Undergraduate Topics in Mathematics (4)
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

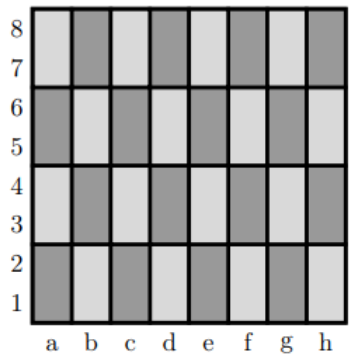
  Definition. A perfect cover of an $m \times n$ board with 2×1 dominoes is an arrangement of those dominoes on the chessboard with no squares left uncovered, and no dominoes stacked or left hanging off the end.

  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. ■

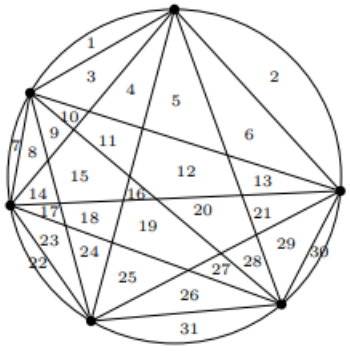
- **Naming Results**: So far, all of our results have been called "propositions." Here's the run-down on the naming of results:
  - A theorem is an important result that has been proved.
  - A proposition is a result that is less important than a theorem. It has also been proved.
  - A lemma is typically 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 principle**

Principle. The principle 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 principle?

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 principle 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 principle (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

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

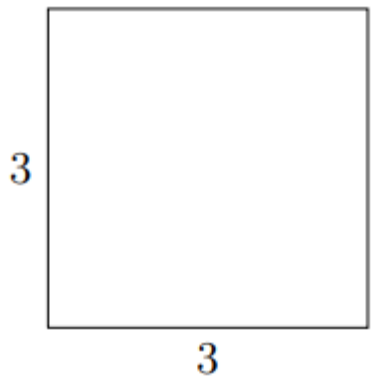
  And of course $8 + 1, 7 + 2, ..$ etc. We see we have four sums, we choose these sums as our boxes. If each of the four sums is a box, and each number is an object, then we are placing five objects into four boxes

  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 principle, 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 $\sqrt{2}$ ■

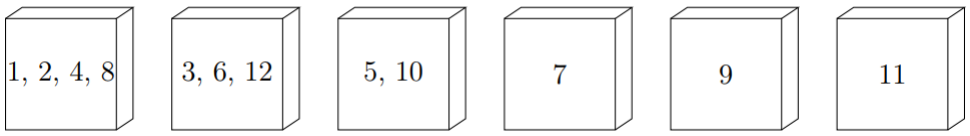
- **Another pingeonhole example**:

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 principle 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 principle (Principle 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**

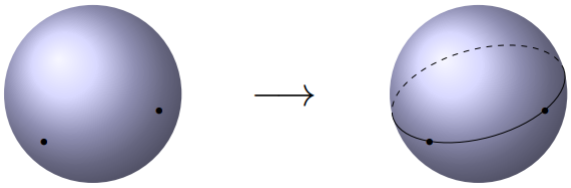
  Proposition. Suppose G is a graph with $n \geqslant 2$ vertices. Then G contains two vertices which have the same degree.

  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 principle, 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 principle. 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:

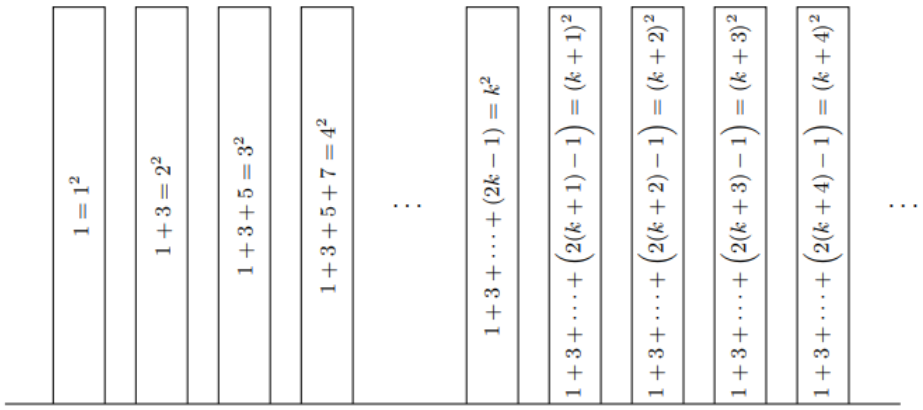
  1. If you give the first domino a push, it will fall (in particular, it will fall into the second domino, knocking it over).
  2. Moreover, every domino, when it's knocked over, falls into the next one and knocks it over.

  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 *principle 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*

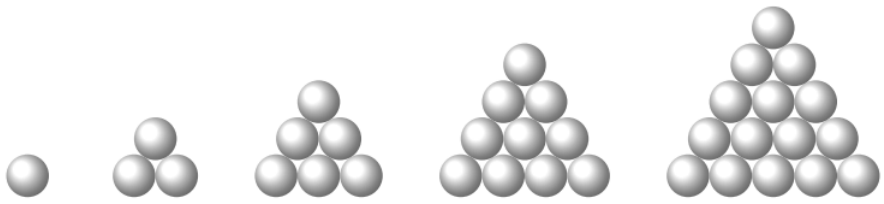
  Base case. $\langle\langle$Demonstration that S_1 is true$\rangle\rangle$

  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

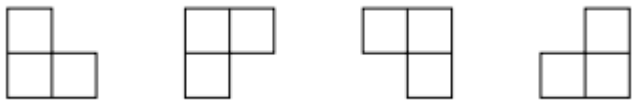
$$\begin{aligned} 1 \cdot 3 \cdot 5 \cdot ... \cdot (2k-1) \cdot (2k+1) &= \frac{(2k)!}{2^k k!}(2k+1) \\ &= \frac{(2k)!(2k+1)}{2^k k!} \\ &= \frac{(2k+1)!}{2^k k!} \\ &= \frac{(2k+1)!}{2^k k!} \cdot \frac{(2k+2)}{(2k+2)} \\ &= \frac{(2k+2)!}{2^k k!(2k+2)} \\ &= \frac{(2k+2)!}{2^k k! \cdot 2(k+1)} \\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

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:

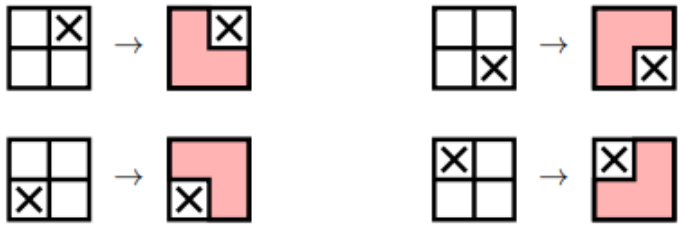


Since the proposition refers to something being true "for every $n \in \mathbb{N}$," that's a pretty good indication that induction is the way to proceed. The base case (when $n = 1$) will be fine. For the inductive hypothesis, we will be assuming that any $2^k \times 2^k$ board, with one square removed,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

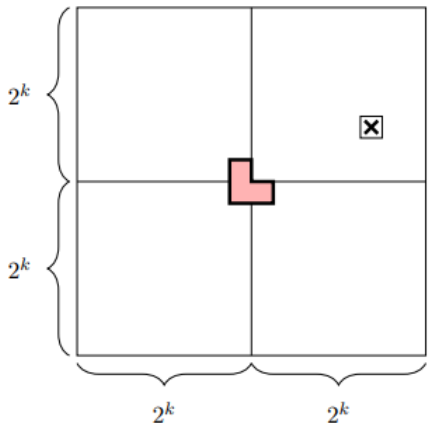
Base Case. The base case is when $n = 1$, and among the four possible squares that one can remove from a 2×2 chessboard, each leaves a chessboard which can be perfectly covered by a single L-shaped tile:



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 $a1$ breaks to completely break it up, and the bar with b breaks will require $b1$ breaks. Therefore, to break up the $m \times n$ bar, we must make a first break, followed by $(a1) + (b1)$ additional breaks. The total number of breaks is then

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

And $mn1$ 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.

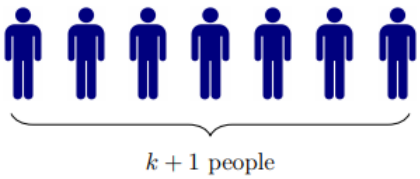
Conclusion. Therefore, by strong induction, every integer $n \geqslant 11$ can be written as the proposition asserts. ■

- **False proofs with induction**:

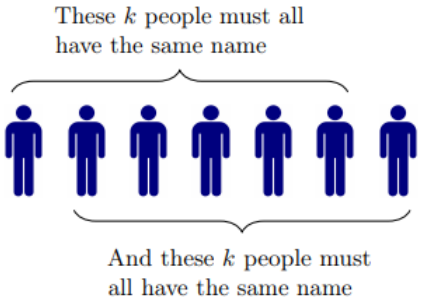
  Proposition. Everyone on Earth has the same name

  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,

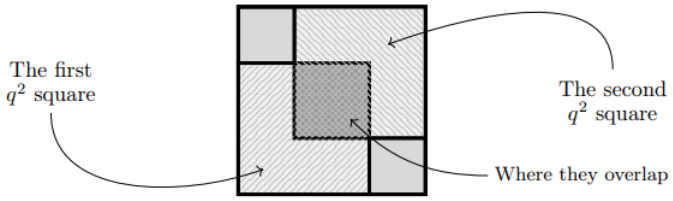
$$p^2 = q^2 + q^2.$$

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.

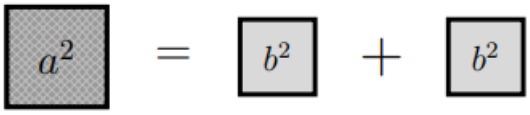
$$\begin{array}{ccccccccccc}
 & & & & \binom{0}{0} & & & & & & \\
 & & & \binom{1}{0} & & \binom{1}{1} & & & & & \\
 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & & \\
 & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & & \\
\binom{4}{0} & & \binom{4}{1} & & \binom{4}{2}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

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.

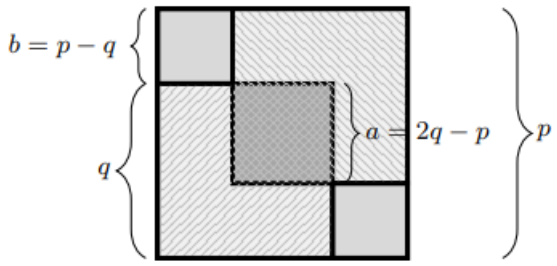
Getting back to the squares above, we are now going to imagine each square is a piece of paper and we are going to place the two q^2 squares on top of the p^2 square. If one q^2 square is placed in the lower-left, and the other is placed in the upper-right, this happens



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 principle**:

  Theorem 8.10 (The func-y pigeonhole principle): 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 principle 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

$$f(X) = \{y \in B : y = f(x) \text{ for some } x \in X\},$$

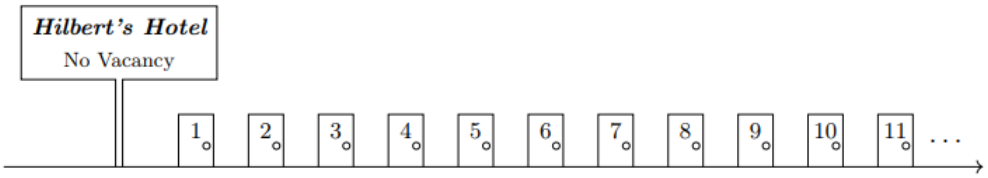
and the *inverse image* of Y is

$$f^{-1}(Y) = \{x \in A : f(x) \in Y\}.$$

- **The bijection principle**:

Principle (*The bijection principle.*) 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.

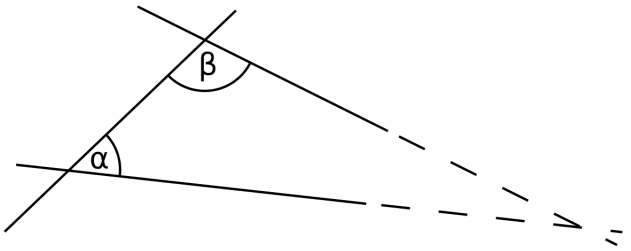
  For centuries, mathematicians like Proclus, Ptolemy, and others tried to prove the 5th postulate as a theorem based on the other four postulates. These attempts were unsuccessful, as the postulate is independent.

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

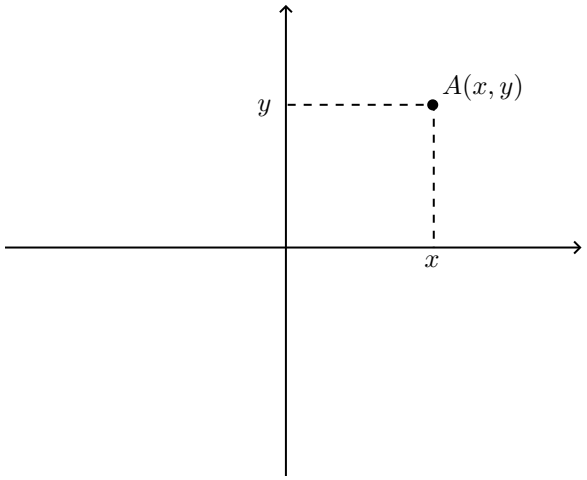
  1. **Flat Surface:** The Euclidean plane is flat, meaning it has no curvature.
  2. **Points and Lines:** It consists of an infinite set of points. Straight lines can be drawn to connect any two points, and these lines extend infinitely in both directions.
  3. **Distance and Angles:** Distance between 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

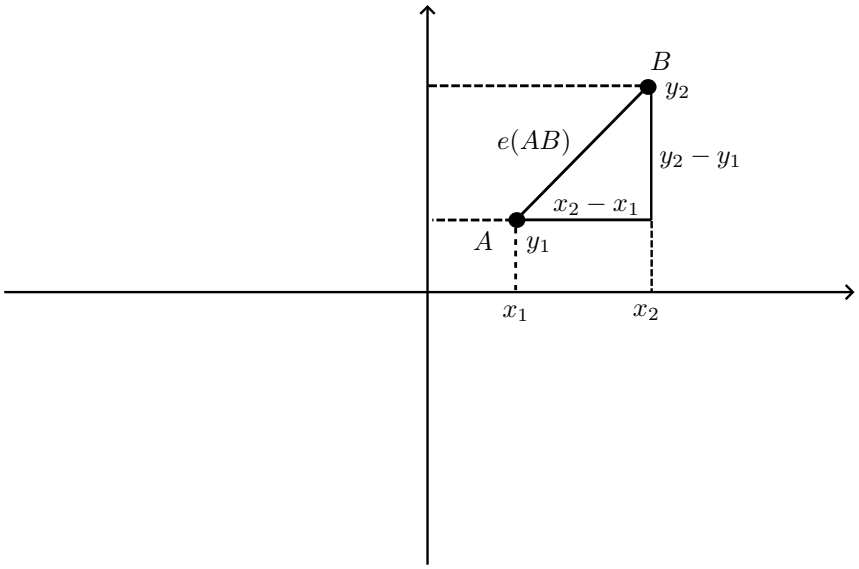
$$m = \frac{y_2 - y_1}{x_2 - x_1} \quad (\text{if } x_2 \neq x_1)$$

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

$$\text{arc length } PQ = \frac{\theta}{2\pi} \cdot 2\pi r = \theta r$$

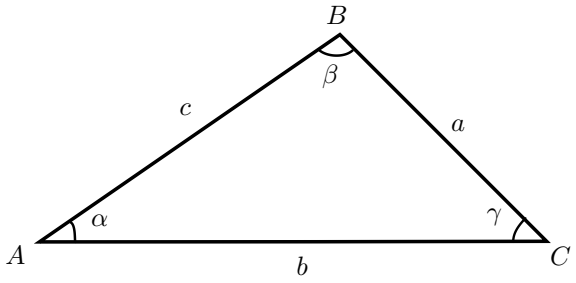
An explicit formula for the spherical distance between two points, in terms of their coordinates, is given next. It follows from the distance formula for three-dimensional space and the Law of Cosines.

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:

$$c^2 = a^2 + b^2 - 2ab\cos\gamma,$$

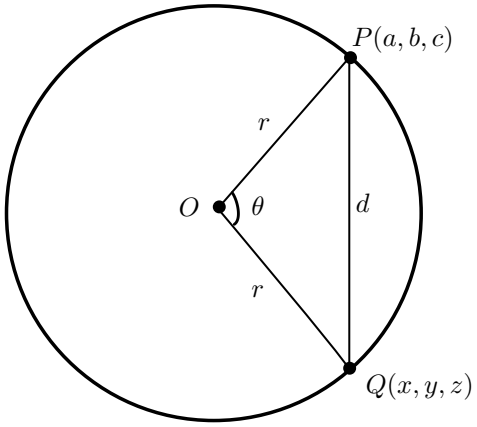
$$a^2 = b^2 + c^2 - 2bc\cos\alpha,$$
$$b^2 = a^2 + c^2 - 2ac\cos\beta.$$

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.

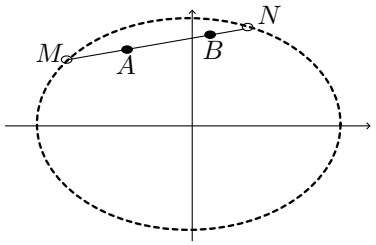
- **The chords of a circle**: A chord of a circle is a straight line segment whose endpoints lie on the circle. In other words, it is a line segment that connects two points on the circumference of a circle
- **The hyperbolic plane (Poincare disk model)**: Let $\mathbb{H}$ denote the hypebolic plane, 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

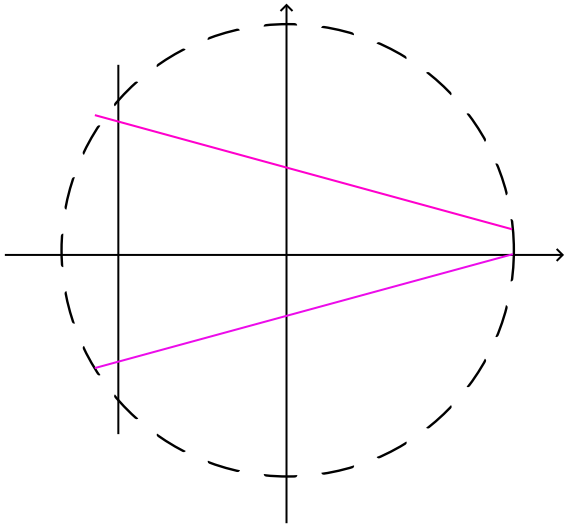
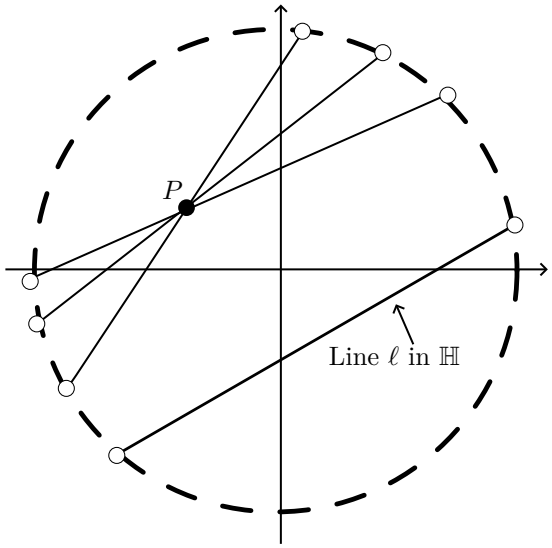


Figure 1: *Euclids fifth postulate does not hold on the hyperbolic plane*

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

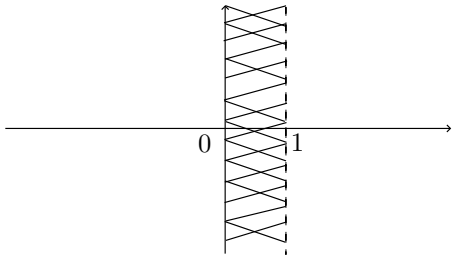
Also,



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

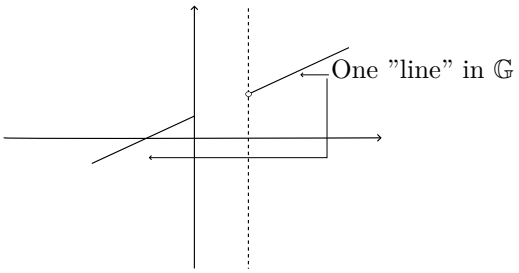
- **The gap plane**: Let $\mathbb{G}$ denote the *gap*, or *missing strip* plane. The points of $\mathbb{G}$ are all those of $\mathbb{E}$ except those (x, y) with $0 < x \leqslant 1$

So the y-axis is part of $\mathbb{G}$, but the line $x = 1$ is not (and neither is any vertical line



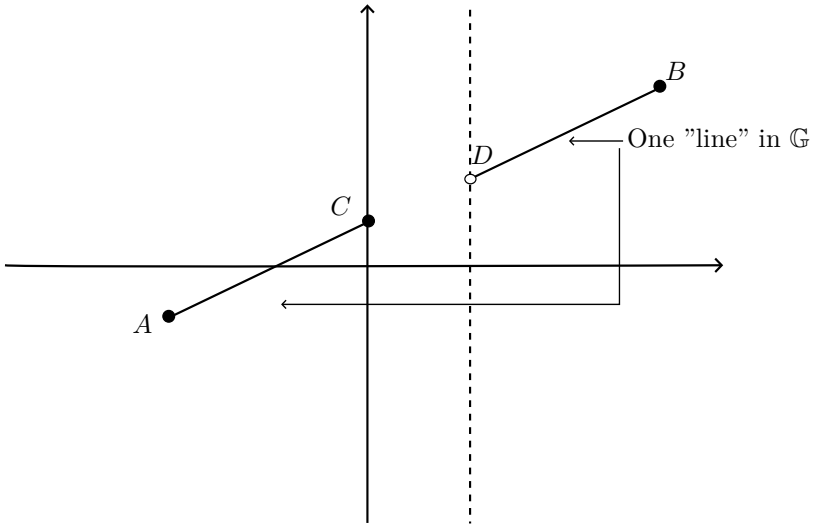
$x = a$ for $0 < a < 1$)

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

$$d_{\mathbb{G}}(AB) = \begin{cases} e(AB) & \text{for } A, B \text{ on the same side of the gap} \\ e(AB) - e(CD) & \text{for } A, B \text{ on the opposite sides of the gap} \end{cases}$$

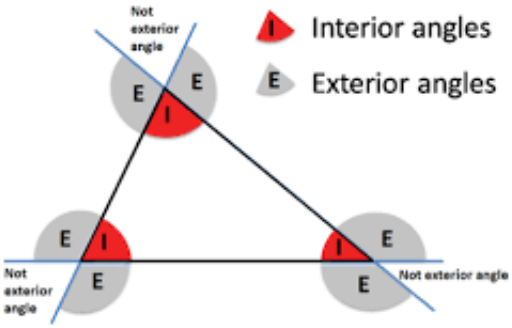
- **Interior and Exterior angles**: Interior angles are the angles inside the triangle. Each vertex of the triangle has one interior angle. The sum of the interior angles of a triangle is always 180°

  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**:

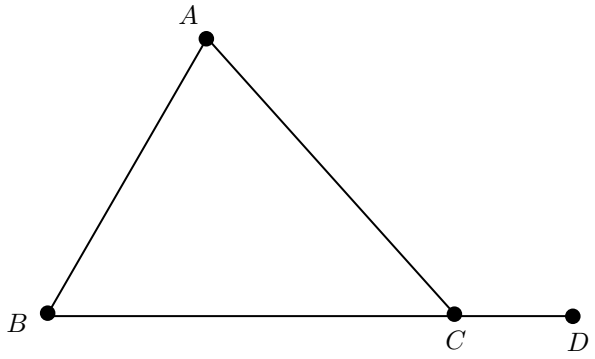
Theorem (*Exterior angle inequality*): An exterior angle of a triangle is greater than either remote interior angle. That is, if $\triangle ABC$ is a any triangle, and point D is on the extension of segment $\overline{BC}$ through C, then

$$\angle ACD > \text{ both } \angle A \text{ and } \angle B$$

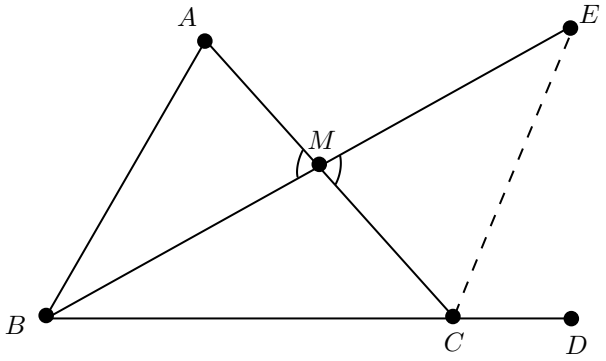
Background facts that are ok for both $\mathbb{E}$ and $\mathbb{S}$

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

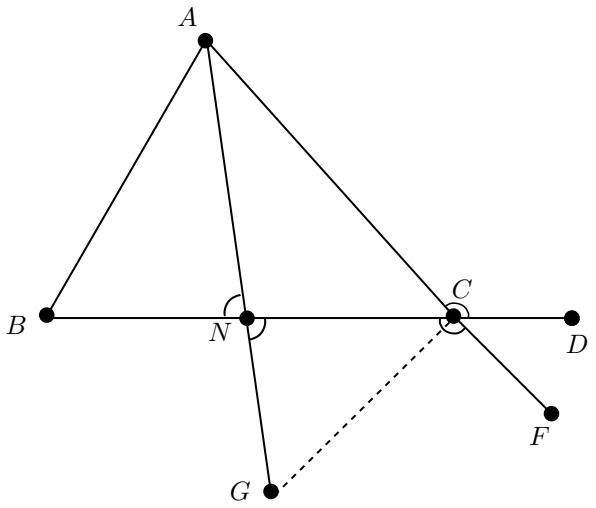
1. $AM = CM$
2. $MB = ME$
3. $\angle AMB = \angle CME$

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. $BN = CN$

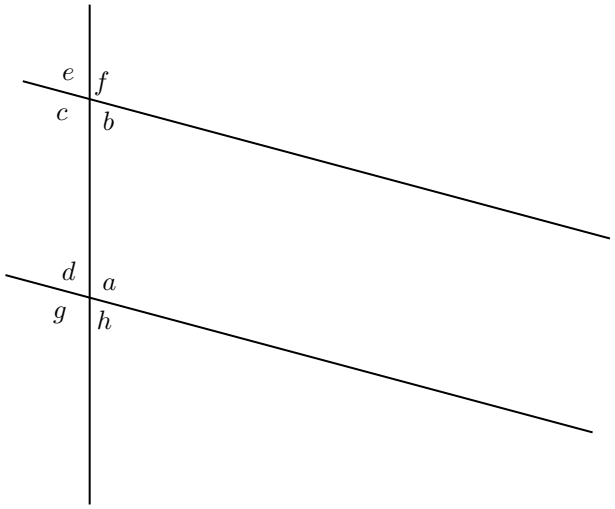
We have congruence, $\triangle ANB \cong \triangle CNG$. Thus, $\angle ABN = \angle NCG$. Therefore,

$$\begin{aligned}\angle ACD = \angle BCF &= \angle BCG + \angle GCF \\ &= \angle NCG + \angle GCF \\ &= \angle ABN + \angle GCF > \angle ABN = \angle B\end{aligned}$$

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, $c=a$

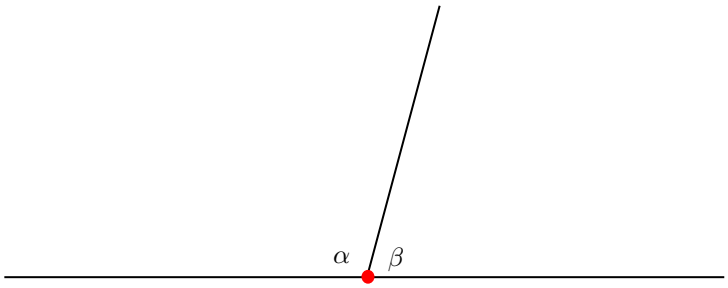
Therefore, we conclude that if $a+b=180$, $b=d$ and $c=a$ ■

- **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.

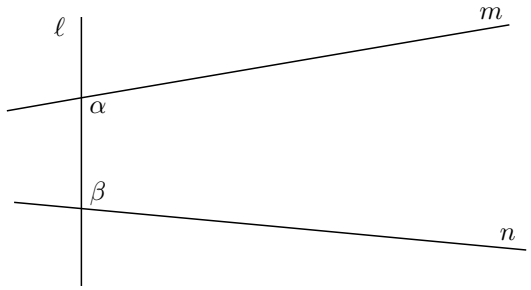
  Two triangles are congruent if their corresponding sides and angles are equal (e.g., by SSS, SAS, ASA, or AAS congruence criteria).

  We generally use the side-angle-side criterion to determine congruent triangles.

Also, recall the exterior angle theorem proved above.

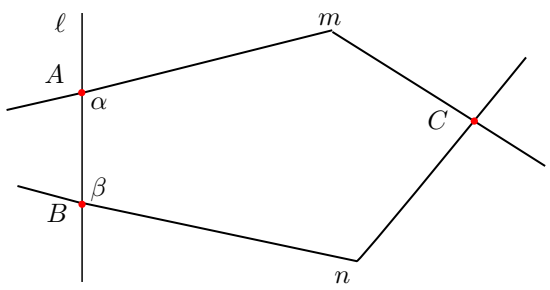
- **Example of proof by contradiction**: Suppose we are on Euclid's plane without the fifth postulate

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°



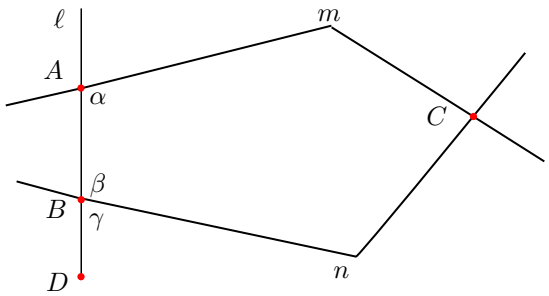
Then, m, n do not meet on that side of ℓ

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).

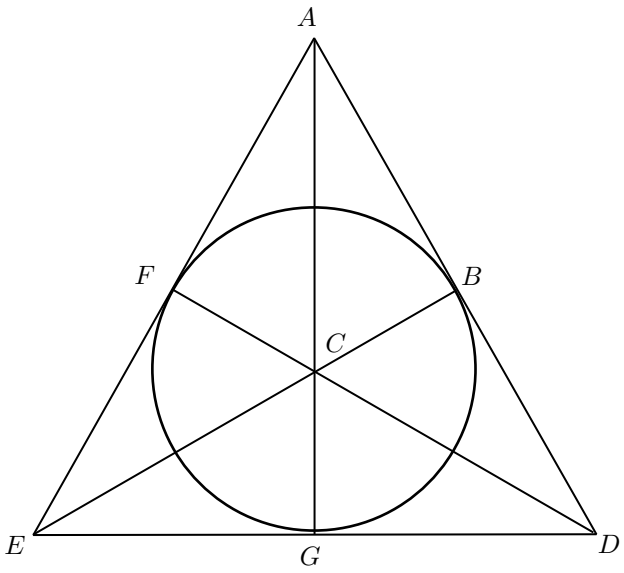
  It includes Euclidean, affine, and projective geometries as special cases.
- **The Fano plane**: The Fano plane is a *projective plane of order two.*

  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 $\mathbb{P}$

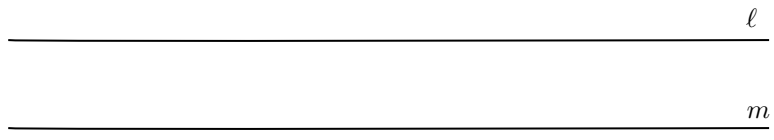
  Define distance as follows: For all $x \neq y \in \mathbb{P}$,

  $$\begin{cases} xy & = 1 \\ xx & = 0 \end{cases}$$

  The seven axioms are true for the TDM. We have $\mathbb{D} = \{0, 1\}$, thus $\omega = 1$

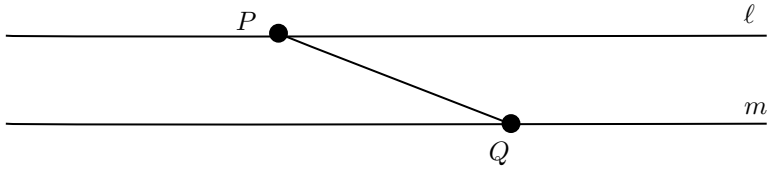
  For example, if $\mathbb{P} = \{A, B, C\}$, which implies $\mathbb{L} = \{A, B\}, \{A, C\}, \{B, C\}$, which forms a triangle where all sides are of length one.

- **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 $\mathbb{D} = [0, \infty),\ \omega = \infty$

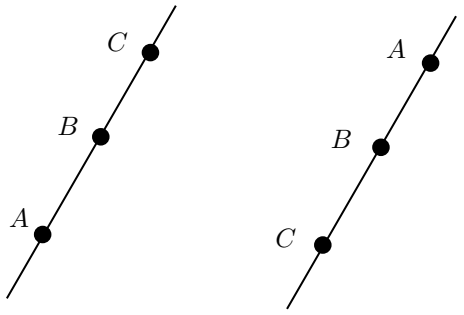


4.5 Betweenness, segements, and rays

- **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

$$AB = 1, \quad BD = 1, \quad AD = 2$$

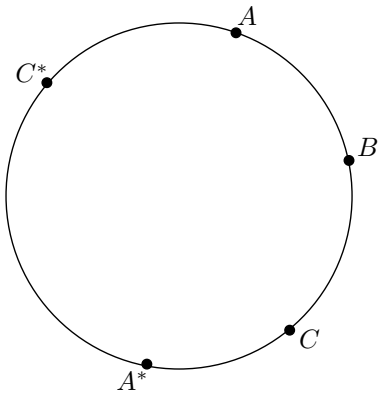
We see $AB + BD = AD$. Thus, $A - B - D$

- **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,

$$d_{\mathbb{S}}(AB) + d_{\mathbb{S}}(BC) = d_{\mathbb{S}}(AC)$$

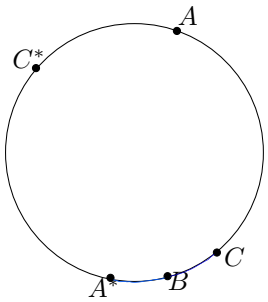
  Thus, $A - B - C$



  Next, let B be on the minor arc $\widehat{A^*C}$. Observe that

$$d_{\mathbb{S}}(AC) + d_{\mathbb{S}}(CB) = d_{\mathbb{S}}(AB)$$

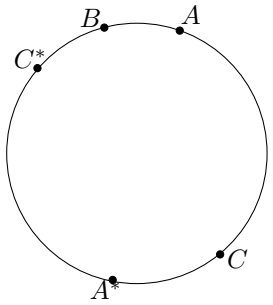
  Thus, $A - C - B$



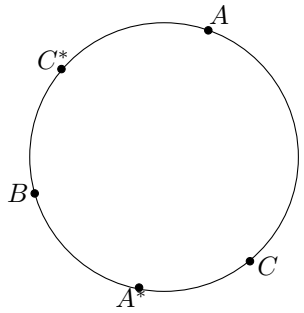
Next, let B be on the minor arc $\overset{\frown}{C^*A}$. Observe that

$$d_{\mathbb{S}}(BA) + d_{\mathbb{S}}(AC) = d_{\mathbb{S}}(BC)$$

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

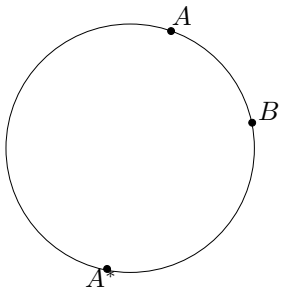
$$d_{\mathbb{S}}(AB) + d_{\mathbb{S}}(BA^*) = \pi r = d_{\mathbb{S}}(AA^*)$$

Thus, $A - B - A^*$

- **Betweenness theorem 1**:

Proposition. 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: If $A - B - C$ then $B - A - C$ and $A - C - B$ are false.

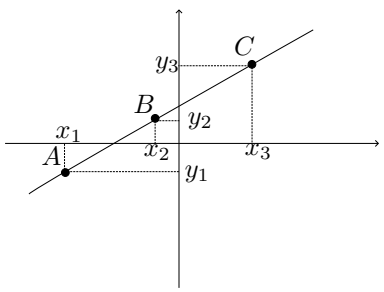
  Proof. Assume $A-B-C$, then A, B, C are different, collinear, and $AB+BC = AC$. We know by distance axioms 1 and 2 that each of AB, BC, and AC are greater than zero. Thus,

  $$AC > AB \quad \text{and } AC > BC$$

  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

$$\begin{aligned}A\text{-}B\text{-}C \text{ in } \mathbb{E} &\iff \text{A-B-C in } \mathbb{M}\\ A\text{-}B\text{-}C \text{ in } \mathbb{E} &\iff \text{A-B-C in } \mathbb{G}\\ A\text{-}B\text{-}C \text{ in } \mathbb{E} &\iff \text{A-B-C in } \mathbb{H}\end{aligned}$$

- **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

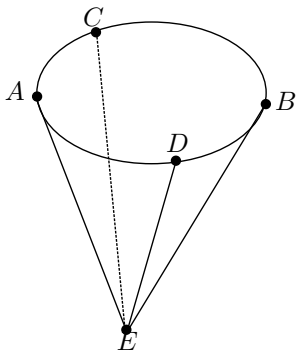
	A	B	C	D	E
A	0	3	1	2	4
B	3	0	2	1	4
C	1	2	0	3	4
D	2	1	3	0	4
E	4	4	4	4	0

The seven axioms hold, $\mathbb{D} = \{0, 1, 2, 3, 4\}, \omega = 4$, and all betweenness occurs for points on ℓ

$$A - C - B \quad A - D - B \quad C - A - D \quad C - B - D$$

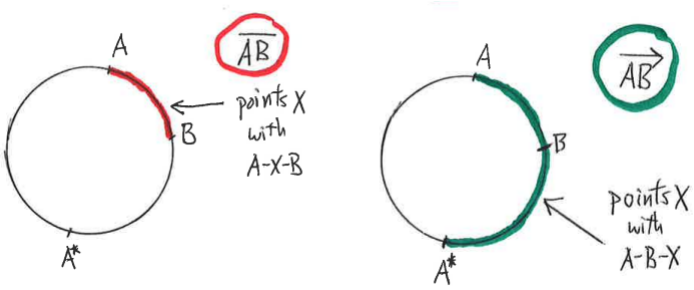
- **Segments and rays**: Let $A \neq B$ be points in $\mathbb{P}$ with $AB < \omega$. Then, there is a unique line through A, B, call it $\overleftrightarrow{AB}$
  - **The segment** $\overline{\{}AB\} = \{A, B\} \cup \{X : A - X - B\}$
  - **The ray** $\overrightarrow{\{}AB\} = \{A, B\} \cup \{X : A - X - B\} \cup \{X : A - B - X\}$

  Note: $\{X : A - X - B\} \cup \{X : A - B - X\} = \varnothing$



Notation: $\overline{AB}, \overrightarrow{AB}, \overleftrightarrow{AB}$ denote sets of points, with $\{A, B\} \subseteq \overline{AB} \subseteq \overrightarrow{AB} \subseteq \overleftrightarrow{AB}$

- **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: 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**

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, $CD = AB$ ■

- **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. ■

4.6 Three axioms for the line

- **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

  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

$$AB + BC \leqslant AC$$

  Note: Don't worry about why the word triangle is in the name. Also, the triangle inequality is not necessarily true without Ax.BP

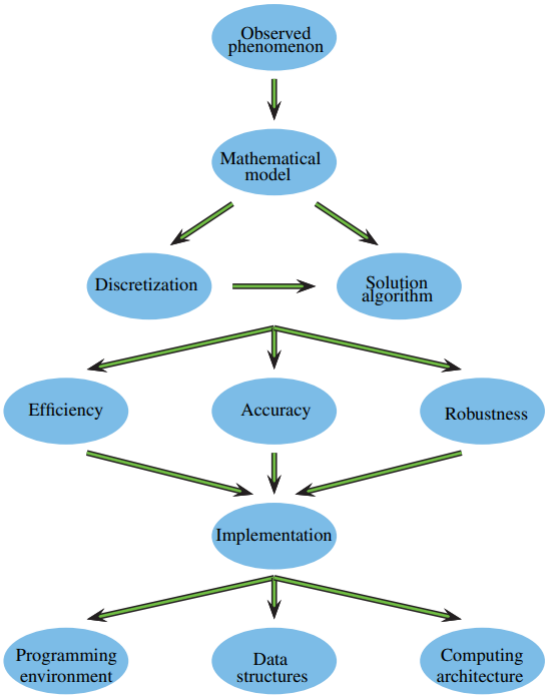
Numerical analysis with Julia

5.1 Numerical algorithms, roundoff errors, and nonlinear equations in one variable

- **Scientific computing**: Scientific computing is a discipline concerned with the development and study of **numerical algorithms** for solving mathematical problems that arise in various disciplines in science and engineering.

  Typically, the starting point is a given **mathematical model** which has been formulated in an attempt to explain and understand an observed phenomenon in biology, chemistry, physics, economics, or any other scientific or engineering discipline. We will concentrate on those mathematical models which are continuous (or piecewise continuous) and are difficult or impossible to solve analytically; this is usually the case in practice

  In order to solve such a model approximately on a computer, the continuous or piecewise continuous problem is approximated by a discrete one. Functions are approximated by finite arrays of values. Algorithms are then sought which approximately solve the mathematical problem efficiently, accurately, and reliably. This is the heart of scientific computing. **Numerical analysis** may be viewed as the theory behind such algorithms



- **Relative and absolute errors**: There are in general two basic types of measured error. Given a scalar quantity u and its approximation v:
  - **Absolute error**: The *absolute error* in v is

$$|u - v|$$

– **Relative error**: The *relative error*, assuming $u \neq 0$ is

$$\frac{|u - v|}{|u|}$$

Note: If we take the absolute error, $|u - v|$. Then it is clear it will be some percentage of u. In other words, some scaled version of u. Thus, we have $|u - v| = p\,|u|$.

The relative error is usually a more meaningful measure. This is especially true for errors in floating point representation. For example, we record absolute and relative errors for various hypothetical calculations in the following table

u	v	Absolute error	Relative error
1	0.99	0.01	0.01
1	1.01	0.01	0.01
-1.5	-1.2	0.3	0.2
100	99.99	0.01	0.0001
100	99	1	0.01

We expect the approximation in the last row of the above table to be similar in quality to the one in the first row. This expectation is borne out by the value of the relative error but is not reflected by the value of the absolute error

When the approximated value is small in magnitude, things are a little more delicate, and here is where relative errors may not be so meaningful. But let us not worry about this at this early point.

- **The sterling approximation**: The quantity

$$v = S_n = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Is called sterling's approximation and is used to approximate $u = n!$ for large n

- **Error types**: Knowing how errors are typically measured, we now move to discuss their source. There are several types of error that may limit the accuracy of a numerical calculation.

  1. **Errors in the problem to be solved**: These may be approximation errors in the mathematical model

     Another typical source of error in the problem is error in the input data. This may arise, for instance, from physical measurements, which are never infinitely accurate. Thus, it may be that after a careful numerical simulation of a given mathematical problem, the resulting solution would not quite match observations on the phenomenon being examined.

     At the level of numerical algorithms, which is the focus of our interest here, there is really nothing we can do about the above-described errors. Nevertheless, they should be taken into consideration, for instance, when determining the accuracy (tolerance with respect to the next two types of error mentioned below) to which the numerical problem should be solved.

  2. **Approximation errors**: Such errors arise when an approximate formula is used in place of the actual function to be evaluated.

We will often encounter two types of approximation errors:

(a) **Discretization errors** arise from discretizations of continuous processes, such as interpolation, differentiation, and integration.

(b) **Convergence errors** arise in iterative methods. For instance, nonlinear problems must generally be solved approximately by an iterative process. Such a process would converge to the exact solution in infinitely many iterations, but we cut it off after a finite (hopefully small!) number of such iterations. Iterative methods in fact often arise in linear algebra.

3. **Roundoff errorss:** Any computation with real numbers involves roundoff error. Even when no approximation error is produced (as in the direct evaluation of a straight line, or the solution by Gaussian elimination of a linear system of equations), roundoff errors are present. These arise because of the finite precision representation of real numbers on any computer, which affects both data representation and computer arithmetic.

- **Digits of accuracy**: If p is the relative error when v approximates u, then the digits of accuracy in the approximation v can be found with

$$\log_{10}\left(\frac{1}{p}\right) = -\log_{10}(p)$$

- **Catastrophic cancellation**: A numerical phenomenon that occurs when subtracting two nearly equal numbers in floating-point arithmetic. The result of this subtraction can lose significant digits, leading to a dramatic loss of precision in the computed result.

  In floating-point representation, numbers are stored with a finite number of significant digits (or bits). When two numbers are nearly equal, their leading digits cancel each other out during subtraction. The result is dominated by the remaining, less significant digits, which are more prone to rounding errors.

  Suppose we want to compute $x - y$, where $x = 1.0000001$, and $y = 1.0000000$. The true result is

$$x - y = 0.0000001$$

  However, if x and y are represented with only 7 significant digits in a floating-point system, $x = 1.000000$, and $y = 1.000000$. Then, their subtraction gives

$$x - y = 0.000000$$

  The true value is completely lost because the subtraction eliminates all significant digits.

- **Numerical noise**: Numerical noise refers to small errors or inaccuracies that arise in numerical computations due to the limitations of floating-point arithmetic. These errors are often very small, but they can accumulate or become significant in certain situations, especially when the computations involve many steps or operations sensitive to precision.

  Computers represent real numbers in a finite number of bits (e.g., 64 bits for double-precision floats).

This representation cannot store every real number exactly, so numbers are rounded to the nearest representable value. These small rounding errors introduce "noise" into computations.

- **Machine epsilon**: The machine epsilon is the smallest positive number ε such that

$$1 + \varepsilon > 1$$

in a given floating-point system. It represents the upper bound on the relative error due to rounding in floating-point arithmetic.

  - **Single Precision (32-bit):**

$$\varepsilon_{\text{machine}} \approx 2^{-23} \approx 1.19 \times 10^{-7}$$

  - **Double Precision (64-bit):**

$$\varepsilon_{\text{machine}} \approx 2^{-52} \approx 2.22 \times 10^{-16}$$

- **Approximation Error (approximating the derivative)**: Consider the formula for the derivative of a differentiable function $f\colon \mathbb{R} \to \mathbb{R}$ at x_0:

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

It is therefore reasonable to approximate $f'(x_0)$ using

$$\frac{f(x_0 + h) - f(x_0)}{h}$$

for some small positive h. The error in this approximation is

$$\left| f'(x_0) - \frac{f(x_0 + h) - f(x_0)}{h} \right|$$

and is called a **discretization error**.

For example, consider $f(x) = \sin x$ at $x_0 = 1$. Note that $f'(x) = \cos x$. We have

$$f'(x_0) = \cos 1 = 0.5403023058681398\ldots$$

Let's now approximate this with

$$f'(x_0) \approx \frac{f(x_0 + h) - f(x_0)}{h}$$

For $h = 10^{-1}, 10^{-2}, ..., 10^{-16}$. The following Julia code

```
using Printf

function deriv_approx(f, x0, fp)
    @printf("%6s %24s %12s %10s %8s\n", "h", "fpapprox",
↪   "abserr", "relerr", "digits")

    for k in 1:16
        h = 10.0^(-k)
        fpapprox = (f(x0+h) - f(x0))/h
        abserr = abs(fp - fpapprox)
        relerr = abserr/abs(fp)
        digits = -log10(relerr)
        @printf("%6.0e %24.16e %12.4e %10.2e %8.1f\n", h,
↪   fpapprox, abserr, relerr, digits)
    end

    return nothing
end

deriv_approx(f, 1.0, cos(1.0))
```

Yields the following output

h	fpapprox	abserr	relerr	digits
1e-01	4.9736375253538911e-01	4.2939e-02	7.95e-02	1.1
1e-02	5.3608598101186888e-01	4.2163e-03	7.80e-03	2.1
1e-03	5.3988148036032690e-01	4.2083e-04	7.79e-04	3.1
1e-04	5.4026023141862112e-01	4.2074e-05	7.79e-05	4.1
1e-05	5.4029809850586474e-01	4.2074e-06	7.79e-06	5.1
1e-06	5.4030188512133037e-01	4.2075e-07	7.79e-07	6.1
1e-07	5.4030226404044868e-01	4.1828e-08	7.74e-08	7.1
1e-08	5.4030230289825454e-01	2.9699e-09	5.50e-09	8.3
1e-09	5.4030235840940577e-01	5.2541e-08	9.72e-08	7.0
1e-10	5.4030224738710331e-01	5.8481e-08	1.08e-07	7.0
1e-11	5.4030113716407868e-01	1.1687e-06	2.16e-06	5.7
1e-12	5.4034554608506369e-01	4.3240e-05	8.00e-05	4.1
1e-13	5.3956838996782608e-01	7.3392e-04	1.36e-03	2.9
1e-14	5.4400928206632670e-01	3.7070e-03	6.86e-03	2.2
1e-15	5.5511151231257827e-01	1.4809e-02	2.74e-02	1.6
1e-16	0.0000000000000000e+00	5.4030e-01	1.00e+00	-0.0

Notice that when h is decreased by a factor of ten, the absolute error decreases by a factor of ten.

Further, notice that when $h = 10^{-k}$, for $k \in \{9, 10, 11, ..., 16\}$, the absolute error gets worse instead of better. Also, when $h = 10^{-16}$, we notice that the approximation reads zero. This is a result of round-off errors and the limitations of floating-point arithmetic in computers.

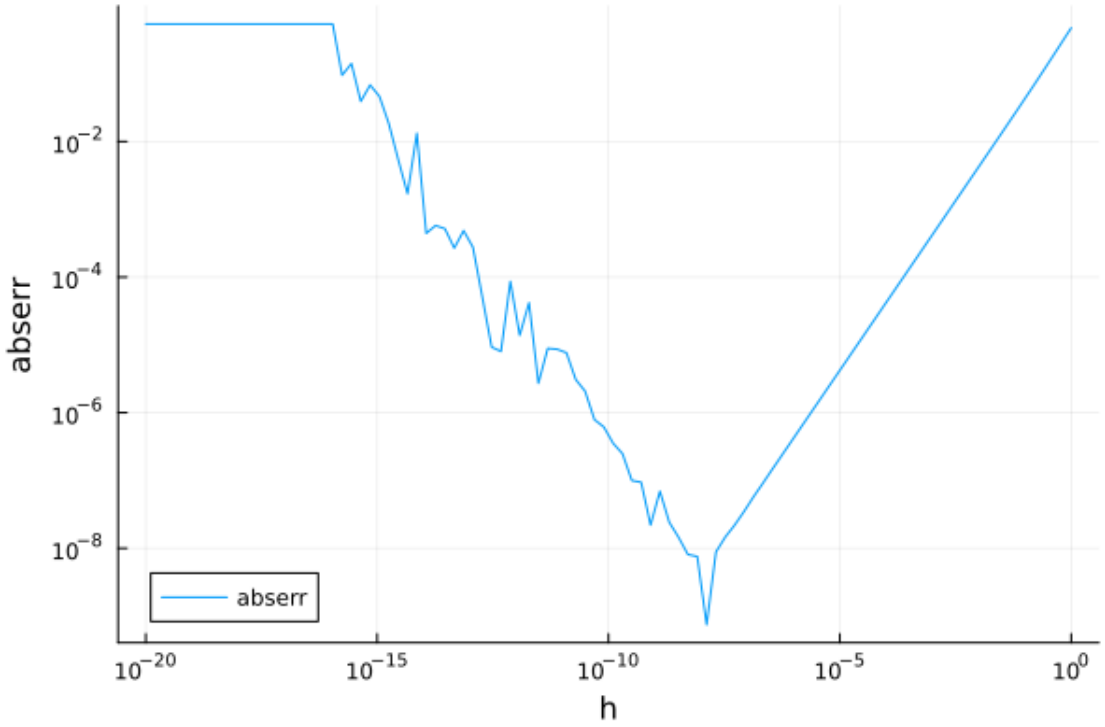
– **Catastrophic Cancellation:** The approximation formula involves subtracting two very close values $f(x_0 + h)$ and $f(x_0)$, for very small h

When h becomes very small, the values of $f(x_0 + h)$ and $f(x_0)$ are nearly identical. In floating-point arithmetic, this subtraction loses precision because the significant digits cancel out, leaving only the less accurate lower-order bits

– **Floating-Point Precision:** Floating-point numbers have limited precision. For typical 64-bit double-precision floating-point numbers, the relative precision is about 10^{-16} (The choice of h going up to 10^{-16} was no coincidence)

When h is smaller than 10^{-8}, the differences $f(x_0+h)-f(x_0)$ approach the limits of floating-point precision. Consequently, the computation becomes dominated by numerical noise, which introduces errors.

Observe the plot of h versus the absolute error in this approximation



- **Taylor series**: Assume that f is a function that is $(k+1)$-differentiable on an interval containing x_0 and $x_0 + h$. Then

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2}f''(x_0) + \cdots + \frac{h^k}{k!}f^{(k)}(x_0) + \frac{h^{k+1}}{(k+1)!}f^{(k+1)}(\xi),$$

for some $\xi \in (x_0, x_0 + h)$.

- **Proof that the discretization error decreases at the same rate as**: Solving for $f'(x_0)$ in the Taylor series expansion, we get

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \left(\frac{h}{2}f''(x_0) + \frac{h^2}{6}f'''(x_0) + \cdots + \frac{h^{k-1}}{k!}f^{(k)}(\xi)\right).$$

Therefore,

$$\left| f'(x_0) - \frac{f(x_0+h)-f(x_0)}{h} \right| = \left| \frac{h}{2} f''(x_0) + \frac{h^2}{6} f'''(x_0) + \cdots + \frac{h^{k-1}}{k!} f^{(k)}(\xi) \right|.$$

If $f''(x_0) \neq 0$ and h is small, then the right-hand-side is dominated by $\frac{h}{2} f''(x_0)$. Thus,

$$\left| f'(x_0) - \frac{f(x_0+h)-f(x_0)}{h} \right| \approx \frac{h}{2} \left| f''(x_0) \right| = \mathcal{O}(h). \quad \blacksquare$$

Recall that $f(x) = \sin x$. Thus, $f''(x) = -\sin x$. Therefore,

$$\frac{|f''(x_0)|}{2} = \frac{-\sin 1}{2} = 0.42073549240394825\ldots$$

- **Roundoff error**: Numbers are stored in the computer using a finite precision representation. Roughly 16 digits of precision are possible using the 64-bit floating point format.

  Whenever an arithmetic operation takes place, the result must be rounded to roughly 16 digits of precision. Such an error is called roundoff error.

- **Accuracy**: As we have seen above, it is easy to write mathematically correct code that produces very inaccurate results.

  Accuracy is affected by the following two conditions:

  1. **Problem conditioning**: Some problems are highly sensitive to small changes in the input: we call such problems ill-conditioned. A problem that is not sensitive to small changes in the input is called well-conditioned. For example, computing $\tan(x)$ for x near $\frac{\pi}{2}$ is an ill-conditioned problem (Example 1.5 in Ascher-Greif).
  2. **Algorithm stability**: An algorithm is called stable if it is guaranteed to produce an exact answer to a slightly perturbed problem. (Example 1.6 in Ascher-Greif gives an example of an unstable algorithm).

     A "slightly perturbed problem" means a problem that has been altered by a small amount. For example, this could be small changes in the input data due to measurement errors or rounding errors.

     The algorithm is said to be stable if it provides the exact solution to this slightly perturbed problem. In other words, the output corresponds to what would happen if you solved the slightly modified problem exactly, rather than the original unmodified problem.

     A stable algorithm ensures that the effects of small input errors or numerical approximations (like rounding) do not grow uncontrollably during computations.

- **Unstable algorithm example**: Let

$$y_n = \int_0^1 \frac{x^n}{x+10} dx.$$

Then

$$y_n + 10y_{n-1} = \int_0^1 \frac{x^n + 10x^{n-1}}{x+10} dx = \int_0^1 x^{n-1} dx = \frac{1}{n} x^n \Big|_0^1 = \frac{1}{n}$$

and

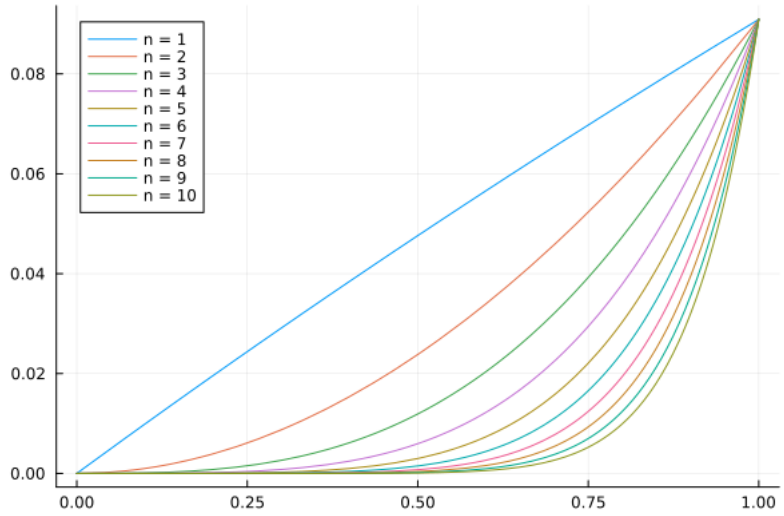
$$y_0 = \int_0^1 \frac{1}{x+10} dx = \ln|x+10| \Big|_0^1 = \ln(11) - \ln(10).$$

Then use these formulas to numerically compute y_{30}.

First, let's look at the functions $\frac{x^n}{x+10}$ as n grows. Using the Julia code

```
plot()
for n in 1:10
    plot!(x -> x^n/(x+10), 0, 1, label="n = $n")
end
plot!()
```

We get the graphs



So it appears the area under the curve is approaching zero. First, let's look at the integral result and error from a known stable algorithm (for $n = 30$)

```
using QuadGK

n = 30
integral, error = quadgk(x -> x^n/(x + 10), 0, 1)

#(0.002940928704861327, 8.45119305817703e-12)
```

The Julia code that uses the derived algorithm gives

```
        y0 = log(11) - log(10)

        yvals = zeros(30)
        yvals[1] = 1 - 10*y0
        for n = 2:30
            yvals[n] = 1/n - 10*yvals[n-1]
        end
        yvals

        # Out
        30-element Vector{Float64}:
       0.04689820195675232
       0.031017980432476833
       0.023153529008564988
       0.01846470991435012
       0.015352900856498819
       0.013137658101678468
       0.011480561840358172
       0.010194381596418278
       0.009167295146928323
       0.00832704853071678
       0.007638605601923115
       0.006947273141021765
       0.007450303782055162

     916.9927348292546
   -9169.877348292546
   91698.82110197308
 -916988.1655651854
       9.169881699130116e6
      -9.169881694963449e7
       9.169881695363449e8
      -9.169881695324987e9
       9.169881695328691e10
      -9.169881695328334e11
       9.16988169532837e12
      -9.169881695328366e13
```

Thus, This algorithm is *very* **unstable**. The reason is the computation of y_0 using the log function. The log function introduces some roundoff error. Continuously using the results of the previous introduces more and more roundoff error.

- **Efficiency**: The efficiency of a code is affected by many factors:
  1. the rate of convergence of the method
  2. the number of arithmetic operations performed
  3. how the data in memory is accessed
- **Robustness (Reliability)**: We want to ensure that our code works under *all possible inputs*, and generates the clear warnings when it is not possible to produce an accurate result for some input.

5.2 Roundoff errors

- **Real numbers stored on a computer**: Real numbers are stored on a computer following the IEEE floating-point standard:

  1. **half precision:** using 16 bits (Julia type: 'Float16')
  2. **single precision:** using 32 bits (Julia type: 'Float32')
  3. **double precision:** using 64 bits (Julia type: 'Float64')

  Julia also has an *arbitrary precision* floating-point data type called 'BigFloat'. It is excellent if you need more precision, but it is also much slower.

  Julia has the type *AbstractFloat*, which is a subtype of the class Real, and is an abstract supertype for all floating point numbers.

```
AbstractFloat <: Real

> subtypes(AbstractFloat)
 5-element Vector{Any}:
 BigFloat
 Core.BFloat16
 Float16
 Float32
 Float64
```

- **Description of the IEEE Float64**: Suppose x is a floating-point number stored in the following 64-bits:

1	2	$\cdots$	12	13	$\cdots$	64
s	e_{10}	$\cdots$	e_0	f_1	$\cdots$	f_{52}

  Where

  – 1 bit s represents the **sign**

  – 11 bits $e_{10}\cdots e_0$ represent the **exponent**

  – 52 bits $f_1\cdots f_{52}$ represent the **fraction** (a.k.a. the mantissa or significand)

  Then

$$x = (-1)^s[1.f_1\cdots f_{52}]_2 \times 2^{(e-1023)}.$$

  Notes:

  – x is **normalized** to have its first digit nonzero.

  – $e = [e_{10}\cdots e_0]_2 = e_{10}2^{10} + \cdots + e_1 2^1 + e_0 2^0 \in \left[0, 2^{11}-1\right] = [0, 2047]$

  – $e = 0$ and $e = 2047$ are reserved for special floating-point values, so

$$e \in [1, 2046]$$

  The "-1023" in the exponent is called the **bias**: $e - 1023 \in [-1022, 1023]$

  Also,

$$[1.f_1\cdots f_{52}]_2 = 1 + \frac{f_1}{2^1} + \frac{f_2}{2^2} + \cdots + \frac{f_{52}}{2^{52}}$$

For example, suppose

$$\begin{aligned} x &= -[1.101101]_2 \times 2^{(1026-1023)} \\ &= -[1.101101]_2 \times 2^3 \\ &= -[1101.101]_2 \\ &= -\left(1 \cdot 8 + 1 \cdot 4 + 0 \cdot 2 + 1 \cdot 1 + 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{8}\right) \\ &= -13.625 \end{aligned}$$

Even if a number can be represented exactly in base-10 with a finite number of digits, it may require an infinite number of digits in base-2.

$$0.1 = [0.000110011001\ldots]_2 = \left[1.\overline{1001}\right]_2 \times 2^{-4}$$

Therefore, 0.1 cannot be represented exactly as a floating-point number.

- **16-bit and 32-bit IEEE representation**
  - **16-bit (half-precision)**: The IEEE 754 half-precision floating-point format consists of:
    - 1 bit for the sign (s)
    - 5 bits for the exponent (e_4, e_3, e_2, e_1, e_0)
    - 10 bits for the fraction/mantissa ($f_1, f_2, \ldots, f_{10}$)

1	2	3	4	5	6	$\cdots$	16	
s	e_4	e_3	e_2	e_1	e_0	f_1	$\cdots$	f_{10}

  - **32-bit (Single precision)**: The IEEE 754 single-precision floating-point format consists of:
    - 1 bit for the sign (s)
    - 8 bits for the exponent ($e_7, e_6, e_5, e_4, e_3, e_2, e_1, e_0$)
    - 23 bits for the fraction/mantissa ($f_1, f_2, \ldots, f_{23}$)

1	2	3	4	5	6	7	8	9	10	$\cdots$	32
s	e_7	e_6	e_5	e_4	e_3	e_2	e_1	e_0	f_1	$\cdots$	f_{23}

- **Bias calculation**: The bias used above is calculated as

$$\text{Bias} = 2^{(E-1)} - 1$$

where E is the number of bits allocated to the exponent field.

  - **16-bit half precision**: Exponent is allowed 5 bits, thus

$$\text{Bias} = 2^{5-1} - 1 = 15$$

  - **32-bit single precision**: Exponent is allowed 8 bits, thus

$$\text{Bias} = 2^7 - 1 = 127$$

  - **64-bit double precision**: Exponent is allowed 11 bits, thus

$$\text{Bias} = 2^{10} - 1 = 1023$$

- **Convert real to binary representation**: So we now know how to convert a binary representation of a float to its decimal representation.

Consider the base ten real -13.625. To convert a base ten real into its binary representation, we first

Convert the integer part to binary: Following the standard algorithm to convert 13 to binary

$$\begin{aligned} 13 &= 2(6) + 1 : \ 1_2 \\ 6 &= 2(3) + 0 : \ 0_2 \\ 3 &= 2(1) + 1 : \ 1_2 \\ 1 &= 2(0) + 1 : \ 1_2 \end{aligned}$$

13_{10} is therefore 1101_2

Convert the fractional part: Convert the fractional part (0.625) by repeatedly multiplying by 2 and recording the whole number parts. Stop when the fractional part becomes 0. We build the resulting representation in the opposite way of the integer algorithm (top down)

$$\begin{aligned} 0.625 \cdot 2 &= 1.25 : \ 1_2 \\ 0.25 \cdot 2 &= 0.5 : \ 0_2 \\ 0.5 \cdot 2 &= 1 : \ 1_2 \end{aligned}$$

The result is therefore 101_2

Combine the integer and fractional parts: We have

$$13.625_{10} = 1101.101_2$$

Normalize the Binary Number: Normalize the binary number to the form

$$1.\text{mantissa} \cdot 2^{\text{exponent}}$$

For 1101.101_2, shift the decimal point left by three places to get 1.101101_2. The exponent is therefore three because

$$1101.101 = 1.101101 \cdot 2^3$$

Determine the Sign Bit: The sign bit is:

- 0 for positive numbers.
- 1 for negative numbers.

Since -13.625 is negative, the sign bit is 1

Encode the Exponent: The exponent is stored in "biased" form, for 64-bit double precision the bias is 1023. We add the bias to the actual exponent to get the biased to get the biases exponent

$$3 + 1023 = 1026$$

Then, convert the biased exponent to binary

$$1026_{10} = 10000000010_2$$

Encode the Mantissa: The mantissa is the fractional part of the normalized binary number (1.mantissa), excluding the leading 1.

From 1.101101_2, the mantissa is 101101. We pad with zeros to make it the required 52 bits

Combine the parts: The IEEE 754 representation is formed by combining

- **Sign bit (1 bit):** 1
- **Exponent (11 bits):** 10000000010
- **Mantissa (52 bits):** $1011010000000000000000...0_{52}$

- **Finding decimal value**: In a binary string of all ones, we have a mathematical formula to compute its value... Specifically

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

For example (unsigned), $1111 = 2^{3+1} - 1 = 15$. But what about for binary decimals numbers, where we instead divide by two to find its value. Consider the binary string

$$0.11111_2 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + ...$$

Notice this a geometric series

$$S = \sum_{k=1}^{n} \frac{1}{2^k}$$

The sum of the geometric series is given by

$$S = \frac{a(1 - r^n)}{1 - r}$$

Where a is the first term, r is the common ratio, and n is the number of terms. Therefore, for $a = \frac{1}{2}, r = \frac{1}{2}$, we have

$$\begin{aligned} S &= \frac{1}{2} \cdot \frac{1 - \left(\frac{1}{2}\right)^n}{1 - \left(\frac{1}{2}\right)} \\ &= 1 - \frac{1}{2^n} \end{aligned}$$

Thus, consider 0.1111, where $n = 4$. Note that n is not 3, like in the formula described for integers, this is because the sum starts at $n = 1$, instead of $n = 0$ like the integer formula. Thus,

$$0.1111 = 0 + 1 - \frac{1}{2^n} = 1 - \frac{1}{2^4} = 0.9375$$

And

$$1.1111 = 1 + 1 - \frac{1}{2^n} = 1 - \frac{1}{2^4} = 1 + 0.9375 = 1.9375$$

- **More on normalized float**: To be normalized in the context of IEEE 754 floating-point numbers means that the number is represented in a standardized form where:
  - The leading digit of the significand (mantissa) is always 1 (except for special cases like subnormal numbers).
  - This representation ensures there is only one unique way to represent each number.

This is akin to scientific notation, where we always write numbers like 3.25×10^2 instead of 32.5×10^1 or 0.325×10^3. In IEEE 754, normalized numbers always take the form:

$$1.\text{fraction} \times 2^{(\text{exponent}-\text{bias})}$$

- **Exponent of all ones**: All ones in the exponent is reserved for infinity and NaN. Thus, the largest exponent to work with in calculations is $11111111110 = 2046$
- **Limits of floating point numbers**: The largest Float64 is $(2 - 2^{-52}) \times 2^{1023} \approx 1.97769 \times 10^{308} \approx 2 \times 10^{308}$

  The largest float is when the sign bit is zero, all exponent bits are one, and all fraction bits are one. The exponent is then $2^{10+1} - 1 - 2047$, However, this value is reserved for infinity (and NaN). The largest finite exponent is 2046. The exponent bias is 1023, so the largest actual exponent $2046 - 1023 = 1023$. is The fractional part is $1.111...1_{52} = 1 + 1 - \frac{1}{2^{52}} = 2 - \frac{1}{2^{52}} = 2$. Thus, we get

  $$(-1)^0 \cdot 2 \cdot 2^{1023} = 2^{1024} \approx 2 \times 10^{308}$$

  Thus, the largest possible float64 is

  $$0\ 11111111110\ 111...1_{64}$$

  The smallest positive possible normalized float64 is $2^{-1022} \approx 2 \times 10^{-308}$, and it occurs when the sign bit is zero, the exponent is 00000000001 (all zeros reserved), and the fractional part is $1.000000...0_{52} = 1.0_{10}$. Thus, we get

  $$(-1)^0 \cdot 1.0 \cdot 2^{1-1023} = 2^{-1022} \approx 2.225 \times 10^{-308} \approx 2 \times 10^{-308}$$

  The smallest negative possible normalized float64 is then when the sign bit is one, we have

  $$(-1)^1 \cdot 1.0 \cdot 2^{-1022} \approx -2.225 \times 10^{-308} \approx -2 \times 10^{-308}$$

- **Finding these values in julia**: In julia, we have the functions

```
floatmax(T = Float64)
floatmin(T = Float64)
typemax(T)
```

- **Float overflow and underflow**: Floating-point overflow occurs when a calculation produces a number larger than the maximum representable value in the floating-point system. In IEEE 754 double precision (Float64), the largest finite number is $\approx 1.79769 \times 10^{308} \approx 2 \times 10^{308}$

  If an operation results in a number greater than this limit, IEEE 754 rules dictate that the result is represented as positive infinity

If the result is negative, it becomes negative infinity

- **De-normalized (subnormal) floating-point numbers**: The IEEE floating-point standard also allows de-normalized numbers that are smaller than $\pm 2^{-1022}$. De-normalized floats are represented by $e = 0$. Also note that subnormal floats have mantissa non-zero. Subnormal is therefore represented as

$$(-1)^s \cdot [0.f_1 f_2 ...]_2 \cdot 2^{-1022}$$

Note that the exponent is -1022 instead of -1023. This is due to the IEE convention for subnormal numbers to insure there is no gap between the largest subnormal number and the smallest normal number

The smallest positive subnormal float that is not zero is therefore

$$0\ 00000000000\ 000...01$$

And is equal to

$$(-1)^0 \frac{1}{2^{52}} \cdot 2^{-1022} \approx 4.94 \times 10^{-324} \approx 5 \times 10^{-324}$$

- **Other special floats**:
  - **0.0 and -0.0**:

$$e_{10}...e_0 = 0...0 \text{ and } f_1...f_{52} = 0...0$$

    If a very small negative number is rounded to zero, then it becomes -0.0. If a very small positive number rounds to zero, it becomes 0.0
  - **Inf and -Inf**:

$$e_{10}...e_0 = 1...1 \text{ and } f_1...f_{52} = 0...0$$

  - **Nan**:

$$e_{10}...e_0 = 1...1 \text{ and } f_1...f_{52} \neq 0$$

Note: $0.0, -0.0, \infty, -\infty, NaN$ are neither normal or subnormal

Also, from Julia

  - Finite numbers are ordered in the usual manner.
  - Positive zero is equal but not greater than negative zero.
  - Inf is equal to itself and greater than everything else except NaN.
  - -Inf is equal to itself and less then everything else except NaN.
  - NaN is not equal to, not less than, and not greater than anything, including itself.

- **Summary (Float64)**
  - **0.0 and -0.0**:

$$0\ 00000000000\ 000...01\ 00000000000\ 000...0$$

- **Smallest positive normal**

$$0\ 00000000001\ 000...00$$

And has value

$$(-1)^0[1.000...0]_2 \cdot 2^{1-1023} = 1.0 \cdot 2^{-1022} \approx 2.225 \cdot 10^{-308}$$

- **Largest positive normal**: Occurs when

$$0\ 11111111110\ 111...11$$

And has value

$$(-1)^0[1.111...11]_2 \cdot 2^{2^{10+1}-1-1-1023}$$
$$= 1 + 1 - \frac{1}{2^{52}} \cdot 2^{2046-1023} = 2 - \frac{1}{2^{52}} \cdot 2^{1023} \approx 1.797 \times 10^{308}$$

- **Smallest positive subnormal**

$$0\ 00000000000\ 000...01$$

And has value

$$(-1)^0 \cdot \frac{1}{2^{52}} \cdot 2^{-1022} \approx 4.94 \times 10^{-324}$$

- **Largest positive subnormal**

$$0\ 00000000000\ 111..11$$

$$(-1)^0 \cdot 1 - \frac{1}{2^{52}} \cdot 2^{-1022} \approx 2.225 \times 10^{-308}$$

- ∞ **and** $-\infty$

$$0\ 11111111111\ 000...00$$
$$1\ 11111111111\ 000...00$$

- **NaN**: Any sign bit, exponent all ones, any nonzero combination of fractional bits.

Therefore, we can also derive the largest and smallest negative normals and subnormals

- **Largest negative normal**

$$1\ 00000000001\ 000...00$$

And has value

$$(-1)^1[1.000...0]_2 \cdot 2^{1-1023} = (-1)1.0 \cdot 2^{-1022} \approx -2.225 \cdot 10^{-308}$$

- **Smallest negative normal**: Occurs when

$$1\ 11111111110\ 111...11$$

And has value

$$(-1)^1[1.111...11]_2 \cdot 2^{2^{10+1}-1-1-1023}$$
$$= (-1)1 + 1 - \frac{1}{2^{52}} \cdot 2^{2046-1023} = (-1)2 - \frac{1}{2^{52}} \cdot 2^{1023} \approx -1.797 \times 10^{308}$$

- **Largest negative subnormal**

  $$1\ 00000000000\ 000...01$$

  And has value

  $$(-1)^1 \cdot \frac{1}{2^{52}} \cdot 2^{-1022} \approx -4.94 \times 10^{-324}$$

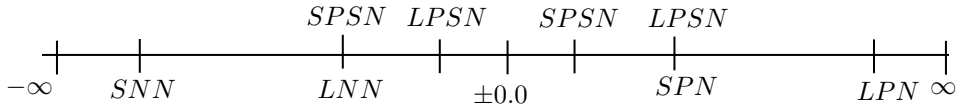
- **Smallest negative subnormal**

  $$1\ 00000000000\ 111..11$$

  $$(-1)^1 \cdot 1 - \frac{1}{2^{52}} \cdot 2^{-1022} \approx -2.225 \times 10^{-308}$$

  Notes: Float underflow takes you to -inf, float overflow takes you to +inf. A negative number rounded to zero will be -0.0, a positive number rounded to zero will be 0.0

  Behold



- **nextfloat and prevfloat in Julia**: In Julia, we can find the next float and the previous float with

```
nextfloat(f)
prevfloat(f)
```

- **Bounding of significand in a normalized number**: The significand is always in the range

  $$1 \leqslant 1.b_1 b_2 b_3 ... b_{t-1} < 2$$

  since the leading bit is always 1 in a normalized number.

  This means that x satisfies

  $$2^e \leqslant |x| < 2^{e+1}$$

  We sometimes take $|x| \approx 2^e$

- **Machine epsilon**: In the IEEE 754 floating-point standard, machine epsilon (denoted as ϵ_{mach}) is the smallest positive number that, when added to 1, results in a different representable floating-point number. It represents the upper bound on relative error due to rounding in floating-point arithmetic. That is

  $$1 + \epsilon \neq 1$$

For a floating-point system with t bits in the significand (mantissa) (including the implicit leading 1), the machine epsilon is:

$$e_{\text{mach}} = 2^{-(t-1)}$$

For the IEEE 754 double precision (64-bit) format, the mantissa has 52 fractional bits and one implicit leading bit. Thus, the machine epsilon is

$$2^{-(53-1)} = 2^{-52} \approx 2.2204 \times 10^{-16}$$

In Julia, we can find the machine epsilon with

```
eps(Float64)
```

Any rounding error in IEEE floating-point arithmetic is bounded above by ϵ_{mach}

- **Understanding machine epsilon**: In an IEEE 754 floating-point number with t bits in the significand (including the implicit 1), the machine epsilon is defined as

$$\epsilon_{\text{mach}} = 2^{-(t-1)} = \frac{1}{2^{t-1}}$$

A floating-point number is stored in normalized form

$$x = 1.b_1 b_2 b_3 ... b_{t-1} \times 2^e$$

The gap between two consecutive representable floating-point numbers is determined by the last bit of the significand

$$\text{Unit gap } = 2^{e-(t-1)}$$

This is because the smallest possible difference in the mantissa is $2^{-t(t-1)}$, which gets scaled by 2^e

When rounding to the nearest floating-point number, the maximum error occurs when x falls exactly between two consecutive representable numbers.

Since the gap between two consecutive numbers is

$$2^{e-(t-1)}$$

the maximum absolute rounding error is

$$\frac{1}{2} \times 2^{e-(t-1)}$$

because the number is rounded to the nearest representable value.

The relative error is given by

$$\frac{\text{max absolute rounding error}}{|x|}$$

Since $|x| \approx 2^e$ in normalized form

$$\frac{\frac{1}{2}2^{e-(t-1)}}{e} = \frac{1}{2} \times 2^{-(t-1)}$$

Since $\epsilon_{\text{mach}} = 2^{-(t-1)}$, we conclude

$$\text{Relative rounding error} \leqslant \frac{1}{2}\epsilon_{\text{mach}}$$

- **Unit roundoff** η: We define the unit roundoff $\eta = \frac{\epsilon}{2.0}$, and it is the largest possible relative error due to roundoff

$$\eta = 2^{-53} \approx 1.1 \times 10^{-16}$$

The unit roundoff represents the maximum rounding error in a floating-point system when using round-to-nearest mode

(the default in IEEE 754). This is because rounding introduces an error at most half the distance between two consecutive representable floating-point numbers.

- **Roundoff error example**: Suppose we are using a base-10 floating-point system with 4 significant digits, using 'RoundNearest':

$$\begin{aligned}\left(1.112 \times 10^1\right) \times \left(1.112 \times 10^2\right) &= 1.236544 \times 10^3 \\ &\to 1.237 \times 10^3 = 1237\end{aligned}$$

The absolute error is $1237 - 1236.544 = 0.456$.

The relative error is

$$\frac{0.456}{1236.544} \approx 0.0004 = 0.04\%$$

The default rounding mode is 'RoundNearest' (round to the nearest floating-point number). This implies that

$$\frac{|x - \text{fl}(x)|}{|x|} \leqslant \eta.$$

If 'RoundToZero' is used (a.k.a. **chopping**), then

$$\frac{|x - \text{fl}(x)|}{|x|} \leqslant 2\eta.$$

'RoundNearest' is used since it produces smaller roundoff errors.

- **Roundoff error accumulation**: When performing arithmetic operations on floats, extra **guard digits** are used to ensure **exact rounding**. This guarantees that the relative error of a floating-point operation (**flop**) is small. More precisely, for floating-point numbers x and y, we have

$$\begin{aligned}\text{fl}(x \pm y) &= (x \pm y)(1 + \varepsilon_1) \\ \text{fl}(x \times y) &= (x \times y)(1 + \varepsilon_2) \\ \text{fl}(x \div y) &= (x \div y)(1 + \varepsilon_3)\end{aligned}$$

where $|\varepsilon_i| \leqslant \eta$, for $i = 1, 2, 3$, where η is the unit roundoff.

Although the relative error of each flop is small, it is possible to have the roundoff error accumulate and create significant error in the final result. If E_n is the error after n flops, then:

  - **Linear roundoff error accumulation** is when $E_n \approx c_0 n E_0$
  - **Exponential roundoff error accumulation** is when $E_n \approx c_1^n E_0$, for some $c_1 > 1$

In general, linear roundoff error accumulation is unavoidable. On the other hand, exponential roundoff error accumulation is not acceptable and is an indication of an **unstable algorithm**. (See Example 1.6 in Ascher-Greif for an example of exponential roundoff error accumulation, and see Exercise 5 in Section 1.4 for a numerically stable method to accomplish the same task.)

- **General advice**:
  1. Adding $x + y$ when $|x| \gg |y|$ can cause the information in y to be "lost" in the summation.
  2. Dividing by very small numbers or multiplying by very large numbers can **magnify error**.
  3. Subtracting numbers that are almost equal produces **cancellation error**.
  4. An **overflow** occurs when the result is too large in magnitude to be representable as a float. The result will become either `Inf` or `-Inf`. Overflows should be avoided.
  5. An **underflow** occurs when the result is too small in magnitude to be representable as a float. The result will become either `0.0` or `-0.0`.
- **Information lose example**: This example shows that the summation order can make a difference. Consider the sum

$$s = \sum_{n=1}^{1,000,000} \frac{1}{n}$$

Let's do this sum in two different ways. First, from largest to smallest, then from smallest to largest.

Largest to smallest

```
sum = 0
for i in 1:1000000
    sum+=1/i
end
# 14.392726722864989
```

Smallest to largest

```
sum = 0
for i in 1000000:-1:1
    sum += 1/i
end
# 14.392726722865772
```

We can do the computation with a BigFloat to see which one is more precise

```
bfsum::BigFloat = 0
for i in 1:1000000
    sum+=BigFloat(1)/i
end
# 14.392726722865723631381127493188587676644800013744311653418433045812958507511 94
```

So we see the smallest to largest sum is slightly more accurate. Why is this? When summing a sequence of numbers, the order in which the numbers are added affects how rounding errors accumulate. Adding smaller numbers to a large sum can cause the smaller values to be "swallowed" due to the limitations of floating-point precision.

When summing from large terms to small terms, the large values dominate early in the computation. Because floating-point numbers have limited precision, adding much smaller numbers later may result in those numbers being effectively ignored (i.e., they contribute little due to rounding errors).

When summing from small terms to large terms, the intermediate sum stays small for longer, allowing more precise accumulation of the smaller values before reaching the larger ones. This reduces the impact of rounding errors.

- **Cancellation error example**: We consider

$$\ln\left(x-\sqrt{x^2-1}\right) = -\ln\left(x+\sqrt{x^2-1}\right)$$

First, we show that these expressions are actually equivalent

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

Thus,

$$\begin{aligned} \ln\left(x-\sqrt{x^2-1}\right) &= \ln\left(\left(x+\sqrt{x^2-1}\right)^{-1}\right) \\ &= -\ln\left(x+\sqrt{x^2-1}\right) \end{aligned}$$

Let's see which one is more accurate in numerical computations.

```
x = 1e6
fl = log(x-sqrt(x^2 -1))
fr = -log(x+sqrt(x^2 -1))
fl,fr

# (-14.50865012405984, -14.508657738523969)
```

If we examine the quantities x, and $\sqrt{x^2-1}$, we see that the are very close to each other

```
x, sqrt(x^2-1)
# (1.0e6, 999999.9999995)
```

The first expression $\ln\left(x-\sqrt{x^2-1}\right)$ gives cancellation error, which happens when two nearly equal floating-point numbers are subtracted, leading to significant loss of precision.

When two very close numbers are subtracted, the leading digits cancel out, leaving only a small result with much fewer significant digits. This makes the result inaccurate.

- **Example: Avoiding overflow**: Overflow is possible when squaring a large number. This needs to be avoided when computing the Euclidean norm (a.k.a. the 2-norm) of a vector x:

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

If some x_i is very large, it is possible that x_i^2 will overflow, causing the final result to be **Inf**. We can avoid this as follows.

Let

$$\bar{x} = \max_{i=1:n} |x_i|.$$

Then

$$\|x\|_2 = \bar{x}\sqrt{\left(\frac{x_1}{\bar{x}}\right)^2 + \left(\frac{x_2}{\bar{x}}\right)^2 + \cdots + \left(\frac{x_n}{\bar{x}}\right)^2}.$$

Since $|x_i/\bar{x}| \leqslant 1$ for all i, no overflow will occur. Underflow may occur, but this is harmless.

5.3 Non linear equations in one variable

- **Julia LaTeX strings**: In a Julia REPL or jupyter notebooks, we can include the LatexStrings package. This package allows us to create strings that contain LaTeX and format them. We do this by creating a string like

  L"string contents"

  For example

  L"$\int f(x)\ dx$"

```
using LatexStrings
```

- **Intro**: In many applications, one needs the solution to a **nonlinear equation** for which there is no closed formula.

  Suppose you do not have a cube-root function, but only the operations $+$, $-$, $\times$, $\div$

  Polynomials with degree at least five have no general algebraic solution

  Some nonlinear equations may not be solved analytically, for example

  $$10\cosh\left(\frac{x}{4}\right) = x \quad \text{and } 2\cosh\left(\frac{x}{4}\right) = x$$

  Recall the hyperbolic sine, cosine, and tangent functions are defined as

  $$\sinh(t) = \frac{e^t - e^{-t}}{2}$$
  $$\cosh(t) = \frac{e^t + e^{-t}}{2}$$
  $$\tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}}$$

  Also, $\tanh(t) = \frac{\sinh(t)}{\cosh(t)}$, $\frac{d}{dt}\sinh(t) = \cosh(t)$, and $\frac{d}{dt}\cosh(t) = \sinh(t)$

- **Problem statement: roots**: Given $f \in C[a,b]$ (i.e., a *continuous* function $f\colon [a,b] \to \mathbb{R}$) and we want to find $x^* \in [a,b]$ such that

  $$f(x^*) = 0.$$

  The solution x^* is called a **root** or **zero** of the function f. There could be exactly one root, many roots, or no roots at all.

- **The Julia Roots package**: In Julia, the package *Roots* gives us functions like `find_zero` to find or approximate the roots of an equation

```
using Pkg
Pkg.add("Roots")
using Roots
```

Consider the functions

```
f(x) = 10cosh(x/4) - x
g(x) = 2cosh(x/4) - x
```

We can first plot the functions to get an tight interval that contains a root

```
plot(axes_style=:zerolines, xlims=[-2,12], xlabel=L"x",
↪  ylabel=L"y")
plot!(f, -2, 12, label=L"y = 10\cosh(x/4) - x")
plot!(g, -2, 12, label=L"y = 2\cosh(x/4) - x")
```

Using the function `find_zero`, passing in a function and an interval, we can find approximate or exact roots

```
x1 = find_zero(g, (2,3))
# 2.357551053877402

g(x1) # 0.0
```

- **Iterative methods**: Often there is no closed formula for a root x^* of the function f. Instead of using a formula to compute a root x^*, we will start with an **initial guess** x_0 and generate a **sequence of iterates**

$$x_1, x_2, x_3, \ldots, x_k, \ldots$$

that we hope **converges** to x^*; i.e.,

$$\lim_{k\to\infty} x_k = x^*$$

Note: Different initial guesses x_0 may generate sequences of iterates that converge to different roots. We will see how to deal with this issue.

- **Iterative methods: When to stop**: Since the sequence of iterates is infinite, we must decide when we are close enough to a root x^* . However, we do not know , so how can we decide when we are close enough?

Stop options are to stop when

1. The function value is small:
$$|f(x_k)| < \texttt{ftol}.$$
A problem with this test is that $|f(x_k)|$ may be very small although x_k is still very far from a root.
2. Consecutive iterates are very close to each other:
$$|x_k - x_{k-1}| < \texttt{atol}.$$
A problem with this test is that *atol* must take into account the magnitude of the iterates.

3. Consecutive iterates are **relatively** close to each other:

$$|x_k - x_{k-1}| < \texttt{rtol}\,|x_k|\,.$$

Usually this is more robust than the above absolute test.

Often a combination of the above conditions is used. For example, items 2 and 3 can be combined:

$$|x_k - x_{k-1}| < \texttt{tol}(1 + |x_k|).$$

- **Intermediate value theorem**: If $f \in C[a,b]$ and $f(a) \leqslant s \leqslant f(b)$, then there exists a real number $c \in [a,b]$ such that $f(c) = s$.
- **Bisection method**: Suppose $f \in C[a,b]$ and that $f(a)$ and $f(b)$ have opposite signs; i.e.,

$$f(a) \cdot f(b) < 0.$$

Recall the IVT from calculus

If $f \in C[a,b]$ and $f(a) \leqslant s \leqslant f(b)$, then there exists a real number $c \in [a,b]$ such that $f(c) = s$.

Since f changes sign over $[a,b]$, the Intermediate Value Theorem implies that there is some $x^* \in [a,b]$ such that $f(x^*) = 0$. The **bisection method** searches for a root of f in $[a,b]$ as follows.

1. Let $p = \frac{a+b}{2}$ be the **midpoint** of $[a,b]$.
2. If $f(a) \cdot f(p) < 0$, then there is a root in $[a,p]$.
3. If $f(a) \cdot f(p) = 0$, then p is a root.
4. If $f(a) \cdot f(p) > 0$, then there is a root in $[p,b]$.

Each time we apply the above, we get a subinterval that contains a root that is **half the size** of the interval $[a,b]$.

Consider the Julia code for the bisection method

```
function bisect(f, a, b; maxiters=1000, tol=1e-6)
    fa, fb = f(a), f(b)

    if fa * fb > 0
        error("f(a) and f(b) must have opposite signs")  # Ensure root exists
    end

    for i in 1:maxiters
        p = (a + b) / 2
        fp = f(p)

        if abs(fp) < tol  abs(b - a) < tol  # Stop if function value is small or interval is tiny
            return p
        elseif fa * fp < 0
            b, fb = p, fp
        else
            a, fa = p, fp
        end
    end

    return (a + b) / 2  # Return best approximation if maxiters is reached
end

f(x) = 2cosh(x/4) - x
a, b = 5.0, 10.0

p = bisect(f, a, b, tol=1e-6)
p, f(p)
```

The example

```
f(x) = 2cosh(x/4) - x
a, b = 5.0, 10.0

p = bisect(f, a, b, tol=0.0)
p, f(p)

# (8.507199570713027, 1.7763568394002505e-15)
```

Shows that we get a pretty good approximation

- **Analyzing the bisection method**: Initially, we know a root x^* is somewhere in the interval $[a, b]$. If we let x_k be the midpoint of the kth subinterval, then

$$|x^* - x_0| \leqslant \frac{b-a}{2}.$$

In the next iteration,

$$|x^* - x_1| \leqslant \frac{b-a}{4},$$

and in the following iteration,

$$|x^* - x_2| \leqslant \frac{b-a}{8},$$

and so on, each time reducing our error bound by a factor of 2. In general,

$$|x^* - x_k| \leqslant \frac{b-a}{2} \cdot 2^{-k}, \qquad \text{for } k = 0, 1, 2, \ldots.$$

Suppose we want to compute x_k such that

$$|x^* - x_k| \leqslant \texttt{atol}.$$

Then we just need to find the smallest positive integer k such that

$$\frac{b-a}{2} \cdot 2^{-k} \leqslant \texttt{atol}.$$

That is,

$$\frac{b-a}{2\texttt{atol}} \leqslant 2^k,$$

which gives us

$$\log_2\left(\frac{b-a}{2\texttt{atol}}\right) \leqslant k,$$

so we just need the first integer k that is larger than $\log_2\left(\frac{b-a}{2\texttt{atol}}\right)$. Therefore,

$$k = \left\lceil \log_2\left(\frac{b-a}{2\texttt{atol}}\right) \right\rceil.$$

- **Pros and cons of the bisection method**:
- **Pros:**

  1. **Simple:** The bisection method only requires function values, is easy to understand and implement, and it is easy to analyze.
  2. **Robust:** The bisection method is guaranteed to work, provided that f is continuous and changes sign on the interval $[a, b]$.

  Cons:

  1. **Slow to converge:** The bisection method often requires many function evaluations.
  2. **Does not generalize:** The bisection method only applies to solving equations involving one variable; it does not generalize to solving equations involving multiple variables.

- **Fixed point iteration**: Another simple approach to solving

  $$f(x) = 0$$

  is to re-write it as

  $$x = g(x)$$

  for some continuous function g. We call a point x a **fixed-point** of g if $x = g(x)$.

For example, If we let

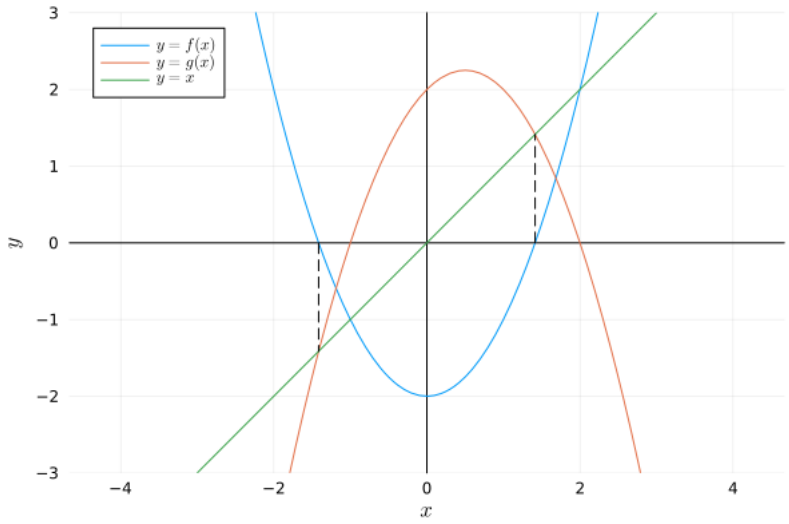
$$g(x) = x - f(x),$$

then

$$x = g(x) \quad \Rightarrow \quad x = x - f(x) \quad \Rightarrow \quad f(x) = 0.$$

Let's plot these functions using $f(x) = x^2 - 2$.

```
f(x) = x^2 - 2
g(x) = x-f(x)
a, b = -3.0, 3.0

plot(axes_style=:zerolines, aspect_ratio=:equal, legend=:topleft, ylims=[-3,3])
plot!(f, a, b, label=L"y = f(x)", c=1)
plot!(g, a, b, label=L"y = g(x)", c=2)
plot!(x -> x, a, b, label=L"y = x", c=3)
plot!([-sqrt(2), -sqrt(2)], [0, -sqrt(2)], linestyles=:dash, color=:black, label=:none)
plot!([sqrt(2), sqrt(2)], [0, sqrt(2)], linestyles=:dash, color=:black, label=:none)
xlabel!(L"x"); ylabel!(L"y")
```



We see that $f(x) = 0$ precisely when $g(x) = x$ (notice when the orange curve intersects the line $y = x$, it traces back up to the root of the blue curve)

- **Fixed point iteration: Choices of** g: There are many possible choices for g:
  - $g(x) = x - f(x)$
  - $g(x) = x + cf(x)$, for some nonzero constant c
  - $g(x) = x - f(x)/f'(x)$

  Some choices for g will be better than others.

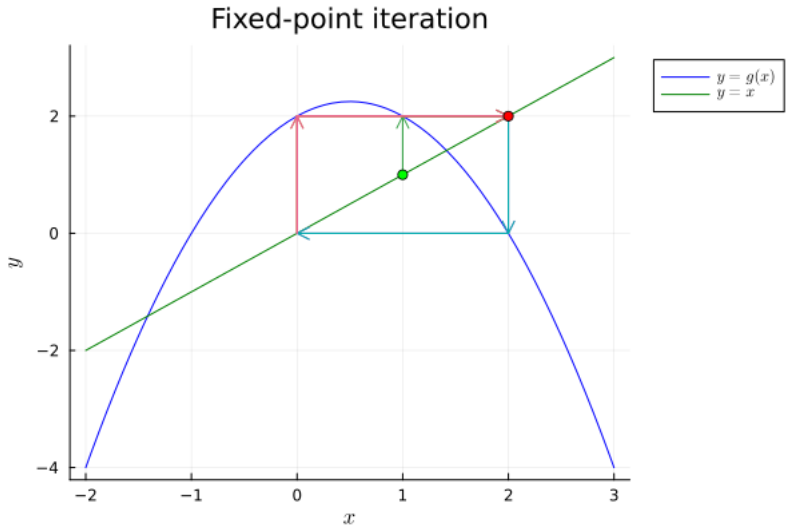
- **Iterations with fixed point iteration**: Given some initial guess x_0, we can use the function g to generate a sequence of iterates as follows:

$$x_{k+1} = g(x_k), \qquad k = 0, 1, 2, \ldots.$$

If the sequence $\{x_k\}$ converges to some point x^*, then we must have $x^* = g(x^*)$, so $f(x^*) = 0$.

Consider the Julia function that runs the above algorithm for $f(x) = x^2 - 2$, and $g(x) = x - f(x)$

```
function fixedPointPlot(g, a, b, x0; num=5, usequiver=true)

    plt = plot(g, a, b, label=L"y = g(x)", color=:blue)
    plot!(x -> x, a, b, label=L"y = x", color=:green)

    x = x0
    for i = 1:num
        if usequiver
            quiver!([x, x], [x, g(x)],
                quiver=([0, g(x)-x], [g(x)-x, 0]))
        else
            plot!([x, x], [x, g(x)], color=i, label=:none)
            plot!([x, g(x)], [g(x), g(x)], color=i,
↪   label=:none)
        end
        x = g(x)
    end

    scatter!([x0], [x0], label=:none, color=:lime)
    scatter!([x], [x], label=:none, color=:red)

    xlabel!(L"x")
    ylabel!(L"y")
    plot!(legend=:outertopright)
    title!("Fixed-point iteration")

    return plt
end

g1(x) = x - f(x)
fixedPointPlot(g1, -2, 3, x0, num=5)
```



We can examine k, and x_k

```
using Printf
x0 = 1.0; x = x0
@printf("%4s %12s\n", "k", "xk")
for k = 1:20
    x = g1(x)
    @printf("%4d %12.4e\n", k, x)
end
```

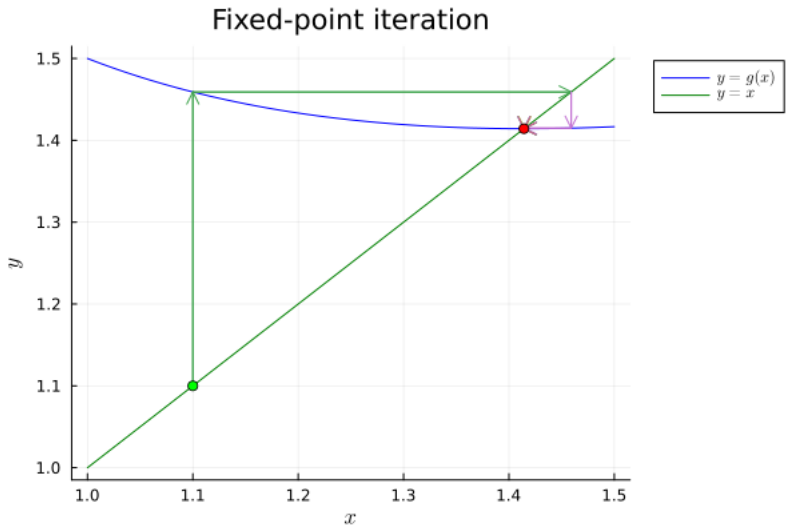
Which gives the table

k	x_k
1	2.0000e+00
2	0.0000e+00
3	2.0000e+00
4	0.0000e+00
5	2.0000e+00
6	0.0000e+00
7	2.0000e+00
8	0.0000e+00
9	2.0000e+00
10	0.0000e+00
11	2.0000e+00
12	0.0000e+00
13	2.0000e+00
14	0.0000e+00
15	2.0000e+00
16	0.0000e+00
17	2.0000e+00
18	0.0000e+00
19	2.0000e+00
20	0.0000e+00

It does not seem to be converging to anything. Now let's try $g(x) = x - f(x)/f'(x)$.

```
fixedPointPlot(x -> x-f(x)/2x, 1.0, 1.5, 1.1, usequiver=true)
```

Gives the plot



We can also examine the absolute errors

```
x = 1.0; xs = sqrt(2)

@printf("%4s %12s\n", "k", "error")
for k = 1:5
    x = (x->x-f(x)/2x)(x)
    @printf("%4d %12.4e\n", k, x - xs)
end
```

k	error
1	8.5786e-02
2	2.4531e-03
3	2.1239e-06
4	1.5947e-12
5	0.0000e+00

It converges very rapidly! We will see later why this is happening.

- **Mean value theorem**: If $f \in C[a,b]$ and f is differentiable on the open interval (a,b), then there exists a number $c \in (a,b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

Which means there exists some point c in which the tangent line at c is equal to the secant line drawn connecting points a and b

- **Existence and uniqueness of a fixed point**: A fixed point may not exist in $[a, b]$, and if it does, it may not be unique.

  Fixed point theorem: Let $g \in C[a, b]$ such that one of the two following conditions hold:

  1. $g(a) \geqslant a$ and $g(b) \leqslant b$;
  2. $g(a) \leqslant a$ and $g(b) \geqslant b$.

  Then $\exists x^* \in [a, b]$ such that $g(x^*) = x^*$. In addition, if g is differentiable on the open interval (a, b) and

  $$|g'(x)| \leqslant \rho, \quad \forall x \in (a, b),$$

  for some $\rho < 1$, then x^* is the *unique* fixed point in $[a, b]$.

  Proof. Suppose $g(a) \geqslant a$ and $g(b) \leqslant b$. If $g(a) = a$ or $g(b) = b$, then we are done. Otherwise we have $g(a) > a$ and $g(b) < b$. Let

  $$\phi(x) = g(x) - x.$$

  Then $\phi(a) > 0$ and $\phi(b) < 0$. Thus, since ϕ is continuous, the **Intermediate Value Theorem** tells us that there is an $x^* \in [a, b]$ such that $\phi(x^*) = 0$. Thus $x^* = g(x^*)$.

  The other case of $g(a) \leqslant a$ and $g(b) \geqslant b$ can be proven similarly.

  Now suppose g is differentiable and there is a $\rho < 1$ such that $|g'(x)| \leqslant \rho$ for all $x \in (a, b)$. Suppose, **for the sake of contradiciton**, that x^* is not the only fixed point of g in $[a, b]$. Then, there is a $y^* \in [a, b]$ such that $g(y^*) = y^*$ and $y^* \neq x^*$.

  By the **Mean Value Theorem**, there is a ξ strictly between x^* and y^* such that

  $$g'(\xi) = \frac{g(x^*) - g(y^*)}{x^* - y^*} = \frac{x^* - y^*}{x^* - y^*} = 1.$$

  Note that $\xi \in (a, b)$. This contradicts our assumption that $|g'(x)| \leqslant \rho < 1$, for all $x \in (a, b)$. Therefore, the fixed point of g in $[a, b]$ must be unique. ■

- **Convergence**: We have seen that the fixed point iteration does not always converge.

  Theorem: (Convergence of the Fixed Point Iteration): Let $g \in C[a, b]$. If

  - $a \leqslant g(x) \leqslant b$, for all $x \in [a, b]$, and
  - there is a $\rho < 1$ such that $|g'(x)| \leqslant \rho$ for all $x \in (a, b)$,

  then the iteration

  $$x_{k+1} = g(x_k), \qquad k = 0, 1, 2, \ldots$$

  converges to the unique fixed point $x^* \in [a, b]$ starting from any $x_0 \in [a, b]$.

  Proof. First of all, since $g(x) \in [a, b]$ for all $x \in [a, b]$, and since $x_0 \in [a, b]$, we have $x_k \in [a, b]$, for all $k = 0, 1, 2, \ldots$. Moreover, by the **Fixed Point Theorem**, our assumptions imply that there is a unique fixed point $x^* \in [a, b]$.

Let $k \in \{1, 2, \ldots\}$. If $x_{k-1} = x^*$, then we have already converged to the fixed point. Otherwise, suppose that $x_{k-1} \neq x^*$. By the **Taylor Series Theorem** (could also use the **Mean Value Theorem** like above), there exists a ξ strictly between x_{k-1} and x^* such that

$$g(x_{k-1}) = g(x^* + (x_{k-1} - x^*)) = g(x^*) + g'(\xi)(x_{k-1} - x^*).$$

Note that $\xi \in (a, b)$. Thus,

$$|x_k - x^*| = |g(x_{k-1}) - g(x^*)| = |g'(\xi)(x_{k-1} - x^*)| = |g'(\xi)|\,|x_{k-1} - x^*| \leqslant \rho\,|x_{k-1} - x^*|\,.$$

So $|x_k - x^*| \leqslant \rho\,|x_{k-1} - x^*|$ for $k = 1, 2, \ldots$, which implies that

$$0 \leq |x_k - x^*| \leqslant \rho\,|x_{k-1} - x^*| \leqslant \rho^2\,|x_{k-2} - x^*| \leqslant \cdots \leqslant \rho^k\,|x_0 - x^*|\,.$$

Since $\rho < 1$, the right-hand-side converges to 0 as $k \to \infty$. Therefore,

$$\lim_{k\to\infty} |x_k - x^*| = 0,$$

so x_k converges to x^*. ■

- **Contraction factor**: We call ρ the *contraction factor*, the smaller ρ is, the faster the convergence
- **Convergence example**: The first g we considered was

$$g(x) = x - x^2 + 2$$

which has the fixed points $x_1^* = -\sqrt{2}$ and $x_2^* = \sqrt{2}$. Note that

$$g'(x) = 1 - 2x$$

and that

$$g'(x_1^*) = 1 + 2\sqrt{2} = 3.8284271247461903\ldots,$$
$$g'(x_2^*) = 1 - 2\sqrt{2} = -1.8284271247461903\ldots.$$

So, $|g'(x_i^*)| > 1$ for $i = 1, 2$, which explains why the fixed point iteration would not converge to either fixed point.

The second g we considered was

$$g(x) = x - \frac{x^2 - 2}{2x} = \frac{x}{2} + \frac{1}{x},$$

which has the fixed points $x_1^* = -\sqrt{2}$ and $x_2^* = \sqrt{2}$. Now the derivative is

$$g'(x) = \frac{1}{2} - \frac{1}{x^2},$$

and so

$$g'(x_i^*) = 0, \quad i = 1, 2.$$

Thus, for $i = 1, 2$, we have $|g'(x_i^*)| < \rho$, for any $\rho \in (0, 1)$. This explains why the fixed point iteration converged rapidly to x_2^* from $x_0 = 1.1$; we also expect rapid convergence to x_1^* from suitable x_0.

- **Newton's method**: Let
$$f \in C^2[a, b].$$
That is, f is a **twice-continuously differentiable** function over $[a, b]$, which means that the **first** and **second** derivatives of f **exist** and are **continuous** on the open interval (a, b). **Newton's method** is defined as:
$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots.$$
This is the fixed point iteration using the function $g(x) = x - f(x)/f'(x)$.

- **Formulating Newton's method**: Suppose that $f(x^*) = 0$ and that we are at the iterate x_k. By the **Taylor Series Theorem**, we have
$$f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\xi)}{2}(x^* - x_k)^2,$$
for some point ξ between x^* and x_k. If x_k is already fairly close to x^*, then $(x^* - x_k)^2$ will be very small, so we have
$$0 \approx f(x_k) + f'(x_k)(x^* - x_k).$$
Solving for x^*, we obtain
$$x^* \approx x_k - \frac{f(x_k)}{f'(x_k)}.$$
Therefore, it makes sense to define our next iterate x_{k+1} using this approximation.

- **Another formulation for Newton's method**: Another way to obtain Newton's method is as follows. Consider the **first-order (linear) approximation** of f around the point x_k:
$$f(x) \approx f(x_k) + f'(x_k)(x - x_k), \quad \text{for all } x \approx x_k.$$
Suppose that x_k is close to x^*, and that $f(x^*) = 0$. Then
$$f(x^*) \approx f(x_k) + f'(x_k)(x^* - x_k),$$
which implies that
$$x^* \approx x_k - \frac{f(x_k)}{f'(x_k)}.$$
Therefore, our next iterate should be
$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

- **Newton's method example: The Babylonian method for computing $\sqrt{a}$**: Let $f(x) = x^2 - a$. Newton's method gives us the iteration:
$$x_{k+1} = x_k - \frac{x_k^2 - a}{2x_k} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right).$$

- **Speed of convergence**: If $x_k \to x^*$, we can measure the speed of the convergence as follows.
  - **Linear convergence** means there is a constant $0 < \rho < 1$ such that
  $$|x_{k+1} - x^*| \leqslant \rho\,|x_k - x^*|\,, \quad \text{for all } k \text{ sufficiently large;}$$
  that is,
  $$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = \rho < 1.$$

– **Superlinear convergence** means there is a sequence $\rho_k \to 0$ such that

$$|x_{k+1} - x^*| \leqslant \rho_k |x_k - x^*|, \quad \text{for all } k \text{ sufficiently large;}$$

that is,

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = 0.$$

– **Quadratic convergence** means there is a constant M such that

$$|x_{k+1} - x^*| \leqslant M |x_k - x^*|^2, \quad \text{for all } k \text{ sufficiently large;}$$

that is,

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^2} = M < \infty.$$

Note that **quadratic convergence** is an example of **superlinear convergence** with $\rho_k = M|x_k - x^*|$.

- **Quadratic convergence of Newton's method**

Theorem: Let $f \in C^2[a,b]$. If f has a root $x^* \in (a,b)$ such that $f'(x^*) \neq 0$, then there is a $\delta > 0$ such that Newton's method **converges quadratically** to x^* from any $x_0 \in [x^* - \delta, x^* + \delta]$.

Proof. Since

– $f \in C^2[a,b]$
– $x^* \in (a,b)$
– $f'(x^*) \neq 0$

there are positive constants δ_1, ε, and M such that

– $|f'(x)| \geqslant \varepsilon$
– $|f''(x)| \leqslant M$

for all $x \in [x^* - \delta_1, x^* + \delta_1] \subset (a,b)$.

Suppose $x_k \in [x^* - \delta_1, x^* + \delta_1]$. Then, there is a ξ_k between x^* and x_k such that

$$f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\xi_k)}{2}(x^* - x_k)^2.$$

Using the fact that $f(x^*) = 0$, we have

$$0 = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\xi_k)}{2}(x^* - x_k)^2.$$

Also, x_{k+1} satisfies

$$0 = f(x_k) + f'(x_k)(x_{k+1} - x_k).$$

Subtracting these equations, we obtain

$$0 = f'(x_k)(x^* - x_{k+1}) + \frac{f''(\xi_k)}{2}(x^* - x_k)^2.$$

Since $f'(x_k) \neq 0$, we have

$$x^* - x_{k+1} = -\frac{f''(\xi_k)}{2f'(x_k)}(x^* - x_k)^2.$$

Thus,

$$|x^* - x_{k+1}| = \left|\frac{f''(\xi_k)}{2f'(x_k)}\right| |x^* - x_k|^2 \leqslant \frac{M}{2\varepsilon} |x^* - x_k|^2,$$

so if $x_k \to x^*$, then the **convergence will be quadratic**.

We just need to find $\delta > 0$ so that if $x_0 \in [x^* - \delta, x^* + \delta]$, then $x_k \to x^*$. Let

$$\delta = \min\left\{\frac{\varepsilon}{M}, \delta_1\right\}.$$

Suppose that $x_k \in [x^* - \delta, x^* + \delta]$. Then

$$\begin{aligned} |x^* - x_{k+1}| &\leqslant \frac{M}{2\varepsilon} |x^* - x_k|^2 \\ &\leqslant \frac{M}{2\varepsilon}\delta |x^* - x_k| \\ &\leqslant \frac{1}{2} |x^* - x_k| \\ &< \delta, \end{aligned}$$

so $x_{k+1} \in [x^* - \delta, x^* + \delta]$ as well. Thus, if $x_0 \in [x^* - \delta, x^* + \delta]$, we have $x_k \in [x^* - \delta, x^* + \delta]$ for $k = 0, 1, 2, \ldots$.

Moreover,

$$0 \leqslant |x^* - x_k| \leqslant \frac{1}{2} |x^* - x_{k-1}| \leqslant \frac{1}{4} |x^* - x_{k-2}| \leqslant \cdots \leqslant \frac{1}{2^k} |x^* - x_0|.$$

Since $\frac{1}{2^k} |x^* - x_0| \to 0$ as $k \to \infty$, we conclude that $x_k \to x^*$. Thus, if $x_0 \in [x^* - \delta, x^* + \delta]$ then x_k converges to x^* quadratically. ■

- **Pros and cons of Newton's method**:

  Pros:

  - **Fast to converge:** Newton's method enjoys quadratic convergence near the root when $f'(x^*) \neq 0$.
  - **Generalizes to multiple variables:** Let $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^n$. Newton's method for solving

    $$\mathbf{F}(\mathbf{x}) = \mathbf{0}$$

    (i.e., n nonlinear equations with n unknowns) is

    $$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{J}(\mathbf{x}_k)^{-1}\mathbf{F}(\mathbf{x}_k),$$

    where $\mathbf{J}(\mathbf{x})$ is the $n \times n$ **Jacobian** of $\mathbf{F}$:

    $$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial x_1} & \cdots & \frac{\partial F_n}{\partial x_n} \end{bmatrix}$$

Cons

- **Requires the derivative:** We must give Newton's method both the function f and its derivative f'. This may not always be possible or easy.
- **Need to start close to x^*:** Newton's method is a **local method**. When x_0 is far from x^*, Newton's method may not converge to x^*, or may require many iterations before quadratic convergence begins.

- **Secant method**: Sometimes it is not possible to evaluate the derivative f':
  - f' is unknown or difficult to obtain
  - evaluating f' takes too much time

  Instead, we can use the **secant approximation** of the derivative. When $x_k \approx x_{k-1}$, we have
  $$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$
  Plugging this approximation into the formula for Newton's method, we get:
  $$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}$$

  The secant method is an example of a *Quasi-Newton method* method since we are replacing f' with an approximation of f'.

  When $f'(x^*) \neq 0$, the secant method will converge **superlinearly**, so it may not be as fast as Newton's method.

- **The case of a multiple root**: When $f'(x^*) = 0$, we are no longer guaranteed to obtain superlinear convergence of the secant method, nor quadratic convergence of Newton's method. In this case, both methods will be merely **linearly convergent**.

- **Minimizing a function in one variable**: We can use the root-finding methods described above to find the **minimum** or **maximum** value of a function $\phi \in C^2[a, b]$. Recall that $x^* \in (a, b)$ is a **critical point** of ϕ if
  $$\phi'(x^*) = 0.$$
  We can find x^* by applying Newton's method to this nonlinear equation to obtain:
  $$x_{k+1} = x_k - \frac{\phi'(x_k)}{\phi''(x_k)}.$$

- **Another interpretation**: We can also obtain this by considering the **second-order (quadratic) approximation** of ϕ around the point x_k:
  $$\phi(x) \approx \phi(x_k) + \phi'(x_k)(x - x_k) + \frac{\phi''(x_k)}{2}(x - x_k)^2, \quad \text{for all } x \approx x_k.$$
  If x_k is close to x^*, we expect the minimum/maximum of ϕ to be near the minimum/-maximum of the **quadratic approximation** of ϕ:
  $$q(x) = \phi(x_k) + \phi'(x_k)(x - x_k) + \frac{\phi''(x_k)}{2}(x - x_k)^2.$$
  We should choose x_{k+1} to be the critical point of q, so we want to find x_{k+1} such that $q'(x_{k+1}) = 0$. Note that
  $$q'(x) = \phi'(x_k) + \phi''(x_k)(x - x_k).$$
  Thus $q'(x_{k+1}) = 0$ gives us
  $$x_{k+1} = x_k - \frac{\phi'(x_k)}{\phi''(x_k)}.$$