditions are satisfied

  1. A is symmetric. That is, $A = A^\top$
  2. $x^\top Ax > 0$ for all $x \neq 0$

- **Cholesky decomposition and the Cholesky Factor**: Let $A \in \mathbb{R}^{n \times n}$ be p.d, then $A = R^\top R$ where R is upper triangular with $r_{ii} > 0$. The matrix R is called the **Cholesky factor**.

If $A = R^\top R$, then $Ax = b$ can be written as

$$R^\top Rx = b$$

where

$$\begin{cases} Rx & = y \quad \text{(Lower triangular)} \\ R^\top y & = b \quad \text{(Upper triangular)} \end{cases}$$

and since these new systems are triangular, they can be solved quickly with forward or backward substitution.

- **Banded matrix**: A banded matrix is a sparse matrix whose nonzero entries are confined to a diagonal band, consisting of the main diagonal and a fixed number of diagonals on either side of it.

  Let $A \in \mathbb{R}^{m \times n}$. Then A is called a **banded matrix** if there exist nonnegative integers p, q (called the *lower* and *upper bandwidths*) such that

  $$a_{ij} = 0 \quad \text{whenever } i - j > p \text{ or } j - i > q.$$

  - The *lower bandwidth* p is the number of subdiagonals (below the main diagonal) that may contain nonzero entries.
  - The *upper bandwidth* q is the number of superdiagonals (above the main diagonal) that may contain nonzero entries.

  The *total bandwidth* is sometimes defined as $p + q + 1$, counting the main diagonal as well.

- LU **decomposition**: Consider a matrix $A \in \mathbb{R}^{n \times n}$. If we can factor A as $A = LU$, for L lower triangular, U upper triangular, then the system $Ax = b$, for vectors $x, b \in \mathbb{R}^n$ turns into

  $$LUx = b.$$

  We can then split this system as follows

  $$\begin{cases} Ly & = b \\ Ux & = y \end{cases}$$

  First, we solve $Ly = b$ with forward substitution to find y. We can then solve $Ux = y$ with backward substitution to find the target x.

- **More definiteness**: Let $A \in \mathbb{R}^{n \times n}$ be symmetric. Then,
  - A is **positive semidefinite** if $x^\top Ax \geqslant 0$ for all $x \in \mathbb{R}^n$
  - A is negative definite if $x^\top Ax < 0$ for all nonzero $x \in \mathbb{R}^n$
  - A is negative semidefinite if $x^\top Ax \leqslant 0$ for all $x \in \mathbb{R}^n$

- **Definiteness notation**: Let $A \in \mathbb{R}^{n \times n}$ by symmetric.
  - $A \succ 0$ means A is positive definite.
  - $A \succeq 0$ means A is positive semidefinite
  - $A \prec 0$ means A is negative definite
  - $A \preceq 0$ means A is negative semidefinite

- **Ill-conditioned problem**: A problem P is **ill-conditioned** if tiny variations in the information of P leads to large variations of the solution of P

Consider a linear system. If A, b is the information of the problem, then for a slightly perturbed A and b, call them $\bar{A}, \bar{b}$, where $\bar{A} - A$ and $\bar{b} - b$ tiny, an ill-conditioned problem will have

$$\bar{x} - x$$

large, if x is the solution to $Ax = b$, and $\bar{x}$ is the solution to the system with $\bar{A}$ and $\bar{b}$.

If the problem is **well-conditioned**, $\bar{x} - x$ is also tiny.

- **Permutation matrix**: A permutation matrix is a special kind of square matrix that represents a permutation of elements. Formally: It is obtained from the identity matrix by rearranging its rows (or equivalently, its columns).

  Each row and each column has exactly one entry equal to 1, and all other entries are 0.

  Multiplying a vector (or another matrix) by a permutation matrix reorders its entries.

  Suppose P is formed by taking I and interchanging rows one and two. Then,

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

  Then,

  $$P \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} b \\ a \\ c \end{pmatrix}.$$

  So, it swaps the first two entries.

6.7.2 Properties

- **Eigenvalues of a triangular or diagonal matrix**: The eigenvalues are simply the entries on the main diagonal. Suppose $A \in \mathbb{R}^{n \times n}$ is triangular or diagonal. Then, A is one of

$$A = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ 0 & a_{22} & ... & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & a_{nn} \end{bmatrix},$$

$$\begin{bmatrix} a_{11} & 0 & ... & 0 \\ a_{21} & a_{22} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & ... & a_{nn} \end{bmatrix},$$

$$\begin{bmatrix} a_{11} & 0 & ... & 0 \\ 0 & a_{22} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & a_{nn} \end{bmatrix}.$$

  Then, $\lambda_1 = a_{11},\ \lambda_2 = a_{22},\ ...,\ \lambda_n = a_{nn}$

- **Properties of a nonsingular matrix**: The following are equivalent, if any one holds, they all hold
  - $Ax = b$ has a unique solution
  - $\det(A) \neq 0$
  - A^{-1} exists
  - There is no nonzero vector $y \in \mathbb{R}^m$ such that $Ay = 0$
  - The columns of A are linearly independent
  - The rows of A are linearly independent
  - Given any vector b, there is exactly one vector x such that $Ax = b$

  If any one of the following are true, they all are true, and A is nonsingular

- **More properties of nonsingular matrices**:
  1. The product of nonsingular matrices is nonsingular.
  2. The inverse of a nonsingular matrix is nonsingular (obvious)
  3. The sum of two nonsingular matrices **may not be** nonsingular
  4. A nonsingular matrix scaled by a nonzero scalar is nonsingular

- **Triangular matrices**: Triangular matrices are invariant under multiplication, transposition, and inversion
  - Upper triangular $\times$ upper triangular = upper triangular
  - Lower triangular $\times$ lower triangular = lower triangular
  - The transpose of an upper triangular matrix is a lower triangular matrix
  - The transpose of an lower triangular matrix is a upper triangular matrix
  - The inverse of a lower triangular matrix is lower triangular, and the inverse of an upper triangular matrix is upper triangular

- **Properties of positive definite (p.d) matrices**:
  1. If A is p.d then A is *nonsingular*

Note: Since A is nonsingular there is no $y \in \mathbb{R}^n$, $y \neq 0$ such that $Ay = 0$

2. If $A = M^\top M$ for some M nonsingular than A is p.d
3. If A is p.d than $\det(A) > 0$
4. If A is p.d then all principal submatrices are p.d
5. If A is p.d then $a_{ii} > 0$ for $i = 1, 2, ..., n$. So, if any $a_{ii} \leqslant 0$, A is not p.d.
6. A is p.d if and only if all leading principal minors are positive
7. A is p.d if and only if there exists a unique upper triangular matrix R such that $A = R^\top R$ (Cholesky factorization described below)
8. A is p.d if and only if all eigenvalues of A are positive

   Recall that λ is an eigenvalue of A if there exists $x_\lambda \neq 0$ such that $Ax_\lambda = \lambda x_\lambda$

Note: Property two is a key property.

- **Relationship between definiteness and eigenvalue signs**:

Definiteness	**Eigenvalue signs**	**Name**
$A \succ 0$	all $\lambda_i > 0$	Positive definite
$A \succeq 0$	all $\lambda_i \geqslant 0$	Positive semidefinite
$A \prec 0$	all $\lambda_i < 0$	Negative definite
$A \preceq 0$	all $\lambda_i \leqslant 0$	Positive definite

Note: Eigenvalue signs are a necessary and sufficient condition for definiteness. Also, recall that in all cases, A is symmetric, so all eigenvalues are real.

- **Cholesky factor R in a diagonal matrix**: If $A = D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$ then

$$R = \begin{bmatrix} \sqrt{a_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{a_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{a_{nn}} \end{bmatrix}$$

- *LU* **factorization in a symmetric matrix**:
- **Properties of a permutation matrix**:

1. **Orthogonal:** $P^T = P^{-1}$
2. **Determinant:**: $\det(P) = \pm 1$, depending on whether the permutation is even or odd.
3. **Action on vectors:**: Px permutes the coordinates of x
4. **Action on matrices:** Left multiplication permutes rows; right multiplication permutes columns.

6.7.3 Theorems

- **Theorem 1.1.19**: Let $A \in \mathbb{R}^{m\times n}$,$X \in \ltimes \times \imath$ and $B \in \mathbb{R}^{m\times p}$ be partitioned as follows

$$A = \begin{matrix} & \begin{matrix} m_1 & m_2 \end{matrix} \\ \begin{matrix} n_1 \\ n_2 \end{matrix} & \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \end{matrix}, \quad \begin{cases} m_1 + m_2 = m \\ n_1 + n_2 = n \end{cases}$$

$$X = \begin{matrix} & \begin{matrix} p_1 & p_2 \end{matrix} \\ \begin{matrix} n_1 \\ n_2 \end{matrix} & \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \end{matrix}, \quad \begin{cases} p_1 + p_2 = p \\ n_1 + n_2 = n \end{cases}$$

$$B = \begin{matrix} & \begin{matrix} p_1 & p_2 \end{matrix} \\ \begin{matrix} m_1 \\ m_2 \end{matrix} & \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \end{matrix}, \quad \begin{cases} p_1 + p_2 = p \\ m_1 + m_2 = m \end{cases}$$

  Then $AX = B$ if and only if

$$A_{i1}X_{1j} = B_{ij} \quad \text{for } i, j = 1, 2$$

- **Theorem 1.1.24**: Make a finer partition of A into r block rows and s block columns.

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1s} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rs} \end{bmatrix}, \quad \begin{matrix} n_1 + \cdots + n_r = n \\ m_1 + \cdots + m_s = m \end{matrix}$$

  Then partition X *conformably* with A; that is, make the block row structure of X identical to the block column structure of A.

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1t} \\ \vdots & \ddots & \vdots \\ X_{s1} & \cdots & X_{st} \end{bmatrix}, \quad \begin{matrix} m_1 + \cdots + m_s = m \\ p_1 + \cdots + p_t = p \end{matrix}$$

  Finally, partition the product B conformably with both A and X.

$$B = \begin{bmatrix} B_{11} & \cdots & B_{1t} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rt} \end{bmatrix}, \quad \begin{matrix} n_1 + \cdots + n_r = n \\ p_1 + \cdots + p_t = p \end{matrix}$$

  Theorem: Let A, X and B be partitioned like they are above. Then, $AX = B$ if and only if

$$B_{ij} = \sum_{k=1}^{s} A_{ik}X_{kj} \quad i = 1, ..., r, \ j = 1, ..., t$$

- **Theorem 1.2.3**: Let A be a square matrix. The following six conditions are equivalent; that is, if any one holds, they all hold.

  (a) A^{-1} exists.

  (b) There is no nonzero y such that $Ay = 0$.

  (c) The columns of A are linearly independent.

  (d) The rows of A are linearly independent.

(e) $\det(A) \neq 0$.

(f) Given any vector b, there is exactly one vector x such that $Ax = b$.

- **Theorem 1.3.1**: Let G be a triangular matrix. Then G is **nonsingular** if and only if $g_{ij} \neq 0$ for $i = 1, ..., n$
- **Theorem 1.4.2**: If A is positive definite, then A is nonsingular
- **Corollary 1.4.3**: If A is positive definite, the linear system $Ax = b$ has exactly one solution.
- **Theorem 1.4.4**: Let M be any $n \times n$ nonsingular matrix, and let $A = M^\top M$. Then A is positive definite.
- **Theorem 1.4.7 (*Cholesky Decomposition Theorem*)**: Let A be positive definite. Then A can be decomposed in exactly one way into a product

$$A = R^\top R$$

such that R is upper triangular and has all main diagonal entries r_{ii} positive. R is called the Cholesky factor of A.

We have

$$r_{ii} = +\sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2},$$
$$r_{ij} = \frac{\left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}\right)}{r_{ii}} \quad j = i+1, ..., n$$

Note: R is upper triangular, so we do not have to calculate entries r_{ij} for $i > j$

- **Theorem 1.5.7**: Let A be positive definite, and let R be the Cholesky factor of A. Then R and A have the same envelope.
- **Theorem**: Let A be p.d, if R is the Cholesky factor of A, then

$$\text{colenv}\{R\} = \text{colenv}\{A\}$$

- **Theorem**: Let $A \in \mathbb{R}^{n \times n}$ be nonsingular. Then, we can solve the system $Ax = b$, $b \in \mathbb{R}^n$ using Gaussian Elimination without row interchanges if and only if all landing principal sub-matrices of A are nonsingular.
- **Theorem**: Let $A \in \mathbb{R}^{n \times n}$. Then A admits an LU factorization

$$A = LU,$$

where $L \in \mathbb{R}^{n \times n}$ is unit lower triangular and $U \in \mathbb{R}^{n \times n}$ is upper triangular, **without row interchanges**, if and only if all leading principal submatrices of A are nonsingular.

- **Theorem 1.7.19** ***(LU Decomposition Theorem)***: Let A be an $n \times n$ matrix whose leading principal submatrices are all nonsingular. Then, A can be decomposed in exactly one way into a product $A = LU$ such that L is unit lower triangular and U is upper triangular.

6.7.4 Propositions

- **Proposition 1.1.6**: If $b = Ax$, then b is a linear combination of the columns of A.

  If we let A_j denote the jth column of A, we have

  $$b = \sum_{j=1}^{m} A_j x_j.$$

- **Proposition 1.4.24**: Cholesky's algorithm (Row-oriented inner product formals) applied to an $n \times n$ matrix performs about $\frac{n^3}{3}$ flops.

- **Proposition 1.4.51**: If A is positive definite, then $a_{ii} > 0$ for $i = 1, 2, .., n$

- **Proposition 1.4.53**: Let A be positive definite, and consider a partition

  $$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

  in which A_{11} and A_{22} are square. Then A_{11} and A_{22} are positive definite.

- **Proposition 1.4.55** If A and X are $n \times n$, A is positive definite, and X is nonsingular then the matrix $B = X^\top AX$ is also positive definite.

- **Proposition 1.7.1**: If $\hat{A}x = \hat{b}$ is obtained from $Ax = b$ by an elementary operation of type 1,2, or 3, then the systems $Ax = b$ and $\hat{A}x = \hat{b}$ are equivalent.

- **Proposition 1.7.3**: Suppose $\hat{A}$ is obtained from A by an elementary row operation of type 1, 2, or 3. Then $\hat{A}$ is nonsingular if and only if A is.

6.7.5 Algorithms and complexities

- **Matrix multiplication**: $\mathcal{O}(n^3)$
- **Row oriented forward substitution of a lower triangular matrix**: Consider the system

$$\begin{bmatrix} \ell_{11} & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

  In general, we have

$$x_i = \frac{b_i - \sum_{j=1}^{i-1} \ell_{ij} x_j}{\ell_{ii}}$$

  for $i = 1, 2, ..., n$. This method is called **Forward Substitution**.

  The row oriented forward substitution algorithm requires $\mathcal{O}(n^2)$ flops.

- **Column oriented forward substitution (recursive algorithm)**: Suppose we have $Lx = b$ when L is lower triangular, we split the matrix into the following blocks

$$\begin{bmatrix} \ell_{11} & 0 \\ \hat{\ell} & \hat{L} \end{bmatrix} \begin{bmatrix} x_1 \\ \hat{x} \end{bmatrix} = \begin{bmatrix} b_1 \\ \hat{b} \end{bmatrix}.$$

  With $\hat{\ell} \in \mathbb{R}^{n-1}$, $\hat{L} \in \mathbb{R}^{n-1 \times n-1}$, $\hat{x} \in \mathbb{R}^{n-1}$, $\ell_{11}, x_1, b_1 \in \mathbb{R}$. Note that $\hat{L}$ is also lower triangular.

  1. Compute $x_1 = \frac{b_1}{\ell_{11}}$
  2. Compute $\hat{b} - \hat{\ell} x_1 = \tilde{b} \in \mathbb{R}^{n-1}$
  3. Find $\hat{L}x = \tilde{b}$
  4. Run the algorithm on $\hat{L}$, $\tilde{b}$. That is, $\text{Alg}(\hat{L}, \tilde{b})$

  The recursive column oriented forward substitution algorithm requires $\mathcal{O}(n^2)$ flops.

- **Row oriented backward substitution of an upper triangular matrix**: Let $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^n$, with

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

  In general, we have that

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{ij} x_j}{a_{ii}}, \quad i = n, n-1, ..., 1$$

  The row oriented backward substitution algorithm requires $\mathcal{O}(n^2)$ flops.

- **Column-oriented backward substitution**: Let $U \in \mathbb{R}^{n \times n}$ be upper triangular, $x \in \mathbb{R}^n$, and $b \in \mathbb{R}^n$ which gives the system

$$\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Split the system into the following block decomposition

$$\begin{bmatrix} \hat{U} & u \\ 0^\top & u_{nn} \end{bmatrix} \begin{bmatrix} \hat{x} \\ x_n \end{bmatrix} = \begin{bmatrix} \hat{b} \\ b_n \end{bmatrix}$$

Then,

$$\hat{U}\hat{x} + ux_n = \hat{b} \implies \hat{U}\hat{x} = \hat{b} - ux_n = \tilde{b},$$
$$u_{nn}x_n = b_n \implies x_n = \frac{b_n}{u_{nn}}$$

Thus, the column-oriented backward substitution algorithm is defined by the following steps

1. Compute $x_n = \frac{b_n}{u_{nn}}$
2. Compute $\tilde{b} = \hat{b} - ux_n$
3. Run the algorithm on $\hat{U}, \tilde{b}$. That is, $\text{Alg}(\hat{U}, \tilde{b})$

The non-recursive pseudocode in the spirit of 1.3.5 and 1.3.13 is

```
for i = n, ..., 1
    if U[i, i] = 0, set error flag, exit

    b[i] = b[i]/U[i, i]

    for j = i − 1, ..., 1
        b[j] = b[j] − U[j, i] · b[i]
    end
end
```

Requires $\mathcal{O}(n^2)$ flops.

- **Flops required to solve triangular systems**: Let $Ax = b$ be a system of linear equations where A is nonsingular. If A is upper triangular or lower triangular we can solve the system in roughly n^2 flops.
- **Inner product formulas to compute R (Cholesky factor)**: We have the formulas

$$r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2} \quad i = 1, 2, ..., n$$
$$r_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} r_{ki}r_{kj}}{r_{ii}} \quad j = i+1, ..., n$$

The inner product formulas to compute R requires $\mathcal{O}(n^3)$ flops.

- **Recursive column oriented method to find the Cholesky factor R (Outer product method)**: Let $A \in \mathbb{R}^{n\times n}$. Assume that A is positive definite, so $A = A^\top$, and $A = R^\top R$ for a unique upper triangular matrix R. We have,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} r_{11} & 0 & \cdots & 0 \\ r_{12} & r_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \cdots & r_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}$$

We then perform a matrix decomposition

$$\begin{bmatrix} a_{11} & a^\top \\ a & \hat{A} \end{bmatrix} = \begin{bmatrix} r_{11} & 0^\top \\ r & \hat{R}^\top \end{bmatrix} \begin{bmatrix} r_{11} & r^\top \\ 0 & \hat{R} \end{bmatrix}.$$

Where $\hat{A} = \hat{A}^\top \in \mathbb{R}^{n-1\times n-1}$, $a \in \mathbb{R}^{n-1}$, $\hat{R}^\top \in \mathbb{R}^{n-1\times n-1}$ lower triangular, and $\hat{R} \in \mathbb{R}^{n-1\times n-1}$ upper triangular. Further,

The recursive column oriented algorithm to compute the Cholesky factor R is given by the following steps

1. $r_{11} = \sqrt{a_{11}}$
2. $r = \frac{a}{r_{11}}$
3. $\tilde{A} = \hat{A} - rr^\top$
4. $\text{Alg}(\tilde{A}) = \hat{R}$

The recursive column oriented algorithm to compute the Cholesky factor R requires $\mathcal{O}(n^3)$ flops.

- **Cholesky's algorithm**: Cholesky's algorithm applied to an $n \times n$ matrix performs about $\frac{n^3}{3}$
- **Bordered form of Choleskys method**: Suppose $A \in \mathbb{R}^{n\times n}$ is positive definite. Then, A admits a decomposition $A = R^\top R$, for a unique upper triangular matrix R called the Cholesky factor, with $r_{ii} > 0$ for $i = 1, 2, ..., n$. So,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} r_{11} & 0 & \cdots & 0 \\ r_{12} & r_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \cdots & r_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}$$

We then perform a matrix decomposition

$$\begin{bmatrix} \hat{A} & a \\ a^\top & a_{nn} \end{bmatrix} = \begin{bmatrix} \hat{R}^\top & 0 \\ r^\top & r_{nn} \end{bmatrix} \begin{bmatrix} \hat{R} & r \\ 0 & r_{nn} \end{bmatrix}.$$

So,

$$\begin{aligned} \hat{A} &= \hat{R}^\top \hat{R}, \\ a &= \hat{R}^\top r, \\ a_{nn} &= r^\top r + r_{nn}^2 \implies r_{nn} = \sqrt{a_{nn} - r^\top r}. \end{aligned}$$

So, the steps for the algorithm are

1. Recurse $\hat{A}$ until $A \in \mathbb{R}^{1\times 1}$
2. Solve the lower triangular system $\hat{R}^\top r = a$ by forward substitution
3. Compute $r_{nn} = \sqrt{a_{nn} - r^\top r}$
4. Return the step two on the previous call

The above algorithm is postorder recursion and requires $\mathcal{O}(n^3)$ flops.

- **Row oriented algorithm to compute LU factorization**: Let $A \in \mathbb{R}^{n\times n}$. If A can be factored into products LU, for $U \in \mathbb{R}^{n\times n}$ upper triangular, $L \in \mathbb{R}^{n\times n}$ unit lower triangular, then

$$A = LU$$

implies

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{1n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}.$$

The formulas are

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} \ell_{ik} u_{kj} \quad j = i, i+1, ..., n, \tag{1}$$

$$\ell_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} u_{kj}}{u_{jj}} \quad i = j+1, j+2, ..., n. \tag{2}$$

To use these formulas to find each u_{ij} we first need to plug $i = 1$ into (1), then after we get the first row of U, we can plug in $j = 1$ into (2) to get the first column of L, and so on.

- **Column oriented recursive algorithm to find the LU factorization**: Assume $A \in \mathbb{R}^{n\times n}$ admits an LU factorization for $L \in \mathbb{R}^{n\times n}$ unit lower triangular, U upper triangular. Then,

$$A = LU$$

implies

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}.$$

Decompose $A = LU$ into the blocks

$$\begin{bmatrix} a_{11} & b^\top \\ a & \hat{A} \end{bmatrix} = \begin{bmatrix} 1 & 0^\top \\ \ell & \hat{L} \end{bmatrix} \begin{bmatrix} u_{11} & u^\top \\ 0 & \hat{U} \end{bmatrix}.$$

The recursive algorithm is defined by the following steps.

1. $u_{11} = a_{11}$ (zero flops)
2. $u^\top = b^\top$ (zero flops)
3. $\ell = \frac{a}{u_{11}}$ ($n-1$ flops)
4. $\tilde{A} = \hat{L}\hat{U} = \hat{A} - \ell u^\top$ ($2(n-1)^2$ flops)
5. Alg($\tilde{A}$)

The number of flops required for the recursive outer product method to find the LU factorization is $\frac{2}{3}n^3 + \mathcal{O}(n^2)$

- **Bordered form LU decomposition algorithm**:

- **Flops required to find LU decomposition**: Suppose A is a matrix that admits an LU decomposition $A = LU$ for L unit lower triangular, and U upper triangular. The flops required to find this decomposition is roughly $\frac{2}{3}n^3$.

Thus, to solve the system $Ax = b$, we have

$$Ax = b \iff LUx = b.$$

Let

$$\begin{cases} Ly &= b \quad (n^2 \text{ flops}) \\ Ux &= y \quad (n^2 \text{ flops}) \end{cases}.$$

So, in total, we have $\frac{2}{3}n^3 + n^2 + n^2 = \frac{2}{3}n^3 + 2n^2 = \mathcal{O}(n^3)$ flops to solve the system $Ax = b$ with an LU decomposition.

- **Flops required to perform Gaussian Elimination without row interchanges (pivoting)**: For a system $Ax = b$, reducing the system to $Ux = b'$, where U is upper triangular requires roughly $\frac{2}{3}n^3 + n^2 + 2n = \mathcal{O}(n^3)$ flops

  Consider step k of the process. We need to eliminate all entries below the pivot in column k, there are $(n-k)$ of them. Then, we need to update the $(n-k) \times (n-k)$ submatrix that doesn't include column k and row k. The operation on each element a_{ij} in the $(n-k) \times (n-k)$ submatrix is

  $$a_{ij} \to a_{ij} - m_{ik}a_{kj}$$

  where $m_{ik} = a_{ik}/a_{kk}$ is the multiplier. Thus, updating the $(n-k) \times (n-k)$ submatrix requires $2(n-k)^2$ flops.

  We require an additional $(n-k)$ flops to each multiplier, and an additional $2(n-k)$ flops to eliminate each entry below the pivot.

  In total, we have

  $$\sum_{k=1}^{n-1} 2(n-k)^2 + (n-k) + 2(n-k) = \sum_{k=1}^{n-1} 2(n-k)^2 + 3(n-k).$$

  Let $j = n - k$. When $k = 1$, $j = n - 1$, and when $k = n - 1$, $j = 1$. So, we have

  $$\begin{aligned} \sum_{j=1}^{n-1} 2j^2 + 3j &= 2 \cdot \frac{(n-1)n(2n-1)}{6} + 3 \cdot \frac{(n-1)n}{2} \\ &= \frac{2}{3}n^3 + \frac{1}{2}n^2 - \frac{7}{6}n \approx \frac{2}{3}n^3 + \mathcal{O}(n^2) = \mathcal{O}(n^3). \end{aligned}$$

- **Flops required to solve $Ax = b$ with A^{-1}**: Suppose we have a system $Ax = b$, where A is nonsingular. How many flops would it take to find A^{-1}, and then solve $x = A^{-1}b$.

  If A^{-1} exists, then it is a matrix X such that $AX = I$. So,

  $$A \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

  Let $\text{col}_j(X)$ denote the j^{th} column of X, and e_j denote the j^{th} column of I. Then,

  $$A\text{col}_1(X) = e_1, \quad A\text{col}_2(X) = e_2, \quad ..., \quad A\text{col}_n(X) = e_n.$$

In total we need to solve n systems. If we solved all n systems with Gaussian Elimination, it would require

$$n\left(\frac{2}{3}n^3 + \mathcal{O}(n^2)\right) = \frac{2}{3}n^4 + \mathcal{O}(n^3)$$

flops. But, we can do better. Instead if we find the LU decomposition for A, which would require $\frac{2}{3}n^3$ flops, we would have the n systems

$$\begin{array}{cccc} Ly_1 = e_1 & Ly_2 = e_2 & \cdots & Ly_n = e_n \\ U\text{col}_1(X) = y_1 & U\text{col}_2(X) = y_2 & \cdots & U\text{col}_n(X) = y_n. \end{array}$$

We see that each system would require $2n^2$ flops. So, in total $2n^3$ flops to solve all systems. For the whole process,

$$\frac{2}{3}n^3 + 2n^3 = \frac{8}{3}n^3$$

flops are required. So, 4 times the flops of LU decomposition and 8 times the flops of Choleksy decomposition.

6.8 Chapter 2: Sensitivity of linear systems

6.8.1 Definitions

- **Unit ball**: A unit ball is the set of all points whose distance from the origin is less than or equal to 1

  Given a norm $\|\cdot\|$ on a vector space (say $\mathbb{R}^n$), the unit ball is the set of all points that are within distance 1 of the origin under that norm:

  $$B_{\|\cdot\|}(0,1) = \{x \in \mathbb{R}^n : \ \|x\| \leqslant 1\}.$$

  Where

  - **B "Ball"**: the set of all points within a certain distance (radius)
  - **0**: The center of the ball (here, the origin)
  - **1**: The radius of the ball
  - $\|\cdot\|$: The norm used to measure distance

  It's the region that's "1 unit away" from the origin according to the norm.

  The boundary of this set, where $\|x\| = 1$, is called the **unit sphere** (even though it might not look like a sphere geometrically).

  It's the set of all vectors you can "reach" from the origin if your allowed length is 1 (according to your chosen norm).

- **A more general ball**: In general, a ball with center at a, and radius r is defined as

  $$B_{\|\cdot\|}(a,r) = \{x \in \mathbb{R}^n : \ \|x - a\| \leqslant r\}.$$

- **Vector norms**:
  - **Euclidean norm (2-norm)**: The standard Euclidean distance. For $x \in \mathbb{R}^n$,

    $$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}.$$

  - **Manhattan norm (1-norm)**: Denoted L^1, and also called **Taxicab norm**. For $x \in \mathbb{R}^n$,

    $$\|x\|_1 = |x_1| + |x_2| + ... + |x_n|.$$

  - **L-Infinity (max) norm (∞-norm)**: Denoted L^∞. for $x \in \mathbb{R}^n$,

    $$\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i| = \max\{|x_1|, x_2, ..., x_n\}.$$

  - **p-norm**: In $\mathbb{R}^n$, A more general norm is

    $$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} = (|x_1^p| + |x_2|^p + ... + |x_n|^p)^{\frac{1}{p}}$$

    for $1 \leqslant p < \infty$. The general p-norm satisfies all three properties of a norm only when $p \geqslant 1$. For smaller p, the triangle inequality does not hold.

- **Frobenius norm**

  $$\|A\|_2 = \|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}}.$$

Note: The Frobenius norm is not an induced norm. We have for the identity matrix I,

$$\|I\|_F = \sqrt{n} \neq 1.$$

- **Induced (operator) matrix norms**: Induced (or operator) matrix norms tell us exactly how much a matrix can stretch a vector under a given vector norm. Given a vector norm $\|\cdot\|$ on $\mathbb{R}^n$, the induced matrix norm of $A \in \mathbb{R}^{n\times n}$ is defined as

$$\|A\| := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|..$$

So it's the **largest possible magnification factor** of the matrix A acting as a linear transformation.

We know A is a map that sends vectors to new vectors. Each vector x has a direction and a length:

  - $\|x\|$ = its original length.
  - $\|Ax\|$ = its new length after transformation.

The ratio $\dfrac{\|Ax\|}{\|x\|}$ tells you **how much** A stretches or shrinks that vector.

The induced norm picks out the **maximum stretching** over all possible directions. So $\|A\|$ represents the largest factor by which A can stretch any vector.

If the induced norm is defined as

$$\|A\| = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|},$$

write $x = \|x\|\, y$, so $y = \frac{x}{\|x\|}$. Then,

$$\|A\| = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|} = \max_{\|y\|=1} \frac{\|A(\|x\|\, y)\|}{\|x\|} = \max_{\|y\|=1} \frac{\|x\|\, \|Ay\|}{\|x\|} = \max_{\|y\|=1} \|Ay\|\,.$$

- **Induced matrix norms special cases**:

p	**Name**	**Explicit formula**
1	Maximum column sum	$\|A\|_1 = \max\limits_{1\leqslant j\leqslant n} \sum\limits_{i=1}^{m} \|a_{ij}\|$
2	Spectral norm	$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$
∞	Maximum row sum	$\|A\|_\infty = \max\limits_{1\leqslant i\leqslant m} \sum\limits_{j=1}^{n} \|a_{ij}\|$

Note: Recall that the eigenvalues of a 2×2 matrix are

$$\lambda(A) = \lambda^2 - \mathrm{Tr}(A)\lambda + \det(A) = \frac{a+d}{2} \pm \sqrt{\left(\frac{a-d}{2}\right)^2 + bc}.$$

- **Singular values**: For $A \in \mathbb{R}^{m\times n}$, its **singular values** are the numbers

$$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0,$$

defined as the square roots of the eigenvalues of A^TA

$$\sigma_i(A) = \sqrt{\lambda_i(A^TA)}.$$

If A has rank r, then there are r positive singular values. The rest are zero.

- **Singular values of A^{-1}**: The singular values of A^{-1} are the reciprocals of the singular values of A.

$$\sigma_1(A^{-1}) = \frac{1}{\sigma_1(A)}, \sigma_2(A^{-1}) = \frac{1}{\sigma_2(A)}, \cdots, \sigma_n(A^{-1}) = \frac{1}{\sigma_n(A)}.$$

- **Eigenvalues of A^TA**:

$$\lambda_i(A^TA) = \sigma_i(A)^2.$$

Where $\sigma_i(A)$ is the ith singular value for A.

- **Singular values of A^TA**:

$$\sigma_i(A^TA) = \sqrt{\lambda_i((A^TA)^T(A^TA))} = \sqrt{\lambda_i(A^TA)^2}.$$

But, to get the eigenvalues for the square of a matrix you square the eigenvalues for the matrix. So, $\lambda_i(A^TA)^2 = (\lambda_i(A^TA))^2$. Thus,

$$\sigma_i(A^TA) = \sqrt{(\lambda_i(A^TA))^2} = \lambda_i(A^TA) = \sigma_i(A)^2.$$

So, the singular values for A^TA are the squares of the singular values of A. Thus, the set of eigenvalues for A^TA is the same as the set of singular values for A^TA

- **Spectral norm**: We have

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \sigma_1 = \sqrt{\lambda_{\max}(A^TA)},$$

$$\left\|A^{-1}\right\|_2 = \frac{1}{\sigma_n} = \frac{1}{\sqrt{\lambda_{\min}}(A^TA)}.$$

- **Relative error**: The relative error in $\hat{x}$ is given by

$$\frac{\|\hat{x} - x\|}{\|x\|} = \frac{\|\delta x\|}{\|x\|}$$

where $\hat{x} = x + \delta x$, which implies $x = \hat{x} - \delta x$.

- **Perturbation**: $\hat{A}$ and $\hat{b}$ are called perturbed if they are modified versions of the original. If $\hat{A}$ is a perturbed matrix A, and $\hat{b}$ is a perturbed vector b, then

$$\hat{A} = A + \delta A,$$
$$\hat{b} = b + \delta b.$$

Note: When we say δA or δb, we do not mean some constant δ times some matrix A or some vector b, They represent changes (perturbations), we have

$$\delta A = \hat{A} - A,$$
$$\delta b = \hat{b} - b.$$

They are differences, not scaled versions. If $\hat{A}$ is a perturbed A, then δA is simply the matrix of entrywise differences:

$$(\delta A)_{ij} = \hat{a}_{ij} - a_{ij}.$$

- **Condition number** $\kappa(A)$: The condition number of a matrix A measures how sensitive the solution of a linear system $Ax = b$ is to small changes in b (or in A).

$$\kappa(A) = \left\|A^{-1}\right\| \|A\|.$$

  We see that as $\kappa(A) \to \infty$, the relative error in x grows without bound.

- **Condition number** ($\kappa_2(A)$):

$$\kappa_2(A) = \left\|A^{-1}\right\|_2 \|A\|_2 = \sigma_{\max} \cdot \frac{1}{\sigma_{\min}} = \frac{\sigma_{\max}}{\sigma_{\min}}.$$

- **Condition number** ($\kappa_2(A^T A)$):

$$\kappa_2(A^T A) = \frac{\sigma_{\max}(A^T A)}{\sigma_{\min}(A^T A)} = \frac{\sigma_{\max}(A)^2}{\sigma_{\min}(A)^2} = \kappa_2^2(A).$$

- **Numerical stability vs conditioning**
  - **Conditioning**
    - A property of the **mathematical problem** itself.
    - Measures the **sensitivity** of the true solution to small input changes.
    - Example: if $Ax = b$, then

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\delta b\|}{\|b\|}.$$

      A large $\kappa(A)$ indicates an **ill-conditioned problem**.
  - **Numerical Stability**
    - A property of the **algorithm** used to solve the problem.
    - Measures how much **round-off and truncation errors** the algorithm introduces or amplifies.
    - A **stable algorithm** gives the exact solution to a *nearby problem*:

$$(A + \delta A)\hat{x} = b + \delta b, \quad \|\delta A\|,\ \|\delta b\| \text{ small.}$$

  Note: If an algorithm is **numerically unstable**, then the perturbations δA and/or δb required to explain its result might be **large**:

$$\|(A + \delta A) - A\| \text{ is not small.}$$

  Hence, the computed $\hat{x}$ is the exact solution to a *far-away problem*:

$$(A + \delta A)\hat{x} = b + \delta b,$$

  where $\|\delta A\|$ or $\|\delta b\|$ are no longer small compared to $\|A\|$ or $\|b\|$.

  This means the algorithm's result may not correspond meaningfully to the original system at all.

- **Ball of guaranteed nonsingularity (neigborhood of nonsingularity)**: The ball in matrix space with center A and radius $r = \|A\| / \kappa(A)$ is

$$\mathcal{B}(A, r) = \{A + \delta A : \ \|\delta A\| < r\},$$

  where $r = \frac{\|A\|}{\kappa(A)}$. Any matrix within the ball is nonsingular, and any matrix outside the ball or on the boundary is singular.

- **Backward stable algorithm**: A backward stable algorithm is one that produces the exact solution to a slightly perturbed version of the original problem.

6.8.2 Properties

- **Norms**: $\|\cdot\|$ is a norm if and only if the following properties are satisfied
  1. $\|x\| = 0 \iff x = 0$
  2. $\|\alpha x\| = |\alpha| \|x\|$
  3. $\|x + y\| \leqslant \|x\| + \|y\|$ (triangle inequality)
- **Matrix norms**: A matrix norm is a function

$$\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}_+ : A \mapsto \|A\| .$$

- **Properties of matrix norms**: Matrix norms satisfy the three required properties of norms.
  1. $\|A\| = 0 \iff A = 0$
  2. $\|\alpha A\| = |\alpha| \|A\|$
  3. $\|A + B\| \leqslant \|A\| + \|B\|$ (Triangle inequality)
- **Additional properties of matrix norms**;
  1. $\|A\| < \infty$ for any finite matrix A
- **Properties of induced matrix norms**
  - **Sub-multiplicativity**: $\|AB\|_p \leqslant \|A\|_p \|B\|_p$
  - **Consistency**: $\|Ax\|_p \leqslant \|A\|_p \|x\|_p$
  - **Normalization**: $\|I\|_p = 1$
- **Additional induced matrix norm properties**
  1. If A is singular, then A^{-1} does not exists, we define

$$\left\|A^{-1}\right\| = \infty.$$

  2. For any $A \in \mathbb{R}^{m\times n}$,

$$\|A\| = \left\|A^T\right\| .$$

     Except for $\|A\|_1$ and $\|A\|_\infty$. In this case,

$$\left\|A^T\right\|_1 = \|A\|_\infty .$$

- **Properties of the spectral norm**
  1. $\|I\|_2 = 1$ (Since the eigenvalues are all one)
  2. $\left\|A^T\right\|_2 = \|A\|_2$
  3. $\|Q\| = \left\|Q^T\right\| = \left\|Q^{-1}\right\| = 1$ for Q orthogonal
- **Properties of the condition number**: Let A be a matrix, and $\kappa(A)$ be the condition number that measures the system $Ax = b$. The following two properties hold
  1. $\kappa(A) \geqslant 1$
  2. $\kappa(I) = 1$
  3. $\kappa(A) = \kappa(A^{-1})$
- **Additional properties of the condition number**:

1. Since $\|A\| = \left\|A^T\right\|$,

$$\kappa(A) = \kappa(A^T).$$

- **Well-conditioned and ill-conditioned in terms of** $\kappa(A)$: If $\kappa(A)$ is large, then (P) is ill-conditioned. If $\kappa(A)$ is small (close to one), then (P) is well-conditioned.
- **The condition number also measures how close** A **is to singularity**: The condition number

$$\kappa(A) = \left\|A^{-1}\right\| \|A\|$$

measures how close a matrix A is to being **singular**. If A is singular, then A^{-1} does not exist, meaning that

$$\left\|A^{-1}\right\| = \infty.$$

Therefore,

$$\kappa(A) = \infty \quad \text{if and only if} \quad A \text{ is singular,}$$

and $\kappa(A) \to \infty$ as A approaches singularity.

6.8.3 Theorems

- **Theorem *(Relative Error Bound I)***: Let A be nonsingular, $b \neq 0$, and $Ax = b$. If $A(x + \delta x) = b + \delta b$, then

$$\frac{\|\delta x\|}{\|x\|} \leqslant \kappa(A) \frac{\|\delta b\|}{\|b\|}$$

where $\kappa(A) = \left\|A^{-1}\right\| \|A\|$. According to this bound.

Analysis: As $\kappa(A) \to \infty$,

$$\frac{\|\delta x\|}{\|x\|} \leqslant \kappa(A) \frac{\|\delta b\|}{\|b\|} \to \infty.$$

So, as $\kappa(A) \to \infty$, for a fixed nonzero relative perturbation in b, the bound allows the relative error in x to become arbitrarily large. If the relative error in b is zero, then

$$0 \leqslant \frac{\|\delta x\|}{\|x\|} \leqslant 0 \implies \frac{\|\delta x\|}{\|x\|} = 0.$$

As $\kappa(A) \to 1$,

$$\frac{\|\delta x\|}{\|x\|} \leqslant \kappa(A) \frac{\|\delta b\|}{\|b\|} \to \frac{\|\delta b\|}{\|b\|}.$$

Thus, perturbations in b produce at most proportional perturbations in x.

- **Theorem (*Singularity of perturbed* A)**: If

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)}$$

then $A + \delta A$ is nonsingular.

- **Theorem *(Relative error bound II)***: Let A be nonsingular, $b \neq 0$, and $Ax = b$. If $(A + \delta A)(x + \delta x) = b$, and

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)},$$

then

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{\kappa(A) \frac{\|\delta A\|}{\|A\|}}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}}.$$

- **Theorem *(Relative error bound III)***: Let A be nonsingular, $b \neq 0$, and $Ax = b$. If $(A + \delta A)(x + \delta x) = b + \delta b$, and

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)},$$

then

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{\kappa(A) \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}}.$$

6.8.4 Algorithms and complexities

-

6.9 Chapter 3: The least squares problem and orthogonal matrices

6.9.1 Definitions

- **Over-determined systems**: If a system $Ax = b$ has more equations than unknowns ($m > n$), we call the system **over-determined**
- **under-determined systems**: If a system $Ax = b$ has less equations than unknowns ($m < n$), we call the system **under-determined**
- **Determined system**: If a system $Ax = b$ has the same number of equations as unknowns ($m = n$), we call the system **determined**.
- **Well-determined and degenerate**: If a system $Ax = b$ has a unique solution for a given b, then the system is said to be **well-determined**. If the system has no solution or infinitely many, the system is **degenerate**.
- **The discrete least squares problem**: Let $A \in \mathbb{R}^{m\times n}$, where $m \geqslant n$, and $b \in \mathbb{R}^m$. The discrete least squares problem is finding

$$\min_{x\in\mathbb{R}^n} \|r\|_2^2,$$

  where $r = b - Ax$. If $b \notin \text{Im}(L) = \text{col}(A)$, then no solution to $Ax = b$ exists, and we can instead solve the discrete least squares problem to get the best approximation.

  Recall that since the closest point in a subspace to a vector is its orthogonal projection, minimizing $\|r\|_2^2$ is equivalent to finding the projection of b onto the column space of A.

- **Orthogonal matrices**: A matrix $Q \in \mathbb{R}^{n\times n}$ is orthogonal if $QQ^T = Q^TQ = I$, so $Q^T = Q^{-1}$.

  An orthogonal matrix is a matrix whose columns form a set of orthonormal vectors, each has length one and is perpendicular to the others.

- **Definition of orthogonal matrices**: $Q \in \mathbb{R}^{n\times n}$ is orthogonal if the columns of Q satisfy
  1. $\|q_i\| = 1$ for $i = 1, 2, ..., n$
  2. $\langle q_i, q_j\rangle = 0$ if $i \neq j$
- **The discrete LSP with QR**

$$\min_{x\in\mathbb{R}^n} \|b - Ax\|_2^2 = \min_{x\in\mathbb{R}^n} \left\|\hat{c} - \hat{R}x\right\|_2^2 + \|\bar{c}\|_2^2,$$

  .

  x is found by solving the system $\hat{R}x = \hat{c}$, and the residual is $\|\bar{c}\|_2^2$

- **QR factorization of a square matrix**: Let $A \in \mathbb{R}^{n\times n}$, then $Q \in \mathbb{R}^{n\times n}$, $R = \hat{R} \in \mathbb{R}^{n\times n}$, $Q^Tb = c = \hat{c}$, and

$$Ax = b \implies QRx = b \implies Rx = Q^Tb \implies \hat{R}x = Q^Tb = \hat{c}.$$

  So, since $\hat{R}$ is upper triangular, we can solve the system using backward substitution.

- **Givens rotations**: A Givens rotation is an orthogonal transformation that acts only in a 2-dimensional coordinate plane — say the plane spanned by the i-th and j-th coordinate axes. It's the identity matrix except for a 2×2 rotation block:

$$G(i,j,\theta) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c & & s & & \\ & & & 1 & & & \\ & & -s & & c & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}, \quad c = \cos(\theta),\ s = \sin(\theta).$$

  Everywhere else its the identity matrix I. Only entries $(i,i),(i,j),(j,i),(j,j)$ are modified.

- QR **with Householder reflectors**: Let $A \in \mathbb{R}^{m \times n}$, where

$$A = \begin{bmatrix} | & | & \cdots & | \\ x^1 & x^2 & \cdots & x^n \\ | & | & \cdots & | \end{bmatrix},$$

  where x^j denotes the j^{th} column of A. For each column j of A, we have the following steps

  1. $\tau_j = \operatorname{sgn}(x_1) \left\| x^j \right\|_2$
  2. $\gamma_j = \frac{\tau_j + x_1}{\tau_j}$
  3. $u_j = \begin{pmatrix} 1 \\ x_2/(\tau_j + x_1) \\ \vdots \\ x_m/(\tau_j + x_1) \end{pmatrix}$
  4. $Q_j = I - \gamma_j u_j u_j^T$

  Then, just like with givens rotations, $Q = Q_k \cdots Q_j \cdots Q_2 Q_1$, and $R = QA = Q_k \cdots Q_j \cdots Q_2 Q_1 A$.

  We can avoid forming Q with

$$Qx = x - \gamma(u^T x)u.$$

- **Normal equations**: The equations

$$A^T A x = A^T b$$

  that solve the least squares problem

$$\min_{x \in \mathbb{R}^n} \|b - Ax\|_2^2$$

  are called the **normal equations**. They characterize the vector x for which the residual $b - Ax$ is orthogonal to the column space of A.

6.9.2 Properties

- **Properties of orthogonal matrices**
  1. $Q^T Q = QQ^T = I$
  2. $Q^{-1} = Q^T$
  3. $\det(Q) = \pm 1$
- **Additional properties of orthogonal matrices**:
  1. The product of orthogonal matrices is orthogonal
  2. If Q is orthogonal, so is Q^T
- **Properties of Givens rotation**: Let Q be the givens rotation matrix
  1. $Q^T Q = QQ^T = I$ (Q is orthogonal)
- **Properties of Householder reflection matrices**
  1. **Symmetry**: $Q = Q^T$
  2. **Orthogonality**: $Q^T Q = QQ^T = I$
- **Properties of the normal matrix**: Let $A \in \mathbb{R}^{m \times n}$
  1. $A^T A \succeq 0$
  2. $A^T A \succ 0 \iff \text{rank}(A) = n$

6.9.3 Theorems

- **Projection onto a line theorem**: Let ℓ be a line spanned by a vector q, and v be a vector not on ℓ. Project v down onto ℓ, call this projection p. The residual (error) vector $v - p$ is orthogonal to the line ℓ.
- **Orthogonal projection theorem**: Let W be a subspace of $\mathbb{R}^n$, and x be a vector in $\mathbb{R}^n$. The closest point in W to x is its orthogonal projection onto W, denoted $\text{proj}_W(x) = w$. That is,

$$\min_{w \in W} \|x - w\| = \text{proj}_W(x).$$

- **Theorem**: If $Q \in \mathbb{R}^{n \times n}$ is orthogonal, then
  1. $\langle Qx, Qy \rangle = \langle x, y \rangle$
  2. $\|Qx\|_2 = \|x\|_2$

6.9.4 Propositions

-

6.9.5 Algorithms and complexities

- **The discrete LSP algorithm**: Let $A \in \mathbb{R}^{m \times n}$, for $m > n$. Define

$$(P): \min_{x \in \mathbb{R}^n} \|x - Ax\|_2^2 .$$

To solve discrete least squares, we take the following steps

1. $c = Q^T b = \begin{bmatrix} \hat{c} \\ \bar{c} \end{bmatrix}$, $c \in \mathbb{R}^m$, $\hat{c} \in \mathbb{R}^n$, $\bar{c} \in \mathbb{R}^{m-n}$
2. Solve $(\hat{P})$: $\hat{R}x = \hat{c}$, solution of $(\hat{P})$ is the solution of (P).
3. $\min_{x \in \mathbb{R}^n} \|b - Ax\|_2^2 = \|\bar{c}\|_2^2$

Note: For now we want to assume that $\text{rank}(A) = n$, so all columns are linearly independent. In this case, $\hat{r}_{ii} \neq 0$.

6.10 Chapter 1 Proofs

6.10.1 Positive definite (1)

- **Theorem n.1.1**: Let $A \in \mathbb{R}^{n\times n}$, $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ be a positive definite matrix. Then $A_{11},\ A_{22}$ are positive definite.

  Note that $A_{11} \in \mathbb{R}^{n_1\times n_1}$, $A_{12} \in \mathbb{R}^{n_1\times n_2}$, $A_{21} \in \mathbb{R}^{n_2\times n_1}$, and $A_{22} \in \mathbb{R}^{n_2\times n_2}$. So, $n = n_1 + n_2$

  Proof. Assume $A \in \mathbb{R}^{n\times n}$, $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, A positive definite.

  First, we show that $A_{11} = A_{11}^\top$, and $A_{22} = A_{22}^\top$. Since A p.d,

  $$A = A^\top \implies \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11}^\top & A_{21}^\top \\ A_{12}^\top & A_{22}^\top \end{bmatrix}.$$

  So, we see that

  $$\begin{aligned} A_{11} &= A_{11}^\top, \\ A_{22} &= A_{22}^\top. \end{aligned}$$

  Next, we show that $x^\top A_{11}x > 0$, for $x \in \mathbb{R}^{n_1}$, $x \neq 0$ and $x^\top A_{22}x > 0$ for $x \in \mathbb{R}^{n_2}$, $x \neq 0$

  Let $\bar{x} \in \mathbb{R}^{n_1}$, $\bar{x} = \begin{pmatrix} x_1 \\ 0 \end{pmatrix}$, $x_1 \neq 0$. Note that $\bar{x}^\top = \begin{pmatrix} x_1^\top & 0 \end{pmatrix}$. We observe

  $$\begin{aligned} 0 < \bar{x}^\top Ax &= \begin{pmatrix} x_1^\top & 0 \end{pmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{pmatrix} x_1 \\ 0 \end{pmatrix} \\ &= x_1^\top A_{11} x_1 > 0. \end{aligned}$$

  Similarly, $\bar{x} \in \mathbb{R}^n$, $\bar{x} = \begin{pmatrix} 0 \\ x_2 \end{pmatrix}$, $x_2 \neq 0 \in \mathbb{R}^{n_2}$ reveals $x_2^\top A_{22} x_2 > 0$.

  Therefore, $A_{11},\ A_{22}$ are positive definite. ■

- **Theorem n.1.2**: Let $A \in \mathbb{R}^{n\times n}$ be a positive definite matrix, then $a_{ii} > 0$ for $i = 1, 2, ..., n$

  Proof. Assume that $A \in \mathbb{R}^{n\times n}$ is a positive definite matrix. Define e_i as the set of vectors with all zeros except for a one at the i^{th} position. Since A is positive definite,

  $$e_i^\top A e_i = a_{ii} > 0 \tag{1}$$

  for $i = 1, 2, ..., n$ ■

- **Theorem n.1.3**: Let $A \in \mathbb{R}^{n\times n}$ be a positive definite matrix and $X \in \mathbb{R}^{n\times n}$ be nonsingular. Then, $B = X^\top AX$ is positive definite.

Remark. If A, B, C are matrices, then

$$(ABC)^\top = C^\top B^\top A^\top$$

Proof. Assume that $A \in \mathbb{R}^{n\times n}$ is positive definite, $X \in \mathbb{R}^{n\times n}$ nonsingular, and $B = X^\top AX$.

First, we show that $B = B^\top$. We have

$$\begin{aligned} B^\top &= (X^\top AX)^\top = X^\top A^\top (X^\top)^\top \\ &= X^\top A^\top X. \end{aligned}$$

But, A is positive definite, and is therefore symmetric. So,

$$B^\top = X^\top AX = B.$$

Thus, B is symmetric.

Next, we show that $x^\top Bx > 0$, for $x \in \mathbb{R}^n$, $x \neq 0$. We have

$$\begin{aligned} x^\top Bx &= x^\top (X^\top AX)x = (x^\top X^\top)A(Xx) \\ &= (Xx)^\top A(Xx) \end{aligned}$$

Let $y = Xx$. Thus,

$$(Xx)^\top A(Xx) = y^\top Ay$$

Note that since $x \neq 0$, and X nonsingular, $y = Xx \neq 0$. Since $y \neq 0$, and A p.d, $y^\top Ay > 0$.

Therefore, B is positive definite. ■

- **Theorem n.1.4**: Let A be positive definite. Then, $\det(A) > 0$

Proof. Assume A is a positive definite matrix.

Since A p.d, $A = R^\top R$ for a unique upper triangular matrix R. Further, $r_{ii} > 0$. We have

$$\begin{aligned} A &= R^\top R \\ \implies \det(A) &= \det(R^\top R) \\ &= \det(R^\top)\det(R) \\ &= \det(R)\det(R) \\ &= \det(R)^2 \\ &= (r_{11} \cdot r_{12} \cdot \cdots \cdot r_{1n})^2 \\ &= r_{11}^2 \cdot r_{12}^2 \cdot \cdots \cdot r_{1n}^2 > 0 \end{aligned}$$

.

Therefore $\det(A) > 0$ ■

6.11 Chapter 2 Proofs

- **The vector 2-norm is a norm**: To prove that this is true, we need to show that $\|x\|_2$ satisfies the following properties

  1. $\|x\|_2 = 0 \iff x = 0$
  2. $\|\alpha x\|_2 = |\alpha| \|x\|_2$
  3. $\|x + y\|_2 \leqslant \|x\|_2 + \|y\|_2$

  Proof. (1) If $\|x\|_2 = 0$, then

$$\begin{aligned} \|x\| = 0 &\implies \left(x_1^2 + x_2^2 + ... + x_n^2\right)^{\frac{1}{2}} = 0 \\ &\iff x_1^2 + x_2^2 + ... + x_n^2 = 0 \\ &\iff x_1 = x_2 = ... = x_n = 0. \end{aligned}$$

  Conversely, if $x = 0$, then

$$\|0\|_2 = \left(0^2 + 0^2 + ... + 0^2\right)^{\frac{1}{2}} = 0^{\frac{1}{2}} = 0.$$

  (2)

$$\begin{aligned} \|\alpha x\|_2 &= \left((\alpha x_1)^2 + (\alpha x_2)^2 + ... + (\alpha x_n)^2\right)^{\frac{1}{2}} \\ &= \left(\alpha^2 x_1^2 + \alpha^2 x_2^2 + ... + \alpha^2 x_n^2\right)^{\frac{1}{2}} \\ &= \left(\alpha^2 (x_1^2 + x_2^2 + ... + x_n^2)\right)^{\frac{1}{2}} \\ &= (\alpha^2)^{\frac{1}{2}} (x_1^2 + x_2^2 + ... + x_n^2)^{\frac{1}{2}} \\ &= |\alpha| \|x\|_2 . \end{aligned}$$

  As desired.

  (3)

$$\begin{aligned} \|x + y\|_2^2 &= (x + y)^T (x + y) = x^T x + x^T y + y^T x + y^T y \\ &= x^T x + 2x^T y + y^T y = \|x\|_2^2 + 2x^T y + \|y\|_2^2 . \end{aligned}$$

  By Cauchy Schwarz, $x^T y \leqslant \|x\|_2 \|y\|_2$. So,

$$\|x\|_2^2 + 2x^T y + \|y\|_2^2 \leqslant \|x\|_2^2 + 2 \|x\|_2 \|y\|_2 + \|y\|_2^2 = \left(\|x\|_2 + \|y\|_2\right)^2 .$$

  Since $\|x\|_2^2 + 2x^T y + \|y\|_2^2 = \|x + y\|^2$, we have

$$\|x + y\|_2^2 \leqslant \left(\|x\|_2 + \|y\|_2\right)^2 ,$$

  which implies $\|x + y\|_2 \leqslant \|x\|_2 + \|y\|_2$ ■

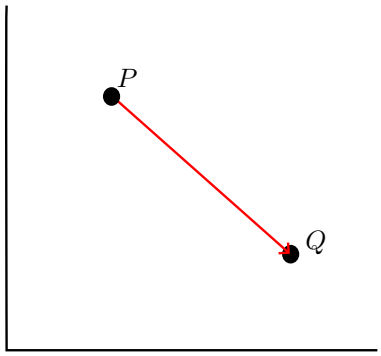
Geometric linear algebra

7.1 Vectors in $\mathbb{R}^n$, projections, and parallelepipeds

- **Displacement vector**: Let P, Q be points in $\mathbb{R}^n$, the vector that goes from P to Q is denoted

$$\overrightarrow{PQ} = Q - P.$$

This vector represents the **displacement** needed to move from P to Q. So, $\overrightarrow{PQ}$ is called the **displacement vector** and represents the displacement from P to Q.



Recall that the position of a vector is arbitrary. Vectors only encode two things, length and direction.

Vectors in $\mathbb{R}^n$ are always defined by two points, a starting point and an ending point. When a vector in $\mathbb{R}^n$ begins at the origin, then $P - O = P$, since $O = (0, 0, 0, ..., 0)$. Thus, the end point P is what determines the length and direction, and so the coordinates of the vector are precisely the coordinates of P.

Notice that we can flip the vector by starting at P and arriving at O, this gives $O - P$, or $-P$.

So, the displacement vector from P_{start} to P_{end} is always

$$P_{\text{end}} - P_{\text{start}}.$$

Consider a displacement vector from P to Q in $\mathbb{R}^2$. If $P = (2, 3)$, and $Q = (4, 1)$, then

$$v_{PQ} = \overrightarrow{PQ} = Q - P = \begin{pmatrix} 4-2 \\ 1-4 \end{pmatrix} = \begin{pmatrix} 2 \\ -3 \end{pmatrix}.$$

If we were to position this vector v_{PQ} starting from the origin, the length and direction would remain the same, but the destination would loose meaning, as it would arrive at a different point. The destination would be $(2, -3)$.

- **Vector acting at a point**: Suppose we had some point $P \in \mathbb{R}^n$, and a vector $v \in \mathbb{R}^n$. Then,

$$P + v$$

is the point you reach by starting at P and moving in the direction of v. So, a vector **based at** P is represented as

$$P \xrightarrow{v} P + v.$$

We can see this by using displacement vectors. Starting from P and moving along v, we will reach some end point Q, if P and v are known, and Q is unknown, then we can find Q by solving for v using the displacement from P to Q

$$Q - P = v \implies Q = P + v.$$

- **Projection onto a line theorem**: Let ℓ be a line spanned by a vector q, and v be a vector not on ℓ. Project v down onto ℓ, call this projection p. The residual (error) vector $v - p$ is orthogonal to the line ℓ.

Proof. The projection of v onto ℓ is given by

$$\text{proj}_\ell(v) = \frac{v^T q}{q^T q} q.$$

Thus, the error $v - p$ is

$$r = v - \frac{v^T q}{q^T q} q.$$

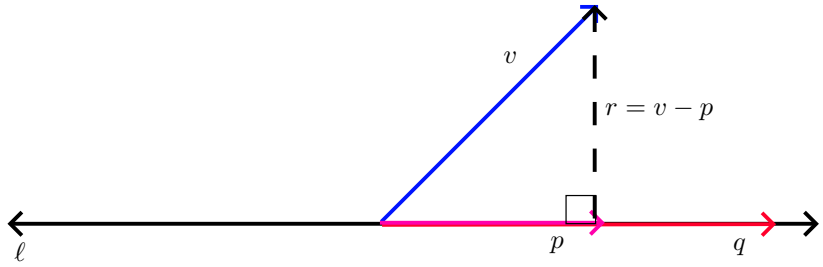
So,

$$\begin{aligned} r^T q &= \left(v - \frac{v^T q}{q^T q} q \right)^T q \\ &= v^T q - \left(\frac{v^T q}{q^T q} q \right)^T q \\ &= v^T q - q^T \left(\frac{v^T q}{q^T q} \right)^T q \\ &= v^T q - \frac{v^T q}{q^T q} q^T q \\ &= v^T q - v^T q = 0. \end{aligned}$$

So, $r \perp q$.

In general, the residual in the projection of a vector onto a space is orthogonal to the space.

- **Projection onto a line derivation**: Consider a line ℓ, a vector a on ℓ, and a vector b. Define the projection of b onto a as p. That is, $p = \text{proj}_a(b)$. Since p and a lie on the same line, they are parallel. So, $p = xa$, for $x \in \mathbb{R}$.



Define the residual as $r = b - p$. Since the residual r is orthogonal to p, we have $r \perp p$, so

$$\begin{aligned} r^T p &= 0 \\ \implies (b-p)^T p &= 0 \\ \implies (b - xa)^T xa &= 0 \\ \implies b^T xa - xa^T xa &= 0 \\ \implies b^T xa &= x^2 a^T a \\ \implies b^T a &= x a^T a \\ \implies x &= \frac{b^T a}{a^T a} = \frac{a^T b}{a^T a}. \end{aligned}$$

Thus,

$$\text{proj}_a(b) = p = xa = \frac{a^T b}{a^T a} a.$$

Notice that we can move some things around, and get

$$p = a \frac{a^T b}{a^T a} = \frac{1}{a^T a} a a^T b.$$

Notice that $\frac{1}{a^T a} a a^T$ is a matrix. Call this matrix the projection matrix P. We therefore have that

$$\text{proj}_a(b) = p = Pb,$$

where $P = \frac{1}{a^T a} a a^T$ is the projection matrix that projects vectors onto a (the line ℓ).

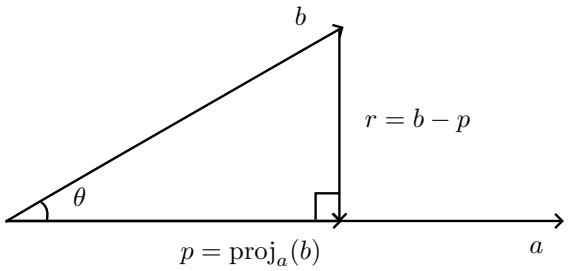
- **Norm of the projection onto a line**: Let $a, b \in \mathbb{R}^n$. From above,

$$p = \text{proj}_a(b) = \frac{a^T b}{a^T a} a.$$

Thus,

$$\|p\| = \left\| \frac{a^T b}{a^T a} a \right\| = \left| \frac{a^T b}{a^T a} \right| \|a\| = \frac{a^T b}{\|a\|^2} \|a\| = \frac{a^T b}{\|a\|}.$$

- **Geometric inner product**: Let $a, b \in \mathbb{R}^n$. By projecting one onto the other, we create a right triangle since the residual is always orthogonal to the vector being projected onto.



Thus, a right triangle is formed. This fact is true regardless of the dimension n. Notice that

$$p = \frac{a^T b}{a^T a} a, \quad \|p\|_2 = \frac{a^T b}{\|a\|_2},$$

and

$$\cos(\theta) = \frac{\|p\|_2}{\|b\|_2} = \frac{a^T b}{\|a\|_2} \left(\frac{1}{\|b\|_2} \right).$$

Thus,

$$a^T b = \|a\|_2 \|b\|_2 \cos(\theta), \quad \theta = \cos^{-1} \left(\frac{a^T b}{\|a\|_2 \|b\|_2} \right).$$

- **Sine of the angle between a and b**: Consider the figure above, we have that

$$\sin(\theta) = \frac{\|r\|_2}{\|b\|_2}.$$

Since $r = b - p$,

$$r = b - \frac{a^T b}{a^T a} a = \frac{(a^T a)b - (a^T b)a}{a^T a}.$$

So,

$$\|r\|_2 = \frac{1}{a^T a} \left\| (a^T a)b - (a^T b)a \right\|_2.$$

Thus,

$$\sin(\theta) = \frac{\left\| (a^T a)b - (a^T b)a \right\|_2}{\|a\|_2^2 \|b\|_2} = \frac{\|q\|_2}{\|a\|_2^2 \|b\|_2},$$

where $q = (a^T a)b - (a^T b)a$

- **Area of the parallelogram spanned by a and b**: Consider a parallelogram with sides spanned by $a, b \in \mathbb{R}^n$

Recall that the area of such a shape is base $\times$ height, so

$$A = \|b\|_2 \cdot h.$$

Notice that

$$\sin(\theta) = \frac{h}{\|a\|_2} \implies h = \|a\|_2 \sin(\theta).$$

But, θ is precisely the angle between a and b, and we derived a formula for the sine of that angle above. That formula is

$$\sin(\theta) = \frac{\left\|(a^T a)b - (a^T b)a\right\|_2}{\|a\|_2^2 \|b\|_2}.$$

So, the height of the parallelogram is given by

$$h = \|a\|_2 \sin(\theta) = \|a\|_2 \frac{\left\|(a^T a)b - (a^T b)a\right\|_2}{\|a\|_2^2 \|b\|_2} = \frac{\left\|(a^T a)b - (a^T b)a\right\|_2}{\|a\|_2 \|b\|_2}.$$

With this information, the area of the parallelogram spanned by two vectors $a, b \in \mathbb{R}^n$ is given by

$$\begin{aligned} A = \|b\|_2 h = \|a\|_2 \|b\|_2 \sin(\theta) &= \|a\|_2 \|b\|_2 \frac{\left\|(a^T a)b - (a^T b)a\right\|_2}{\|a\|_2^2 \|b\|_2} \\ &= \frac{\left\|(a^T a)b - (a^T b)a\right\|_2}{\|a\|_2}. \end{aligned}$$

- **Gram determinants form of area of a parallelogram spanned by a and b and the sine of the angle between them**: Let $a, b \in \mathbb{R}^n$, consider the form above for $\sin(\theta)$,

$$\sin(\theta) = \frac{\left\|(a^T a)b - (a^T b)a\right\|_2}{\|a\|_2^2 \|b\|_2}.$$

Let $\alpha,\ \beta,\ \gamma \in \mathbb{R}$ with

$$\alpha = a^T a = \|a\|_2^2, \quad \beta = a^T b, \quad \gamma = b^T b = \|b\|_2^2,$$

and define

$$v := (a^T a)b - (a^T b)a.$$

Thus,

$$v = \alpha b - \beta a.$$

Compute $\|v\|_2^2$,

$$\begin{aligned}\|v\|_2^2 = v^T v = (\alpha b - \beta a)^T(\alpha b - \beta a) = (\alpha b^T - \beta a^T)(\alpha b - \beta a)\\ = \alpha^2 b^T b - \alpha\beta b^T a - \alpha\beta a^T b + \beta^2 a^T a = \alpha^2 b^T b - 2\alpha\beta a^T b + \beta^2 a^T a\\ = \alpha^2\gamma - 2\alpha\beta^2 + \beta^2\alpha = \alpha^2\gamma - \alpha\beta^2 = \alpha(\alpha\gamma - \beta^2).\end{aligned}$$

Thus,

$$\|v\|_2 = \sqrt{\|v\|_2^2} = \sqrt{\alpha(\alpha\gamma - \beta^2)} = \sqrt{\alpha}\sqrt{\alpha\gamma - \beta^2}.$$

So,

$$\begin{aligned}\sin(\theta) = \frac{\|v\|_2}{\|a\|_2^2 \|b\|_2} = \frac{\sqrt{\alpha}\sqrt{\alpha\gamma - \beta^2}}{\alpha\sqrt{\gamma}} = \frac{\sqrt{\alpha\gamma - \beta^2}}{\sqrt{\alpha}\sqrt{\gamma}}\\ = \frac{\sqrt{(a^T a)(b^T b) - (a^T b)^2}}{\|a\|_2 \|b\|_2} = \frac{\sqrt{\|a\|_2^2 \|b\|_2^2 - (a^T b)^2}}{\|a\|_2 \|b\|_2},\end{aligned}$$

and the height is therefore given by

$$h = \|a\|_2 \sin(\theta) = \|a\|_2 \frac{\sqrt{\|a\|_2^2 \|b\|_2^2 - (a^T b)^2}}{\|a\|_2 \|b\|_2} = \frac{\sqrt{\|a\|_2^2 \|b\|_2^2 - (a^T b)^2}}{\|b\|_2}.$$

Thus, the area of the parallelogram is

$$A = \|b\|_2 h = \|a\|_2 \|b\|_2 \sin(\theta) = \sqrt{\|a\|_2^2 \|b\|_2^2 - (a^T b)^2}.$$

Note: Notice that the quantity $\sqrt{(a^T a)(b^T b) - (a^T b)^2}$ can be expressed with

$$\sqrt{\det\begin{pmatrix} a^T a & a^T b \\ a^T b & b^T b \end{pmatrix}}.$$

Thus,

$$A = \sqrt{\det\begin{pmatrix} a^T a & a^T b \\ a^T b & b^T b \end{pmatrix}} = \|a\|_2 \|b\|_2 \sin(\theta).$$

That determinant you see above is called the **Gram determinant**, or the **Gramian determinant** of the pair $\{a, b\}$, for $a, b \in \mathbb{R}^n$

- **Gram matrix and Gram determinant**: Given vectors $v_1, ..., v_k$ in an inner product space, their **Gram matrix** is

$$G = (v_i^T v_j)_{i,j=1}^k,$$

and the determinant $\det(G)$ is the **Gram** determinant, with interpretations

  - It equals the squared k-dimensional volume of the parallelepiped spanned by the vectors $v_1, ..., v_k$
  - It is always nonnegative (Gram matrix is positive semidefinite).
  - It vanishes exactly when the vectors are linearly dependent.

- **Area of parallelogram spanned by two vectors a, b in 2-d space**: Consider $a, b \in \mathbb{R}^2$ with

$$a = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \quad b = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}.$$

From the derivation above,

$$A = \sqrt{\det \begin{pmatrix} a^T a & a^T b \\ a^T b & b^T b \end{pmatrix}}.$$

Since

$$a^T a = x_1^2 + y_1^2, \quad a^T b = x_1 x_2 + y_1 y_2, \quad b^T b = x_2^2 + y_2^2,$$

we have

$$\begin{aligned} A &= \sqrt{(x_1^2 + y_1^2)(x_2^2 + y_2^2) - (x_1 x_2 + y_1 y_2)^2} \\ &= \sqrt{(x_1 x_2)^2 + (x_1 y_2)^2 + (x_2 y_1)^2 + (y_1 y_2)^2 - ((x_1 x_2)^2 + 2x_1 y_1 x_2 y_2 + (y_1 y_2)^2)} \\ &= \sqrt{(x_1 x_2)^2 + (x_1 y_2)^2 + (x_2 y_1)^2 + (y_1 y_2)^2 - (x_1 x_2)^2 - 2x_1 y_1 x_2 y_2 - (y_1 y_2)^2} \\ &= \sqrt{(x_1 y_2)^2 - 2x_1 y_2 x_2 y_1 + (x_2 y_1)^2} = \sqrt{(x_1 y_2 - x_2 y_1)^2} = |x_1 y_2 - x_2 y_1|. \end{aligned}$$

Thus, the area of the parallelogram spanned by $a, b \in \mathbb{R}^2$ is the absolute value of the determinant of the matrix formed by a, b as the columns. Specifically,

$$A = |\det(M)| = \left| \det \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} \right|.$$

- **Parallelograms in n-dimensional space**: A parallelogram can be formed by two vectors in any dimension $\mathbb{R}^n$. Two vectors always span a 2-dimensional plane inside $\mathbb{R}^n$, and in that plane you can form a parallelogram exactly as in 2D.

  Let $v, w \in \mathbb{R}^n$. If they are not multiples of each other, the set

$$\text{span}\{v, w\}$$

  is a 2-dimensional plane inside $\mathbb{R}^n$. Inside that plane, the geometry is exactly like ordinary 2D geometry, and the two vectors form a parallelogram the same way they do in the plane.

  If v and w are multiples, then they lie on a line and the parallelogram collapses to a segment (area 0).

  Take two vectors $v, w \in \mathbb{R}^n$. Consider the set of all linear combinations:

$$\text{span}\{v, w\} = \{sv + tw : \ s, t \in \mathbb{R}\}.$$

  This set contains all vectors you can reach by moving in directions v and w

  There are two possibilities

  1. v and w are multiples. Then,

$$w = \lambda v.$$

     Everything in the span lies on a single line. So the span is 1-dimensional.
  2. If v and w are not multiples, then the two directions are independent. You can move forward / backward along v, and forward / backward along w, and by combing these motions, you generate a flat 2-dimensional plane inside $\mathbb{R}^n$.

This plane behaves exactly like $\mathbb{R}^2$, but embedded in higher-dimensional space.

Thus,

$$\dim(\text{span}\{v, w\}) = 2.$$

So, even if the vectors have 5 or 100 coordinates, $v, w \in \mathbb{R}^{100}$, they still span a 2D plane (inside 100-dimensional space), so long as they are not multiples.

The key idea is that k independent vectors in n-dimensional space define k unique directions, so the span of these k independent vectors in n-space define a k-dimensional space embedded in k-space.

This is why

$$A = \sqrt{(v^T v)(w^T w) - (v^T w)^2}$$

even when $v, w \in \mathbb{R}^n$

- **Unit square**: The unit square is the set

$$Q = \{(c_1, c_2 \in \mathbb{R}^2 : \ 0 \leqslant c_1, c_2 \leqslant 1)\},$$

which is intrinsically a 2-dimensional object. Its defining basis vectors are

$$e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and they live in $\mathbb{R}^2$, not $\mathbb{R}^n$.

When we talk about parallelograms in higher-dimensional spaces, the unit square does not move into $\mathbb{R}^n$. Instead, it is mapped into $\mathbb{R}^n$ by a linear transformation.

- **Describing a parallelogram using a matrix**: Let $v_1, v_2 \in \mathbb{R}^n$ define a parallelogram. Define the matrix

$$A = \begin{bmatrix} v_1 & v_2 \end{bmatrix}.$$

These vectors define the set

$$P = \{av_1 + bv_2 : \ 0 \leqslant a, b \leqslant 1\},$$

which is the parallelogram that lives in the 2-dimensional subspace

$$W = \text{span}\{v_1, v_2\} \subset \mathbb{R}^n.$$

Although the ambient space is $\mathbb{R}^n$, all geometry is intrinsically planar. The extra dimensions are irrelevant except for how the plane is embedded.

Note: "**Planar geometry**" refers to shapes and arrangements confined to a single flat plane (2D).

The matrix A defines

$$A : \ \mathbb{R}^2 \to \mathbb{R}^n, \quad (c_1, c_2) \mapsto c_1 v_1 + c_2 v_2.$$

Consider the unit square, defined by basis vectors $e_1, e_2 \in \mathbb{R}^2$. Under A,

$$Ae_1 = v_1, \quad Ae_2 = v_2.$$

Thus, the unit square in $\mathbb{R}^2$ is mapped to the parallelogram defined by v_1 and v_2.

Consider $u \in \mathbb{R}^2$, with $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$

$$Au = u_1 v_1 + u_2 v_2.$$

If $0 \leqslant u_1, u_2 \leqslant 1$, then Au lies inside the parallelogram (or on its boundary). If one or the coordinates is negative or exceeds 1, then Au lies in the same plane, but outside the parallelogram. Thus, Au parameterizes position relative to the parallelograms edges.

Consider the map A^T,

$$A^T = \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix}.$$

So,

$$A^T A = \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \begin{pmatrix} v_1 & v_2 \end{pmatrix} = \begin{pmatrix} v_1^T v_1 & v_1^T v_2 \\ v_1^T v_2 & v_2^T v_2 \end{pmatrix},$$

which is precisely the Gram matrix. Thus,

$$\mathcal{A}(P) = \sqrt{\det(A^T A)},$$

where $\mathcal{A}(P)$ is the area of the parallelogram spanned by v_1 and v_2.

- **Parallelograms using a point** P: Take any point $P \in \mathbb{R}^n$, define one side as

$$P \to P + v,$$

and the adjacent side

$$P \to P + w.$$

For two vectors $v, w \in \mathbb{R}^n$. Then, the parallelogram is

$$\{P + sv + tw : \ 0 \leqslant s \leqslant 1, \ 0 \leqslant t \leqslant 1\}.$$

This describes a true parallelogram lying in the 2D subspace embedded in n-space

- **Cross product for area**: Recall that for two vectors $v, w \in \mathbb{R}^3$, the cross product $v \times w$ gives the third vector $u \in \mathbb{R}^3$ orthogonal to v and w, and the area of the parallelogram formed by v, w is given by

$$\|v \times w\|_2 = \|v\|_2 \|w\|_2 \sin(\theta).$$

In higher dimensions, the cross product does not exist as a vector, but the same geometric quantity exists:

$$\text{Area} = \|v\|_2 \|w\|_2 \sin(\theta).$$

This is valid in any inner product space, including $\mathbb{R}^n$. Alternatively, one can compute the Gram determinant.

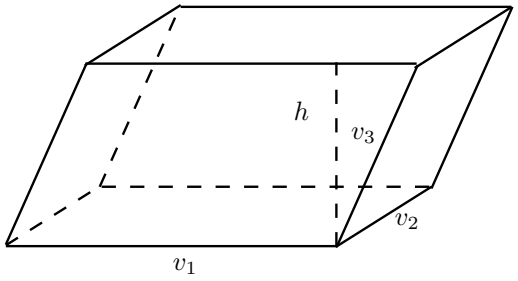
- **Unit cube**: Just like how $e_1, e_2 \in \mathbb{R}^2$ defines the **unit square** in 2-dimensional space ($\mathbb{R}^2$), the basis vectors e_1, e_2, e_3 define the **unit cube** in $\mathbb{R}^3$. The unit cube is the set of points

$$Q = \{(c_1, c_2, c_3) \in \mathbb{R}^3 : \ 0 \leqslant c_1, c_2, c_3 \leqslant 1\}.$$

- **Parallelepiped**: A parallelepiped is a 3D geometric shape made of six parallelograms, like a slanted box, where opposite faces are parallel and congruent, related to a parallelogram as a cube is to a square.

A parallelepiped is the solid generated by three vectors $v_1, v_2, v_3 \in \mathbb{R}^3$, and consists of all points of the form

$$\{av_1 + bv_2 + cv_3 : \ 0 \leqslant a, b, c \leqslant 1\}.$$



Equivalently, it is the image of the unit cube under the linear map whose columns are v_1, v_2, v_3. A parallelepiped has the following properties

- 8 vertices
- 12 edges
- 6 faces, each a parallelogram
- Opposite faces are parallel and congruent
- All edges come in three parallel families corresponding to v_1, v_2, v_3
- It has $3C2 = 3$ distinct parallelograms, the remaining three faces are congruent to one of the three distinct ones.

It need not have right angles; a rectangular box is a special case.

The parallelepiped has three distinct parallelogram faces, each defined by choosing two of the three vectors that define it. Thus,

$$\begin{aligned} P_1 &= \{a_1v_1 + b_1v_2 : \ 0 \leqslant a_1, b_1 \leqslant 1\}, \\ P_2 &= \{a_2v_1 + b_2v_2 : \ 0 \leqslant a_2, b_2 \leqslant 1\}, \\ P_3 &= \{a_3v_1 + b_3v_2 : \ 0 \leqslant a_3, b_3 \leqslant 1\}. \end{aligned}$$

Let $\mathcal{A}(P_i)$ be the area of such faces. Then, the surface area of the parallelepiped is

$$\mathcal{S} = 2(\mathcal{A}(P_1) + \mathcal{A}(P_2) + \mathcal{A}(P_3)).$$

Let $A_1 = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$, $A_2 = \begin{bmatrix} v_1 & v_3 \end{bmatrix}$, $A_3 = \begin{bmatrix} v_2 & v_3 \end{bmatrix}$, for each distinct face P_i,

$$\mathcal{A}(P_i) = \sqrt{\det\left(A_i^T A_i\right)}.$$

Thus,

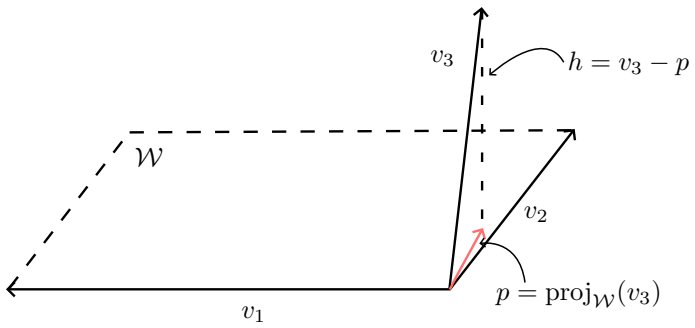
$$\mathcal{S} = 2\left(\sqrt{\det\left(A_1^T A_1\right)} + \sqrt{\det\left(A_2^T A_2\right)} + \sqrt{\det\left(A_3^T A_3\right)}\right).$$

The volume of the solid is given by the area of base times the height. If the base is the parallelogram spanned by v_1 and v_2, then the height of the solid orthogonal to these sides.

The area of the base defined by v_1 and v_2 is

$$\mathcal{A}(v_1, v_2) = \sqrt{\det\left(\begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix}\begin{pmatrix} v_1 & v_2 \end{pmatrix}\right)} = \sqrt{(v_1^T v_1)(v_2^T v_2) - (v_1^T v_2)^2}.$$

To find the height of the solid h, we can project v_3 onto the plane spanned by v_1 and v_2, then the height h is the residual $v_3 - \text{proj}_{\text{span}\{v_1,v_2\}}(v_3)$.



Where $\mathcal{W} = \text{span}\{v_1, v_2\}$, $p \in \mathcal{W}$, and $v_3 - p = h \perp \mathcal{W}$. Thus, let $V = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$, then

$$p = \text{proj}_{\mathcal{W}}(v_3) = V(V^T V)^{-1} V^T v_3.$$

Note that we can instead find the projection by solving the normal equations

$$V^T V \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = V^T v_3$$

for α and β, then finding p with

$$p = \alpha v_1 + \beta v_2.$$

Then, the height vector is given by $v_3 - p$, and so the height of the solid is

$$\|h\|_2 = \|v_3 - p\|_2 = \left\| v_3 - V\left(V^T V\right)^{-1} V^T v_3 \right\|_2.$$

If we compute the square of the norm of h, we see

$$\|h\|_2^2 = h^T h = (v_3 - p)^T (v_3 - p) = v_3^T v_3 - 2v_3^T p + p^T p.$$

But,

$$p^T p = v_3^T V (V^T V)^{-1} V^T v_3 = v_3^T p,$$

so

$$\begin{aligned} \|h\|_2 &= \sqrt{\|v_3\|_2^2 - v_3^T p} = \sqrt{\|v_3\|_2^2 - v_3^T V (V^T V)^{-1} V^T v_3} \\ &= \sqrt{v_3^T (I - P) v_3}. \end{aligned}$$

Then, with the area of the base and the height, we can compute the volume $\mathcal{V}$ of the parallelepiped spanned by $v_1,\ v_2$, and v_3

$$\mathcal{V} = \mathcal{A}(v_1, v_2)\,\|h\|_2 = \sqrt{(v_1^T v_1)(v_2^T v_2) - (v_1^T v_2)^2}\sqrt{\|v_3\|_2^2 - v_3^T V (V^T V)^{-1} V^T v_3}.$$

With some work, one can show that

$$v_3^T V (V^T V)^{-1} V^T v_3 = \frac{v_3^T v_1 (v_2^T v_2 v_1^T v_3 - v_1^T v_2 v_2^T v_3) + v_3^T v_2 (v_1^T v_1 v_2^T v_3 - v_1^T v_2 v_1^T v_3)}{v_1^T v_1 v_2^T v_2 - (v_1^T v_2)^2}.$$

Thus,

$$\begin{aligned}
\|h\|_2 &= \sqrt{v_3^T v_3 - \frac{v_3^T v_1 (v_2^T v_2 v_1^T v_3 - v_1^T v_2 v_2^T v_3) + v_3^T v_2 (v_1^T v_1 v_2^T v_3 - v_1^T v_2 v_1^T v_3)}{v_1^T v_1 v_2^T v_2 - (v_1^T v_2)^2}} \\
&= \sqrt{\frac{v_3^T v_3 (v_1^T v_1 v_2^T v_2 - (v_1^T v_2)^2) - v_3^T v_1 (v_2^T v_2 v_1^T v_3 - v_1^T v_2 v_2^T v_3) - v_3^T v_2 (v_1^T v_1 v_2^T v_3 - v_1^T v_2 v_1^T v_3)}{v_1^T v_1 v_2^T v_2 - (v_1^T v_2)^2}} \\
&= \sqrt{\frac{v_3^T v_3 (v_1^T v_1 v_2^T v_2 - (v_1^T v_2)^2) + v_3^T v_1 (v_1^T v_2 v_2^T v_3 - v_2^T v_2 v_1^T v_3) - v_3^T v_2 (v_1^T v_1 v_2^T v_3 - v_1^T v_2 v_1^T v_3)}{v_1^T v_1 v_2^T v_2 - (v_1^T v_2)^2}}.
\end{aligned}$$

Let G be the Gram matrix, and A be the matrix whose columns are v_1, v_2, v_3. That is, $A = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \in \mathbb{R}^{n \times 3}$

$$G = \begin{bmatrix} V^T V & V^T v_3 \\ v_3^T V & v_3^T v_3 \end{bmatrix} = \begin{bmatrix} v_1^T v_1 & v_1^T v_2 & v_1^T v_3 \\ v_2^T v_1 & v_2^T v_2 & v_2^T v_3 \\ v_3^T v_1 & v_3^T v_2 & v_3^T v_3 \end{bmatrix} = A^T A.$$

Notice that the numerator of $\|h\|_2$ is precisely the determinant of G (expand along the bottom row). So,

$$\|h\|_2 = \sqrt{\frac{\det(G)}{v_1^T v_1 v_2^T v_2 - (v_1^T v_2)^2}} = \sqrt{\frac{\det(G)}{\det(V^T V)}}.$$

Thus, we finally arrive at the volume of the parallelepiped spanned by three vectors $v_1, v_2, v_3 \in \mathbb{R}^n$, where v_1 and v_2 define the base parallelogram.

$$\begin{aligned}
\mathcal{V} &= \sqrt{(v_1^T v_1)(v_2^T v_2) - (v_1^T v_2)^2}\sqrt{\frac{\det(G)}{\det(V^T V)}} \\
&= \sqrt{\det(V^T V)}\sqrt{\frac{\det(G)}{\det(V^T V)}} = \sqrt{\det(G)}.
\end{aligned}$$

Note: If $v_1, v_2, v_3 \in \mathbb{R}^3$, then $A = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \in \mathbb{R}^{3 \times 3}$. So $\det(A)$ exists, and

$$G = A^T A \implies \det(G) = \det(A^T A) = \det(A^T)\det(A) = \det(A)^2.$$

Therefore, for $v_1, v_2, v_3 \in \mathbb{R}^3$,

$$\mathcal{V} = \sqrt{\det(G)} = \sqrt{(\det(A))^2} = |\det(A)|\,.$$

- **Projection onto a plane**: Consider a plane spanned by linearly independent vectors $v, w \in \mathbb{R}^n$, so

$$\mathcal{P} = \text{span}\{v, w\}.$$

Observe that this is a 2-dimensional plane embedded in the ambient space $\mathbb{R}^n$. Suppose that $b \in \mathbb{R}^n$ is a vector not in the plane $\mathcal{P}$. We wish to project this vector b onto $\mathcal{P}$. That is, we project b onto $\text{span}\{v, w\}$. Call this projection p. The residual is then $b - p$, which is orthogonal to all vectors in $\mathcal{P}$, namely v and w.

Projecting b onto $\mathcal{P}$ yields $p \in \mathcal{P}$, so

$$p = \alpha v + \beta w.$$

$b - p$ is orthogonal to both v and w, so

$$\begin{aligned} v^T(b-p) = 0 &\implies v^T(b - (\alpha v + \beta w)) = 0, \\ w^T(b-p) = 0 &\implies w^T(b - (\alpha v + \beta w)) = 0. \end{aligned}$$

Thus,

$$\begin{aligned} \alpha v^T v + \beta v^T w &= v^T b, \\ \alpha w^T v + \beta w^T w &= w^T b. \end{aligned}$$

Let $V = \begin{bmatrix} v & w \end{bmatrix} \in \mathbb{R}^{n \times 2}$, so $V^T = \begin{bmatrix} v^T \\ w^T \end{bmatrix} \in \mathbb{R}^{2 \times n}$, and

$$V^T V = \begin{bmatrix} v^T v & v^T w \\ v^T w & w^T w \end{bmatrix} \in \mathbb{R}^{2 \times 2}.$$

So, our **projection system**, which yields the **projection equations**, also called the **normal equations** is

$$V^T V c = V^T b,$$

where $c = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ is the coefficient vector containing the coefficients to $p \in \mathcal{P}$, the projection of b onto the space spanned by v and w.

Note: Notice that $p = Vc$, and $c = (V^T V)^{-1} V^T b$. Thus,

$$p = \text{proj}_{\mathcal{P}}(b) = V(V^T V)^{-1} V^T b.$$

Define the projection matrix $V(V^T V)^{-1} V^T = P$, then

$$p = \text{proj}_{\mathcal{P}}(b) = Pb.$$

The 2-norm of p is given by

$$\|p\|_2 = \left\| V(V^T V)^{-1} V^T b \right\|_2.$$

If we compute the square of the norm,

$$\begin{aligned} \|p\|_2 = p^T p &= \left(V(V^T V)^{-1} V^T b\right)^T \left(V(V^T V)^{-1} V^T b\right) \\ &= \left(V V^{-1} V^{-T} V^T b\right)^T \left(V V^{-1} V^{-T} V^T b\right) \\ &= b^T V V^{-1} V^{-T} V^T V V^{-1} V^{-T} V^T b \\ &= b^T V V^{-1} V^{-T} V^T b = b^T V (V^T V)^{-1} V^T b. \end{aligned}$$

Thus,

$$\|p\|_2 = \sqrt{\|p\|_2^2} = \sqrt{b^T V (V^T V)^{-1} V^T b}.$$

- **Projection onto a space**: Let v be a vector in some space V, and W be a subspace of V. If W is spanned by a single vector b (W is one-dimensional), then the projection is the one we know,

$$\text{proj}_W(v) = \text{proj}_b(v) = \frac{v^T b}{b^T b} b.$$

If W is spanned by an orthonormal basis $\mathcal{B} = \{q_1, ..., q_k\}$, then

$$\text{proj}_W(v) = \sum_{i=1}^{k} (v^T q_i) q_i,$$

and if Q is the matrix formed by the orthonormal basis $\mathcal{B}$, $Q = \begin{bmatrix} q_1 & \cdots & q_k \end{bmatrix}$, then

$$\text{proj}_W(v) = QQ^T v.$$

If W is spanned by a basis $\mathcal{B} = \{a_1, ..., a_k\}$ that may not be orthonormal, and $A = \begin{bmatrix} a_1 & ... & a_k \end{bmatrix}$, then the projection of v onto the subspace W is given by

$$\text{proj}_W(v) = A(A^T A)^{-1} A^T v.$$

The matrix $P = A(A^T A)^{-1} A^T$ is called the **projection matrix**, so $\text{proj}_W(v) = Pv$

- **Orthogonal projection theorem**: Let W be a subspace of $\mathbb{R}^n$, and x be a vector in $\mathbb{R}^n$. The closest point in W to x is its orthogonal projection onto W, denoted $\text{proj}_W(x)$. That is,

$$\min_{w \in W} \|x - w\|_2 = \text{proj}_W(x).$$

Proof. Let $w \in W$, and p be the projection of x onto the subspace W. We have

$$x - w = x - w + p - p = (x - p) + (p - w).$$

Observe that $x - w \notin W$, $x - p \notin W$, $p - w \in W$. So, three noncollinear points, which form a triangle. Since $x - p \perp W$, and $p - w \in W$, $x - w$ is the hypotenuse, and by the Pythagorean Theorem,

$$\|x - w\|_2^2 = \|x - p\|_2^2 + \|p - w\|_2^2 .$$

Now, since $\|p - w\|_2 \geqslant 0$, $\|x - w\|_2^2 \geqslant \|x - p\|_2^2$, and $\|x - w\|_2^2 = \|x - p\|_2^2$ when $w = p$.

This fact is only true in inner-product space, where distance is defined using the norm induced by that inner product, so

$$\langle x, x \rangle = \sqrt{\|x\|}.$$

- **n-dimensional parallelepiped**: Let

$$v_1, v_2, \ldots, v_k \in \mathbb{R}^n, \quad k \leqslant n.$$

The **parallelepiped** spanned by $v_1, \ldots, v_k$ is the set

$$P(v_1, \ldots, v_k) = \left\{ \sum_{i=1}^{k} t_i v_i : \ 0 \leqslant t_i \leqslant 1 \right\}.$$

  - For $k = 1$: a line segment
  - For $k = 2$: a parallelogram
  - For $k = 3$: a 3D parallelepiped
  - For $k = n$: an n-dimensional parallelepiped (sometimes called a parallelotope)

This definition works in any ambient dimension n.

An n-dimensional parallelepiped has

- n edge vectors $v_1, \ldots, v_k$
- 2^n vertices
- All faces are lower-dimensional parallelepipeds
- Opposite faces are parallel and congruent

Each $(n1)$-dimensional face of an n-parallelepiped is itself an $(n1)$-dimensional parallelepiped, with volume given by determinants of submatrices (or Gram minors).

The n-dimensional volume (hypervolume) of the parallelepiped spanned by $v_1, \ldots, v_k$ is given by

$$\mathcal{V}(v_1, \ldots, v_k) = \sqrt{\det\left(\text{Gram}(v_1, \ldots, v_k)\right)}.$$

Where $\text{Gram}(v_1, \ldots, v_k)$ is the Gram matrix

$$\text{Gram}(v_1, \ldots, v_k) = G = \left(v_i^T v_j\right)_{i,j=1}^{n} = A^T A,$$

with $A = \begin{bmatrix} v_1 & v_2 & \cdots & v_k \end{bmatrix}$. If $k = n$, then $\det(A)$ is defined, and

$$\det(G) = \det(A^T A) = \det(A^T)\det(A) = (\det(A))^2.$$

So,

$$\mathcal{V}(v_1, \ldots, v_n) = \sqrt{(\det(A))^2} = |\det(A)|.$$

If the set of vectors $v_1, \ldots, v_k$ is orthonormal, then $v_i v_j = 0$ for $i \neq j$, and $\|v_i\|_2 = 1$. So,

$$G = (v_i^T v_j)_{i,j=1}^{k} = I_k.$$

Thus,

$$\mathcal{V}(v_1, \ldots, v_k) = \sqrt{\det(G)} = \sqrt{\det(I_k)} = 1.$$

So, if the edges vectors form an orthonormal set, the volume of k-dimensional parallelepiped is one.

Suppose that a k-dimensional parallelepiped with edge vectors $v_1, \ldots, v_k$ is embedded in the ambient space $\mathbb{R}^n$, so $v_1, \ldots, v_k \in \mathbb{R}^n$. But, suppose that the set of edge vectors is not linearly independent.

Recall that the Gram matrix $G = V^T V$, where $V = \begin{bmatrix} v_1 & \cdots & v_k \end{bmatrix}$. So, $G \in \mathbb{R}^{k \times k}$, and $V \in \mathbb{R}^{n \times k}$. Let $u \in \mathbb{R}^k$ such that $Gu = 0$, so $u \in \ker(G)$,

$$\begin{aligned} Gu = 0 &\iff V^T V u = 0 \iff u^T V^T V u = 0 \iff (Vu)^T (Vu) = 0 \\ &\iff \|Vu\|_2^2 = 0 \iff Vu = 0. \end{aligned}$$

Thus, $u \in \ker(V)$, and $\ker(G) = \ker(V)$. Consider the mapping of G and V,

$$\begin{aligned} G &: \mathbb{R}^k \to \mathbb{R}^k, \\ V &: \mathbb{R}^k \to \mathbb{R}^n. \end{aligned}$$

By the rank-nullity theorem,

$$\begin{aligned} k &= \text{rank}(G) + \dim(\ker(G)), \\ k &= \text{rank}(V) + \dim(\ker(V)). \end{aligned}$$

But, since $\dim(\ker(G)) = \dim(\ker(V))$, $\text{rank}(G) = \text{rank}(V)$

If the $v_1, \ldots, v_k$ is linearly dependent, then $\text{rank}(V) < k$, so $\text{rank}(G) < k$. Thus,

$$\det(G) = 0.$$

Since the determinant is zero,

$$\mathcal{V}(v_1, \ldots, v_k) = \sqrt{\det(G)} = 0.$$

Thus, the volume is zero.

Note: The parallelepiped does exist, but it is degenerate. "Volume zero" does not mean "nonexistent"; it means the object has collapsed into a lower-dimensional shape.

Collapsed means that the set defined by the vectors still exists as a set of points, but it no longer occupies the full dimension you are trying to measure. One or more independent directions have disappeared. A direction that was supposed to provide thickness or height becomes zero. The shape lies entirely in a lower-dimensional subspace.

- In $\mathbb{R}^3$: three dependent vectors span a plane or a line, not a volume-filling solid.
- In $\mathbb{R}^2$: two dependent vectors give a line segment, not an area.
- In $\mathbb{R}^n$: k dependent vectors span a subspace of dimension $r < k$

Algebraically, let $V = \begin{bmatrix} v_1 & v_2 & \cdots & v_k \end{bmatrix}$, for $v_1, \ldots, v_k \in \mathbb{R}^n$ where $k \leqslant n$. If $\text{rank}(V) = r < k$, then the linear map

$$x \mapsto Vx$$

maps the k-dimensional unit cube into an r-dimensional set, so the "collapse" is exactly the loss of injectivity of this map.

- **Understanding the cross product**: In $\mathbb{R}^3$, the cross product is a binary operation that takes two vectors and returns a third vector that is orthogonal to both. It is a geometric construction specific to three-dimensional Euclidean space.

Let $a, b \in \mathbb{R}^3$, with

$$a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

The cross product $a \times b$ is

$$a \times b = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_1 b_3 - a_3 b_1 \\ a_1 b_2 - a_2 b_1 \end{pmatrix},$$

which can also be expressed as a determinant

$$a \times b = \det \begin{pmatrix} i & j & k \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix}.$$

The cross product satisfies the following key properties

- **Orthogonality**: $(a \times b) \cdot a = 0$, $(a \times b) \cdot b = 0$
- **Bilinearity**: Linear in each argument separately

– **Anti-commutativity**: $a \times b = -(b \times a)$
– **Zero condition**: $a \times b = 0 \iff a$ and b are linearly dependent

Observe $(a \times b)^T a$ and $(a \times b)^T b$,

$$\begin{aligned}(a \times b)^T a &= \begin{pmatrix} a_2b_3 - a_3b_2 \\ a_1b_3 - a_3b_1 \\ a_1b_2 - a_2b_1 \end{pmatrix}^T \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \\ &= a_1(a_2b_3 - a_3b_2) + a_2(a_1b_3 - a_3b_1) + a_3(a_1b_2 - a_2b_1) \\ &= \det \begin{pmatrix} a_1 & a_2 & a_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} = 0.\end{aligned}$$

The same goes for $(a \times b)^T b$,

$$\begin{aligned}(a \times b)^T b &= \begin{pmatrix} a_2b_3 - a_3b_2 \\ a_1b_3 - a_3b_1 \\ a_1b_2 - a_2b_1 \end{pmatrix}^T \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \\ &= b_1(a_2b_3 - a_3b_2) + b_2(a_1b_3 - a_3b_1) + b_3(a_1b_2 - a_2b_1) \\ &= \det \begin{pmatrix} b_1 & b_2 & b_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} = 0.\end{aligned}$$

Thus,

$$a \times b = \det \begin{pmatrix} i & j & k \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix}$$

forms a vector orthogonal to both a and b.

Next, from the logic above we see that for $a, b, c \in \mathbb{R}^3$, we have

$$(a \times b)^T c = (a \times b) \cdot c = \det \begin{pmatrix} c_1 & c_2 & c_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} = \det \left(\begin{bmatrix} a & b & c \end{bmatrix} \right).$$

Notice that we made two row swaps, which has equal determinant, then took the transpose, which also has equal determinant.

Using this fact, we can see that

$$\|a \times b\|_2^2 = (a \times b)^T (a \times b) = \det(\begin{bmatrix} a & b & a \times b \end{bmatrix}).$$

From here, we can use the Gram determinant identity $\det(G) = (\det(A))^2$

$$\left(\|a \times b\|_2^2 \right)^2 = \left(\det \begin{bmatrix} a & b & a \times b \end{bmatrix} \right)^2 = \det \begin{pmatrix} a^T a & a^T b & a^T (a \times b) \\ b^T a & b^T b & b^T (a \times b) \\ (a \times b)^T a & (a \times b)^T b & (a \times b)^T (a \times b) \end{pmatrix}.$$

But, $a \times b$ is orthogonal to a and b, so

$$\|a \times b\|_2^4 = \det \begin{pmatrix} a^T a & a^T b & 0 \\ b^T a & b^T b & 0 \\ 0 & 0 & (a \times b)^T (a \times b) \end{pmatrix}.$$

Notice that we can expand along the bottom row to get

$$\begin{aligned}\|a\times b\|_2^4 &= \det\begin{pmatrix} a^Ta & a^Tb & 0\\ b^Ta & b^Tb & 0\\ 0 & 0 & (a\times b)^T(a\times b)\end{pmatrix}\\ &= (a\times b)^T(a\times b)\left((a^Ta)(b^Tb)-(a^Tb)^2\right)\\ &= \|a\times b\|_2^2\left((a^Ta)(b^Tb)-(a^Tb)^2\right)\\ &= \|a\times b\|_2^2\det\begin{pmatrix} a^Ta & a^Tb\\ a^Tb & b^Tb\end{pmatrix}.\end{aligned}$$

Thus,

$$\|a\times b\|_2^2 = \det\begin{pmatrix} a^Ta & a^Tb\\ a^Tb & b^Tb\end{pmatrix} = \det\left(\mathrm{Gram}(a,b)\right).$$

Now, recall that the area of the parallelogram spanned by a and b is precisely $\sqrt{\det \mathrm{Gram}(a,b)}$. Thus, we can conclude that

$$\|a\times b\|_2 = \sqrt{\det\left(\mathrm{Gram}(a,b)\right)} = \mathcal{A}(a,b) = \|a\|_2\|b\|_2\sin(\theta).$$

From this fact, we see that if $a\times b = 0$, then $\|a\times b\| = 0$, and so the area of the parallelogram spanned by a and b is zero, and so a and b are linearly dependent.

Bilinearity and anti-commutativity comes from the multilinearity and alternating nature of the determinant.

- **Summary of the cross product**: Let $a,b\in\mathbb{R}^3$, the cross product $a\times b$ produces a vector $c\in\mathbb{R}^3$ orthogonal to both a and b, with length equal to the area of the parallelogram spanned by a and b.

  The cross product $a\times b$ is defined as

$$a\times b = \det\begin{pmatrix} i & j & k\\ a_1 & a_2 & a_3\\ b_1 & b_2 & b_3\end{pmatrix} = \begin{pmatrix} a_2b_3-a_3b_2\\ a_1b_3-a_3b_1\\ a_1b_2-a_2b_1\end{pmatrix},$$

  with

$$\|a\times b\|_2 = \|a\|_2\|b\|_2\sin(\theta) = \sqrt{\det\left(\mathrm{Gram}(a,b)\right)},$$

  and

$$(a\times b)^Tc = \det\left(\begin{bmatrix} a & b & c\end{bmatrix}\right)$$

  for all $c\in\mathbb{R}^3$.

  The cross product is bilinear in a and b and anti-commutative. Also, $a\times b = 0$ is a necessary and sufficient condition to a and b being linearly dependent.

- **Triple scalar product**: The triple scalar product is a scalar quantity associated with three vectors in $\mathbb{R}^3$. It combines the cross product and the dot product and is the standard algebraic representation of oriented volume.

For $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$, the triple scalar product is defined as

$$[\mathbf{a}, \mathbf{b}, \mathbf{c}] = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \det \begin{bmatrix} \mathbf{a} & \mathbf{b} & \mathbf{c} \end{bmatrix}.$$

By the alternating nature of the determinant,

$$\det \begin{bmatrix} \mathbf{a} & \mathbf{b} & \mathbf{c} \end{bmatrix} = \det \begin{bmatrix} \mathbf{b} & \mathbf{c} & \mathbf{a} \end{bmatrix} = \det \begin{bmatrix} \mathbf{c} & \mathbf{a} & \mathbf{b} \end{bmatrix}.$$

Thus,

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}).$$

Also, by the nature of the dot product,

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{b} \times \mathbf{c}) \cdot \mathbf{a}.$$

Since

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \det \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix},$$

the volume of the parallelepiped spanned by $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ is given by

$$\mathcal{V}(\mathbf{a}, \mathbf{b}, \mathbf{c}) = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|.$$

7.2 Geometric operations

7.2.1 Position vectors and Translations

- **Positions vs directions**: Vectors play two roles

  1. **Direction / displacement**: Free vectors
  2. **Location / position**: Points, identified with position vectors

  So, either a vector is a **direction vector** or a **position vector**. Direction vectors are sometimes called **free vectors**.

- **Direction (free) vectors**: A direction vector represents a change.

  - It can be freely added to other direction vectors.
  - It can be scaled.
  - It represents "how to move," not "where you are."

  For example,

  - Elements of $\ker(A)$
  - Basis vectors
  - Velocity, force, displacement in physics

  Formally, these live in **vector spaces**. Direction vectors by convention start at the origin, their position in space is not of concern. These vectors encode a direction and a length (magnitude).

  The sum of two direction vectors is also a direction vector, and so it also begins at the origin by convention.

- **Position vectors**: A position vector represents a location relative to a chosen origin.

  - You do not add two positions meaningfully.
  - You can subtract two positions to get a direction.
  - You can add a direction to a position.

  For example,

  - A particular solution x_0
  - A point on an affine line or plane
  - Coordinates of a physical location

  Formally, these live in **affine spaces**. Position vectors also start at the origin by convention.

  When we add a position vector to a direction vector, the starting point is at the position vector, and the direction is

- **Position and direction**: Suppose x_0 is a position vector, which represents a point in space, and v is a direction vector, with length and direction.

  $$x_0 + v$$

  means start at x_0, and move in the direction of v for the length of v. The norm $\|v\|$ gives us the distance to travel from x_0.

Let $x_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, and $v = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Then,

$$x_0 + v = \begin{pmatrix} 0+1 \\ 1+0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

We do not interpret $x_0 + v$ as a direction vector, it is a new position vector. Thus, length and direction is not of concern. It is a new point, not a new direction.

- **Translations**: Fix a direction vector $a \in \mathbb{R}^n$. The translation by a is the map

$$T_a : \mathbb{R}^n \to \mathbb{R}^n, \quad T_a(x) = x + a,$$

where the input x is a point in space, a position. Since the domain of this translation is $\mathbb{R}^n$, every point in $\mathbb{R}^n$ is moved by the displacement a.

In $\mathbb{R}^2$,

7.3 Geometry of linear equations

- **Direct sum**: A direct sum is a way of saying that a vector space can be decomposed into smaller subspaces in a manner that is both complete and non-overlapping.

  Let V be a vector space and let $U, W \subseteq V$ be subspaces. We write

  $$V = U \oplus W$$

  if and only if both of the following hold.

  1. **Every vector can be written as a sum**:

     $$\forall v \in V, \quad v = u + w \quad \text{with } u \in U,\ w \in W.$$

  2. **The representation is unique**:

     $$U \cap W = \{0\}.$$

  Suppose v has two different decompositions,

  $$v = u_1 + w_1 = u_2 + w_2$$

  for $u_1, u_2 \in U,\ w_1, w_2 \in W$. Then, subtracting the two equations gives

  $$0 = u_1 + w_1 - u_2 - w_2.$$

  So,

  $$u_1 - u_2 = -(w_1 - w_2).$$

  Thus, this vector is common to both subspaces. If we enforce that $U \cap W = \{0\}$, then this vector must be zero. So,

  $$\begin{aligned} u_1 - u_2 &= 0, \\ -w_1 + w_2 &= 0. \end{aligned}$$

  Therefore, $u_1 = u_2$, $w_1 = w_2$, and the decomposition is unique.

  Equivalent characterizations are

  - $V = U \oplus W$
  - $V = U + W$ and $U \cap W = \{0\}$
  - Every $v \in V$ has a **unique** decomposition $v = u + w$.

  So, the $\oplus$ notation when used on subspaces explains that an ambient space can be decomposed into subspaces, and the subspaces only share the zero vector.

  Without the word *direct*, the sum

  $$V = U + W$$

  only means that vectors can be written as sums, but not uniquely. In this case, $V \cap W \neq \{0\}$, and some vectors admit multiple decompositions, so the subspaces overlap.

If

$$V = U \oplus W,$$

then

$$\dim(V) = \dim(U) + \dim(W).$$

This definition extends naturally. For subspaces $U_1, \ldots, U_k$,

$$V = U_1 \oplus \cdots \oplus U_k.$$

So, every vector in V can be uniquely expressed as a sum $v = u_1 + \cdots + u_k$, for $u_i \in U_i$

Geometrically, each subspace provides an independent "direction of freedom", no direction is counted twice, together they span the entire space.

As an example, consider standard coordinates in $\mathbb{R}^2$,

$$\mathbb{R}^2 = \text{span}\{(1,0)\} \oplus \text{span}\{(0,1)\}.$$

Every vector (x, y) decomposes uniquely as

$$(x, y) = (x, 0) \oplus (0, y).$$

- **Linear functionals and the dual space**: A linear functional is a map

$$\ell : V \to \mathbb{F},$$

where V is a vector space over a field $\mathbb{F}$. In $\mathbb{R}^n$, every linear functional has the form

$$\ell(x) = a_1x_1 + a_2x_2 + \cdots + a_nx_n = a^Tx$$

for a fixed vector $a \in \mathbb{R}^n, a \neq 0$.

The set of all linear functionals on V forms a vector space, called the **dual space $\mathbf{V}^*$**. In finite dimensions,

$$\dim(v^*) = \dim(V).$$

A linear functional measures a signed component of a vector along a fixed direction. For $\ell(x) = a^Tx$, the vector a acts as a **normal vector**. The value $\ell(x)$ is proportional to the projection of x onto a. This makes linear functionals fundamental tools for describing orientation, constraints, and projections.

- **Nonzero functional**: If V is a vector space over a field $\mathbb{F}$, then a linear functional

$$\ell : V \to \mathbb{F}$$

is called **nonzero** if

$$\ell(v) \neq 0 \quad \text{for at least one } v \in V$$

- **The zero functional**: The zero functional is the map

$$\ell_0(v) = 0 \quad \text{for all } v \in V.$$

It is linear, but geometrically degenerate. Only nonzero functional produce meaningful geometry.

- **Rank and nullity of a linear functional**: For $\mathbb{R}^n$ over $\mathbb{R}$, the linear functional

$$\ell : V \to \mathbb{F}, \quad \ell(x) = a_1x_1 + \cdots + a_nx_n = Ax$$

Has rank one, since A has one row and the functional maps to $\mathbb{R}$. By the rank-nullity theorem,

$$n = \dim(\ker(\ell)) + \dim(\text{Im}(\ell)) = \text{nullity}(A) + \text{rank}(A).$$

Since $\text{rank}(A) = \dim(\ker(\ell)) = 1$,

$$n = \dim(\ker(\ell)) + 1.$$

So, the dimension of the kernel is $n - 1$. The kernel is the set of vectors $v \in \mathbb{R}^n$ such that

$$Av = 0.$$

So, it is the set of vectors orthogonal to a, where a is the first (only) row of A, the normal vector that describes the orientation.

$$a_1v_1 + \cdots + a_nv_n = 0.$$

Which describes the $(n-1)$-dimensional hyperplane through the origin.

Thus,

$$\mathbb{R}^n = \ker(\ell) \oplus \text{span}\{a\}.$$

So, every vector in $\mathbb{R}^n$ can be uniquely expressed as a sum of two components, one lying in the kernel of ℓ, and one lying in the direction of a.

Take any vector $x \in \mathbb{R}^n$, let $x_\parallel$ be the projection of x onto the span of a. So,

$$x_\parallel = \frac{a^Tx}{a^Ta}a.$$

Now, observe that $a^Tx = a^Tx_\parallel$. That is the amount of x in the direction of a is the same as the amount of $x_\parallel$ is the direction of a, because the amount of x in the direction of a is expressed in $x_\parallel$.

We want a vector $x_\perp$ such that $a^Tx_\perp = 0$. Notice that we have two quantities equal to a^Tx. So,

$$a^Tx_\perp = 0 = a^Tx - a^Tx = a^Tx - a^Tx_\parallel.$$

So, define

$$x_\perp := x - x_\parallel.$$

Then,

$$a^Tx_\perp = a^T(x - x_\parallel) = a^Tx - a^Tx_\parallel = a^Tx - a^Tx = 0.$$

Thus, $x_\perp$ is orthogonal to a. Therefore,

$$x = x_\perp + x_\parallel.$$

This shows that every vector is a sum of an element in $\ker(\ell)$ and an element of $\text{span}\{a\}$.

Recall that

$$\ker(\ell) = \{v \in \mathbb{R}^n : \ a^T v = 0\},$$
$$\text{span}\{a\} = \{\lambda a : \ \lambda \in \mathbb{R}\}.$$

Suppose that $\lambda a \in \ker(\ell)$. Then,

$$a^T(\lambda a) = \lambda a^T a = \lambda \|a\|_2^2 = 0.$$

But, recall that in the definition of a linear functional, $a \neq 0$. Thus, $\lambda = 0$, which implies that $\lambda a = 0$, which is the zero vector. So, since a member of the span of a being in the kernel of ℓ implies the zero vector, it must be that

$$\ker(\ell) \cap \text{span}\{a\} = \{0\}.$$

Thus, the sum is direct.

- **Level sets and hyperplanes**: For a fixed scalar b, the equation

$$\ell(x) = b$$

defines a **level set**. This set is an $(n-1)$-dimensional **affine hyperplane** in $\mathbb{R}^n$. All points on the hyperplane have the same dot product with a. Changing b translates the hyperplane parallel to itself.

When $b = 0$, the level set is the **kernel** of ℓ, a linear subspace of dimension $n-1$.

- **Kronecker delta**: For integers i and j,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \end{cases}.$$

The Kronecker delta acts as an index selector. It allows you to write statements that "pick out" a specific component while annihilating the others.

For example, the identity matrix $I_n \in \mathbb{R}^{n \times n}$ can be expressed with the Kronecker delta as follows,

$$(I_n)_{ij} = \delta_{ij}.$$

- **The dual space**: Let V be a vector space over a field $\mathbb{F}$. The **dual space** of V, denoted V^*, is defined as

$$V^* := \{\ell : \ V \to \mathbb{F} \ \mid \ \ell \text{ is linear}\}.$$

So, V^* is the vector space of **all linear functionals** on V.

The dual space is itself a vector space, with operations defined pointwise:

  - $(\ell_1 + \ell_2)(v) = \ell_1(v) + \ell_2(v)$
  - $(\alpha \ell)(v) = \alpha \ell(v)$

- **The dual space is a morphism space**: Let V be a vector space over a field $\mathbb{F}$. The dual space of V is defined as

$$V^* := \text{Hom}(V, \mathbb{F}).$$

That is, the space of all linear morphisms from V to the base field.

- **The dual basis**: Let V^* be the dual space of V. If $\mathcal{B} = \{b_1, \ldots, b_n\}$ is a basis of V, then there exists a unique **dual basis**

$$\{\varphi^1, \ldots, \varphi^n\} \subset V^*$$

such that

$$\varphi^i(b_j) = \delta_{ij}.$$

Consider some vector $v \in V$. Since $\mathcal{B}$ is a basis for V,

$$v = \sum_{j=1}^{n} x_j b_j.$$

If we apply φ^i, we get

$$\varphi^i(v) = \varphi^i\left(\sum_{j=1}^{n} x_j b_j\right).$$

Notice that φ^i is a linear functional $\varphi^i : V \to \mathbb{F}$, and each basis vector $b_j \in V$. Thus, by linearity,

$$\varphi^i\left(\sum_{j=1}^{n} x_j b_j\right) = \sum_{j=1}^{n} x_j \varphi^i(b_j) = \sum_{j=1}^{n} x_j \delta_{ij} = x_i$$

Therefore,

$$\varphi^i(v) = x_i.$$

So, the i^{th} dual basis applied to a vector $v \in V$ extracts the i^{th} entry of v.

- **Dual basis example**: Let $V = \mathbb{R}^2$, and $\mathbb{F} = \mathbb{R}$. Then, the dual space $\left(\mathbb{R}^2\right)^*$ is

$$(\mathbb{R}^2)^* = \{\ell : \mathbb{R}^2 \to \mathbb{R}\}.$$

So, $\ell \in \left(\mathbb{R}^2\right)^*$ implies

$$\ell(x) = a_1 x_1 + a_2 x_2.$$

Let $\mathcal{B}$ be the standard basis for $\mathbb{R}^2$, so

$$\mathcal{B} = \{b_1, b_2\} = \left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}.$$

The unique dual basis is then $\Phi = \{\varphi^1, \varphi^2\}$, defined by

$$\varphi^i b_j = \delta_{ij},$$

where

$$\varphi^i = \varphi^i_1 x_1 + \varphi^i_2 x_2.$$

Thus,

$$\varphi^1\begin{pmatrix}1\\0\end{pmatrix} = 1, \quad \varphi^1\begin{pmatrix}0\\1\end{pmatrix} = 0,$$
$$\varphi^2\begin{pmatrix}1\\0\end{pmatrix} = 0, \quad \varphi^2\begin{pmatrix}0\\1\end{pmatrix} = 1.$$

So, we have

$$\varphi_1^1(1) + \varphi_2^1(0) = \varphi_1^1 = 1, \quad \varphi_1^1(0) + \varphi_2^1(1) = \varphi_2^1 = 0,$$
$$\varphi_1^2(1) + \varphi_2^2(0) = \varphi_1^2 = 0, \quad \varphi_1^2(0) + \varphi_2^2(1) = \varphi_2^2 = 1.$$

Hence,

$$\varphi^1 = x_1, \quad \varphi^2 = x_2.$$

Suppose instead that we use $\mathcal{B} = \left\{ \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$ as our basis for $\mathbb{R}^2$. Then, using

$$\varphi^1 \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1, \quad \varphi^1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0,$$
$$\varphi^2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 0, \quad \varphi^2 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1,$$

we find that

$$\varphi^1 = \frac{1}{2}x_1 - \frac{1}{2}x_2 = \frac{1}{2}(x_1 - x_2),$$
$$\varphi^2 = \frac{1}{2}x_1 + \frac{1}{2}x_2 = \frac{1}{2}(x_1 + x_2).$$

Let $v \in \mathbb{R}^2$, $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$. Sending this vector to our basis $\mathcal{B}$ yields

$$v_{\mathcal{B}} = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} v_1 - v_2 \\ v_1 + v_2 \end{pmatrix}.$$

If we instead use the fact that $v_i = \varphi^i(v)$, we see that

$$v_1 = \varphi^1 \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \frac{1}{2}(v_1 - v_2),$$
$$v_2 = \varphi^2 \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \frac{1}{2}(v_1 + v_2).$$

Thus,

$$v_{\mathcal{B}} = \frac{1}{2} \begin{pmatrix} v_1 - v_2 \\ v_1 + v_2 \end{pmatrix}.$$

- **Linear equations**: A single linear equation

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$$

describes a hyperplane in $\mathbb{R}^n$

  - In $\mathbb{R}^2$: A line
  - In $\mathbb{R}^3$: A plane
  - In $\mathbb{R}^n$: A $(n-1)$ dimensional object

The normal vector $\begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix}^\top$ is perpendicular to this hyperplane. Changing b translates the hyperplane without changing its orientation.

A single linear equation in n variables describes an $(n-1)$-dimensional hyperplane because we can solve for one variable in terms of the others. Observe

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b \implies x_n = \frac{b - (a_1x_1 + a_2x_2 + \cdots + a_{n-1}x_{n-1})}{a_n}.$$

Thus, there are exactly $(n-1)$ free parameters, so the solution set is $(n-1)$-dimensional

- **Linear maps vs equations**: When we write

  $$\ell(x) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

  we are defining a linear functional $\ell : \mathbb{R}^n \to \mathbb{R}$. At this stage, we are not yet describing a geometric object. We are defining a function.

  The geometry appears only when we impose a constraint

  $$\ell(x) = b.$$

  This equation is asking for the **preimage** of the scalar b under ℓ,

  $$\ell^{-1}(\{b\}) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \ a_1x_1 + \cdots + a_nx_n = b\}.$$

  This set is the geometric object, the $(n-1)$-dimensional hyperplane.

- **Geometry of linear functionals**: A linear functional $\ell : \mathbb{R}^n \to \mathbb{R}$ does not by itself describe a geometric object. Instead, it describes a **family of geometric objects**: its level sets,

  $$\{\ell^{-1}(\{b\}) : \ b \in \mathbb{R}\}.$$

  A linear functional ℓ is a rule assigning a scalar value to each vector. As an object, ℓ lives in the dual space $(\mathbb{R}^n)^*$, not in $\mathbb{R}^n$ itself. So it is not a subset of space, and hence not a geometric object in the sense of a set of points.

  ℓ determines all level sets at once,

  $$\mathbb{R}^n = \bigsqcup_{b \in \mathbb{R}} \ell^{-1}(\{b\}).$$

  Every vector lies in exactly one level set of ℓ, different level sets do not intersect. Together, they partition the space.

- **Family of level sets versus the graph**: Consider a linear functional

  $$\ell : \mathbb{R}^3 \to \mathbb{R}, \quad \ell(x) = a_1x_1 + a_2x_2 + a_3x_3.$$

  Then, ℓ describes the family of level sets

  $$\mathcal{L} = \{\ell^{-1}(\{d\}) : \ d \in \mathbb{R}\}.$$

  Each level set in $\mathcal{L}$ is a 2-dimensional affine hyperplane in $\mathbb{R}^3$. This family of parallel planes partition $\mathbb{R}^3$.

  $$\mathbb{R}^3 = \bigsqcup_{d \in \mathbb{R}} \ell^{-1}(\{d\}).$$

  We can rename ℓ to

  $$\ell(x, y, z) = ax + by + cz.$$

  Then, the graph of ℓ, $\mathcal{G}(\ell)$ is

  $$\mathcal{G}(\ell) = \{(x, y, z, d) \in \mathbb{R}^4 : \ d = \ell(x, y, z)\}.$$

  This is a 3-dimensional affine subspace of $\mathbb{R}^4$. Although these two objects look the same, there is difference in what they mean.

A level set is defined by fixing an output values $d \in \mathbb{R}$ and collecting all inputs that map to it,

$$\ell^{-1}(\{d\}) = \{(x, y, z) \in \mathbb{R}^3 : \ ax + by + cz = d\}.$$

The inputs of ℓ are points in $\mathbb{R}^3$, a level set is a subset of the domain. Therefore, every level set is a subset of $\mathbb{R}^3$. The family is simply a collection of subsets of $\mathbb{R}^3$. Nothing leaves $\mathbb{R}^3$, we are simply slicing it.

The graph of ℓ is defined as

$$\mathcal{G}(\ell) = \{(x, y, z, d) \in \mathbb{R}^3 \times \mathbb{R} : \ d = \ell(x, y, z)\}.$$

A graph records both input and output simultaneously. Inputs live in $\mathbb{R}^3$, outputs live in $\mathbb{R}$. Therefore, ordered pairs (input, output) live in

$$\mathbb{R}^3 \times \mathbb{R} \cong \mathbb{R}^4.$$

So, the graph must live in $\mathbb{R}^4$, because it is a set of input-output pairs.

The graph is defined by a single linear equation in four variables,

$$ax + by + cz - d = 0.$$

Therefore, the solution set has dimension $4 - 1 = 3$. So, $\mathcal{G}(\ell)$ is a 3-dimensional affine subspace of $\mathbb{R}^4$.

So, the graph is the disjoint union of all level sets, indexed by d, but lifted into one higher dimension. Formally,

$$\mathcal{G}(\ell) \cong \bigsqcup_{d \in \mathbb{R}} \left(\ell^{-1}(\{d\}) \times \{d\}\right).$$

We say that the graph is the **total space**, and the level sets are its **fibers**.

7.4 Geometry of systems of linear equations

- **Linear systems**: Consider a system

$$Ax = b,$$

where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$. Each row of $Ax = b$ represents a hyperplane in $\mathbb{R}^n$. The solution set is the intersection of these hyperplanes.

Three fundamental cases arise:

1. **Unique solution**: The hyperplanes intersect at exactly one point.
2. **Infinitely many solutions**: The hyperplanes intersect along a line, plane, or higher-dimensional affine subspace.
3. **No solution**: The hyperplanes have no common intersection (they are inconsistent).

- **Homogeneous systems and the null space**: A homogeneous system has the form

$$Ax = 0.$$

Its solution set is always a linear subspace of $\mathbb{R}^n$, called the null space of A.

- The zero vector is always a solution
- If x and y are solutions, then so is any linear combination $\alpha x + \beta y$
- The dimension of the solution space equals $n - \text{rank}(A)$

Geometrically, solutions form a hyperplane through the origin.

- **Nonhomogeneous systems and translation**: For a nonhomogeneous system $Ax = b$, assume at least one solution x_0 exists. Then, every solution can be written as

$$x = x_0 + v, \quad \text{with } v \in \text{null}(A).$$

Proof. Assume that $Ax = b$ is a nonhomogeneous system with at least one solution x_0. Suppose that x is any other solution, so $Ax = b$. If we subtract these two equations,

$$Ax - Ax_0 = A(x - x_0) = b - b = 0.$$

Thus, $x - x_0 \in \ker(A)$. Call this vector v. So, $v = x - x_0 \in \ker(A)$. But, this implies that

$$x = x_0 + v, \quad v \in \ker(A).$$

This shows that **every** solution differs from x_0 by a null-space vector. ■

Thus, the solution set is an **affine subspace**, a translation of the null space by a vector x_0. We say that the solution set $\mathcal{S} := \{x \in \mathbb{R}^n : \ Ax = b\}$ is

$$\mathcal{S} = \{x_0 + v : \ v \in \ker(A)\} = x_0 + \ker(A)$$

for a fixed solution x_0. So, the dimension of the solution space is given by the dimension of the kernel, which is given by

$$\ker(A) = n - \text{rank}(A).$$

- **Dimension of the solution set**: So, either the system is homogeneous or nonhomogeneous. In any case, the dimension of the solution set $\mathcal{S}$ is given by the dimension of the kernel. Thus,
  - $\dim(\ker(A)) = 0$ implies $\mathcal{S}$ is a single point.
  - $\dim(\ker(A)) = 1$ implies $\mathcal{S}$ forms a 1-dimensional line
  - $\dim(\ker(A)) = 2$ implies $\mathcal{S}$ forms a 2-dimensional plane
  - $\dim(\ker(A)) = k$ implies $\mathcal{S}$ forms a k-dimensional subspace of $\mathbb{R}^n$

  Since the domain is $\mathbb{R}^n$, the ambient space is $\mathbb{R}^n$, and the solution set is therefore embedded in $\mathbb{R}^n$.

7.4.1 Eigenvalues and eigenvectors

- **Geometry of a linear transformation**: Let $A \in \mathbb{R}^{n \times n}$, so

$$A : \mathbb{R}^n \to \mathbb{R}^n.$$

  Geometrically, A can

  - stretch or compress space,
  - reflect it,
  - shear it,
  - rotate it,
  - or combine several of these effects.

  Most vectors change both direction and length under A. Eigenvectors are the exception.

- **Invariant directions**: A nonzero vector v is an eigenvector of A if

$$Av = \lambda v$$

  for some scalar λ. The vector v lies along a direction that is preserved by the transformation, applying A does not rotate or shear v; it only rescales it.

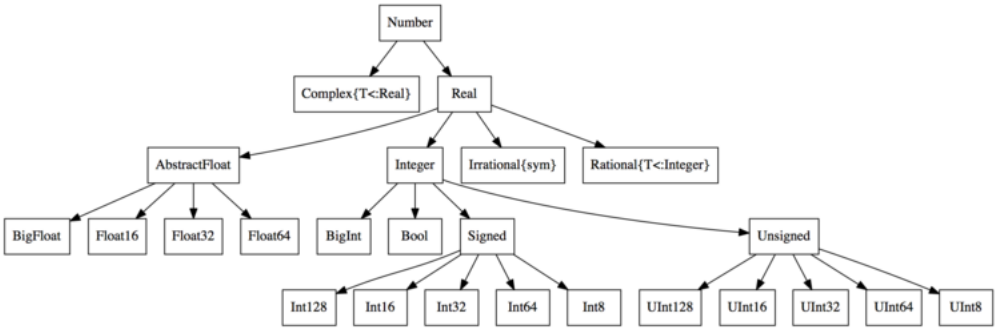
- **Eigenvalues**: The eigenvalue λ tells you how the transformation acts along its eigenvector.

  - $|\lambda| > 1$: stretching along that direction
  - $0 < |\lambda| < 1$: compression along that direction
  - $\lambda = 1$: direction unchanged
  - $\lambda = 0$: collapse onto a lower-dimensional space
  - $\lambda < 0$: reflection plus scaling

  In short, eigenvalues measure the factor by which space is expanded or contracted along the associated eigenvector.

Julia

8.1 Types



- **Subtype constraint <:** $A <: B$ means A is a subtype of B

```
Int <: Number #true
```

8.2 Functions

-

8.3 Linear Algebra

8.3.1 Matrix creation and operations

- **Array constructors**:
  - Array{T}(undef, dims...)
  - Vector{T}(undef, n)
  - Matrix{T}(undef, m, n)
- **Zeros/ones/fills**
  - zeros(n), zeros(m,n)
  - ones(n), ones(m,n)
  - fill(x, dims...)
- **Uniform ranges**:
  - collect(1:n) → vector
  - collect(1:m, 1:n) → matrix (grid)

Derivations

9.1 Series

- **Finite geometric series**: A series with the form

$$\sum_{k=0}^{n} ar^k$$

is called geometric. If we list the terms in this series, we see

$$S_n = a + ar + ar^2 + ar^3 + \ldots + ar^{n-1} + ar^n.$$

If $r \neq 1$, notice that if we multiply this sum by r, we get

$$r_n = ar + ar^2 + ar^3 + ar^4 + \ldots + ar^n + ar^{n+1}.$$

If we subtract rS_n from S_n, all terms will cancel except for a and ar^{n+1}. So,

$$\begin{aligned} S_n - rS_n = S_n(1-r) &= a + ar + ar^2 + ar^3 + \ldots + ar^{n-1} + ar^n \\ &\quad - \left(ar + ar^2 + ar^3 + ar^4 + \ldots + ar^n + ar^{n+1}\right) \\ &= a + ar^{n+1}. \end{aligned}$$

Thus,

$$S_n = \frac{S_n(1-r)}{1-r} = \frac{a + ar^{n+1}}{1-r} = \frac{a(1-r^{n+1})}{1-r}.$$

If the sum has the form $\sum_{k=1}^{n} ar^{k-1}$, notice that it has almost the same terms as $\sum_{k=0}^{n} ar^k$, except it does not have the ar^n term. So,

$$\sum_{k=1}^{n} ar^{k-1} = -ar^n + \sum_{k=0}^{n} ar^k = \sum_{k=0}^{n-1} ar^k = \frac{a(1-r^{(n-1)+1})}{1-r} = \frac{a(1-r^n)}{1-r}.$$

Similarly, if the sum has the form $\sum_{k=1}^{n} ar^k$ it has the same terms as $\sum_{k=0}^{n} ar^k$, except for the initial $ar^0 = a$ term. So,

$$\begin{aligned}\sum_{k=1}^{n} ar^k &= -a + \sum_{k=0}^{n} ar^k = \frac{a(1-r^{n+1})}{1-r} - a = \frac{a - ar^{n+1} - a(1-r)}{1-r} \\ &= \frac{a - ar^{n+1} - a + ar}{1-r} = \frac{ar - ar^{n+1}}{1-r} = \frac{ar(1-r^n)}{1-r}.\end{aligned}$$

In all cases, we assume that $r \neq 1$. If $r = 1$, then

$$\sum_{k=0}^{n} ar^k = \sum_{k=0}^{n} a = a(n+1).$$

If $r = -1$, then we have

$$\sum_{k=0}^{n} a(-1)^k = \frac{a(1-(-1)^{n+1})}{1-(-1)} = \frac{a(1-(-1)^{n+1})}{2}.$$

The sum depends on the parity of n. The number of terms in the sum is $n-0+1 = n+1$. If the sum has an even number of terms, all terms cancel and the sum is zero. If the sum has an add number of terms, all terms will cancel except for one, and so the sum is a.

Therefore, if n is even, then $n+1$ is odd, so the sum is a. If n is odd, then $n+1$ is even, so the sum is zero.

For example, let $r = -1$, and $n = 2$. The sum is

$$\sum_{k=0}^{2} a(-1)^n = a(-1)^0 + a(-1)^1 + a(-1)^2 = a - a + a = a.$$

If we instead let $n = 3$, then the sum is

$$\sum_{k=0}^{3} a(-1)^n = a(-1)^0 + a(-1)^1 + a(-1)^2 + a(-1)^3 = a - a + a - a = 0.$$

- **Infinite geometric series**: Suppose $r \neq 1$. If we examine what happens to $\sum_{k=0}^{n} ar^k$ as $n \to \infty$,

$$\lim_{n\to\infty} \sum_{k=0}^{n} ar^k = \lim_{n\to\infty} \frac{a(1-r^{n+1})}{1-r}.$$

We see that the behavior depends on r^{n+1}. The behavior is

$$\lim_{n\to\infty} r^{n+1} = \begin{cases} 0 & \text{if } |r| < 1 \\ \pm\infty & \text{if } |r| > 1 \end{cases}.$$

So,

$$\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r} \quad \text{if } |r| < 1.$$

If $r = 1$, then

$$\sum_{k=0}^{\infty} a(1)^k = \sum_{k=0}^{\infty} a,$$

which diverges to infinity. If $r = -1$, then

$$\sum_{k=0}^{\infty}.$$

9.2 Quantities

- **Time taken to travel**: Suppose you are traveling k miles at ℓ miles per hour (mph). The time taken in minutes is given by

$$M = \frac{1}{\ell}\ \frac{hr}{m} \cdot k\ m = \frac{k}{\ell}\ hr \cdot \frac{60\ \text{min}}{1\ hr} = \frac{60k}{\ell}\ \text{min}.$$

  Notice that if we set $f_k(\ell) = \frac{1}{\ell}(60k)$ be the number of minutes it takes to travel k miles at ℓ miles per hour, we see there is an inverse relationship between the speed traveled and the time it takes. In order to half the time it takes to travel k miles, we need to double the speed.

  By looking at a graph of $f_k(\ell)$, we notice that the time saved by traveling faster approaches 0 as $\ell \to \infty$.

  The derivative is

$$f_k'(\ell) = -\frac{1}{\ell^2}(60k).$$

  Notice

  - $f_k'(\ell) < 0$ for all ℓ. So, increasing speed always decreases time.
  - $|f_k'(ell)|$ decreases as ℓ increases because of the ℓ^2 term. This means each additional unit of speed saves less and less time that the previous one.

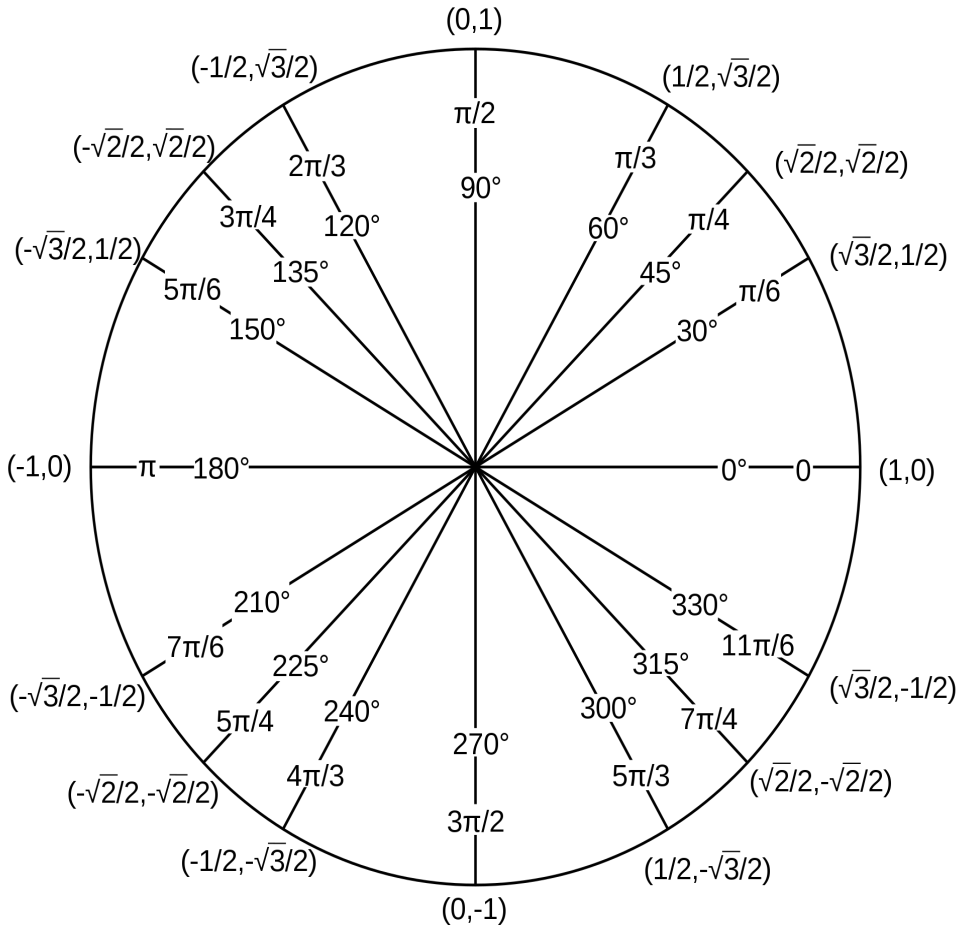
    For example, increasing speed from 5 mph to 10 mph cuts time in half. But increasing from 100 mph to 105 mph barely changes anything.

Ordinary differential equations

10.1 Review

10.1.1 Trig

- **Unit circle**



- **Periodicity**:
  - **Sin, cos, csc, sec**: Period 2π
  - **Tan, cot**: Period π
- **Even / odd trig functions**:
  - **Sin, tan, csc, cot**: Even
  - **Cos, sec**: Odd
- **Transformations of trig functions**: A trig function can be transformed by four things

1. **Amplitude** A: The amplitude is the maximum distance the graph moves above or below its midline. Controls vertical stretch or compression.

   If $A < 0$, the graph is reflected across the midline
2. **Period** $T = \frac{2\pi}{\omega}$: The period is the horizontal length of one full cycle. Controls horizontal stretching and compression
3. **Phase (horizontal) shift** c: The phase shift moves the graph left or right. The phase shift is given by $\frac{c}{\omega}$
4. **Vertical shift** d: Moves the entire graph up or down. The new midline becomes

$$y = d.$$

A transformed trig function is of the form (using sin as an example)

$$A\sin(\omega x - c) + d.$$

- **Pythagorean identities**

$$\begin{aligned}\sin^2(\theta) + \cos^2(\theta) &= 1,\\ \tan^2(\theta) + 1 &= \sec^2(\theta),\\ \sec^2(\theta) - 1 &= \tan^2(\theta).\end{aligned}$$

- **Product to sum**:

$$\begin{aligned}\sin A \sin B &= \frac{1}{2}\left[\cos(A-B) - \cos(A+B)\right],\\ \cos A \cos B &= \frac{1}{2}\left[\cos(A-B) + \cos(A+B)\right],\\ \sin A \cos B &= \frac{1}{2}\left[\sin(A+B) + \sin(A-B)\right],\\ \cos A \sin B &= \frac{1}{2}\left[\sin(A+B) - \sin(A-B)\right].\end{aligned}$$

- **Sum to product**:

$$\begin{aligned}\sin A + \sin B &= 2\sin\left(\frac{A+B}{2}\right)\cos\left(\frac{A-B}{2}\right),\\ \sin A - \sin B &= 2\cos\left(\frac{A+B}{2}\right)\sin\left(\frac{A-B}{2}\right),\\ \cos A + \cos B &= 2\cos\left(\frac{A+B}{2}\right)\cos\left(\frac{A-B}{2}\right),\\ \cos A - \cos B &= -2\sin\left(\frac{A+B}{2}\right)\sin\left(\frac{A-B}{2}\right).\end{aligned}$$

- **Sum and difference**:

$$\begin{aligned}\sin(A \pm B) &= \sin A \cos B \pm \cos A \sin B,\\ \cos(A \pm B) &= \cos A \cos B \mp \sin A \sin B,\\ \tan(A \pm B) &= \frac{\tan A \pm \tan B}{1 \mp \tan A \tan B}.\end{aligned}$$

- **Double angle**:

$$\begin{aligned}\sin(2x) &= 2\sin x\cos x,\\ \cos(2x) &= \cos^2 x - \sin^2 x,\\ \cos(2x) &= 2\cos^2 x - 1,\\ \cos(2x) &= 1 - 2\sin^2 x,\\ \tan(2x) &= \frac{2\tan x}{1-\tan^2 x}.\end{aligned}$$

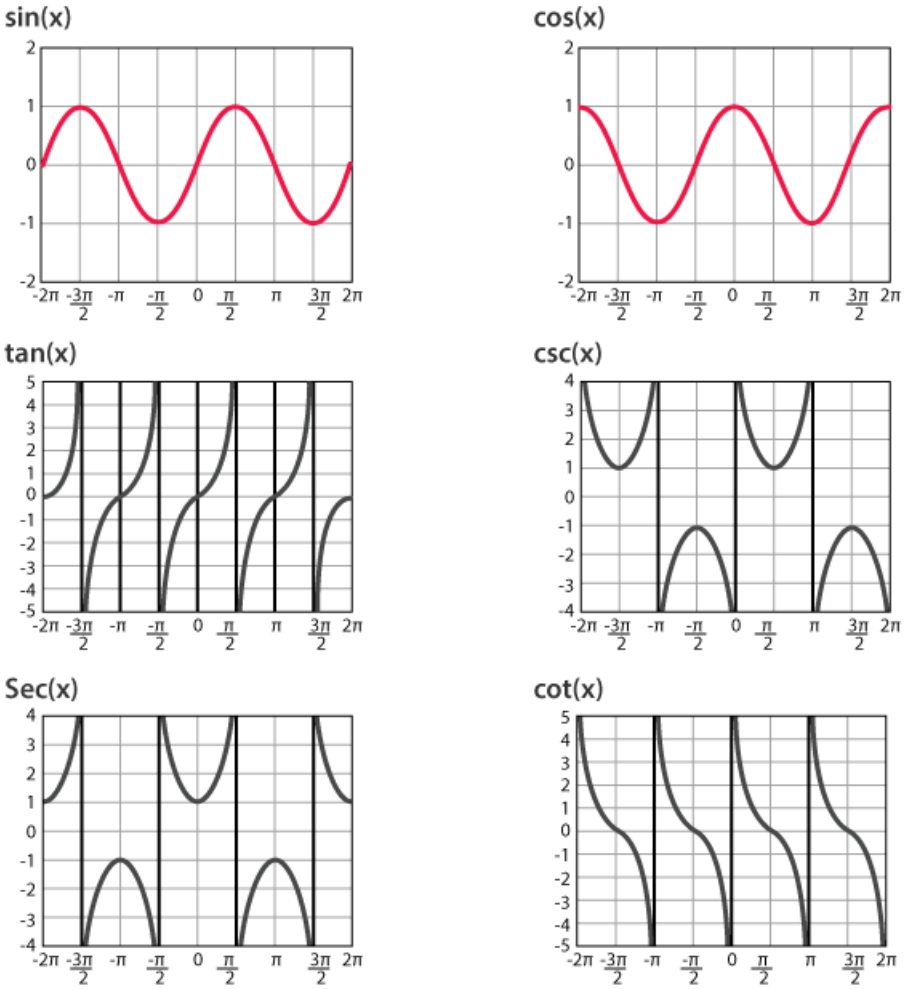
- **Half angle**:

$$\begin{aligned}\sin^2\left(\frac{x}{2}\right) &= \frac{1-\cos x}{2},\\ \cos^2\left(\frac{x}{2}\right) &= \frac{1+\cos x}{2},\\ \tan\left(\frac{x}{2}\right) &= \pm\sqrt{\frac{1-\cos x}{1+\cos x}},\\ \tan\left(\frac{x}{2}\right) &= \frac{\sin x}{1+\cos x} = \frac{1-\cos x}{\sin x}.\end{aligned}$$

- **Domain and range of trig functions**:

Function	**Domain**	**Range**
$y = \sin(x)$	$(-\infty, \infty)$	$[-1, 1]$
$y = \cos(x)$	$(-\infty, \infty)$	$[-1, 1]$
$y = \tan(x)$	$(-\infty, \infty) \setminus \left\{(2k+1)\frac{\pi}{2} \mid k \in \mathbb{Z}\right\}$	$(-\infty, \infty)$
$y = \csc(x)$	$(-\infty, \infty) \setminus \{k\pi \mid k \in \mathbb{Z}\}$	$(-\infty, -1] \cup [1, \infty)$
$y = \sec(x)$	$(-\infty, \infty) \setminus \left\{(2k+1)\frac{\pi}{2} \mid k \in \mathbb{Z}\right\}$	$(-\infty, -1] \cup [1, \infty)$
$y = \cot(x)$	$(-\infty, \infty) \setminus \{k\pi \mid k \in \mathbb{Z}\}$	$(-\infty, \infty)$

- **Asymptotes of trig functions**: Only Tan, Secant, cosecant and cotangent have Asymptotes, and they occur at:
  - **Tan:** When $\cos\theta = 0$ at $\{(2k+1)\frac{\pi}{2} \mid k \in \mathbb{Z}\}$
  - **Cosecant:** When $\sin\theta = 0$ at $\{\pi k \mid k \in \mathbb{Z}\}$
  - **Secant:** When $\cos\theta = 0$ at $\{(2k+1)\frac{\pi}{2} \mid k \in \mathbb{Z}\}$
  - **Cotangent:** When $\sin\theta = 0$ at $\{k\pi \mid k \in \mathbb{Z}\}$
- **Graphs of trig functions**:



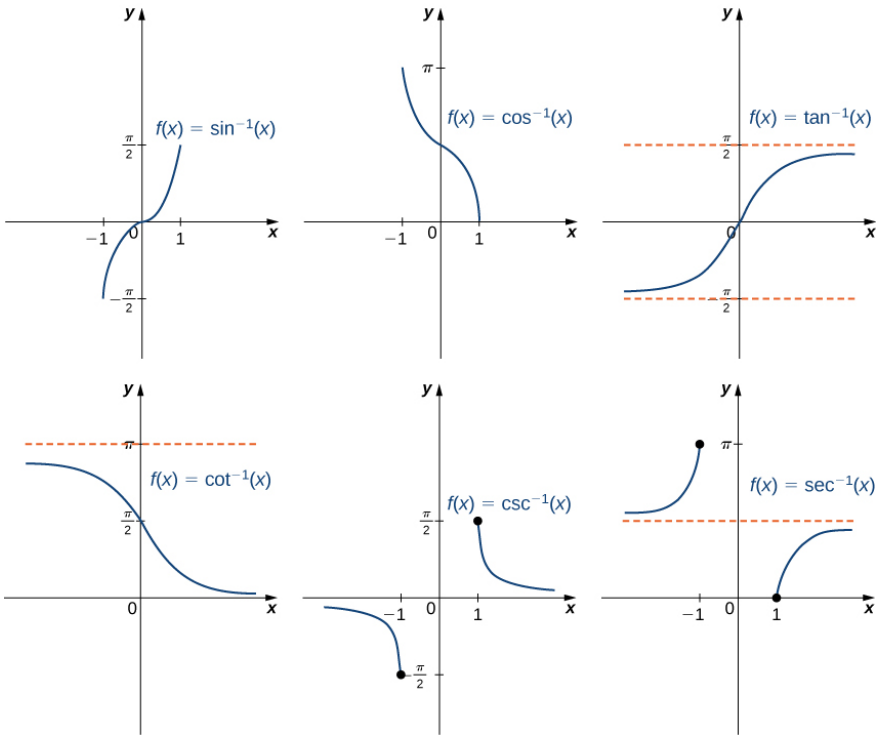
- **Domain and range of inverse trig functions**: Trigonometric functions are not one-to-one over their natural domains, so they cannot be inverted unless we restrict their domains. Trigonometric functions are not one-to-one over their natural domains, so they cannot be inverted unless we restrict their domains.

Function	**Domain**	**Range**
$\arcsin(x) = \sin^{-1}(x)$	$[-1, 1]$	$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$
$\arccos(x) = \cos^{-1}(x)$	$[-1, 1]$	$[0, \pi]$
$\arctan(x) = \tan^{-1}(x)$	$(-\infty, \infty)$	$\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$
$(x) = \sec^{-1}(x)$	$(-\infty, -1] \cup [1, \infty)$	$[0, \pi],\ y \neq \frac{\pi}{2}$
$(x) = \csc^{-1}(x)$	$(-\infty, -1] \cup [1, \infty)$	$\left[-\frac{\pi}{2}, 0\right) \cup \left(0, \frac{\pi}{2}\right]$
$(x) = \cot^{-1}(x)$	$(-\infty, \infty)$	$(0, \pi)$

- **Asymptotes of inverse trig functions**:

Function	Horizontal Asymptotes	Vertical Asymptotes
$\tan^{-1} x$	$y = -\frac{\pi}{2}, \frac{\pi}{2}$	None
$\csc^{-1} x$	$y = 0$	None
$\sec^{-1} x$	$y = \frac{\pi}{2}$	None
$\cot^{-1} x$	$y = 0, \pi$	None

- **Graphs of inverse trig functions**:



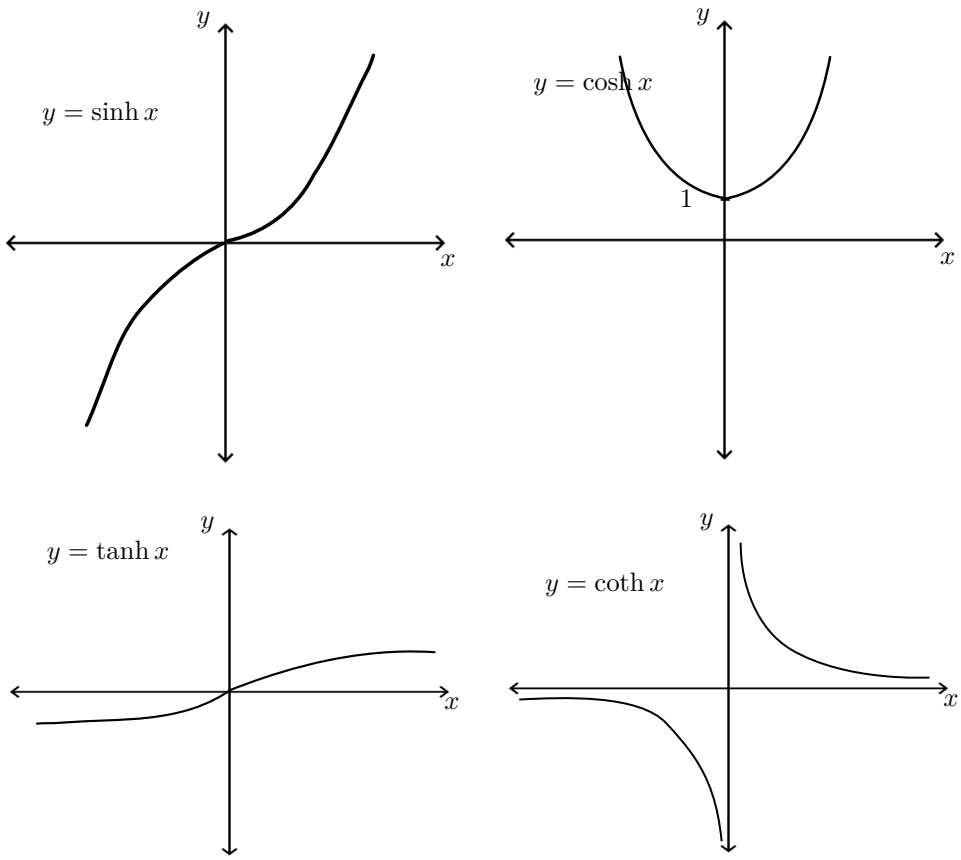
- **Hyperbolic trig**

$$\sinh x = \frac{e^x - e^{-x}}{2}, \quad \cosh x = \frac{e^x + e^{-x}}{2}, \tag{2}$$

$$\tanh(x) = \frac{\sinh x}{\cosh x}, \quad \operatorname{csch}(x) = \frac{1}{\sinh x}, \tag{3}$$

$$\operatorname{sech}(x) = \frac{1}{\cosh x}, \quad \coth(x) = \frac{\cosh x}{\sinh x}. \tag{4}$$

- **Graphs of hyperbolic trig functions**:



10.1.2 Calculus I

- **Limits**: A limit describes the value that a function approaches as the input approaches a certain number.

$$\lim_{x \to a} f(x) = L.$$

As x gets close to a, $f(x)$ gets close to L.

- **One sided limits**: The limit

$$\lim_{x \to a^-} f(x)$$

is the value of the function that is approached as x approaches a from the left. Similarly,

$$\lim_{x \to a^+} f(x)$$

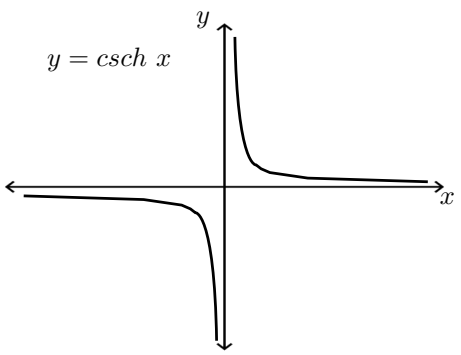
is the value of the function that is approached as x approaches a from the right

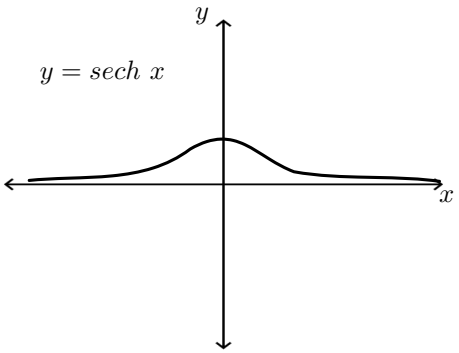
The limit exists if and only if

$$\lim_{x \to a^-} f(x) = \lim_{x \to a^+} f(x).$$

- **Limit laws**: Let

$$\lim_{x \to a} f(x) = L, \quad \lim_{x \to a} g(x) = M.$$





Then,

$$\lim_{x\to a} c = c, \quad \lim_{x\to a} (f(x) + g(x)) = L + M,$$
$$\lim_{x\to a} cf(x) = cL, \quad \lim_{x\to a} f(x)g(x) = LM,$$
$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{L}{M}, \quad M \neq 0,$$
$$\lim_{x\to a} (f(x))^n = L^n, \quad \lim_{x\to a} \sqrt[n]{f(x)} = \sqrt[n]{L},$$
$$\lim_{x\to a} x = x, \quad \lim_{x\to a} x^n = a^n.$$

- **Horizontal asymptotes**: If either $\lim_{x\to\infty} f(x) = L$ or $\lim_{x\to-\infty} f(x) = L$, there is a horizontal asymptote at $y = L$
- **Continuity**: Let $f(x)$. For f to be continuous at a, the following must be true
  1. $f(x)$ is defined at a
  2. $\lim_{x\to a} f(x)$ exists
  3. $\lim_{x\to a} f(x) = f(a)$

  Item three is a necessary and sufficient condition for f to be continuous at a.
- **One sided continuity**: Continuity from the right implies

$$\lim_{x\to a^+} f(x) = f(a).$$

  Similarly, continuity from the left implies that

$$\lim_{x\to a^-} f(x) = f(a).$$

- **Properties of continuity**: If f and g are continuous at a, then $f + g$, $f - g$, fg, $\frac{f}{g}$ for $g \neq 0$, cf, and cg are all continuous at a.
- **Differentiability implies continuity**: If $f(x)$ is differentiable on its domain, then it is continuous on its domain.
- **Basic facts**:
  - **Power Rule**

$$\frac{d}{dx}x^n = nx^{n-1}$$

- **Exponential Functions**

$$\frac{d}{dx}e^x = e^x,$$
$$\frac{d}{dx}a^x = a^x \ln a.$$

- **Logarithmic Functions**

$$\frac{d}{dx}\ln x = \frac{1}{x},$$
$$\frac{d}{dx}\ln |x| = \frac{1}{x}.$$

- **Trigonometric Functions**

$$\frac{d}{dx}\sin x = \cos x,$$
$$\frac{d}{dx}\cos x = -\sin x,$$
$$\frac{d}{dx}\tan x = \sec^2 x,$$
$$\frac{d}{dx}\sec x = \sec x \tan x,$$
$$\frac{d}{dx}\csc x = -\csc x \cot x,$$
$$\frac{d}{dx}\cot x = -\csc^2 x.$$

- **Chain Rule**

$$\frac{d}{dx}f(g(x)) = f'(g(x))\,g'(x)$$

- **Product Rule**

$$\frac{d}{dx}(uv) = u'v + uv'$$

- **Quotient Rule**

$$\frac{d}{dx}\left(\frac{u}{v}\right) = \frac{u'v - uv'}{v^2}$$

- **Indefinite integral identities**:
  - **Power Rule**

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C, \quad n \neq -1$$

  - **Logarithmic Integral**

$$\int \frac{1}{x}\,dx = \ln |x| + C$$

  - **Exponential Functions**

$$\int e^x\,dx = e^x + C,$$
$$\int a^x\,dx = \frac{a^x}{\ln a} + C.$$

– **Trigonometric Functions**

$$\int \sin x\,dx = -\cos x + C,$$
$$\int \cos x\,dx = \sin x + C,$$
$$\int \tan x\,dx = \ln|\sec x| + C,$$
$$\int \sec^2 x\,dx = \tan x + C,$$
$$\int \csc^2 x\,dx = -\cot x + C,$$
$$\int \sec x \tan x\,dx = \sec x + C,$$
$$\int \csc x \cot x\,dx = -\csc x + C.$$

– **Inverse Trigonometric Forms**

$$\int \frac{1}{1+x^2}\,dx = \arctan x + C,$$
$$\int \frac{1}{\sqrt{1-x^2}}\,dx = \arcsin x + C.$$

- **FTC**:

$$\frac{d}{dx}\left(\int_a^x f(t)\,dt\right) = f(x),$$
$$\int_a^b f(x)\,dx = F(b) - F(a).$$

- **Chain rule in detail**: The chain rule is the rule for differentiating a composition of functions. If a quantity changes because it depends on another quantity, which itself depends on something else, then the overall rate of change is the product of the intermediate rates of change.

  Formally, if

$$y = f(u) \quad \text{and} \quad u = g(x),$$

  then y is a function of x via the composition $y = f(g(x))$, and

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}.$$

  Equivalently,

$$(f \circ g)'(x) = f'(g(x)) \cdot g'(x).$$

  In $f(g(x))$, $g(x)$ is a function of x, and f depends on this function. Therefore, f is indirectly a function of x. f is **not directly** a function of x. But, when you compose it with $g(x)$, the output of f does become a function of x. This distinction is why the chain rule exists.

Suppose that

$$f(u) = u^3, \quad g(x) = x^2.$$

f depends of u, while g depends on x. At this stage, f has nothing to do with x. If we compose the two functions,

$$h(x) = f(g(x)) = f(x^2) = (x^2)^3 = x^6.$$

At this point, f is being fed a quantity that depends on x. So, the output of f now depends on x. A function does not need to explicitly mention x to depend on x. It only needs to depend on something that depends on x.

When we differentiate

$$\frac{d}{dx} f(g(x))$$

we are asking "how fast is the output of f changing as x changes". But, the change happens in two stages,

1. x changes affects $g(x)$
2. $g(x)$ changes affects $f(g(x))$

So, the total change must account for both steps. This is why

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

Consider $f(u) = u^3$, $g(x) = 2x + 1$. The composition is then

$$f(g(x)) = (2x + 1)^3.$$

So, the derivative of f with respect to g is

$$\frac{df}{dg} = 3g(x)^2 = 3(2x + 1)^2,$$

while the derivative of g with respect to x is

$$\frac{dg}{dx} = 2.$$

Thus,

$$(f(g(x)))' = \frac{df}{dg} \cdot \frac{dg}{dx} = 3(2x + 1)^2 \cdot 2 = 6(2x + 1)^2.$$

Thus, for

$$\frac{d}{dx} y^3$$

if we let $f(u) = u^3$, and $y = g(x)$, then

$$f(g(x)) = f(y) = y^3.$$

Since

$$\frac{df}{dg} = 3g(x)^2, \quad \frac{dg}{dx} = \frac{dy}{dx},$$

$$(f(g(x)))' = 3g(x)^2 \cdot \frac{dy}{dx} = 3y^2 \cdot \frac{dy}{dx}.$$

Note: Another way to think about why $\frac{d}{dx}y^3 = 3y^2$ is by recalling that $y = f(x)$, so

$$\frac{d}{dx}y^2 = \frac{d}{dx}(f(x))^2 = 2(f(x)) \cdot f'(x) = 2yy'$$

by the chain rule.

- **Implicit differentiation**: Implicit differentiation is a technique used to differentiate equations in which the dependent variable y is not isolated as a function of the independent variable x. Instead of having $y = f(x)$, we have

$$F(x, y) = 0.$$

So, we differentiate both sides with respect to x. The fundamental idea is that whenever you differentiate a term containing y, multiply by $\frac{dy}{dx}$. This fact follows from the chain rule, since $y = y(x)$.

Suppose that we want to differentiate y^3 with respect to x. So, we want to find

$$\frac{d}{dx}y^3.$$

If we let $f(u) = u^3$, and $u = y(x)$, then

$$f(y(x)) = y^3,$$

with

$$\frac{d}{dx}f(y(x)) = \frac{df}{du} \cdot \frac{dy}{dx}.$$

Thus,

$$\frac{d}{dx}f(y(x)) = 3u^2 \cdot \frac{dy}{dx}.$$

But, notice that we defined $u = y(x)$. So,

$$3u^2 \cdot \frac{dy}{dx} = 3y^2 \cdot \frac{dy}{dx}.$$

Or, more conventionally, we would set $f(u) = u^3$, and $y = g(x)$. Then,

$$f(g(x)) = f(y) = y^3,$$

and

$$\begin{aligned}\frac{d}{dx}y^3 &= (f(g(x)))' = f'(g(x)) \cdot g'(x) \\ &= 3u^2 \cdot \frac{dy}{dx} = 3y^2 \cdot \frac{dy}{dx}.\end{aligned}$$

Notice that we changed $3u^2$ to $3y^2$, since what we really have is $f'(g(x))$, and $g(x) = y$.

Now, as an example, we will implicitly differentiate

$$x^3y^3 = x^3 + 1.$$

So,

$$\begin{aligned} \frac{d}{dx}x^3y^3 &= \frac{d}{dx}x^3 + 1 \\ \implies 3x^2y^3 + 3x^3y^2 \cdot \frac{dy}{dx} &= 3x^2. \end{aligned}$$

Now, we solve for $\frac{dy}{dx}$. So,

$$\begin{aligned} 3x^2y^3 + 3x^3y^2 \cdot \frac{dy}{dx} &= 3x^2 \\ \implies y^3 + xy^2\frac{dy}{dx} &= 1 \\ \implies \frac{dy}{dx} &= \frac{1-y^3}{xy^2}. \end{aligned}$$

- **Implicit differentiation with logarithms**: We can also implicitly differentiate using logarithms and their properties. Consider again

$$x^3y^3 = x^3 + 1.$$

So, we can take the natural log of both sides,

$$\begin{aligned} \ln\left(x^3y^3\right) &= \ln\left(x^3+1\right) \\ \implies 3\ln(x) + 3\ln(y) &= \ln\left(x^3+1\right). \end{aligned}$$

Now, we differentiate,

$$\begin{aligned} \frac{d}{dx}\left(3\ln(x) + 3\ln(y)\right) &= \frac{d}{dx}\ln\left(x^3+1\right) \\ \implies \frac{3}{x} + \frac{3}{y}\cdot\frac{dy}{dx} &= \frac{1}{x^3+1}. \end{aligned}$$

Solving for $\frac{dy}{dx}$ gives

$$\frac{dy}{dx} = y\left(\frac{x^2}{x^3+1} - \frac{1}{x}\right).$$

If we choose, we can solve for y in $x^3y^3 = x^3+1$ and substitute it into the form derived above.

- $\frac{dy}{dx}$ **and differentials**: Formally, the derivative $\frac{dy}{dx}$ is not a ratio, it is a limit

$$\frac{dy}{dx} = \lim_{\Delta x \to 0}\frac{\Delta y}{\Delta x}.$$

$\frac{dy}{dx}$ is a single object. You cannot cancel or rearrange it algebraically at this level.

Once a function is differentiable, we define

$$dy = f'(x)\,dx.$$

This is not a limit, but a linear approximation

$$\Delta y \approx f'(x)\Delta x.$$

At this point:

- dx is treated as an independent variable
- dy is defined in terms of it
- The notation behaves algebraically

Then, algebraically,

$$\frac{dy}{dx} = f'(x).$$

This is why we can treat $\frac{dy}{dx}$ as a fraction.

10.1.3 Calculus II

- **Integration by Power rule**: This is for handling power functions: For any real number n (whether whole number, negative number, rational or irrational number.)

$$\int x^n \, dx = \begin{cases} \frac{x^{n+1}}{n+1} + C & \text{if } n \neq -1 \\ \ln(x) + C & \text{if } n = -1 \end{cases}.$$

- **Integration by u sub**: Integration by substitution (commonly called u-substitution) is a technique used to simplify integrals by reversing the chain rule from differential calculus. The core idea is to transform a complicated integral into a simpler one by changing variables.

  Consider an integral of the form

$$\int f(g(x))g'(x) \, dx.$$

  We let $u = g(x)$, so $du = g'(x) \, dx$. Then, the integral becomes

$$\int f(u) \, du.$$

  After integrating, substitute back $u = g(x)$. For example, consider the integral

$$\int 2x \cos\left(x^2\right) \, dx.$$

  So, we see that if $u = g(x) = x^2$, then $du = 2x \, dx$, and

$$\int 2x \cos\left(x^2\right) \, dx = \int \cos(u) \, du = \sin(u) + C = \sin\left(x^2\right) + C.$$

- **Integration by parts**: Integration by parts is a technique used to evaluate integrals that involve the product of two functions, where direct integration is not convenient. It is derived directly from the product rule for differentiation.

  The product rule states that

$$\frac{d}{dx}(u(x)v(x)) = u'(x)v(x) + u(x)v'(x).$$

  If we integrate both sides,

$$\int \frac{d}{dx}u(x)v(x) = uv = \int u'(x)v(x) \, dx + \int u(x)v'(x) \, dx$$

  .

  Now, notice that $du = u'(x) \, dx$, and $dv = v'(x) \, dx$. So,

$$\begin{aligned} uv &= \int u'(x)v(x) \, dx + \int u(x)v'(x) \, dx \\ &= \int v \, du + \int u \, dv. \end{aligned}$$

  Thus,

$$\int u \, dv = uv - \int v \, du.$$

So, the integrand in question is precisely

$$\int u(x)v'(x)\,dx.$$

The choice in deciding what should be u, and what should be dv used the LIATE rule, which ranks functions by how useful it is to differentiate them.

- **L:** Logarithmic
- **I:** Inverse trig
- **A:** Algebraic
- **T:** Trigonometric
- **E:** Exponential

Choose u to be the function that appears earliest in the list

For example, consider

$$\int xe^x\,dx.$$

Notice that x is algebraic, and e^x is an exponential. Since A comes before E, we choose x to be u, and $e^x\,dx$ to be dv.

With $u = x$, and $dv = e^x\,dx$, we have

$$du = \,dx, \quad v = \int dv = \int e^x\,dx = e^x.$$

Thus,

$$\int u\,dv = xe^x - \int e^x\,dx = xe^x - e^x + C.$$

Notice when finding v from dv, we do not include a constant of integration. The constant of integration belongs to the entire integral, not to intermediate steps.

Note that for definite integrals,

$$\int_a^b u\,dv = uv\,|_a^b - \int_a^b v\,du.$$

- **Integrals of products of trig functions**:

$$\begin{aligned}
\sin(ax)\sin(bx) &= \frac{1}{2}\cos((a-b)x) - \frac{1}{2}\cos((a+b)x),\\
\sin(ax)\cos(bx) &= \frac{1}{2}\sin((a-b)x) + \frac{1}{2}\sin((a+b)x),\\
\cos(ax)\cos(bx) &= \frac{1}{2}\cos((a-b)x) + \frac{1}{2}\cos((a+b)x)
\end{aligned}$$

- **Power reduction formulas**:

$$\int \sin^n x\,dx = -\frac{1}{n}\sin^{n-1}x\cos x + \frac{n-1}{n}\int \sin^{n-2}x\,dx$$

$$\int_0^{\frac{\pi}{2}} \sin^n x\,dx = \frac{n-1}{n}\int_0^{\frac{\pi}{2}} \sin^{n-2}x\,dx.$$

$$\int \cos^n x \; dx = \frac{1}{n} \cos^{n-1} x \sin x + \frac{n-1}{n} \int \cos^{n-2} x \; dx$$

$$\int_0^{\frac{\pi}{2}} \cos^n x \; dx = \frac{n-1}{n} \int_0^{\frac{\pi}{2}} \cos^{n-2} x \; dx.$$

$$\int \sec^n x \; dx = \frac{1}{n-1} \sec^{n-1} x \sin x + \frac{n-2}{n-1} \int \sec^{n-2} x \; dx$$

$$\int \sec^n x \; dx = \frac{1}{n-1} \sec^{n-2} x \tan x + \frac{n-2}{n-1} \int \sec^{n-2} x \; dx$$

$$\int \tan^n x \, dx = \frac{1}{n-1} \tan^{n-1} x - \int \tan^{n-2} x \, dx$$

- **Trigonometric substitution**: Trigonometric substitution is a technique used to evaluate integrals involving expressions of the form

$$\sqrt{a^2 - x^2}, \quad \sqrt{a^2 + x^2}, \quad \sqrt{x^2 - a^2}.$$

The method relies on the Pythagorean identities:

$$\sin^2(\theta) + \cos^2(\theta) = 1,$$
$$1 + \tan^2(\theta) = \sec^2(\theta),$$
$$\sec^2(\theta) - 1 = \tan^2(\theta).$$

If we see $\sqrt{a^2 - x^2}$, use

$$x = a \sin(\theta).$$

Then,

$$dx = a \cos(\theta) \; d\theta.$$

So,

$$\sqrt{a^2 - x^2} = \sqrt{a^2 - (a \sin(\theta))^2} = \sqrt{a^2(1 - \sin^2(\theta))} = \sqrt{a^2 \cos^2(\theta)} = a \cos(\theta).$$

If we see $\sqrt{a^2 + x^2}$, use

$$x = a \tan(\theta).$$

Then,

$$dx = a \sec^2(\theta) \; d\theta.$$

So,

$$\sqrt{a^2 + x^2} = \sqrt{a^2(1 + \tan^2(\theta))} = a \sec(\theta).$$

If we see $\sqrt{x^2 - a^2}$, use

$$x = a \sec(\theta).$$

Then,

$$dx = a \sec(\theta) \tan(\theta) \; d\theta.$$

So,

$$\sqrt{x^2 - a^2} = \sqrt{a^2(\sec^2(\theta) - 1)} = a \tan(\theta).$$

- **Polynomial long division**: Polynomial long division is the algebraic process used to divide one polynomial by another, in the same way that long division is used for numbers. In calculus, it is primarily used to rewrite improper rational functions so that techniques like partial fractions can be applied.

Polynomial division is based on the identity:

$$\frac{P(x)}{Q(x)} = D(x) + \frac{R(x)}{Q(x)}.$$

Where

- $D(x)$ is the quotient,
- $R(x)$ is the remainder, and
- $\deg(R) < \deg(Q)$

- **Integration by partial fractions**: Partial fraction decomposition is a standard technique used primarily to evaluate rational integrals, that is, integrals of the form

$$\int \frac{P(x)}{Q(x)}\, dx$$

where $P(x)$ and $Q(x)$ are polynomials, and

$$\deg(P) < \deg(Q).$$

The core idea is to rewrite a complicated rational expression as a sum of simpler rational terms whose integrals are known. The method relies on a key algebraic fact... Any rational function with a factorable denominator can be expressed as a sum of simpler rational functions. These simpler terms correspond to the factors of the denominator, and each produces an integral you already know how to compute (logarithms or inverse trigonometric functions).

If $\deg(P) \geqslant \deg(Q)$, If this is not true, perform polynomial division first.

To perform partial fraction decomposition, we must ensure that the degree requirement holds. Then, we factor the denominator completely (over the real numbers). The possible factor types are

1. **Distinct linear factors**:

$$(x-a)(x-b).$$

2. **Repeated linear factors**:

$$(x-a)^2, \quad (x-a)^3.$$

3. **Irreducible quadratics**:

$$x^2 + bx + c, \quad (b^2 - 4ac < 0).$$

One we have factored the denominator, we can write in partial fraction form.

- **Distinct linear factors**:

$$\frac{P(x)}{(x-a)(x-b)} = \frac{A}{(x-a)} + \frac{B}{(x-b)}.$$

- **Repeated linear factors**:

$$\frac{P(x)}{(x-a)^n} = \frac{A_1}{(x-a)} + \frac{A_2}{(x-a)^2} + \cdots + \frac{A_n}{(x-a)^n}.$$

– **Irreducible quadratic**:

$$\frac{P(x)}{ax^2+bx+c} = \frac{Ax+B}{ax^2+bx+c}.$$

If repeated,

$$\frac{P(x)}{(ax^2+bx+c)^2} = \frac{Ax+B}{ax^2+bx+c} + \frac{Cx+D}{(ax^2+bx+c)^2}.$$

10.1.4 Calculus III

- **Multivariable chain rule for one independent variable (t)**: For a differentiable function $z(x, y)$ of x and y, where $x = g(t)$, $y = h(t)$ are differentiable functions of t, then $z(x(t), y(t)$ is a differentiable function of t, and

$$\frac{dz}{dt} = \frac{\delta z}{\delta x} \cdot \frac{dx}{dt} + \frac{\delta z}{\delta y} \cdot \frac{dy}{dt}.$$

10.2 First order differential equations and mathematical models

10.2.1 First order differential equations

- **Intro to differential equations**: A differential equation is an equation that relates
  - an unknown function, and
  - one or more of its derivatives.

  Instead of solving for a number, we solve for a **function**. For example,

  $$\frac{dy}{dx} = 3x^2.$$

  This equation asks for a function $y(x)$ whose derivative equals $3x^2$.

- **Independent and dependant variables**: This distinction is fundamental.
  - **Independent variable**: The variable you control or choose freely, does not depend on another variable. For example, the time t, position x, or angle θ
  - **Dependent variable**: The variable whose value depends on the independent variable. Usually written as a function, e.g. $y(x)$. For example, position as a function of time $x(t)$.

  If

  $$y = x^2,$$

  x is the independent variable, and y is the dependent variable. Changing x causes y to change. In a differential equation,

  $$\frac{dy}{dx} = 2x,$$

  you are describing how fast the dependent variable changes with respect to the independent variable.

- **What a differential equation represents**: A differential equation describes a relationship between a quantity and how it changes.

  For example,

  $$\frac{dx}{dt} = v$$

  says the rate of change of position with respect to time is velocity, since the unknown function is $x(t)$, which represents position as a function of time, and $x'(t)$ is the rate of change of position with respect to time, which equals velocity v.

- **The order of differential equations**: The order is determined by the highest derivative appearing.
  - **First order**:

    $$\frac{dy}{dx} = x.$$

  - **Second order**:

    $$\frac{d^2y}{dx^2} + y = 0.$$

- **Third order**:

$$\frac{d^3y}{dx^3} = t.$$

Higher-order equations usually require more initial conditions.

- **General vs particular solutions**:
  - **General solutions**: Contains constants

$$y = x^3 + C.$$

  - **Particular solutions**: Uses initial conditions to find constants

$$y(0) = 2 \implies y = x^3 + 2.$$

This reflects the idea that many functions can satisfy the same differential equation, but only one satisfies a given physical situation. A differential equation describes a family of curves, not a single curve. The constant C is determined only if you specify a condition like:

$$y(1) = 3.$$

- **Goals of differential equations**: The study of differential equations has three principal goals
  1. To discover the differential equation that describes a specified physical situation.
  2. To find—either exactly or approximately—the appropriate solution of that equation.
  3. To interpret the solution that is found.
- **Solutions**: In algebra, we typically seek the unknown *numbers* that satisfy an equation such as

$$x^3 + 7x^2 - 11x + 41 = 0.$$

By contrast, in solving a differential equation, we are challenged to find the unknown *functions* $y = y(x)$ for which an identity such as

$$y'(x) = 2xy(x)$$

holds.

Note: The solution to a differential equation is a continuous function, since it is differentiable.

- **Interval of definition**: Corresponding to any solution of a differential equation is its interval of definition; also called interval of existence, interval of validity or domain of the solution. This can be an open, closed, bounded or unbounded interval.

If a differential equation has a solution

$$y = f(x).$$

then the interval of definition is the interval of x-values where

  - $f(x)$ is defined
  - $f(x)$ is differentiable

- The original differential equation makes sense
- No division by zero or undefined expressions occur

Consider the equation

$$\frac{dy}{dx} = \frac{1}{x}.$$

A solution is

$$y = \ln|x| + C.$$

Then, the **interval of definition** is

$$(-\infty, 0) \cup (0, \infty).$$

Because The function is undefined at $x = 0$, and the solution cannot cross $x = 0$

As a general solution, this is acceptable because:

- It represents two possible families of solutions
- One on $(-\infty, 0)$
- One on $(0, \infty)$

At this stage, no single interval has been chosen yet. So,

$$y = \ln|x| + C$$

is a formal solution that represents **both possibilities**. However, a particular solution chooses a specific function form this family. Since a solution to a differential equation must

- Be defined on an interval
- Be continuous on that interval
- Satisfy the differential equation everywhere on that interval

The differential equation

$$\frac{dy}{dx} = \frac{1}{x}$$

is undefined at $x = 0$. That point splits the real line into two disconnected intervals

$$(-\infty, 0) \quad \text{and} \quad (0, \infty).$$

A particular solution **cannot cross this singularity**.

Suppose that

$$\frac{dy}{dx} = \frac{1}{x}, \quad y(1) = 0.$$

Then,

$$y = \ln|x|.$$

Notice that since $y(1) = 0$, the point $(1, 0)$ must live in the domain of the solution. Thus, this fact forces the domain of the solution to be $(0, \infty)$, since it must be one, but not both. It cannot be both because a particular solution cannot

- cross a point where the DE is undefined,
- be discontinuous,
- or "jump" from one interval to another.

Therefore, once the solution is fixed at $x = 1$, it is locked into $(0, \infty)$. In fact, since this must be the domain, the solution becomes simply

$$y = \ln(x).$$

If instead $y(-1) = 0$, then the particular solution is defined at $(-1, 0)$. Thus, the domain is the left side of the singularity $(-\infty, 0)$. In this case,

$$y = \ln(-x).$$

The key takeaway is that a general solution may span multiple disjoint intervals, but a particular solution must live on one continuous interval where the differential equation is defined. This is why the general solution is

$$y = \ln|x| + C,$$

but a particular solution must be either

$$y = \ln(x) \quad \text{or} \quad y = \ln(-x)$$

depending on the initial condition.

Another key point arises when you recall that

$$y = \ln|x|$$

with domain

$$(-\infty, 0) \cup (0, \infty)$$

is in fact differentiable across the entire domain, since the discontinuity at $x = 0$ is not included in the domain. However, In differential equations, **a solution is not just any differentiable function.** A solution must

- Satisfy the DE
- Be differentiable
- be defined on a single interval
- That interval must be connected

This last condition is key.

- **Why must a particular solution to a DE be defined on a single interval**: A differential equation describes how a function changes locally. That means if you know the value of the function at one point, the equation tells you how it behaves near that point. That only makes sense if "near that point" actually exists. That's why a solution must live on one continuous interval.

Consider again the solution

$$y = \ln|x|\,.$$

Its algebraic domain is

$$(-\infty, 0) \cup (0, \infty).$$

This is two separate pieces. Now ask yourself... If I stand at $x = 1$, what does the differential equation tell me about what happens at $x = -1$? The answer is nothing, there is no path from 1 to -1 that stays inside the domain. So the DE gives no relationship at all between the two sides. This means

- The behavior on the left side is completely independent
- The behavior on the right side is completely independent
- They are not part of the same solution

Imagine a particle moving along a line. A differential equation gives its velocity

$$\frac{dy}{dx} = f(x).$$

Now, suppose the road has a cliff at $x = 0$. You can

- Walk on the left side
- Walk on the right side

But, you **cannot** cross the cliff. So, motion on the left and right sides are **two different journeys**, not one. That's exactly what happens with singularities.

The key fact is that a solution to a DE is not just a differentiable function. The DE determines the function everywhere on its domain by *local behavior*. This only works if the domain is connected.

If disconnected domains were allowed, then:

- Solutions wouldn't be unique
- Initial conditions wouldn't determine behavior
- Existence–uniqueness theorems would fail
- Physical interpretations would break
- You could literally "glue together" unrelated functions and call them a solution

This is why mathematics forbids it.

- **Singularities**: A singularity is a point where a mathematical expression or equation breaks down - meaning it is
  - Undefined,
  - Infinite,
  - Or not well behaved

  In differential equations, a singularity is a point where the right-hand side of the equation is not defined or not continuous.

Recall for

$$\frac{dy}{dx} = \frac{1}{x},$$

the singularity is at $x = 0$, since division by zero is undefined. Hence, the slope is not defined at that point. The equation literally stops making sense there.

- **Families of solutions**: A solution of the first order DE $y = f(x, y)$ containing a single constant is called a one parameter family of solutions. An nth order DE has an n-parameter family of solutions.

- **Implicit solution**: Sometimes it may not be conducive or feasible to write the solution explicitly, as in, $y = f(x)$, Thus a relation $G(x, y) = 0$ between x and y is said to be an implicit solution of a DE if it satisfies the DE. That is, the solution can be given as an equation involving x and y instead of a function y of x.

  For example, we can show that the function $x^3y^3 = x^3 + 1$ is an implicit solution to the DE $xy' + y = y^{-2}$

- **Integrating both sides of an equation**: Consider the equation

  $$f(x) = k.$$

  If we integrate both sides, we get

  $$\int f(x)\, dx = \int k\, dx.$$

  We write the integral with respect to x, which is denoted by dx. We are not adding a dx algebraically, we are simply specifying the variable of integration. If our equation is instead.

  $$\frac{dy}{dx} = f(x).$$

  Then, integrating both sides yields

  $$\int \frac{dy}{dx}\, dx = \int f(x)\, dx.$$

  Since derivatives and integrals are inverse operators,

  $$\int \frac{dy}{dx}\, dx = y + C.$$

  Thus,

  $$\int \frac{dy}{dx}\, dx = \int f(x)\, dx \implies y = \int f(x)\, dx = F(x),$$

  where $F'(x) = f(x)$.

- **Higher order derivatives**: Consider a higher order derivative

  $$\frac{d^3y}{dx^3} = f(x).$$

  To undo the third derivative, we must integrate three times. So,

  $$\int\int\int \frac{d^3y}{dx^3}\, dx\, dx\, dx = \int\int\int f(x)\, dx\, dx\, dx$$

.

Note that

$$\int \frac{d^3y}{dx^3}\,dx = \frac{d^2y}{dx^2}.$$

Thus,

$$\int\int\int \frac{d^3y}{dx^3}\,dx\,dx\,dx = y,$$

and

$$\int\int\int f(x)\,dx\,dx\,dx = H(x) + Ax^2 + Bx + C,$$

where we can say

$$F'(x) = f(x),$$
$$G''(x) = f(x),$$
$$H'''(x) = f(x).$$

- **Integrals as solutions**: A derivative tells you how something changes, an integral tells you what the original quantity must have been.

  If a differential equation tells you

  $$\frac{dy}{dx} = f(x),$$

  then an integral gives you the function $y(x)$ whose derivative equals $f(x)$.

  $$y(x) = \int f(x)dx.$$

  This is why integrals are best understood as solutions to differential equations, not just geometric areas.

  For example, consider the statement "the rate of change of a quantity is proportional to time". So,

  $$\frac{dy}{dt} = kt.$$

  To find the quantity itself, we integrate

  $$y(t) = \int kt\,dt = \frac{k}{2}t^2 + C.$$

  The differential equation describes how the system evolves, the integral gives the actual behavior

- **Integrals for higher order differential equations**: Consider the differential equation

  $$\frac{d^3y}{dx^3} = e^{4x}. \tag{5}$$

  So,

  $$y = \int\int\int e^{4x}\,dx\,dx\,dx.$$

  Since $\int e^{rx}\,dx = \frac{1}{r}e^{rx} + C$,

  $$y = \frac{1}{4^3}e^{4x} + Ax^2 + Bx + C.$$

- **Verifying solutions to differential equations**: The solution to a differential equation is any function that satisfies the differential equation on the given interval

Suppose we want to verify that

$$y(x) = \frac{1}{16}x^4$$

is a solution to the differential equation

$$\frac{dy}{dx} = xy^{\frac{1}{2}}$$

over the interval $(-\infty, \infty)$.

So,

$$\frac{dy}{dx} = \frac{1}{4}x^3.$$

Then,

$$\frac{1}{4}x^3 = xy^{\frac{1}{2}} \implies y = \left(\frac{1}{4}x^2\right)^2 = \frac{1}{16}x^4.$$

Thus, $\frac{1}{16}x^4$ is a solution to the differential equation $\frac{dy}{dx} = xy^{\frac{1}{2}}$. Observe that

$$\frac{dy}{dx} = xy^{\frac{1}{2}} = x\left(\frac{1}{16}x^4\right)^{\frac{1}{2}} = x\left(\frac{1}{4}x^2\right) = \frac{1}{4}x^3.$$

Next, suppose that we want to verify that

$$y(x) = 4\cos(2x) + 6\sin(2x)$$

is a solution to the differential equation

$$y'' + 4y = 0$$

on the interval $(-\infty, \infty)$. So, we have that

$$\begin{aligned} y'(x) &= \frac{d}{dx}\left(4\cos(2x) + 6\sin(2x)\right) = -8\sin(2x) + 12\cos(2x), \\ y''(x) &= \frac{d}{dx}\left(-8\sin(2x) + 12\cos(2x)\right) = -16\cos(2x) - 24\sin(2x). \end{aligned}$$

So, since $y'' = -4y$,

$$-16\cos(2x) - 24\sin(2x) = -4y \implies y = 4\cos(2x) + 6\sin(2x).$$

Thus, it is verified. Alternatively, with $y'' + 4y = 0$, we have

$$\begin{aligned} -16\cos(2x) - 24\sin(2x) + 4\left(4\cos(2x) + 6\sin(2x)\right) &= 0, \\ \implies 0 &= 0. \end{aligned}$$

Again, verified.

- **Separable differential equation**: A differential equation is separable if it can be written in the form

$$\frac{dy}{dx} = g(x)h(y).$$

or equivalently,

$$\frac{1}{h(y)}dy = g(x)dx.$$

This structure allows the variables to be separated and integrated independently. Thus,

$$\int \frac{1}{h(y)}\, dy = \int g(x)\, dx \implies F(y) + C_1 = G(x) + C_2,$$
$$\implies F(y) = G(x) + C_2 - C_1 = G(x) + C.$$

For example, consider

$$\frac{dy}{dx} = xy.$$

So,

$$\frac{1}{y}dy = xdx \implies \int \frac{1}{y}\, dy = \int x\, dx$$
$$\implies \ln|y| + C_1 = \frac{1}{2}x^2 + C_2$$
$$\implies e^{\ln|y|} = e^{\frac{1}{2}x^2 + C_2 - C_1}$$
$$\implies e^{\ln|y|} = e^{\frac{1}{2}x^2 + C}$$
$$\implies |y| = e^{\frac{1}{2}x^2} e^C$$
$$\implies y = \pm e^{\frac{1}{2}x^2} e^C = Ce^{\frac{1}{2}x^2}.$$

A differential equation is separable if

1. It is first order
2. Variables can be isolated on opposite sides
3. No mixed terms remain after separation
4. Integration is straightforward

An example of a DE that is **not** separable is

$$\frac{dy}{dx} = x + y.$$

If we try to treat this as separable, we get

$$dy = (x + y)dx.$$

So,

$$\frac{1}{x + y}dy = dx.$$

Since the left side is a function of both x and y, instead of strictly y, we have not separated variables.

- **Dealing with** $\pm$: When solving differential equations, you often reach a point like:

$$|y| = e^{f(x)}.$$

So,

$$y = \pm e^{f(x)}.$$

Rather than writing $\pm$, we write

$$y = Ce^{f(x)},$$

with

$$C \in \mathbb{R} \setminus \{0\}.$$

- **Newton's law of cooling**: Newton's Law of Cooling describes how the temperature of an object changes over time as it exchanges heat with its surrounding environment

  Newton's Law of Cooling states that the rate of change of the temperature of an object is proportional to the difference between the object's temperature and the ambient (surrounding) temperature.

  If the object is hotter than its surroundings, it cools down. If it is colder, it warms up. The greater the temperature difference, the faster the rate of change.

  Let

  - $T(t)$ = temperature of the object at time t
  - T_a = ambient (constant) temperature
  - $k > 0$ = cooling constant (depends on material and environment)

  Then, the law of cooling states

  $$\frac{dT}{dt} = -k(T - T_a).$$

  The negative sign ensures:

  - Cooling when $T > T_a$
  - Heating when $T < T_a$

  Notice that we can use separation to solve this differential equation. We have

  $$\frac{1}{T - T_a} dT = -k dt.$$

  So,

  $$\begin{aligned} &\int \frac{1}{T - T_a}\, dT = -\int k\, dt \\ \implies &\ln|T - T_a| = -kt + C \\ \implies &T - T_a = e^{-kt+C} = e^{-kt}e^C = Ce^{-kt}. \end{aligned}$$

  Thus,

  $$T(t) = T_a + Ce^{-kt}.$$

  This is the general solution to the cooling problem. If we let $T(0) = T_0$ be the initial temperature of the object at time $t = 0$, then

  $$T_0 = T_a + Ce^0 = T_a + C.$$

Thus,

$$C = T_0 - T_a.$$

Plugging this into the general form yields

$$T(t) = T_a + (T_0 - T_a)e^{-kt}.$$

This is the standard form.

10.2.2 Mathematical models with first order differential equations

:

- **Tools we might need**: A quantity A is proportional to another quantity B if the ratio between them is a constant. We write $A \propto B \iff A = kB$ where k is the proportionality constant.

  Next, if A and B are two groups, then the product AB represents the mathematical model of interaction between the two groups

- **Mathematical model**: The mathematical description of a system or phenomenon is called a mathematical model .

  Mathematical modeling involves the following:

  - Translating a real-world problem into mathematics using differential equations;
  - Analysing or solving the resulting differential equation;
  - Interpreting the results in the context of the original situation, answering the question originally posed and checking whether our answer actually makes logical sense

  In mathematical modeling, we make assumptions and consider the simplest cases to get a 'feasible' model that may only work in very rare circumstances; then we seek to improve the model by modifying some of our assumptions (like modifying the constant growth rate of the exponential model to the more realistic one in the logistic model)

  For example, the basic population model (below) states

  $$\frac{dP}{dt} = kP(t).$$

  This model has some drawbacks because it projects exponential growth, which is unrealistic in several scenarios. It claims a population's per capita (per individual) growth rate remains constant irrespective of population size, making the population grow faster as it gets larger.

  Consequently, The quest for a more realistic model led to the logistic model

  $$\frac{dP}{dt} = kP\left(1 - \frac{P}{M}\right).$$

  which proposes that a population's per capita growth rate gets smaller as population size approaches a maximum imposed by limited resources, where M denotes the environmental carrying capacity (the number of individuals the environment can contain).

- **Basic population model**: Thomas Malthus, the English clergyman and economist proposed the earliest mathematical model of population growth. This model claims that the rate of change of a population ($P(t)$) with time is proportional to the existing population.

  Thus, the model is

  $$\frac{dP}{dt} = kP(t).$$

- **Newton's law of cooling and warming**: Let $T(t)$ be the temperature of an object at time t, and let A be the temperature of the surrounding environment (usually a constant). Newton's law of cooling/warming states that the rate at which the temperature of an object changes is proportional to the difference between the temperature of the object and the temperature of the surrounding environment.

Thus, the model is

$$\frac{dT}{dt} = k(T - A).$$

- **A third model example**: Suppose a student carrying a flu virus returns to an isolated college campus of 1000 students. Determine a differential equation for the number of people $x(t)$ who have contracted the flu if the rate at which the disease spreads is proportional to the number of interactions between the number of students who have the flu and the number of students who have not yet been exposed to it.

  The number of people who have contacted the flu is given by $x(t)$, so the number of people who have **not** contacted the flu is given by

  $$1000 - x(t).$$

  Because we are interested in the interaction between these two quantities, the interaction is their product

  $$x(t)(1000 - x(t)).$$

  Thus, the rate at which the disease spreads is the rate at which the number of people who have contacted the flu $x(t)$ increases with respect to time, $\frac{dx}{dt}$, which is proportional to the interaction. Thus, the differential equation is

  $$\frac{dx}{dt} = kx(t)(1000 - x(t)).$$

  Note that $x(0) = 1$. A single student is infected at time $t = 0$.

10.2.3 Slope fields, solution curves, and the existence / uniqueness theorem

- **Slope fields**: A slope field (also called a direction field) is a graphical tool used in differential equations to visualize the behavior of solutions without explicitly solving the equation.

  A slope field corresponds to a first-order differential equation of the form

  $$\frac{dy}{dx} = f(x, y).$$

  At each point (x, y) in the plane,

  - The value $f(x, y)$ gives the slope of the solution curve at that point.
  - A short line segment with that slope is drawn.
  - Collectively, these small segments form the slope field.

  Each segment shows the direction a solution curve would follow if it passed through that point.

  Slope fields are used to:

  - Visualize solutions without solving the differential equation.
  - Understand qualitative behavior (growth, decay, equilibrium).
  - Estimate solutions given an initial condition.
  - Analyze stability and long-term behavior.

  They are especially useful when:

  - The equation cannot be solved analytically.
  - You want geometric intuition before solving.

- **Solution curves**: A solution curve:

  - Is a smooth curve that follows the slope field everywhere.
  - Is tangent to each segment it passes through.
  - Represents a specific solution $y(x)$.

  An initial condition, such as:

  $$y(0) = 2$$

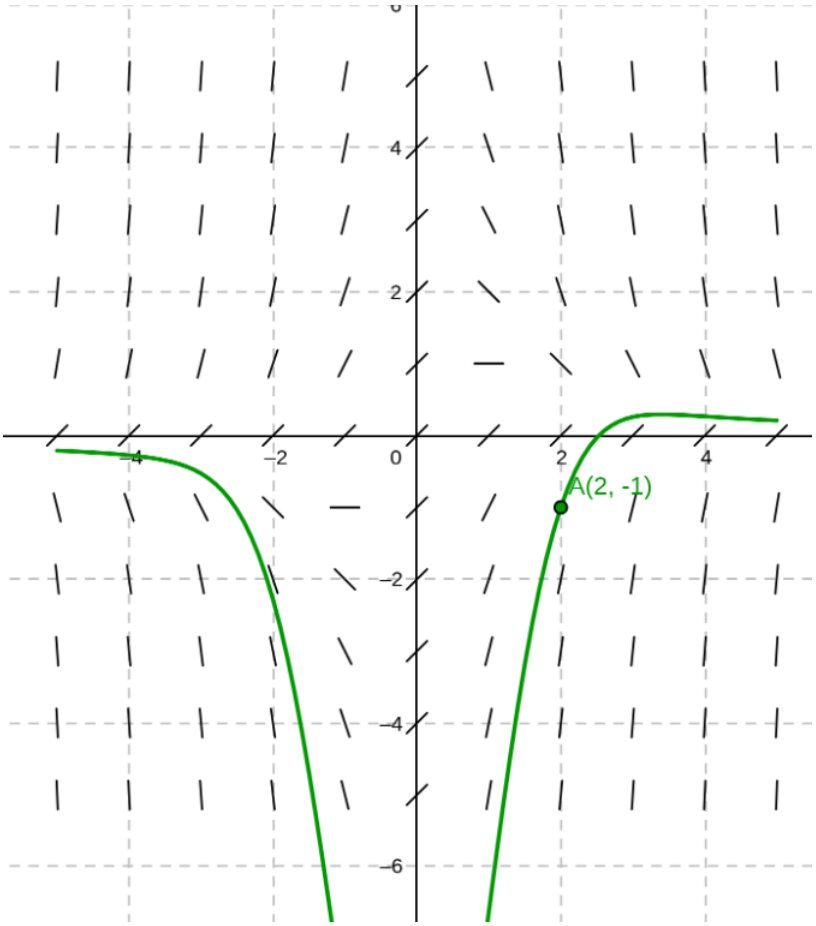
  selects one specific solution from infinitely many possible ones.

- **Sketching solutions with direction fields**: If we write a first order ODE in standard form as $\frac{dy}{dx} = f(x, y)$ Then we can use the fact that derivative at a point gives the slope of the tangent line at that point to sketch line segments depicting these slopes. These line segments gives us a visual depiction of the shape of solutions; thus we can use them to sketch the solution passing through a given point because the graph of the solutions are tangent to the lineal elements or equivalently the direction field.

  For example, consider

  $$\frac{dy}{dx} = 1 - xy.$$

  The direction field is then



Note: To actually draw a curve, we must solve the DE. There are infinitely many possible curves, each one coming from a different initial point.

- **Approximate solution curve**: If we cannot solve the DE, we can instead

  1. Pick a starting point (x_0, y_0)
  2. Use the slope at that point
  3. Move a small step in that direction
  4. Repeat

  This is called **Euler's method**. That is how software draws solution curves when no closed form solution is known.

- **Equilibrium (constant) solutions**: An equilibrium occurs when

$$\frac{dy}{dx} = 0.$$

  This happens when

$$f(x, y = 0).$$

  These appear as horizontal lines in the slope field. For example, if

$$\frac{dy}{dx} = y(1 - y),$$

then the equilibria occurs when

$$y(1-y)=0.$$

So, $y=0$ or $y=1$

These lines often represent:

- Stable equilibrium (solutions approach it)
- Unstable equilibrium (solutions move away)

- **Existence / uniqueness theorem**: Consider the IVP

$$\begin{cases} \frac{dy}{dx} & = f(x,y), \\ y(x_0) & = y_0 \end{cases}.$$

1. If $f(x,y)$ is continuous **around** (x_0,y_0), then the IVP has a solution on an open interval containing x_0
2. If in addition, $\frac{\delta f}{\delta y}$ is continuous around (x_0,y_0), then the solution is unique on that interval.

Note: This is a sufficient but not necessary condition for existence. Also, A solution exists and is unique on some open interval containing the initial point.

Consider an example,

$$\begin{cases} \frac{dy}{dx} & = x\ln(y), \\ y(1) & = 1 \end{cases}.$$

Since the natural log is continuous on its domain $(0,\infty)$, $x\ln(y)$ is continuous around $(1,1)$. Thus, a solution exists.

The partial derivative of $f(x,y)$ with respect to y is

$$\frac{\delta f}{\delta y}=\frac{x}{y},$$

which has a discontinuity at $y=0$. Thus, since $y\neq 0$, $\frac{x}{y}$ is continuous around $(1,1)$.

Therefore, there is a unique solution around x_0.

Next, consider

$$\begin{cases} \frac{dy}{dx} & = \sqrt{x-y}, \\ y(2) & = 2 \end{cases}.$$

Recall that $\sqrt{x-y}$ is only defined for real x,y if $x-y\geqslant 0$. Since $x_0-y_0=2-2=0$, $\sqrt{x-y}$ is not continuous around (x_0,y_0). Thus, no solution exists.

Consider

$$\begin{cases} \frac{dy}{dx} = \frac{y}{x}, \\ y(0) = 0 \end{cases} .$$

So

$$\begin{aligned} \frac{1}{y}dy &= \frac{1}{x}dx \\ \implies \int \frac{1}{y}\, dy &= \int \frac{1}{x}\, dx \\ \implies \ln|y| &= \ln|x| + C_0 \\ \implies |y| &= e^{\ln|x|+C_0} \\ \implies |y| &= C\,|x| \\ \implies y &= \pm Cx = Kx. \end{aligned}$$

Notice that $f(x, y) = \frac{y}{x}$, which is not continuous around $(0, 0)$. Thus, the theorem does not guarantee a solution on an open interval containing $(0, 0)$. In fact,

$$y = Kx$$

satisfies the DE for any $x \neq 0$, and $y(0) = 0$ holds for any value of K. Thus, infinitely many solutions satisfy the DE, so we see that uniqueness fails. The theorem does not apply at $(0, 0)$, so it gives no information about solutions near $x = 0$. However, even though the theorem does not apply, solutions do exist.

10.2.4 Linear differential equations

- **Linear DE**: A DE is linear in y if it is linear in y and its derivatives. Thus, no products of y and itself or its derivatives, and no nonlinear function of y or its derivatives. Consider the following functions

  (a) $y' + y = x^2$: Linear

  (b) $y' + y^2 = x$: Non-linear, problem is y^2

  (c) $2y' = (y^2 + 1)(y + 1)$: Non-linear, problem is y^2 and y^3

  (d) $y'(1 + y) + 3y = 4x$: Non-linear, problem is $y'y$

  (e) $\frac{d^4y}{dx^4} + y^2 = 0$: Non-linear, y^2

  (f) $x^3\frac{d^3y}{dx^3} + x\frac{dy}{dx} - 5y = e^x$: Linear

  (g) $\frac{d^2y}{dx^2} + \sin(y) = 0$: Non-linear, $\sin(y)$

  Notice that the linearity in y of a DE does not say anything about linearity of functions of x, all we care about is y and its derivatives.

- **Linear DEs (formal)**: A linear differential equation is one in which the dependent variable and all its derivatives appear linearly (i.e., no products, powers, or nonlinear functions of the dependent variable). An n^{th} order linear differential equation has the form

$$a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \cdots + a_1(x)y' + a_0(x)y = g(x).$$

  Where

  – $y = y(x)$ is the unknown function

  – $a_i(x), g(x)$ are given functions of x

  – The coefficients depend only on x, not on y or its derivatives

- **First order differential equations**: The standard form is

$$\frac{dy}{dx} + P(x)y = Q(x).$$

  This form is essential because it allows solution via an integrating factor.

  If we take a first order linear equation,

$$a_1(x)y' + a_0(x)y = g(x).$$

  Then, we see that

$$y' + \frac{a_0(x)}{a_1(x)}y = \frac{g(x)}{a_1(x)}.$$

  Now, let $\frac{a_0(x)}{a_1(x)} = P(x)$, and $\frac{g(x)}{a_1(x)} = Q(x)$, this gives the standard form

$$\frac{dy}{dx} + P(x)y = Q(x).$$

  Note that we are assuming here that $a_1(x) \neq 0$ on the interval of interest.

Consider the first order linear differential equation

$$(2x-1)\frac{dy}{dx}+3y=\sin(x).$$

Then,

$$\frac{dy}{dx}+\frac{3}{2x-1}y=\frac{\sin(x)}{2x-1},$$

now it is in standard form.

- **Integrating factor method (First order)**: Define

$$\mu(x)=e^{\int P(x)\,dx}.$$

Multiply the standard form of the differential equation by $\mu(x)$,

$$\mu(x)\frac{dy}{dx}+\mu(x)P(x)y=\mu(x)Q(x).$$

Then, the left hand side becomes

$$\frac{d}{dx}\left(\mu(x)y\right).$$

Thus,

$$\frac{d}{dx}\left(\mu(x)y\right)=\mu(x)Q(x).$$

Then, we integrate

$$\mu(x)y=\int \mu(x)Q(x)\ +C.$$

Finally, solve for y,

$$y=\frac{1}{\mu(x)}\left(\int \mu(x)Q(x)\,dx+C\right).$$

Consider an example,

$$\frac{dy}{dt}-5y=te^{4t}.$$

So, $P(t)=-5$, and $Q(t)=te^{4t}$. Define

$$\mu(t)=e^{\int P(t)\,dt}=e^{\int -5\,dt}=e^{-5t}.$$

Now,

$$y(t)=\frac{1}{\mu(t)}\left(\int \mu(t)Q(t)\,dt+C\right).$$

So,

$$y(t)=\frac{1}{e^{-5t}}\left(\int te^{-5t}e^{4t}\,dt+C\right)=\frac{1}{e^{-5t}}\left(\int te^{-t}\,dt+C\right).$$

Using integration by parts, let $u = t$, so $du = dt$, and $dv = e^{-t}$, so $v = -e^{-t}$. Then,

$$\int te^{-t}\,dt = -te^{-t} - \int -e^{-t}\,dt = -te^{-t} - e^{-t}.$$

Thus,

$$y(t) = \frac{1}{e^{-5t}}\left(-te^{-t} - e^{-t} + C\right).$$

Now, consider

$$\frac{dy}{dt} = -\frac{y}{t} + 7.$$

In standard form,

$$\frac{dy}{dt} + \frac{1}{t}y = 7.$$

Thus, $P(t) = \frac{1}{t}$, and $Q(t) = 7$. Define

$$\mu(t) = e^{\int \frac{1}{t}\,dt} = e^{\ln|t|} = |t|.$$

On any interval not containing zero (because $\frac{1}{t}$ in the differential equation is not defined at $t = 0$), we may take

$$\mu(t) = t \quad \text{or} \quad \mu(t) = -t.$$

Take $\mu(t) = t$. Then,

$$y(t) = \frac{1}{t}\left(\int 7t + C\right) = \frac{1}{t}\left(\frac{7}{2}t^2 + C\right).$$

- **The integrating factor is not unique**: Any nonzero constant multiple of an integrating factor is also an integrating factor. Suppose $\mu(t)$ is an integrating factor. Then, for any $C \neq 0$,

$$C\mu(t)$$

also works. This is because

$$\frac{d}{dt}\left(C\mu(t)y\right) = C\frac{d}{dt}\left(\mu(t)y\right).$$

Consider $k\mu(x)$ for $k \neq 0$, we would then have

$$k\mu(x)\frac{dy}{dx} + k\mu(x)P(x)y = k\mu(x)Q(x) \implies k\frac{d}{dx}\left(\mu(x)y\right) = k\mu(x)Q(x).$$

Thus, we get the same

$$\frac{d}{dx}\left(\mu(x)y\right) = \mu(x)Q(x),$$

which yields the same expression for y,

$$y(x) = \frac{1}{\mu(x)}\left(\int \mu(x)Q(x)\,dx + C\right).$$

This is why for $\mu(t) = |t|$, we can choose either $\mu(t) = t$ or $\mu(t) = -t$, because both integrating factors yield the same solution set.

- **Homogeneous and non-homogeneous linear first order DEs**: A first order linear DE of the form

$$\frac{dy}{dx} + P(x)y = 0$$

is called **homogeneous**, and the solution is

$$y = Ce^{-\int P(x)\,dx}.$$

If $Q(x) \neq 0$, then

$$\frac{dy}{dx} + P(x)y = Q(x)$$

is said to be **non-homogeneous**, with solution

$$y = y_h + y_p,$$

where

  - y_h = solution to homogeneous equation
  - y_p = particular solution

- **Model example with first order linear DE**: A 30-gallon tank initially contains 15 gallons of salt water containing 3 pounds of salt. Suppose salt water containing 1 pound of salt per gallon is pumped into the top of the tank at the rate of 2 gallons per minute, while a well-mixed solution leaves the bottom of the tank at a rate of 1 gallon per minute. How much salt is in the tank when the tank is full?

  Notice that we are interested in the change of salt. If $S(t)$ denotes the amount of salt in the tank at time t, and $V(t)$ denotes the amount of water in the tank at time t, then we are interested in $\frac{dS}{dt}$.

  The initial amount of water in the tank is 15 gallons. The inflow of water into the tank is 2 gallons per minute, while the outflow is 1 gallon per minute. So, the net change in the amount of water per unit time is $2 - 1$ gallons per minute. Thus,

$$V(t) = 15 + \text{net change per unit time} = 15 + (2-1)t = 15 + t.$$

  From this, if t_f is the time when the tank is full (at 30 gallons),

$$V(t_f) = 30 = 15 + t_f.$$

  So, $t_f = 15$ minutes.

  Since $S(t)$ is the unknown function, $S(0) = 3$ is the initial condition. Observe that a well-mixed solution implies

$$\frac{S(t)}{V(t)}\frac{\text{lb}}{\text{gal}}.$$

  We require the rate of salt in minus the rate of salt out, since we are interesting in only the amount of salt at the end. First, we can find the rate of salt into the tank. If 1 lb of salt per gallon is pumped into the tank at the rate of 2 gallons per minute, then the rate of salt in is given by

$$1\frac{\text{lb}}{\text{gal}} \cdot 2\frac{\text{gal}}{\text{min}} = 2\frac{\text{lb}}{\text{min}}.$$

Now for the rate of salt out of the tank. Recall that there are $\frac{S(t)}{V(t)}$ pounds of salt water per gallon of water. So, if the amount of salt water that leaves the tank per minute is one, then the rate of salt out is given by

$$\frac{S(t)}{V(t)}\frac{\text{lb}}{\text{gal}} \cdot 1\frac{\text{gal}}{\text{min}} = \frac{S(t)}{V(t)}\frac{\text{lb}}{\text{min}}.$$

So, the rate of change in salt is given by

$$\frac{dS}{dt} = \text{rate in } - \text{ rate out} = \left(2 - \frac{S(t)}{15+t}\right)\frac{\text{lb}}{\text{min}}.$$

Notice that this is a first order linear differential equation, which has standard form

$$\frac{dS}{dt} + \frac{1}{15+t}S = 2.$$

Now, we can solve. Define

$$\mu(t) = e^{\int \frac{1}{15+t}\,dt} = e^{\ln|15+t|} = 15+t.$$

Thus,

$$S(t) = \frac{1}{15+t}\left(\int (2(15+t))\,dt + C\right) = \frac{1}{15+t}\left(30t + t^2 + C\right).$$

With the initial condition $S(0) = 3$,

$$3 = \frac{1}{15+0}(30(0) + 0^2 + C) \implies C = 45.$$

So,

$$S(t) = \frac{1}{15+t}\left(t^2 + 30t + 45\right) = \frac{t^2+30t+45}{15+t}.$$

When the tank is full, $t = 15$, so the amount of salt in the tank when it is full is given by

$$S(15) = \frac{15^2 + 30(15) + 45}{30} = 24.$$

10.3 Substitution methods and exact solutions

10.3.1 Substitution methods

- **Multivariate polynomials**: When a polynomial depends on more than one independent variable, it is called a multivariate polynomial. Let $x_1, x_2, \ldots, x_n$ be independent variables. A **polynomial** in n variables over a field $\mathbb{F}$ has the form

$$p(x_1, \ldots, x_n) = \sum_{\alpha \in \mathbb{N}^n} c_\alpha x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$$

  where

  - Each $c_\alpha \in \mathbb{F}$
  - The multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$ has only finitely many nonzero coefficients
  - All exponents are non-negative integers

  Each term $c_\alpha x^\alpha$ is called a **monomial**. For a monomial $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, the **total degree** is

$$|\alpha| = \alpha_1 + \cdots + \alpha_n.$$

  The degree of the polynomial is the maximum total degree among its monomials.

  Consider

$$p(x, y) = 3x^2y + 5y^3 - 7.$$

  Here, p has degree 3.

- **Homogeneous polynomials**: A polynomial is **homogeneous of degree** d if *every* monomial has total degree d. For example,

$$q(x, y, z) = x^2 + y^2 + z^2$$

  is homogeneous of degree 2.

- **Homogeneous functions**: f is homogeneous of degree n if

$$f(tx) = t^n f(x).$$

  For example, if $f(x) = x^2$, then $f(tx) = t^2x^2$, so f is homogeneous of degree 2. Linear functions are homogeneous of degree 1.

- **First order DE differential form**: For any first order differential equation, we can write it in the differential form

$$M(x, y)dx + N(x, y)dy = 0.$$

- **Coefficient functions of a differential equation**: In the context of differential equations, coefficient functions are the functions that multiply the unknown function or its derivatives. They play the same structural role as numerical coefficients in algebraic equations, but they are allowed to vary with the independent variable(s). Consider a differential equation for a unknown function $y(x)$. A typical form is

$$a_n(x)y^{(n)}(x) + a_{n-1}(x)y^{(n-1)}(x) + \cdots + a_1(x)y'(x) + a_0(x)y(x) = g(x).$$

  Here, $a_0(x), a_1(x), \ldots, a_n(x)$ are coefficient functions.

- **Homogeneous first order DE**: When written in differential form, we can say that the first order differential equation is homogeneous if the coefficient functions M and N are homogeneous of the same degree. That is, for some real number k,

$$M(\lambda x, \lambda y) = \lambda^k M(x, y), \quad N(\lambda x, \lambda y) = \lambda^k N(x, y)$$

for all $\lambda > 0$. Suppose M and N are homogeneous of the same degree, we have

$$M(x, y)\, dx + N(x, y)\, dy = 0 \implies \frac{dy}{dx} = -\frac{M(x, y)}{N(x, y)}.$$

If we let $F(x, y) = -\frac{M(x,y)}{N(x,y)}$, then we can see that

$$F(\lambda x, \lambda y) = -\frac{M(\lambda x, \lambda y)}{N(\lambda x, \lambda y)} = -\frac{\lambda^k M(x, y)}{\lambda^k N(x, y)} = -\frac{M(x, y)}{N(x, y)} = F(x, y).$$

Thus, F is homogeneous of degree zero. Now, let's let $\lambda = \frac{1}{x}$, then

$$F(\lambda x, \lambda y) = F\left(\frac{x}{x}, \frac{y}{x}\right) = F\left(1, \frac{y}{x}\right).$$

We see that F depends only on the ratio $\frac{y}{x}$, scaling has no effect. Define

$$f(t) := F(1, t),$$

then $f(t) = f\left(\frac{y}{x}\right)$. In a homogeneous first order DE,

$$\frac{dy}{dx} = F(x, y),$$

this structure guarantees that

$$\frac{dy}{dx} = f\left(\frac{y}{x}\right).$$

If we make the substitution $y = xv$, then

$$\frac{dy}{dx} = f\left(\frac{xv}{x}\right) = f(v).$$

So, we can solve using separability. Note that $v = v(x)$ is a function of x, since y is a function of x, and $v = \frac{y}{x}$.

- **Bernoulli first order ODEs**: Consider an ODE of the form

$$\frac{dy}{dx} + P(x)y = Q(x)y^n,$$

where $P(x), Q(x)$ are continuous functions on some interval, $n \in \mathbb{R}$, and $n \neq 0, 1$. If $n = 0$ or $n = 1$, the equation reduces to a first order linear ODE. In the form above, we can divide by y^n to get

$$\frac{1}{y^n}\frac{dy}{dx} + P(x)y^{1-n} = Q(x).$$

Now, let $u = y^{1-n}$, where $u = u(x)$ is a function of x, then

$$\frac{du}{dx} = (1 - n)y^{-n}\frac{dy}{dx} \implies \frac{1}{y^n}\frac{dy}{dx} = \frac{1}{1 - n}\frac{du}{dx}.$$

So,

$$\frac{1}{y^n}\frac{dy}{dx} + P(x)y^{1-n} = Q(x) \implies \frac{1}{1 - n}\frac{du}{dx} + P(x)u = Q(x)$$
$$\implies \frac{du}{dx} + (1 - n)P(x)u = (1 - n)Q(x).$$

Notice that we now have a first-order linear differential equation in u, which can be solved with an integration factor.

Consider an example,

$$y^2\frac{dy}{dx} + 2xy^3 = 6x.$$

The Bernoulli form is then

$$\frac{dy}{dx} + 2xy = 6xy^{-2}.$$

So, $n = -2$, and $u = y^{1-n} = y^{1+2} = y^3$. Thus,

$$\frac{du}{dx} = 3y^2\frac{dy}{dx} \implies \frac{dy}{dx} = \frac{1}{3y^2}\frac{du}{dx}.$$

From this,

$$\frac{dy}{dx} + 2xy = 6xy^{-2} \implies \frac{1}{3y^2}\frac{du}{dx} + 2xy = 6xy^{-2} \implies \frac{du}{dx} + 6xu = 18x.$$

- **Substitution of the dependent variable**: It could be that we can convert a DE to a linear DE by changing the dependent variable. Consider

$$g'(y)y' + p(x)g(y) = f(x).$$

Let $z = g(y(x))$, so y is a function of x and g is a function of y, and z is a function of x, $z(x)$.

$$z' = g'(y(x))y'(x).$$

Thus,

$$y' = \frac{z'}{g'}.$$

Now,

$$g'(y)y' + p(x)g(y) = f(x) \implies z' + p(x)z = f(x),$$

which is linear in z.

- **General substitutions**: We can also make substitutions by choosing a new variable that simplifies the expression appearing in the ODE. Consider

$$\frac{dy}{dx} = 2 + e^{y-2x+7}.$$

If we let $v = y - 2x + 7$, then $\frac{dy}{dx} = v' + 2$, and DE becomes

$$v' + 2 = 2 + e^v \implies v' = e^v \implies \frac{1}{e^v}dv = dx.$$

Which we can solve by integrating both sides.

10.3.2 Exact equations

- **Exact equations**: An exact differential equation is a first-order ODE that arises from the total differential of a scalar potential function. Conceptually, instead of solving directly for $y(x)$, you identify a function $F(x, y)$ whose level curves $F(x, y) = C$ are the solutions.

  An equation written in differential form as

  $$M(x, y)\, dx + N(x, y)\, dy = 0.$$

  where M and N are functions with continuous partial derivatives on some region.

  The equation is exact if there exists a function $F(x, y)$ such that

  $$dF = M\, dx + N\, dy.$$

  By the multivariable chain rule,

  $$dF = \frac{\delta F}{\delta x}\, dx + \frac{\delta F}{\delta y}\, dy.$$

  Therefore, exactness means

  $$M = F_x, \quad N = F_y.$$

- **Necessary and sufficient condition**: If M and N have continuous first partial derivatives on a simply connected region, the equation is exact iff

  $$\frac{\delta M}{\delta y} = \frac{\delta N}{\delta x}.$$

  This follows from equality of mixed partials

  $$F_{xy} = F_{yx}.$$

- **Solving exact equations**: First, verify exactness, so compute

  $$M_y, N_x.$$

  If they match, proceed. Then, integrate M with respect to x.

  $$F(x, y) = \int M(x, y)\, dx + g(y).$$

  where $g(y)$ is an unknown "constant" of integration that may depend on y. All we need to do is determine $g(y)$. So, we differentiate the expression for F with respect to y and match it to N.

  $$F_y(x, y) = N(x, y).$$

  Then, we can solve for $g'(y)$ and integrate to get $g(y)$. The implicit solution is then

  $$F(x, y) = C.$$

- **Why is** $F(x, y) = C$: Notice that since

$$dF = M\ dx + N\ dy,$$

and

$$M(x, y)\ dx + N(x, y)\ dy = 0,$$

we have $dF = 0$. Since the differential of F is zero, the function does not change along the solution curves. Thus,

$$dF = 0 \iff F(x, y) = \text{constant}.$$

Thus, the solution set consists of level curves of F:

$$F(x, y) = C.$$

Consider a solution curve, moving along a solution curve, output does not change. Thus, the curve lies on one contour line $F = C$.

Note: The "constant of integration" that appears while constructing $F(x, y)$ is not the same as the constant that appears in the final solution. They combine into a single arbitrary constant.