Copyright © 2024 Eric Yang Xingyu

PUBLISHED BY PUBLISHER
This LaTeX template is from BOOK-WEBSITE.COM

Licensed under the Creative Commons Attribution-NonCommercial 4.0 License (the "License"). You may not use this file except in compliance with the License. You may obtain a copy of the License at https://creativecommons.org/licenses/by-nc-sa/4.0. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First amendment, March 2024

2.4 Summation 40
2.5 Sequence 48

3 Algorithm and Number 57
3.1 Numbers 57
3.2 Algorithm and Algorithm Analysis 65

4 Inequality 73
4.1 Inequality basics 73
4.2 Solving Quadratic Inequality 75
4.3 important Inequalities 76

5 Complex Number 101
5.1 Algebra of Complex Number 102
5.2 Point representation of Complex Number 105
5.3 Vector and Polar Form 111
5.4 Exponential Form 120
5.5 Finding Complex Roots 126

The first Chapter of this book focus only on the most essential part of mathematics, proofs. Proofs are the very essence of mathematics, serving as the definitive tool for establishing the truth within this discipline of absolute certainty. Unlike empirical sciences, where conclusions are drawn based on observation and experimentation subject to uncertainties, mathematical proofs provide incontrovertible evidence that a statement is true. They are the architects of mathematical theory, constructing a framework of knowledge that is both logical and immutable. Through proofs, we not only validate conjectures but also weave a tapestry of interconnected truths, each supported by the unshakable foundation of previously proven results. This interconnectedness ensures that mathematical knowledge, once proven, becomes a permanent addition to the collective human understanding, transcending time and offering a universal language spoken by all cultures in the language of logic and reason.

1.1 Propositions

The angles in a triangle add up to 180 degrees; the sum of any two even numbers is even. Statement as such is so common in mathematics, which we call proposition.

Definition 1.1 — Proposition. A proposition is a declarative sentence that is either true or false. Propositions are the fundamental building blocks of mathematical reasoning, as they can be clearly judged to be true or false.

Propositions have the following characteristics:

1. **Definiteness:** Propositions must be clear and unambiguous so that they can be definitively judged to be true or false.
2. **Exclusivity:** Propositions admit no middle ground between true and false.
3. **Objectivity:** Propositions represent objective facts, not subjective opinions or questions.

Propositions are essential and crucial for proof, this is because

- **Foundation:** Propositions are the foundation upon which logical reasoning is built.
- **Validity:** Proving the validity of a proposition reinforces the truthfulness of a statement.

- **Interconnectedness:** Propositions are interconnected; proving one can help establish the truth of others.

In mathematical discourse, the clarity and truth of propositions are paramount. Understanding the nature of propositions (truth, falsity, and reasoning) is essential for constructing and understanding mathematical arguments. Commonly, Propositions are categorized by the **truth value**, which we will discuss in Boolean Algebra, and we call a proposition true proposition when the statement is factually correct, while false proposition, vice versa. Most importantly: **If a proposition is true, it can be proven**.

Converse, Inverse, and Contrapositive

In the study of logic, particularly within the context of mathematical reasoning, we come across several important concepts that relate to conditional statements. A conditional statement is typically of the form "If P, then Q", denoted $P \to Q$. Here we define and discuss the converse, inverse, and contrapositive of a conditional statement.

Converse

The converse of a statement flips the hypothesis and the conclusion. For the statement $P \to Q$, the converse is $Q \to P$. It is important to note that the truth of a converse is not necessarily the same as the truth of the original statement.

Inverse

The inverse of a statement negates both the hypothesis and the conclusion. For the statement $P \to Q$, the inverse is $\neg P \to \neg Q$, where $\neg$ denotes negation. Similar to the converse, the truth of the inverse is not dependent on the truth of the original statement.

Contrapositive

The contrapositive of a statement negates and flips the hypothesis and the conclusion. For the statement $P \to Q$, the contrapositive is $\neg Q \to \neg P$. Unlike the converse and the inverse, the truth of the contrapositive is always the same as the truth of the original statement. This property is often used in mathematical proofs, particularly in **proofs by contradiction**, which we will discuss in successive sections.

1.2 Direct Proof

Framework of Direct Proof

A direct proof is a fundamental method in mathematics used to establish the truth of a given statement, typically a theorem or proposition. It is characterized by a straightforward and logical progression from known facts or axioms to the conclusion. The approach of a direct proof is to assume that the premises (initial assumptions or known truths) are correct and then to use logical reasoning and established mathematical principles to demonstrate that the conclusion necessarily follows from these premises. Here's a general structure of how a direct proof works:

1. Start with Known Facts or Assumptions: Begin with what is already known or assumed to be true. These can be definitions, previously proven theorems, or given premises.

2. Logical Argumentation: Use logical reasoning and mathematical operations to derive new information from these known facts. This process often involves applying definitions, using the properties of mathematical operations, and invoking previously established theorems.
3. Arrive at the Conclusion: The final step is to show that the statement you set out to prove is a logical consequence of the initial assumptions. The conclusion should follow naturally and unavoidably from the previous steps.

To illustrate:

■ **Example 1.1** Prove that if n is an odd integer, then n^2 is also odd. ■

Proof. Assume n is an odd integer. By definition, an odd integer can be written as $n = 2k+1$ where k is an integer.

Squaring both sides of this equation, we get

$$n^2 = (2k+1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1.$$

Let $m = 2k^2 + 2k$, which is an integer since it is a sum of integers. Therefore, we can express n^2 as

$$n^2 = 2m + 1.$$

This is the form of an odd integer. Thus, we have shown that if n is an odd integer, then n^2 is also odd. ■

We can see that the idea of direct proof is something everyone can understand, however, what is really challenging is to find the specific know facts to assist the proof. Here are more exercises on direct proof for reference.

1.2.1 Exercises

Exercise 1.1 Prove that if n is an even integer, then n^2 is also even. ■

Hint: follow the procedure of the example of odd number, using the definition of even integer.

Proof. Assume n is an even integer. By definition, an even number can be expressed as $n = 2k$ where k is an integer.

Squaring both sides of this equation, we get:

$$n^2 = (2k)^2 = 4k^2 = 2(2k^2).$$

Since $2k^2$ is an integer (let's call it m), we can express n^2 as $2m$, which is the definition of an even number.

Hence, if n is an even integer, then n^2 is also even. ■

Exercise 1.2 Prove that the sum of two even integers is even. ■

Hint: Use the definition of an even integer.

Proof. Let a and b be even integers. By the definition of even integers, there exist integers k and m such that $a = 2k$ and $b = 2m$.

The sum of a and b is:

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

Let $n = k + m$, which is an integer since both k and m are integers. Hence, $a + b = 2n$. Since $a + b$ is two times an integer, it is even by definition. This concludes the proof. ■

Exercise 1.3 Prove that for any positive integer n, $n^3 + 2n$ is divisible by 3.

Hint: Factorize $n^3 + 2n$ and use the properties of divisibility.

Proof. Consider the expression $n^3 + 2n$. This can be factored as:

$$n^3 + 2n = n(n^2 + 2) = n(n^2 - 1 + 3) = n(n + 1)(n - 1) + 3n.$$

Notice that $n^2 + 2$ can be written as $(n^2 - 1) + 3 = (n + 1)(n - 1) + 3$. The terms $n + 1$, n, and $n - 1$ are three consecutive integers, so one of them must be multiple of 3. Therefore, $n(n + 1)(n - 1)$ can be denoted by $3m$, where m is a positive integer, and the expression is equivalent to $3(m + n)$.

Hence, for any positive integer n, $n^3 + 2n$ is divisible by 3. ■

R Considering there are still concepts that we haven't covered in Boolean algebra and number theory, the exercises in this chapter will be more basic than other chapters. What's really important in this chapter is not to finish difficult proof, but grasp the idea of proving.

1.3 Proof by Cases

Sometimes, to draw a specific conclusion, we must make multiple or even infinite assumptions (The latter we will discuss in **Strong Mathematical induction**). This proof pattern stands alone from direct proof, as when we use this method, the proof itself could still be direct or indirect.

Framework of Proof by Cases

Proof. The proof is by cases. We consider each possible case and show that the theorem holds in each case.

Case 1: [Description of Case 1]

Proof of Case 1. Present the proof for Case 1 here. Use logical reasoning and mathematical principles to demonstrate that the theorem holds under the assumptions of Case 1. ■

Case 2: [Description of Case 2]

Proof of Case 2. Similarly, present the proof for Case 2 here, showing that the theorem is valid in this scenario as well. ■

Final Conclusion: Since the theorem holds in all possible cases, we conclude that the theorem is proved. ■

Here is an example of proof by cases.

■ **Example 1.2** Prove that the sum of any three consecutive integers is divisible by 3. ■

Proof. Let the three consecutive integers be n, $n+1$, and $n+2$, where n is an integer. We consider three cases for n, based on the division by 3.

Case 1: n is of the form $3k$ for some integer k.

Case 2: n is of the form $3k+1$ for some integer k.

Case 3: n is of the form $3k+2$ for some integer k.

In each case, the sum of n, $n+1$, and $n+2$ can be expressed as:

For Case 1: $(3k)+(3k+1)+(3k+2) = 9k+3 = 3(3k+1)$, it is divisible by 3.

For Case 2: $(3k+1)+(3k+2)+(3k+3) = 9k+6 = 3(3k+2)$, it is divisible by 3.

For Case 3: $(3k+2)+(3k+3)+(3k+4) = 9k+9 = 3(3k+3)$, it is divisible by 3.

In each of the three cases, the sum is a multiple of 3. Therefore, we conclude that the sum of any three consecutive integers is divisible by 3. ■

R As mentioned in direct proof, the idea of proving is always simple and easy to understand. For proof by cases, the toughest part is also finding all the cases needed to prove a certain conclusion.

1.3.1 Exercises

Exercise 1.4 Prove that the product of two consecutive integers is always even. ■

Hint: Express the consecutive integers as n and $n+1$.

Proof. Consider two consecutive integers, n and $n+1$, where n is an integer. The product of these two integers is $n(n+1)$.

We have two cases to consider:

Case 1: If n is even, then it can be written as $n = 2k$ for some integer k. The product is then $n(n+1) = 2k(2k+1)$, which is even because it is divisible by 2.

Case 2: If n is odd, then $n+1$ is even. In this case, $n+1 = 2k$ for some integer k. The product is then $n(n+1) = n(2k)$, which is even because it is divisible by 2.

In either case, the product of n and $n+1$ is even. Therefore, the product of two consecutive integers is always even. ■

Exercise 1.5 Prove that if n is an integer, then $n(n+1)$ is even. ■

Hint: Consider the cases where n is even and where n is odd separately.

Proof. We will consider two cases based on the parity of n.

Case 1: n is even.

Let $n = 2k$ for some integer k. Then $n(n+1) = 2k(2k+1)$. Since $2k$ is even, the product $2k(2k+1)$ is also even because the product of an even number and any other integer is even.

Case 2: n is odd.

Let $n = 2k+1$ for some integer k. Then $n(n+1) = (2k+1)(2k+2)$. Here $(2k+2)$ is even, and thus the product $(2k+1)(2k+2)$ is even because the product of an even number and any other integer is even.

In both cases, whether n is even or odd, $n(n+1)$ is even. Hence, we have shown that for any integer n, the product $n(n+1)$ is always even. ■

Exercise 1.6 Prove that for any integer n, n^2+4 cannot be a prime number if $n > 1$. ■

Hint: Consider the cases where n is even and where n is odd.

Proof. Consider two cases based on the parity of n.

Case 1: n is even.

Let $n = 2k$ for some integer k. Then n^2+4 becomes $(2k)^2+4 = 4k^2+4 = 4(k^2+1)$. Since k^2+1 is an integer greater than 1, $4(k^2+1)$ is not prime because it has factors other than 1 and itself, namely 4 and k^2+1.

Case 2: n is odd.

Let $n = 2k+1$ for some integer k. Then n^2+4 becomes $(2k+1)^2+4 = 4k^2+4k+1+4 = 4k^2+4k+5$. This can be factored as $(2k+1)(2k+3)$, which are two consecutive odd numbers. Since both $2k+1$ and $2k+3$ are factors greater than 1 but not the expression itself, the product is not prime. Therefore, in both cases, whether n is even or odd, n^2+4 is not a prime number. ■

Exercise 1.7 Prove that for any integer n, the expression $n^5 - n$ is divisible by 30. ■

Hint: Consider divisibility of 5 and 6 (2 and 3), and prove by cases accordingly.

Proof. To prove that $n^5 - n$ is divisible by 6, it suffices to show that it is divisible by 2, 3, and 5.

First, observe that $n^5 - n = n(n^4 - 1)$.

Divisibility by 2:

- If n is even, then $n^5 - n$ is clearly even since it is a multiple of n.
- If n is odd, $n^4 - 1$ is even because n^4 is odd and $n^4 - 1$ is one less than an odd number, making it even.

Divisibility by 3:

Any integer n is either a multiple of 3, one more than a multiple of 3, or two more than a multiple of 3. That is, n can be written as $3k$, $3k+1$, or $3k+2$ for some integer k.

- If $n = 3k$, then $n^5 - n = (3k)^5 - 3k$ is clearly a multiple of 3.
- If $n = 3k+1$, then $n^5 - n = (3k+1)^5 - (3k+1)$. Expanding $(3k+1)^5$ using the binomial theorem, all terms except the last are multiples of 3, and the last term 1^5 minus $3k+1$ leaves $-3k$, which is a multiple of 3.
- If $n = 3k+2$, a similar expansion shows that $n^5 - n$ is a multiple of 3.

Thus, $n^5 - n$ is divisible by 3.

Divisibility by 5: For any integer n, we can express n in one of the following forms, where k is an integer:

1. $n = 5k$ (where n is a multiple of 5)
2. $n = 5k+1$
3. $n = 5k+2$

4. $n = 5k+3$
5. $n = 5k+4$

We will prove that $n^5 - n$ is divisible by 5 for each case. Since we are working modulo 5, we only need to consider the last digit of n when raised to the fifth power due to the cyclicity of powers modulo 5.

- For $n = 5k$: $n^5 - n = (5k)^5 - 5k$. Clearly, both terms are divisible by 5.
- For $n = 5k+1$: $n^5 - n = (5k+1)^5 - (5k+1)$. When expanded, each term of $(5k+1)^5$ except the last will contain a factor of $5k$ and thus be divisible by 5. The last term $1^5 = 1$, and subtracting $5k+1$ leaves a result which is still divisible by 5.
- For $n = 5k+2$: $n^5 - n = (5k+2)^5 - (5k+2)$. Again, each term in the expansion of $(5k+2)^5$, except the last, will be divisible by 5. The last term will be $2^5 = 32$, which is congruent to 2 modulo 5, and subtracting $(5k+2)$ leaves a result divisible by 5.
- For $n = 5k+3$ and $n = 5k+4$, a similar argument holds as for $n = 5k+1$ and $n = 5k+2$, respectively.

In all cases, $n^5 - n$ is divisible by 5. This completes the proof for Case 3.

Therefore, $n^5 - n$ is divisible by 2, 3, and 5, and hence by 30. ■

1.4 Indirect Proof

An indirect proof is a powerful method in mathematics used to prove a statement by showing that the negation leads to a contradiction or by proving the contrapositive of the statement. We will discuss proof by contradiction and proof by contrapositive separately.

1.4.1 Proof by Contradiction

Proof by contradiction is a critical reasoning technique used extensively in mathematics and logic. Its importance lies in its ability to confirm the truth of a statement by demonstrating that assuming the opposite leads to an illogical or impossible conclusion. This method is particularly valuable because it can sometimes prove assertions that are otherwise difficult to demonstrate directly. It is a cornerstone of mathematical reasoning and is often used to establish fundamental theorems that form the bedrock of mathematical theories. The use of this technique highlights the rigorous nature of mathematical proof and emphasizes the importance of logical consistency within mathematical frameworks.

Here is an abstract template for proof by contradiction:

Suppose, for the sake of contradiction, that the statement we wish to prove is false.

1. State the negation of the theorem or proposition you are trying to prove.
2. Use logical reasoning and established mathematical principles to derive consequences of this assumption.
3. Show that these consequences lead to a contradiction, something that is known to be false or violates a basic principle of mathematics.
4. Conclude that since the assumption leads to a contradiction, the negation must be false, and therefore, the original statement is true.

Contradiction could be used to proof many interesting conclusions in mathematics, here is a classical example that prove the length of a diagonal in a square with side length 1 is irrational number.

▪ **Example 1.3** Prove that the length of a diagonal in a square with side length 1 is irrational number ▪

Proof. Assuming $\sqrt{2}$ is rational, we can express x as $\frac{p}{q}$ where p and q are integers with no common factors, and we have:

R Rational numbers must be able to be expressed in the form of $\frac{p}{q}$, where p and q are both natural numbers with no common factors other than 1.

Therefore, we have $\frac{p^2}{q^2} = 2$. Thus, $p^2 = 2q^2$. This means that

$$\left(\frac{p}{q}\right)^2 = 2. \tag{1.1}$$

This leads to:

$$p^2 = 2q^2. \tag{1.2}$$

This implies that p^2 is an even number since it is twice some integer. Therefore, p must also be even, as the square of an odd number cannot be even. Let $p = 2r$, then equation (1.2) becomes:

$$(2r)^2 = 2q^2, \tag{1.3}$$

which simplifies to:

$$2r^2 = q^2. \tag{1.4}$$

This indicates that q^2 is also even, and hence q must be even. However, this is a contradiction because if both p and q are even, they are not coprime, which violates the initial assumption that p and q have no common factors. Therefore, our initial assumption that $\sqrt{2}$ is rational must be false. Hence, $\sqrt{2}$ is irrational.

Thus, we conclude that the number $\sqrt{2}$ is irrational.

■

With this example, we can see that this is a very rigorous proving method that could be used to prove propositions that we have to prove indirectly. Do take note that using this method requires you to make clear assumption and find contradiction accurately.

Exercises

Exercise 1.8 Prove that $\sqrt{5}$ is irrational. ■

Hint: Refer to the example. Consider the following lemma: If an integer p can be expressed as the product of two integers a and b, that is, $p = ab$, then p is not a prime number.

Proof. We will prove by contradiction that the square root of 5 is an irrational number. Suppose $\sqrt{5}$ is rational, which means it can be expressed as a fraction $\frac{p}{q}$, where p and q are coprime integers, and $q \neq 0$.

1. We express $\sqrt{5}$ as a fraction in the lowest terms, so we have:

$$\sqrt{5} = \frac{p}{q}$$

where p and q have no common factors other than 1.

2. Squaring both sides of the equation, we get:

$$5 = \frac{p^2}{q^2}$$

which implies:

$$p^2 = 5q^2$$

3. Since $p^2 = 5q^2$, p^2 is a multiple of 5, and hence p must also be a multiple of 5, because the square of a number is only a multiple of 5 if the number itself is a multiple of 5. Let $p = 5k$, where k is an integer.
4. Substituting $p = 5k$ into $p^2 = 5q^2$, we get:

$$(5k)^2 = 5q^2$$

which simplifies to:

$$25k^2 = 5q^2$$

Dividing both sides by 5, we find:

$$5k^2 = q^2$$

5. This implies that q^2, and hence q, is also a multiple of 5.
6. Therefore, both p and q are multiples of 5, which contradicts our initial assumption that they have no common factors other than 1. Hence, our initial assumption that $\sqrt{5}$ is rational must be false.

Thus, we conclude that $\sqrt{5}$ cannot be expressed as a fraction, and therefore it is irrational. ■

Exercise 1.9 Prove that there is no smallest positive rational number. ■

Hint: Assume that there is a smallest positive rational number and then show that you can find a smaller one, which leads to a contradiction.

Proof. Assume, for the sake of contradiction, that there exists the smallest positive rational number $\frac{p}{q}$, where p and q are positive integers with no common factors other than 1, and $q \neq 0$.

Consider the rational number $\frac{p}{2q}$. This number is positive since p and q are positive. It is also rational because it is the ratio of two integers. Moreover, $\frac{p}{2q}$ is smaller than $\frac{p}{q}$.

Since q is a positive integer and greater than 1 (because there are no smaller positive integers than 1), the inequality holds true. This means we have found a positive rational number smaller than our assumed smallest positive rational number, which is a contradiction.

Therefore, there cannot exist a smallest positive rational number, and our initial assumption is false. ■

Exercise 1.10 Prove that if n is an integer and n^2 is even, then n is even. ■

Proof. Assume for the sake of contradiction that n is odd. According to the definition of an odd number, n can be expressed as $n = 2k+1$ for some integer k. We square this expression to get:

$$n^2 = (2k+1)^2 = 4k^2 + 4k + 1.$$

This expression can be rewritten as:

$$n^2 = 2(2k^2 + 2k) + 1.$$

Since $2k^2 + 2k$ is an integer, the expression $2(2k^2 + 2k)$ is even, and adding 1 to an even number results in an odd number. Thus, n^2 is odd.

However, this contradicts our initial assumption that n^2 is even. Therefore, our assumption that n is odd must be false, and it follows that n must be even. ■

Exercise 1.11 Prove that there are infinitely many prime numbers. ■

hint: Assume that there are only finitely many primes. Consider the product of all these primes plus one and analyze its prime factors to reach a contradiction.

Proof. Suppose for the sake of contradiction that there are only finitely many prime numbers. Let us list them as $p_1, p_2, \ldots, p_n$. Consider the number P defined by the product of all these primes plus one:

$$P = p_1 p_2 \cdots p_n + 1.$$

By construction, P is greater than any of the listed prime numbers.

Now, P must be divisible by some prime number (as every integer greater than 1 has a prime divisor). If P is divisible by any of the primes $p_1, p_2, \ldots, p_n$, then there would be a remainder of 1, which is a contradiction because a prime number dividing P would leave no remainder. Therefore, P cannot be divided by any of the p_i without leaving a remainder.

This implies that P must have a prime divisor that is not in our list, contradicting the assumption that we have listed all prime numbers. Thus, there must be more prime numbers than those in the list, and since our list was arbitrary, this means there are infinitely many prime numbers. ■

Exercise 1.12 Prove that the sum of an irrational number and a rational number is irrational. ■

Proof. Assume for the sake of contradiction that the sum of an irrational number r and a rational number s is rational, and denote this sum as t. Express t and s as the quotient of integers, such that $t = \frac{a}{b}$ and $s = \frac{c}{d}$, where a, b, c, d are integers and $b, d \neq 0$.

Since t is the sum of r and s, we can write $t = r + s$. Substituting the expressions for t and s gives us $\frac{a}{b} = r + \frac{c}{d}$. Rearranging this equation to isolate r gives us $r = \frac{a}{b} - \frac{c}{d}$. Combining the fractions, we find that $r = \frac{ad-bc}{bd}$.

This expression shows that r is the quotient of two integers, which means r is rational. However, this contradicts our original statement that r is irrational.

Hence, we conclude by contradiction that the sum of an irrational number and a rational number cannot be rational, and therefore it is irrational. ■

1.4.2 Proof by Contrapositive

Proof by contrapositive is a valid form of mathematical proof that is often used when direct proof or proof by contradiction is not as clear or straightforward. It's based on the logical equivalence between an implication and its contrapositive.

To prove a statement of the form If P, then Q by contrapositive, we prove the equivalent statement If not Q, then not P. It works as follows:

1. Assume Q is false.
2. Use logical reasoning to show that under this assumption, P must also be false.
3. Thus, we have shown that If not Q, then not P is true, which by the equivalence of implication, means If P, then Q is true.

▪ **Example 1.4** Suppose $x \in \mathbb{Z}$. If $x^2 - 6x + 5$ is even, then x is odd. ▪

Proof. The original proposition is equivalent to that if x is not odd, then $x^2 - 6x + 5$ is odd.

Thus, when x is even, $x = 2a$ for some integer a. So,

$$x^2 - 6x + 5 = (2a)^2 - 6(2a) + 5$$

$$= 4a^2 - 12a + 5 = 4a^2 - 12a + 4 + 1$$

$$= 2(2a^2 - 6a + 2) + 1$$

Therefore $x^2 - 6x + 5 = 2b + 1$, where b is the integer $2a^2 - 6a + 2$. Consequently $x^2 - 6x + 5$ is odd. Therefore $x^2 - 6x + 5$ is not even.

Hence, for $x \in \mathbb{Z}$, if $x^2 - 6x + 5$ is even, then x is odd. ■

R In short, proving by contrapositive is just a way to prove a given proposition in a more approachable way through its logical equivalence.

1.4.3 Exercises

Exercise 1.13 Prove that for any two integers a and b, if $a \cdot b$ is odd, then both a and b are odd. ▪

Proof. We will prove this by contrapositive. The contrapositive of the given statement is: If either a or b is not odd (that is, at least one of them is even), then $a \cdot b$ is not odd (that is, $a \cdot b$ is even).

Assume that at least one of the integers, without loss of generality say a, is even. Then a can be written as $a = 2k$ for some integer k. The product $a \cdot b$ is then $(2k) \cdot b = 2(k \cdot b)$. Since $k \cdot b$ is an integer, $2(k \cdot b)$ is clearly even. Hence, the product $a \cdot b$ is even.

Since we have shown that the contrapositive statement is true, the original statement must also be true. Therefore, if $a \cdot b$ is odd, then both a and b must be odd. ■

R For contrapositive of the original statement, the negation of "both" will be "either", and the negation for "all" is naturally "not all"

Exercise 1.14 Prove that if n is an integer and $3n+2$ is even, then n is even. ■

Proof. Let us prove the statement by contrapositive. Assume n is not even, which means n is odd. By definition, an odd number can be written as $n = 2k+1$ for some integer k. We can then express $3n+2$ as follows:

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

Since $3k+2$ is an integer, $2(3k+2)$ is even, and adding 1 to an even number results in an odd number, it follows that $3n+2$ is odd.

Thus, we have shown that if n is not even, then $3n+2$ is not even. By contrapositive, this means that if $3n+2$ is even, then n must be even. ■

Exercise 1.15 For any integers a and b, the condition $a+b \geq 15$ implies that $a \geq 8$ or $b \geq 8$. ■

R Note that the negation of the conclusion in the original claim requires changing the logical "or" to an "and".

Proof. The contrapositive of the given claim states that if both a and b are integers less than 8, then their sum $a+b$ is less than 15.

Assume that a and b are such integers with $a < 8$ and $b < 8$. Being integers, the greatest values they can take are $a = 7$ and $b = 7$. Summing these maximal values yields $a+b = 7+7 = 14$, which is less than 15.

Thus, having $a+b < 15$ necessarily implies that both a and b must be less than 8, which completes the proof by contrapositive. Consequently, if $a+b \geq 15$, then it must be that either $a \geq 8$ or $b \geq 8$.

■

1.5 Mathematical Induction

Many mathematical problems involve only integers; computers perform operations in terms of integer arithmetic. The natural numbers enable us to solve problems by working one step at a time. After giving a definition of the natural numbers as a subset of the real numbers, we study the principle of mathematical induction. We use this fundamental technique of proof to solve problems such as the following.

Problem 1.1 The Checkerboard Problem. Counting squares of sizes one-by-one through eighty-by-eighty, an ordinary eight-by-eight checkerboard has 204 squares. How can we obtain a formula for the number of squares of all sizes on an n-by-n checkerboard?

Problem 1.2 The Handshake Problem. Consider n married couples at a party. Suppose that no person shakes hands with his or her spouse, and the $2n-1$ people other than the host shake hands with different numbers of people. With how many people does the hostess shake hands?

Problem 1.3 Sums of Consecutive Integers. Which natural numbers are sums of consecutive smaller natural numbers? For example, 30 = 9 + 10 + 11 and 31 = 15 + 16, but 32 has no such representation.

Problem 1.4 The Coin-Removal Problem. Suppose that n coins are arranged in a row. We remove heads-up coins, one by one. Each time we remove a coin we must flip the coins still present in the (at most) two positions surrounding it. For which arrangements of heads and tails can we remove all the coins? For example, THTHT fails, but THHHT succeeds. Using dots to denote gaps due to removed coins, we remove THHHT via THHT, .H.T, ..HT, ...H, ...

It is noticeable that these problems share something in common, which is actually the scale of the problem. Also, these problems provide us some procedures that is usable in other cases, making it possible to expand our confirmatory conclusion from the minimum scale to the infinity, or rather, all cases.

1.5.1 Framework of MI

The base of MI is what we call **Principle of Induction**.

Theorem 1.1 — Principle of Induction. For each natural number n, let $P(n)$ be a mathematical statement. If properties (a) and (b) below hold, then for each $n \in \mathbb{N}$ the statement $P(n)$ is true.

a) $P(1)$ is true.
b) For $k \in \mathbb{N}$, if $P(k)$ is true, then $P(k+1)$ is true.

Proof. Let $S = \{n \in \mathbb{N} : P(n) \text{ is true}\}$. By definition, $S \subseteq \mathbb{N}$. On the other hand, (a) and (b) here imply that S satisfies (a) and (b) of Definition 3.5. Since $\mathbb{N}$ is the smallest such set, $\mathbb{N} \subseteq S$. Therefore $S = \mathbb{N}$, and $P(n)$ is true for each $n \in \mathbb{N}$. ■

R Set may be not yet a concept known to you as we only discuss this topic in later chapters. You may ignore this piece of proof for now if this idea is not known for you at this moment.

When we proceed to prove something with MI, it follows this pattern:

Proof. The proof is by mathematical induction.
Base Case:
First we prove that $P(n_0)$ holds.
Inductive Assumption/hypothesis:
Then, assume that for $k \geq n_0$, $P(k)$ holds.
Inductive Step:
Finally, with is assumption, prove $P(k+1)$ also holds.
By the principle of mathematical induction, $P(n)$ is true for all integers $n \geq n_0$. ■

To illustrate, we use the sum of the first nth positive integer as an example.

▪ **Example 1.5** Prove, using mathematical induction:

Theorem 1.2 The sum of the first n positive integers is:

$$1+2+3+\ldots+n = \frac{n(n+1)}{2}$$

for all integers $n \geq 1$.

▪

Proof. **Base Case:** For $n = 1$, $1 = \frac{1(1+1)}{2} = 1$, which is true.

Inductive Hypothesis: Assume the statement is true for $n = k$, i.e.,

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

Inductive Step: For $n = k+1$, we have:

$$\begin{aligned} 1 + 2 + \ldots + k + (k+1) &= \frac{k(k+1)}{2} + (k+1) \\ &= \frac{k(k+1) + 2(k+1)}{2} \\ &= \frac{(k+1)(k+2)}{2} \end{aligned}$$

which is exactly $\frac{(k+1)((k+1)+1)}{2}$, and the proof is complete. ■

R Actually, the example only show part of the MI, as in the real practice, we need to find the statement we would like to prove either by examine or reasonable postulation. Do try to get this idea by completing the problem set for this section.

1.5.2 Strong Mathematical Induction

Strong mathematical induction is the other method of proving that a statement holds for all natural numbers greater than or equal to some initial value. The main difference between the two methods is that strong mathematical induction allows for the use of the statement for all natural numbers less than the current value n in the inductive step, while mathematical induction only allows for the use of the statement for the previous value $n-1$. This additional flexibility can make strong mathematical induction a more powerful tool for proving certain types of statements. When we proceed to prove something with MI, it follows this pattern:

Proof. The proof is by mathematical induction.

Base Case:

First we prove that $P(n_0)$ holds.

Inductive Assumption/hypothesis:

Then, assume that the statement holds for all values k such that $n_0 \leq k < n$

Inductive Step:

Finally, with is assumption, prove $P(n)$ also holds.

By the principle of mathematical induction, $P(n)$ is true for all integers $n \geq n_0$. ■

Here is an example:

▪ **Example 1.6** Prove the following theorem:

Theorem 1.3 For all natural numbers n, the sum of the first n odd numbers is equal to n^2.

▪

Proof. **Base Case:** - When $n = 1$, the sum of the first n odd numbers is 1. Also, $1^2 = 1$. Therefore, the statement holds for $n = 1$.

Inductive Hypothesis: - Assume that $P(k)$ is true for some integer $k \geq 1$. That is, assume that the sum of the first k odd numbers is equal to k^2.

Inductive Step: - We need to prove that if $P(k)$ is true, then $P(k+1)$ is also true. That is, we need to show that the sum of the first $k+1$ odd numbers is equal to $(k+1)^2$.

- The sum of the first $k+1$ odd numbers is given by:

$$1+3+5+\cdots+(2k+1)+(2k+3) = \sum_{i=1}^{k+1}(2i-1)$$

- Using the inductive hypothesis, we know that the sum of the first k odd numbers is k^2. Therefore, we have:

$$\sum_{i=1}^{k+1}(2i-1) = k^2+(2k+1) = (k+1)^2$$

- This shows that if $P(k)$ is true, then $P(k+1)$ is also true.

Therefore, by the principle of mathematical induction, we can conclude that for all natural numbers n, the sum of the first n odd numbers is equal to n^2. ■

R Sigma notation is a concise and powerful way to represent summation in mathematics. It is denoted by the Greek letter sigma (Σ). The general form of sigma notation is:

$$\sum_{i=m}^{n} a_i$$

This represents the sum of the terms a_i for i ranging from m to n. Here, i is the index of summation; m is the lower limit of summation, where the summation starts; and n is the upper limit, where the summation ends. Each term a_i in the series is generated by substituting values of i from m to n. For example, the sum of the first n natural numbers can be expressed as:

$$\sum_{i=1}^{n} i = 1+2+3+\cdots+n$$

Sigma notation is particularly useful in dealing with series and sequences in mathematics, allowing complex sums to be written in a more compact and readable form. We will explore its property in detail when we get to know sequence and series later in this book.

1.5.3 Exercises

Exercise 1.16 Prove, using mathematical induction, the following theorem.

Theorem 1.4 — Sum of the first n positive integers. The sum of the squares of the first n positive integers is:

$$1^2+2^2+3^2+\ldots+n^2=\frac{n(n+1)(2n+1)}{6}$$

for all integers $n \geq 1$.

Hint: Follow the same procedure of the example. This is a pure algebraic proof. ■

Proof. **Base Case:** For $n=1$, $1^2=\frac{1(1+1)(2\cdot 1+1)}{6}=1$, which is true.

Inductive Hypothesis: Assume the statement is true for $n=k$, i.e.,

$$1^2+2^2+\ldots+k^2=\frac{k(k+1)(2k+1)}{6}$$

Inductive Step: For $n=k+1$, we have:

$$\begin{aligned}
1^2+2^2+\ldots+k^2+(k+1)^2 &= \frac{k(k+1)(2k+1)}{6}+(k+1)^2 \\
&= \frac{k(k+1)(2k+1)+6(k+1)^2}{6} \\
&= \frac{(k+1)(k(2k+1)+6(k+1))}{6} \\
&= \frac{(k+1)(2k^2+7k+6)}{6} \\
&= \frac{(k+1)(k+2)(2k+3)}{6}
\end{aligned}$$

which completes the proof. ■

Exercise 1.17 For $n \in \mathbb{N}$, prove that

$$\sum_{i=1}^{n}(2i-1)^2=\frac{n(2n-1)(2n+1)}{3}.$$

■

Proof. We aim to prove that for all $n \in \mathbb{N}$, the following formula holds:

$$\sum_{i=1}^{n}(2i-1)^2=\frac{n(2n-1)(2n+1)}{3}.$$

Base Case: Let $n=1$.

$$\sum_{i=1}^{1}(2i-1)^2=(2\cdot 1-1)^2=1^2=1,$$

and

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

Since both sides equal 1, the base case holds.

Inductive Hypothesis: Assume the statement is true for some positive integer k. That is,

$$\sum_{i=1}^{k}(2i-1)^2 = \frac{k(2k-1)(2k+1)}{3}.$$

Inductive Step: We must show that the statement holds for $k+1$. Consider the sum up to $k+1$:

$$\sum_{i=1}^{k+1}(2i-1)^2 = \sum_{i=1}^{k}(2i-1)^2 + (2(k+1)-1)^2.$$

Using the inductive hypothesis, we can write this as:

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

Simplifying the right-hand side, we get:

$$\frac{k(2k-1)(2k+1)+3(2k+1)^2}{3} = \frac{(2k+1)[k(2k-1)+3(2k+1)]}{3}.$$

Expanding the terms inside the brackets gives us:

$$\frac{(2k+1)(2k^2-k+6k+3)}{3} = \frac{(2k+1)(2k^2+5k+3)}{3}.$$

This simplifies to:

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

Notice that $(2k+3)$ is just $2(k+1)+1$, so our expression is equivalent to:

$$\frac{(k+1)(2(k+1)-1)(2(k+1)+1)}{3},$$

which matches the right-hand side of our original equation for $n = k+1$.

Therefore, by the principle of mathematical induction, the given formula is true for all $n \in \mathbb{N}$. ■

Exercise 1.18 prove that for all integers $n \geq 1$:

$$\frac{1}{1\cdot 3} + \frac{1}{3\cdot 5} + \frac{1}{5\cdot 7} + \cdots + \frac{1}{(2n-1)(2n+1)} = \frac{n}{2n+1}$$

■

Proof. **Base Case:** For $n = 1$,

$$\frac{1}{1\cdot 3} = \frac{1}{3} = \frac{1}{2\cdot 1+1}$$

Inductive Step: Assume the statement is true for $n = k$, that is:

$$\sum_{i=1}^{k} \frac{1}{(2i-1)(2i+1)} = \frac{k}{2k+1}$$

We need to show it holds for $n = k+1$:

$$\begin{aligned}
\sum_{i=1}^{k+1} \frac{1}{(2i-1)(2i+1)} &= \sum_{i=1}^{k} \frac{1}{(2i-1)(2i+1)} + \frac{1}{(2(k+1)-1)(2(k+1)+1)} \\
&= \frac{k}{2k+1} + \frac{1}{(2k+1)(2k+3)} \\
&= \frac{k(2k+3)+1}{(2k+1)(2k+3)} \\
&= \frac{2k^2+3k+1}{(2k+1)(2k+3)} \\
&= \frac{(k+1)(2k+1)}{(2k+1)(2k+3)} \\
&= \frac{k+1}{2k+3} \\
&= \frac{k+1}{2(k+1)+1}
\end{aligned}$$

Therefore, by the principle of mathematical induction, the statement is true for all integers $n \geq 1$.

■

Exercise 1.19 prove that for a fixed nonnegative integer q and for all positive integers n, the following equation holds:

$$\sum_{j=1}^{n} j(j+1)(j+2)\dots(j+q) = \frac{n(n+1)(n+2)\dots(n+q)(n+q+1)}{q+2}$$

■

Hint: Find the correlation between the sum of the first n and the first $n+1$ term.

Proof. **Base Case:** For $n = 1$, the sum on the left-hand side is simply $1 \cdot 2 \cdot 3 \dots (1+q)$, which is the product of the first $q+1$ positive integers after 1. The right-hand side is

$$\frac{1 \cdot 2 \cdot 3 \dots (1+q)(1+q+1)}{q+2}$$

which, after cancellation of the similar terms in the numerator and the denominator, also equals the product of the first $q+1$ positive integers after 1. Hence, the statement holds for $n = 1$.

Inductive Hypothesis: Assume the statement is true for $n = k$. That is,

$$\sum_{j=1}^{k} j(j+1)(j+2)\dots(j+q) = \frac{k(k+1)(k+2)\dots(k+q)(k+q+1)}{q+2}$$

Inductive Step: Now we need to show that the statement holds for $n = k+1$. Consider the sum

$$\sum_{j=1}^{k+1} j(j+1)(j+2)\dots(j+q)$$

This sum can be split into two parts: the sum up to k (which we assume is correct by the inductive hypothesis) and the $(k+1)^{st}$ term. So, we have:

$$\sum_{j=1}^{k+1} j(j+1)(j+2)\dots(j+q) =$$

$$\left(\sum_{j=1}^{k} j(j+1)(j+2)\dots(j+q)\right) + (k+1)(k+2)(k+3)\dots(k+q+1) =$$

$$\frac{k(k+1)(k+2)\dots(k+q)(k+q+1)}{q+2} + (k+1)(k+2)(k+3)\dots(k+q+1)$$

To simplify the right-hand side, factor out the common term $(k+1)(k+2)(k+3)\dots(k+q+1)$ from both parts:

$$\begin{aligned}
&= (k+1)(k+2)(k+3)\dots(k+q+1)\left(\frac{k}{q+2}+1\right)\\
&= (k+1)(k+2)(k+3)\dots(k+q+1)\left(\frac{k+q+2}{q+2}\right)\\
&= \frac{(k+1)(k+2)(k+3)\dots(k+q+1)(k+q+2)}{q+2}
\end{aligned}$$

Which is exactly the right-hand side of the original equation for $n = k+1$. Therefore, the inductive step is proven.

■

Exercise 1.20 Prove by Mathematical Induction (MI) that for all Natural Number n,

(a)

$$\sum_{j=0}^{n} (j+1)2^j = n2^{n+1} + 1.$$

(b)

$$\sum_{j=0}^{n} (j+1)3^j = \frac{[2n+1]3^{n+1}+1}{4}.$$

(c)

$$\sum_{j=0}^{n}(j+1)r^j = \frac{[(r-1)n+(r-2)]r^{n+1}+1}{(r-1)^2} \quad \text{for all numbers } r \neq 1.$$

Hint: Examine the expressions. Do we really need 3 proofs?

Proof. We will prove (c) by mathematical induction. This also is also a proof of (a) where $r = 2$ and (b) where $r = 3$. Here $a = 0$ and $P(n)$ is an equation with a LHS and a RHS.

Base Case: If $n = 0$ then LHS $= (0+1)r^0 = 1$,

and RHS $= \frac{[(r-1)0+(r-2)]r^{0+1}+1}{(r-1)^2} = \frac{(r-2)r+1}{(r-1)^2} = 1$. $P(1)$ is True.

Inductive Hypothesis: Assume $\exists k \in \mathbb{N}$ where $P(k)$ is True.

Inductive Step: If $n = k+1$ then in the predicate P

$$\begin{aligned}
\text{LHS} &= \sum_{j=0}^{k+1}(j+1)r^j = \sum_{j=0}^{k}(j+1)r^j + (k+1+1)r^{k+1} \\
&= \frac{[(r-1)k+(r-2)]r^{k+1}+1}{(r-1)^2} + (k+1)r^{k+1} \quad \text{// by Step 2} \\
&= \frac{[(r-1)k+(r-2)]r^{k+1}+1+(k+1)(r-1)^2r^{k+1}}{(r-1)^2} \\
&= \frac{[(r(k+1)-k+(r+2)]+(k+2)r-2r+1]r^{k+1}+1}{(r-1)^2} \\
&= \frac{[(k+1)+(k+2)r-(2k+4)r]r^{k+1}+1}{(r-1)^2} \\
&= \frac{[(k+1)r+(k+2)(k+2)r^2-(k+2)2r+(k+2)3]r^{k+1}+1}{(r-1)^2} \\
&= \frac{[(k+1)+(k+2)r-(2k+4)r]r^{k+1}+1}{(r-1)^2} \\
&= \frac{[(k+1)r+r-k-1-2]r^{k+1}+1}{(r-1)^2} \\
&= \frac{[(r-1)(k+1)+(r-2)]r^{k+1}+1}{(r-1)^2} \\
&= \text{RHS}
\end{aligned}$$

Therefore, $\forall n \in \mathbb{N}, \sum_{j=0}^{n}(j+1)r^j = \frac{[(r-1)n+(r-2)]r^{n+1}+1}{(r-1)^2}$.

Therefore, when $r = 2, \forall n \in \mathbb{N}, \sum_{j=0}^{n}(j+1)2^j = \frac{[(2-1)n+(2-2)]2^{n+1}+1}{(2-1)^2} = n2^{n+1}+1$

and when $r = 3, \forall n \in \mathbb{N}, \sum_{j=0}^{n}(j+1)3^j = \frac{[(3-1)n+(3-2)]3^{n+1}+1}{(3-1)^2} = \frac{[2n+1]3^{n+1}+1}{4}$. ■

2.1 Set

We begin our formal development with basic notions of set theory. Our most primitive notion is that of a set. This notion is so fundamental that we do not attempt to give a precise definition. We think of a set as a collection of distinct objects with a precise description that provides a way of deciding (in principle) whether a given object is in it.

Definition 2.1 — Set. The objects in a set are its elements or members. When x is an element of A, we write $x \in A$ and say "x belongs to A". When x is not in A, we write $x \notin A$. If every element of A belongs to B, then A is a subset of B, and B contains A; we write $A \subseteq B$ or $B \supseteq A$.

R By convention, we use the special characters $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ to name the sets of natural numbers, integers, rational numbers, and real numbers, respectively. Each set in this list is contained in the next, so we write $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$.

Sets Could be expressed in different ways. For sets with limited and few elements, we simply list all the elements in a pair of bracket, such as $A = \{1,2,3,4,5\}$. For sets with more elements, we can also define a set by description: $A = \{x : x \in \mathbb{Z}^+ \text{ and } x \leq 5\}$. For example:

▪ **Example 2.1** The rational number set could be expressed as:

$$\mathbb{Q} = \{x : x \in \mathbb{R}, x = \frac{p}{q} \text{ where } p, q \in \mathbb{Z}, \text{ but } q \neq 0\}$$

▪

Definition 2.2 — Equality of Sets. Sets A and B are equal, written $A = B$, if they have the same elements. The empty set, written $\emptyset$, is the unique set with no elements. A proper subset of a set A is a subset of A that is not A itself. The power set of a set A is the set of all subsets of A. In other words, the complement of A includes everything that is not in A.

Definition 2.3 — Basic Set operations. Intersection, union and exclusion.

- the **intersection** of A and B,

$$A \cap B = \{x : x \in A \text{ and } x \in B\};$$

- the **union** of A and B,

$$A \cup B = \{x : x \in A \text{ or } x \in B\};$$

- the **set difference**, A but not B,

$$A \setminus B = \{x : x \in A \text{ and } x \notin B\}.$$

The set $A \setminus B$ is sometimes called the "relative complement" of B in A.
When $A \cap B = \emptyset$, sets A and B are said to be **disjoint**.

Definition 2.4 — Complement of Set. The complement of a set A, often denoted as $\overline{A}$, $\sim A$ or A^c, is defined with respect to a universal set U, which contains all objects under consideration. The complement $\overline{A}$ consists of all elements in U that are not in A. Formally, if we have a universal set U and a subset $A \subseteq U$, then the complement of A is given by:

$$\overline{A} = \{x \in U \mid x \notin A\}$$

Definition 2.5 — Intervals. Intervals. When $a, b \in \mathbb{R}$ with $a < b$, the closed interval $[a,b]$ is the set $\{x \in \mathbb{R} : a \leq x \leq b\}$. The open interval (a,b) is the set $\{x \in \mathbb{R} : a < x < b\}$.

2.2 Properties of Sets with Proofs

Commutative Laws

Union:

$$A \cup B = B \cup A$$

Proof: The union of sets A and B includes all elements that are in A, in B, or in both. Since the notion of "being in" does not depend on the order, $A \cup B$ and $B \cup A$ represent the same set.

Intersection:

$$A \cap B = B \cap A$$

Proof: The intersection of sets A and B includes all elements that are both in A and in B. The order of A and B does not affect the elements that are shared between them, hence the equality.

Associative Laws

Union:

$$(A \cup B) \cup C = A \cup (B \cup C)$$

Proof: When we take the union of sets, we combine their elements. Grouping does not affect the outcome of the union, thus the union operation is associative.

Intersection:

$$(A \cap B) \cap C = A \cap (B \cap C)$$

Proof: The intersection operation finds common elements. The grouping of sets does not affect the commonality of elements, so the intersection operation is associative.

Distributive Laws

Intersection distributes over union:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

Proof: An element in $A \cap (B \cup C)$ is in A and either B or C. This is the same as the element being in both A and B, or in both A and C.

Union distributes over intersection:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Proof: An element in $A \cup (B \cap C)$ is in A, in both B and C, or in all. This is equivalent to the element being in A or B, and in A or C.

De Morgan's Laws

Complement of the union:

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

Proof: An element not in $A \cup B$ is neither in A nor in B, which means it is in both $\overline{A}$ and $\overline{B}$.

Complement of the intersection:

$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$

Proof: An element not in $A \cap B$ is not in both A and B, which means it is either in $\overline{A}$ or in $\overline{B}$.

Properties of Complements

Union with complement:

$$A \cup \overline{A} = U$$

Proof: The set A together with all elements not in A constitutes the entire universe U.

Intersection with complement:

$$A \cap \overline{A} = \emptyset$$

Proof: No element can be both in set A and not in set A at the same time, hence the intersection is the empty set.

Definition 2.6 — Subset. If A and B are sets, we say A is a subset of B if every element of A is also an element of B. This is denoted as $A \subseteq B$

R Note that for every set, it is a subset to itself.

Definition 2.7 — Proper Subset. If $B \subset A$, then B is a proper subset of A.

For instance, consider the set $A = \{1,2,3\}$ and the set $B = \{1,2\}$. In this case, $B \subset A$, because every element of B is in A, but A contains an additional element 3 that is not in B.

Definition 2.8 — Empty Set. The **empty set** is a unique set that contains no elements. It is denoted as $\varnothing$. This set is important in set theory because it serves as the identity element for the union operation and has properties that are fundamental to the construction of other sets. For example, every set, including the empty set itself, contains the empty set as a subset:

$$\varnothing \subseteq A$$

for any set A. Furthermore, the intersection of any set with the empty set is the empty set itself:

$$A \cap \varnothing = \varnothing$$

This highlights the empty set's role in set operations.

Definition 2.9 — Power Set. If A is any set, the **power set** of A,

$$\mathscr{P}(A) = \{S : S \subseteq A\}$$

// the set of all subsets of A

For example, if $A = \{a,b,c\}$ then

$$\mathscr{P}(A) = \{\varnothing, \{a\}, \{b\}, \{c\}, \{a,b\}, \{a,c\}, \{b,c\}, \{a,b,c\}\}.$$

Definition 2.10 — cardinality. The number of elements in a set S is called the **cardinality** of S and denoted by $|S|$. When this is a finite number, then $|S| \in \mathbb{N}$, and when $|S| = n$, we'll say that S is an n-set.

Definition 2.11 — Partition of Sets. Each element of $A \cup B$ is in exactly one of the sets $A \setminus B$, $B \setminus A$, and $A \cap B$. More generally,

subsets $S_1, S_2, S_3, \ldots, S_k$ of T form a **partition** of T means every element of T belongs to exactly one of the sets S_j.

The sets $S_1 = A \setminus B$, $S_2 = B \setminus A$ and $S_3 = A \cap B$ form a partition of $T = A \cup B$. In general, $S_1 \cup S_2 \cup S_3 \ldots \cup S_k \subseteq T$ because each S_j is a subset of T. $T \subseteq S_1 \cup S_2 \cup S_3 \ldots \cup S_k$ because each element of T is in some subset S_j. Therefore, $T = S_1 \cup S_2 \cup S_3 \ldots \cup S_k$.

The subsets in a partition are **mutually disjoint**; that is, any two are disjoint sets.

If $p \neq q$, $S_p \cap S_q = \emptyset$ because no element of T belongs to more than one S_j.

When $S_1, S_2, S_3, \ldots, S_k$ forms a partition of T, then

$$|T| = |S_1| + |S_2| + |S_3| + \ldots + |S_k|.$$

Theorem 2.1 — The Principle of Inclusion-Exclusion. For any pair of sets,

$$|A \cup B| = |A| + |B| - |A \cap B|,$$

and when A and B are disjoint,

$$|A \cup B| = |A| + |B|. \quad \text{// since } A \cap B = \emptyset$$

Furthermore, we always have

$$|A \cup B| = |A \setminus B| + |B \setminus A| + |A \cap B|.$$

Definition 2.12 — The Cartesian Product. **The Cartesian product** of sets A and B, named for René Descartes (1596–1650), is

$$A \times B = \{(a,b) : a \in A \text{ and } b \in B\},$$

where (a,b) denotes an ordered pair of objects; there is a first entry and a second entry in each ordered pair. *Parentheses* indicate that order matters.

R

$\{0,1\} = \{1,0\}$, but $(0,1) \neq (1,0)$. – in sets, order doesn't matter; in ordered pairs it does.

$\{1,1\} = \{1\}$, but $(1,1) \neq (1)$. – in sets, repetitions don't matter; in ordered pairs they do.

■ **Example 2.2** If $A = \{1,3,5,7\}$ and $B = \{2,3,5\}$, then

$$A \times B = \{(1,2),(1,3),(1,5),(3,2),(3,3),(3,5),(5,2),(5,3),(5,5),(7,2),(7,3),(7,5)\}$$

$$B \times A = \{(2,1),(2,3),(2,5),(2,7),(3,1),(3,3),(3,5),(3,7),(5,1),(5,3),(5,5),(5,7)\}.$$

$$(A \times B) \cap (B \times A) = \{(3,3),(3,5),(5,3),(5,5)\}, \text{ so } A \times B \neq B \times A.$$

■

2.2.1 Exercises

Exercise 2.1 Indicate whether each statement is true or false:
(a) $\{4,0,3,0\} = \{4,4,0,3\}$
(b) $\{4\} \subseteq \{0,3,4\}$
(c) $\{0,3,4\} \subseteq \{4\}$
(d) $\{0,3,4\} \subseteq \{0,3,4\}$
(e) $\emptyset \subseteq \{0,3,4\}$

■

Exercise 2.2 What is $\mathscr{P}(\{0,3,4,7\})$? ■

Exercise 2.3 Let $A = \{1,2,3,4\}$ and $B = \{2,3,5,8\}$. Evaluate each of the following expressions:
(a) $A \cap B$
(b) $A \cup B$
(c) $A \setminus B$
(d) $A \times B$

■

Exercise 2.4 Is $\{1,3\},\{2,3\},\{4\}$ a partition of $\{1,2,3,4\}$? Justify your answer. ■

Exercise 2.5 Consider the set $\{a,b,c,d,e\}$. Construct 3 different partitions of this set. ■

Hint: Each partition satisfies the definition: the subsets are non-empty, they cover the entire original set, and they are mutually exclusive (no element is repeated in the subsets of any given partition).

Partition 1:
- $\{a\},\{b\},\{c\},\{d\},\{e\}$

Partition 2:
- $\{a,b\},\{c,d,e\}$

Partition 3:
- $\{a,e\},\{b,c\},\{d\}$

Exercise 2.6 Proof that $A \cap (B - C) = (A \cap B) - (A \cap C)$ ■

Proof. To prove the lemma, we will show that the left-hand side (LHS) is a subset of the right-hand side (RHS) and vice versa.

Using De Morgan's laws:

$$\begin{aligned} \sim (A \cap C) &= \sim A \cup \sim C \\ \therefore (A \cap B) \sim (A \cap C) &= (A \cap B) \cap (\sim A \cup \sim C) \end{aligned}$$

Distributing the intersection over the union:

$$\begin{aligned} (A \cap B) \cap (\sim A \cup \sim C) &= (A \cap B \cap \sim A) \cup (A \cap B \cap \sim C) \\ &= \varnothing \cup (A \cap B \cap \sim C) \quad (\text{since } A \cap \sim A = \varnothing) \\ &= A \cap B \cap \sim C \end{aligned}$$

Simplifying further:

$$\begin{aligned} A \cap B \cap \sim C &= A \cap (B \cap \sim C) \\ &= A \cap (B - C) \quad (\text{since } B \cap \sim C = B - C) \end{aligned}$$

The original statement is proven, as both the LHS and RHS equal $A \cap (B - C)$. ■

Exercise 2.7 Prove that $(A \cup B) - (A \cap B) = (B - A) \cup (A - B)$. ■

Proof. We start by applying De Morgan's laws:

$$\begin{aligned} (A \cup B) - (A \cap B) &= (A \cup B) \cap \sim (A \cap B) \\ \text{By De Morgan's laws:} \quad \sim (A \cap B) &= \sim A \cup \sim B \\ \therefore (A \cup B) \cap \sim (A \cap B) &= (A \cup B) \cap (\sim A \cup \sim B) \end{aligned}$$

Next, we distribute the intersection over the union:

$$\begin{aligned} (A \cup B) \cap (\sim A \cup \sim B) &= ((A \cup B) \cap \sim A) \cup ((A \cup B) \cap \sim B) \\ &= (\varnothing \cup (A \cap \sim B)) \cup (\varnothing \cup (B \cap \sim A)) \\ &= (A \setminus B) \cup (B \setminus A) \end{aligned}$$

Hence, the original statement is proven. ■

Exercise 2.8 Define the set $S = \{x | x = 12m + 8n, m, n \in \mathbb{Z}\}$, and let $P = \{x | x = 20p + 16q, p, q \in \mathbb{Z}\}$.

Prove that $S = P$. ■

Proof:

Let $x \in S$, then $x = 12m + 8n = 4(3m + 2n) = 20(3m + 2n) + 16(-3m - 2n) \in P$, so $S \subseteq P$;

Now let $x \in P$, then $x = 20p + 16q = 4(5p + 4q) = 12(5p + 4q) + 8(-5p - 4q) \in S$, so $P \subseteq S$; thus, by definition, $S = P$.

2.3 Function: a perspective of Set Theory

This section discusses function from a perspective of set. We clarify this by the relation between sets.

Definition 2.13 — Function. Let A and B be nonempty sets. A function f from A to B is an assignment of exactly one element of B to each element of A. We write $f(a) = b$ if b is the unique element of B assigned by the function f to the element a of A. If f is a function from A to B, we write $f : A \to B$.

R Mapping and transformation are equivalent to function in some context. If f is a function from A to B, we say that A is the *domain* of f and B is the *codomain* of f. If $f(a) = b$, we say that b is the *image* of a and a is a *preimage* of b. The *range*, or *image*, of f is the set of all images of elements of A. Also, if f is a function from A to B, we say that f *maps* A to B.

When we say that two functions are equal, they share the same domain, codomain, and the mapping from the domain to the same codomain.

Think about this problem:
Problem 2.1 Is $f(x) = \frac{1}{x}$ equal to $f(x) = x^{-1}$?

Even someone with limited algebra knowledge could know that, even though $\frac{1}{x}$ could be expressed in the same why, it has a different domain to x, as for $x \neq 0$ for the first function, while $x \in \mathbb{R}$ for the latter. Their domain and codomain are different.

This example helps to distinguish codomain and range:

■ **Example 2.3** Let $f : \mathbb{Z} \to \mathbb{Z}$ assign the square of an integer to this integer. Then, $f(x) = x^2$, where the domain of f is the set of all integers, the codomain of f is the set of all integers, and the range of f is the set of all integers that are perfect squares, namely, $\{0, 1, 4, 9, \ldots\}$. ■

Theorem 2.2 — Function Addition and Multiplication. Let f_1 and f_2 be functions from A to $\mathbb{R}$. Then $f_1 + f_2$ and $f_1 f_2$ are also functions from A to $\mathbb{R}$ defined for all $x \in A$ by

$$(f_1 + f_2)(x) = f_1(x) + f_2(x),$$
$$(f_1 f_2)(x) = f_1(x) f_2(x).$$

Problem 2.2 Let f_1 and f_2 be functions from $\mathbb{R}$ to $\mathbb{R}$ such that $f_1(x) = x^2$ and $f_2(x) = -x^2$. What are the functions $f_1 + f_2$ and $f_1 f_2$?

2.3.1 Injective, Surjective, and Bijective Function

We have known that functions are actually reflection from one set to the other set. We will look into several special mapping.

Definition 2.14 — Injective Function. A function $f : A \to B$ is called *injective* (or *one-to-one*) if every element of the codomain B is mapped by at most one element of the domain A. For example, the function $f(x) = 2x$ from $\mathbb{R}$ to $\mathbb{R}$ is injective because each value of $f(x)$ is produced by exactly one value of x.

Definition 2.15 — Surjective Function. A function is *surjective* (or *onto*) if every element of the codomain B is mapped by at least one element of the domain A. For instance, the function $g(x) = \sin(x)$ from $\mathbb{R}$ to $[-1, 1]$ is surjective because every value in $[-1, 1]$ is the sine of some real number x.

Definition 2.16 — Bijective Function. A function is *bijective* if it is both injective and surjective, which means there is a perfect "pairing" between the sets: every element of A is paired with a unique element of B, and every element of B is paired with a unique element of A. An example of a bijective function is the identity function $i(x) = x$ from $\mathbb{R}$ to $\mathbb{R}$.

Figure 2.1: Examples of Special Mappings

Examine the figure **??**, where one example of each type of function is shown. Fundamentally, when we define these functions, only two properties are involve:

- A: It the mapping retrievable? (for every value in the image, we can find where it is from without any confusion.
- B: Is the mapping is full comparing to the preimage?(Whether every value in the preimage has a mapping to the image)

If A is satisfied, we call it injective function.

If B is satisfied, we call it surjective function.

If both A and B are satisfied, we say the function is bijective.

Using A, B, and C to denote the set of injective, surjective, and bijective function respectively, it is therefore that :

$$A \cap B = C$$

Which means that if a function is injective and surjective in the same time, it is bijective.

And these will be all we need to know about function for now, since the rest of the properties will be discussed in single-variable calculus. This section aims only introduce the idea of injectivity, surjectivity, and bijectivity,

2.3.2 Exercises

Exercise 2.9 Determine the images of the functions $f : \mathbb{R} \to \mathbb{R}$ defined as follows:

a) $f(x) = \frac{x^2}{1+x^2}$.

b) $f(x) = \frac{x}{1+|x|}$.

Solution for a). We analyze the function $f(x) = \frac{x^2}{1+x^2}$:

- This function is defined for all $x \in \mathbb{R}$.
- For $x = 0$, $f(0) = 0$.
- For $x \neq 0$, $f(x)$ is always positive.
- As x approaches infinity, $f(x)$ approaches 1.

Thus, the image of f is $(0,1]$. ■

Solution for b). We analyze the function $f(x) = \frac{x}{1+|x|}$:

- This function is defined for all $x \in \mathbb{R}$.
- For $x > 0$, as x increases, $f(x)$ approaches 1.
- For $x < 0$, as x decreases, $f(x)$ approaches -1.

Thus, the image of f is $(-1,1)$. ■

2.4 Summation

Before we move on to the most important part of this chapter, sequence, we use this section to introduce a prerequisite for studying its properties. We introduce the Sigma sign.

In earlier chapters, we have seen exercises such as finding the expression of the sum of the first nth positive integer:

$$1+2+\cdots+(n-1)+(n) = \frac{n(n+1)}{2}$$

From now on, we will use the sigma notation to deal with the summation of numbers. Such as:

2.4.1 Sigma Notation

▪ Notation 2.1 — Sigma Notation.

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

Where i is quite similar to iterator, or sometimes we also call counter in programming languages, and n refers to the condition of termination. The expression right after the sigma sign is called **summand**.

Actually, this is not the only way to express summation, it is also equivalent to:

$$\sum_{1\leq i\leq n} i = \frac{n(n+1)}{2}$$

Also, like what we do for set, we can also write sigma notation using description, such as:

$$\sum_{1\leq k\leq 100} k^2$$

k is odd

2.4.2 Properties and Techniques of Sigma Notation

This part of the section shows how we can handle summation expressions. One of the greatest convenience of sigma notation is that every expression is adjustable, we can change the variable as what we prefer as in the following example.

■ **Example 2.4**

$$\sum_{1\leq k\leq n} a_k = \sum_{1\leq k+1\leq n} a_{k+1}$$

■

This technique has a significant effect to some mathematical proofs. Another points to keep in mind is that: always make the expression simple in terms of upper and lower boundary.

■ **Example 2.5** Examine this expression:

$$\sum_{k=0}^{n} k(k-1)(k-n)$$

The sum when k equals to 0, 1, and n is 0. In this case we cannot say it is a good expression, as what we want is the sum it self, while 0 does not matter for us. Therefore, we just fine-tune it to:

$$\sum_{k=2}^{n-1} k(k-1)(k-n)$$

This makes it concise and clear. ■

2.4.2.1 Manipulation of Sigma Notation

For a set K, the following summation properties hold. Let c be a constant, and a_k, b_k be sequences indexed by K:

$$\sum_{k\in K} ca_k = c\sum_{k\in K} a_k \quad \text{(Distributive Law)} \tag{2.1}$$

$$\sum_{k\in K} (a_k + b_k) = \sum_{k\in K} a_k + \sum_{k\in K} b_k \quad \text{(Associative Law)} \tag{2.2}$$

$$\sum_{k\in K} a_k = \sum_{p(k)\in K} a_{p(k)} \quad \text{(Commutative Law, as in example ??)} \tag{2.3}$$

The proof is attached below.

Proof of Constant Factor Law:

Let c be a constant and a_k be a sequence indexed by a finite set K. We want to show that $\sum_{k\in K} ca_k = c\sum_{k\in K} a_k$.

By the definition of summation and the distributive property of multiplication over addition, we have:

$$\sum_{k\in K} ca_k = ca_1 + ca_2 + \ldots + ca_n = c(a_1 + a_2 + \ldots + a_n) = c\sum_{k\in K} a_k. \tag{2.4}$$

This concludes the proof of the constant factor law.

Proof of Summation of Sums Law:

Let a_k and b_k be sequences indexed by a finite set K. We want to show that $\sum_{k\in K}(a_k + b_k) = \sum_{k\in K} a_k + \sum_{k\in K} b_k$.

By the definition of summation and the associative and commutative properties of addition, we have:

$$\begin{aligned}\sum_{k\in K}(a_k+b_k) &= (a_1+b_1)+(a_2+b_2)+\ldots+(a_n+b_n)\\ &= (a_1+a_2+\ldots+a_n)+(b_1+b_2+\ldots+b_n)\\ &= \sum_{k\in K} a_k + \sum_{k\in K} b_k\end{aligned} \tag{2.5}$$

This concludes the proof of the summation of sums law.

Proof of Permutation Invariance Law:

Let a_k be a sequence indexed by a finite set K. Let $p : K \to K$ be a bijection, which means p permutes the indices. We want to show that $\sum_{k\in K} a_k = \sum_{p(k)\in K} a_{p(k)}$.

By the definition of summation and the fact that addition is commutative (the order does not matter), we have:

$$\sum_{k\in K} a_k = a_1+a_2+\ldots+a_n = a_{p(1)}+a_{p(2)}+\ldots+a_{p(n)} = \sum_{p(k)\in K} a_{p(k)}. \tag{2.6}$$

This concludes the proof of the permutation invariance law.

2.4.2.2 Multiple Sums

Sometimes we use sigma notation with multiple variables, just like what we can do to write loops in programming languages.

$$\sum_{1\le j,k\le 3} a_j b_k = a_1b_1+a_1b_2+a_1b_3+a_2b_1+a_2b_2+a_2b_3+a_3b_1+a_3b_2+a_3b_3$$

In the context of summation, we often encounter a situation where a sum is taken over a set of pairs. Specifically, let $P(j,k)$ be a property involving the indices j and k, and $a_{j,k}$ be elements corresponding to these indices. The summation over all pairs (j,k) satisfying property P is equivalent to summing over all indices separately:

$$\sum_{P(j,k)} a_{j,k} = \sum_{j,k} a_{j,k}\cdot P(j,k).$$

This notation serves as a shorthand for expressing the sum over a subset of indices determined by the property P. There are also cases where we must use two sigma notation in the same time.

▪ **Example 2.6 — Double Summation.** When considering a double sum over a set of pairs, we often come across the following identity:

$$\sum_j\sum_k a_{j,k}[P(j,k)] \tag{2.7}$$

where $[P(j,k)]$ is an Iverson bracket which equals 1 if the property P holds for the pair (j,k) and 0 otherwise. ▪

By interchanging the order of summation, we observe that:

$$\sum_j \sum_k a_{j,k}[P(j,k)] = \sum_k \sum_j a_{j,k}[P(j,k)]. \tag{2.8}$$

This property allows us to switch the order of summation without changing the result, which can be particularly useful in various mathematical analyses.

Double summation could also be used to simplify a given summation. Considering the expression at the beginning of this section:

$$\sum_{1 \le j,k \le 3} a_j b_k = a_1b_1 + a_1b_2 + a_1b_3 + a_2b_1 + a_2b_2 + a_2b_3 + a_3b_1 + a_3b_2 + a_3b_3$$

■ **Example 2.7 — Converting to Double Summation.**

$$\begin{aligned}
\sum_{1 \le i,j,k \le 3} a_j b_k &= \sum_{\substack{i,j,k \\ 1 \le i,j,k \le 3}} a_j b_k \, [1 \le j \le 3] \, [1 \le k \le 3] \\
&= \sum_j \sum_k a_j b_k \, [1 \le j \le 3] \, [1 \le k \le 3] \\
&= \sum_j a_j \, [1 \le j \le 3] \sum_k b_k \, [1 \le k \le 3] \\
&= \sum_j a_j \, [1 \le j \le 3] \left(\sum_k b_k \, [1 \le k \le 3] \right) \\
&= \left(\sum_j a_j \, [1 \le j \le 3] \right) \left(\sum_k b_k \, [1 \le k \le 3] \right) \\
&= \left(\sum_{j=1}^{3} a_j \right) \left(\sum_{k=1}^{3} b_k \right).
\end{aligned}$$

■

To explicit: In the situation where we perform the same range of summation over each variable, the first two lines' triple summation is:

$$(a_1b_1 + a_1b_2 + a_1b_3) + (a_2b_1 + a_2b_2 + a_2b_3) + (a_3b_1 + a_3b_2 + a_3b_3).$$

Utilizing the distributive property to combine the summation operations into one involving a, since a and each k for $1 \le j \le 3$ are independent, yields (as in the third line):

$$a_1(b_1 + b_2 + b_3) + a_2(b_1 + b_2 + b_3) + a_3(b_1 + b_2 + b_3).$$

Consider a double sum over two independent indices, if the indices are independent, the summation of the product can be split into the product of two summations. For instance, the sum of products of a_j and b_k over j in J and k in K can be expressed as: $(a_1 + a_2 + a_3)(b_1 + b_2 + b_3)$. This can be generalized to an expression:

$$\sum_{j \in J} \sum_{k \in K} a_j b_k = \left(\sum_{j \in J} a_j \right) \left(\sum_{k \in K} b_k \right), \tag{2.9}$$

which is known as the **general distributive law**.

R If you are an agile reader, you must have noticed that this expression is a kind of representation of Cartesian sets in algebra. The general distributive law allows the sum over a function of elements from the Cartesian product of two sets to be expressed as the product of sums over each set if the function is separable into independent factors.

2.4.3 Exercises

Exercise 2.10 Express the triple sum

$$\sum_{1\le i<j<k\le 4} a_{ijk}$$

as a three-fold summation (with three $\sum$'s),

a. summing first on k, then j, then i;
b. summing first on i, then j, then k.

Also write your triple sums out in full without the $\sum$-notation, using parentheses to show what is being added together first. ■

Solution:

(a)

$$\sum_{i=1}^{4}\sum_{j=i+1}^{4}\sum_{k=j+1}^{4} a_{ijk} = \sum_{i=1}^{2}\sum_{j=i+1}^{3}\sum_{k=j+1}^{4} a_{ijk} = ((a_{123}+a_{124})+a_{134})+a_{234}.$$

(b)

$$\sum_{k=1}^{4}\sum_{j=1}^{k-1}\sum_{i=1}^{j-1} a_{ijk} = \sum_{k=3}^{4}\sum_{j=2}^{k-1}\sum_{i=1}^{j-1} a_{ijk} = a_{123}+(a_{124}+a_{134}+a_{234}).$$

Exercise 2.11 Demonstrate your understanding of Σ-notation by writing out the sums

$$\sum_{k=0}^{5} a_k \quad \text{and} \quad \sum_{0\le k^2\le 5} a_{k^2}$$

in full. (Watch outthe second sum is a bit tricky.) ■

Solution:

The first sum is:

$$a_0+a_1+a_2+a_3+a_4+a_5$$

The second sum, $k\in\{-2,-1,0,1,2\}$, therefore:

$$a_4+a_1+a_0+a_1+a_4$$

Exercise 2.12 The general rule for summation by parts is equivalent to

$$\sum_{0\le k<n}(a_{k+1}-a_k)b_k = a_nb_n - a_0b_0 - \sum_{0\le k<n} a_{k+1}(b_{k+1}-b_k), \quad \text{for } n\ge 0.$$

Prove this formula by using the distributive, associative, and commutative laws. ■

Hint: Use Associative Law to LHS, try to make the indices of the two sums as similar as possible.

Proof.

$$\begin{aligned}
\text{LHS} &= \sum_{0\le k<n} a_kb_{k+1} - \sum_{0\le k<n} a_kb_k \\
&= \sum_{0\le k<n} a_kb_{k+1} - \sum_{-1\le k<n-1} a_{k+1}b_{k+1} \\
&= \sum_{0\le k<n} a_kb_{k+1} - \sum_{0\le k<n-1} a_{k+1}b_{k+1} \\
&= \sum_{k=0}^{n-1} a_kb_{k+1} - \sum_{k=0}^{n-2} a_{k+1}b_{k+1} \\
&= a_nb_{n-1} - a_0b_0 + \sum_{k=0}^{n-2} a_kb_k - \sum_{k=0}^{n-2} a_{k+1}b_{k+1} \\
&= a_nb_{n-1} - a_0b_0 + \sum_{0\le k<n-1} a_k(b_k - b_{k+1}) \\
&= a_n(b_n - b_{n-1}) + a_nb_{n-1} - a_0b_0 - \sum_{0\le k<n-1} a_{k+1}(b_{k+1}-b_k) \\
&= a_n(b_n - b_{n-1} + b_{n-1}) - a_0b_0 - \sum_{0\le k<n} a_{k+1}(b_{k+1}-b_k) \\
&= a_nb_n - a_0b_0 - \sum_{0\le k<n} a_{k+1}(b_{k+1}-b_k) \\
&= \text{RHS}
\end{aligned}$$

■

Exercise 2.13 Is the following expression correct or not? Give your reason.

$$\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{j=1}^{n}\frac{1}{a_j}\right) = \sum_{1\le i\le n}\sum_{1\le j\le n}\frac{a_i}{a_j} = \sum_{1\le i\le n}\sum_{1\le i\le n}\frac{a_i}{a_i} = \sum_{i=1}^{n} 1 = n$$

■

Solution:

Consider the expression given by:

$$\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{j=1}^{n}\frac{1}{a_j}\right)$$

and its expansion into a double sum:

$$\sum_{1\le i\le n}\sum_{1\le j\le n}\frac{a_i}{a_j}$$

It is claimed that this is equal to:

$$\sum_{1\leq i\leq n}\sum_{1\leq i\leq n}\frac{a_i}{a_i}=\sum_{i=1}^{n}1=n$$

However, this claim overlooks the fact that the double sum includes terms where $i\neq j$, which are not necessarily equal to 1. Only when $i=j$ does the term $\frac{a_i}{a_j}$ simplify to 1, contributing to the count of n.

Hence, the proper expansion of the double sum should be written as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{a_i}{a_j}=\sum_{i=1}^{n}1+\sum_{\substack{i,j=1\\ i\neq j}}^{n}\frac{a_i}{a_j}$$

where the first sum on the right-hand side counts the n instances where $i=j$, and the second sum accounts for the $n(n-1)$ instances where $i\neq j$.

The claim would only be true if all a_i are equal, which is a special case, not the general case. In the general case, the expression evaluates to something different than n due to the presence of terms where $i\neq j$.

Therefore, the original statement is incorrect unless the condition that all a_i are equal is specified.

Exercise 2.14 Consider the following double summation where $a_i,a_j,b_i,b_j\in\mathbb{R}$.

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(a_ib_j-a_jb_i)$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(a_ib_j-a_jb_i)^2$$

Is there anything special about these expressions? Manage to find all the equivalent expressions of the sum of squares in sigma notation. Also consider, if the order of summand increases to infinity, whether these properties still exist? ■

Solutions:

For $\sum_{i=1}^{n}\sum_{j=1}^{n}(a_ib_j-a_jb_i)$

- It could be seen that the sums are actually symmetrical. When $i=j$, $a_ib_j-a_jb_i=0$.
- If you list several of the first nth term, the term with indices (i,j) will be canceled by (j,i) term, since $(a_ib_j-a_jb_i)+(a_jb_i-a_ib_j)=0$

We can visualize it in a matrix with $n=5$.

Figure 2.2: Visualization of $\sum_{i=1}^{n}\sum_{j=1}^{n}(a_ib_j - a_jb_i)$

With this image, we expand it until n; we can still cancel all elements symmetrical by the diagonal on by one. As the sum on the diagonal is 0, we conclude that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(a_ib_j - a_jb_i) = 0$$

Now consider the summation of squares. $(a_ib_j - a_jb_i)^2 = 0$ still holds for $i = j$, but what about the symmetrical pairs (i, j) and (j, i)? We can figure it out by analysis by expanding the square of sum.

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_i^2b_j^2 - 2a_ia_jb_ib_j + a_j^2b_i^2\right)$$

By associative property of summation, we rearrange it as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i^2b_j^2 + a_j^2b_i^2) - 2\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jb_ib_j$$

When $i \neq j$, each pair of (i, j) and (j, i). The sum of symmetric pair is

$$(a_i^2b_j^2 + a_j^2b_i^2 - 2a_ia_jb_ib_j) + (a_j^2b_i^2 + a_i^2b_j^2 - 2a_ia_jb_ib_j)$$

Rearrange it as:

$$2(a_i^2b_j^2 + a_j^2b_i^2) - 4(a_ia_jb_ib_j)$$

Still, as the sum of (i, j) terms where $i = j$ is 0. We can ignore the diagonal. Hence, we have $n/2$ pairs of $(a_i^2b_j^2 + a_j^2b_i^2) - 4(a_ia_jb_ib_j)$. This could be written as:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}2[(a_i^2b_j^2 + a_j^2b_i^2) - 4(a_ia_jb_ib_j)]$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}[(a_i^2b_j^2+a_j^2b_i^2)-2(a_ia_jb_ib_j)]$$

Notice that, $\sum_{i=1}^{n}\sum_{j=1}^{n}a_i^2b_j^2=\sum_{i=1}^{n}\sum_{j=1}^{n}a_j^2b_i^2$ due to the diagonal symmetry of the summation as illustrated in the graph. We are adding the same term twice. Hence, we have:

$$2\sum_{i=1}^{n}\sum_{j=1}^{n}[a_i^2b_j^2-(a_ia_jb_ib_j)]$$

We can further simplify it by rule out the terms where $i=j$, as the summand is 0 in those cases. We rewrite the sum as:

$$2\sum_{i=1}^{n-1}\sum_{j=2}^{n}[a_i^2b_j^2-(a_ia_jb_ib_j)]$$

or in single summation, we have:

$$2\sum_{i\neq j}[a_i^2b_j^2-(a_ia_jb_ib_j)]=2\sum_{1\leq i<j\leq n}[a_i^2b_j^2-(a_ia_jb_ib_j)]$$

For $\sum_{i=1}^{n}\sum_{j=1}^{n}(a_ib_j-a_jb_i)^n$, the symmetric property of summand still exists. However we cannot prove it for now, as we need to use polynomial theorem to be introduced in Combinatorics.

2.5 Sequence

If everything so far is not a problem for you, congratulations, because you have known everything you need to know sequence. Sequence is an important concept that will be used throughout your journey of learning math. A Sequence is defined as:

Definition 2.17 A sequence is a function $f:\mathbb{N}\to\mathbb{R}$, where $\mathbb{N}$ is the set of natural numbers and $\mathbb{R}$ is the set of real numbers. The value $f(n)$ is the n-th term of the sequence, often denoted as a_n. Therefore, a sequence can be represented as $\{a_n\}_{n=1}^{\infty}$ for an infinite sequence or $\{a_n\}_{n=1}^{N}$ for a finite sequence of length N.

2.5.1 Introduction

Sequences could be either infinite or finite. A *sequence* is defined to be a function S whose domain D is a nonempty interval of integers. S is an *infinite sequence* if D has the form $\{a..\}$.

S is a *finite sequence* if D has the form $\{a..b\}$ where $a\leq b$. When $|D|=n$, we will say that S is an n-sequence. We will take the domain of an n-sequence to be the set $\{1..n\}$.

The (natural) ordering of the domain of a sequence S gives a natural ordering to the ordered pairs in the set S. If S is a 5-sequence, then

$$S=\{(1,S(1)),(2,S(2)),(3,S(3)),(4,S(4)),(5,S(5))\}$$

▪ **Example 2.8** Suppose $D=\{1..10\}$, and we define the function S on D by

$$S(i)=\text{the smallest prime factor of the integer }(1+i).$$

Then D is a finite interval of integers, and so S is the sequence denoted by

$$S = (2,3,2,5,2,7,2,3,2,11).$$

■

Definition 2.18 — Sum of Sequence. If $S = (S_1, S_2, S_3, \ldots, S_n)$ is a finite sequence of numbers, the corresponding *series* is the sum of the entries in S is:

$$S_1 + S_2 + S_3 + \ldots + S_n.$$

2.5.2 Special Sequences

This section introduces common sequences as well as their properties.

2.5.2.1 Algorithmic Sequence

Algorithmic sequences are a fundamental concept in both computer science and mathematics, forming the backbone of algorithm design and analysis. These sequences are typically defined as an ordered set of steps or instructions, aimed at solving a specific problem or accomplishing a particular computation.

Definition 2.19 — Arithmetic Sequence. An arithmetic sequence of the form

$$a, a+d, a+2d, \ldots, a+nd, \ldots$$

where the initial term a and the common difference d are real numbers.

Usually, the notation a_n is used to express the nth term of a sequence (starting from 0). For the example in the definition, we have $a_0 = a$ and $a_n = a_0 + nd$. We also have:

$$a_1 = a_0 + d$$

$$a_2 = a_1 + d$$

$$\ldots$$

$$a_n = a_0 + nd$$

for $n >= 1$, $n \in \mathbb{Z}$:

$$a_n = a_{n-1} + d$$

These formula shows the linking between consecutive terms in an arithmetic sequence. We know that the sum s of the sequence is:

$$S = a_0 + a_1 + \ldots + a_n$$
$$S = a_0 + a_0 + d + \ldots + a_{n-1} + d$$

The sum is expressed in infinite terms, and it is called **open form equation**. Accordingly, there are also **closed form equations**.

Definition 2.20 — Open Form. An *open form* or *non-closed form* expression, on the other hand, does not have a finite standard representation and often requires recursive or iterative methods for evaluation. It may involve summations, integrals, or other operations that are not easily simplified into a finite number of operations.

Definition 2.21 — Closed Form. A *closed form* expression is a mathematical expression that can be evaluated in a finite number of standard operations. It typically involves constants, variables, and operations from algebra, calculus, and other areas of mathematics that can be computed in a finite number of steps. A closed form expression provides a direct way to compute the term of a sequence without the need for recursion.

Is the open form good for calculating the sum of a sequence? Suppose now I want to know S_{100} (The sum of the first 100th terms), with the open form, I still have to calculate 99 terms using the definition of this sequence. So is there a way to make it possible that we get the sum in one step? Think about closed form. The closed form allows us to calculate the sum directly. But is it possible to transform an open expression to closed form? If possible, how?

You may already notice that the open form has a property of infinity, and each step is somewhat related. Isn't it a perfect problem to be solved by mathematical induction? We will leave this proof as a exercise, and here we provide another direct proof by the symmetry of arithmetic sequence.

Theorem 2.3 — Sum of Arithmetic Sequence. For arithmetic sequence $a_0, a_1, \ldots, a_{n-1}$, where each term can be expressed as $a_i = a_0 + id$ and d is the common difference. The sum of the first nth terms is:

$$S = \frac{n}{2}[2a_0 + (n-1)d]$$

or

$$S = \frac{n}{2}(a_0 + a_{n-1})$$

where

$$a_n = a_0 + nd$$

R We are trying to find the sum of the first n terms, and the first term is a_0, so the last term is a_{n-1}.

Proof. Consider an arithmetic sequence $a_0, a_1, \ldots, a_n$, where each term can be expressed as $a_i = a_0 + id$ and d is the common difference.

Write the sum of the sequence in order:

$$S = a_0 + (a_0 + d) + (a_0 + 2d) + \ldots + (a_0 + (n-1)d)$$

Write the sum of the sequence in reverse order:

$$S = (a_0 + (n-1)d) + (a_0 + (n-2)d) + \ldots + a_0$$

Add these two equations together, every pair of terms within the brackets forms:

$$2a_0 + (n-1)d$$

Since each term appears in a pair, there are n such pairs.

The resulting equation is $2S = n[2a_0 + (n-1)d]$.

Solving for S gives us $S = \frac{n}{2}[2a_0 + (n-1)d]$ or $S = \frac{n}{2}(a_0 + a_n)$, where $a_n = a_0 + (n-1)d$. ■

2.5.2.2 Geometric Sequence

Geometric sequence is the other important and common sequence that involved in problem-solving of computer Science. A geometric sequence, also known as a geometric progression, is a sequence of numbers where each term after the first is found by multiplying the previous term by a fixed, non-zero number called the common ratio. Mathematically, a geometric sequence is defined as follows:

Definition 2.22 — Geometric Sequence. Given the first term a_0 (also referred to as a_1 in some texts) and the common ratio r, the n-th term of a geometric sequence a_n can be expressed as:

$$a_n = a_0 \cdot r^n \quad \text{for } n \geq 0$$

where n is a non-negative integer representing the position of the term in the sequence.

The common ratio r can be any real number. If $|r| < 1$, the terms of the sequence will get progressively smaller and approach zero. If $|r| > 1$, the terms will grow progressively larger. If $r = 1$, the sequence is constant, and if $r = -1$, the sequence will alternate between two values.

We can deduce the sum of a specific geometric sequence by direct proof.

Theorem 2.4 — Sum of Geometric Sequence.

Proof. Consider a geometric sequence with the first term a_0 and the common ratio r where $r \neq 1$. The sequence is given by:

$$a_0, a_0r, a_0r^2, \ldots, a_0r^{n-1}$$

The sum of the first n terms of the sequence, denoted by S_n, is:

$$S_n = a_0 + a_0r + a_0r^2 + \ldots + a_0r^{n-1}$$

To find a formula for S_n, multiply the entire sequence by r:

$$rS_n = a_0r + a_0r^2 + a_0r^3 + \ldots + a_0r^n$$

Subtract the original sum S_n from this new sum rS_n to get a telescoping series:

$$rS_n - S_n = a_0r^n - a_0$$

Solving for S_n gives us:

$$S_n = \frac{a_0(1-r^n)}{1-r} =$$

This is the sum formula for the first n terms of a geometric sequence when $r \neq 1$. If $r = 1$, the sequence is constant, and the sum of the first n terms is simply n times the first term a_0. ■

2.5.2.3 characteristic Sequence

Definition 2.23 — Characteristic Sequence. Suppose that U is some given n-set whose elements have been *indexed* (listed in a certain order) so that $U = \{x_1, x_2, \ldots, x_n\}$. If A is a subset of U, the *characteristic sequence* of A is the function whose domain is $\{1..n\}$ defined by

$$X_i^A = X^A(i) = \begin{cases} 1 & \text{if } x_i \in A \\ 0 & \text{if } x_i \notin A \end{cases}$$

■ **Example 2.9** If U is the set of the first 10 odd positive integers, A is the subset of primes in U, and B is the set of multiples of 3 in U, then

$$\begin{aligned} U &= \{1,3,5,7,9,11,13,15,17,19\} \quad //x_i = 2i-1. \\ A &= \{3,5,7,11,13,17,19\} \\ B &= \{3,9,15\} \\ X^A &= (0,1,1,1,0,1,1,0,1,1) \\ X^B &= (0,1,0,0,1,0,0,1,0,0). \end{aligned}$$

■

Characteristic sequences may be used as an implementation model for subsets of any given indexed set U. The set operations may be done on these sequences:

$$X_i^{A\cap B} = X_i^A \times X_i^B;$$

$$X_i^{A\cup B} = X_i^A + X_i^B - X_i^A \times X_i^B;$$

$$X_i^{A\setminus B} = X_i^A - X_i^A \times X_i^B.$$

If $A \subseteq B$ then

$$X_i^A \leq X_i^B \quad \text{for each index } i,$$

and

$$|A| = \sum_{i=1}^{n} X_i^A.$$

2.5.3 Exercises

Exercise 2.15 Find the sum of arithmetic sequence using mathematical induction. Try **NOT** use the conclusion in this section. ■

Hint: Consider the sum of the first nth positive integer. Try to make assumption by taking it as an arithmetic sequence.

Proof. Let $S(n)$ denote the sum of the first n terms of an arithmetic sequence with the first term a_0 and common difference d.

- **Base Case** ($n = 1$): The sum of the sequence with only the first term is the first term itself, $S(1) = a_0$.

- **Inductive Step**: Assume that the sum of the first k terms $S(k)$ is given by a certain formula. We want to show that the sum of the first $k+1$ terms $S(k+1)$ can be expressed using the same formula.

For the base case, we can easily see that:

$$S(1) = a_0$$

As $\sum_1^n i = \frac{n(n+1)}{2}$, which could be taken as an arithmetic sequence with $a_0 = 1$ and $a_n = n$. By this, assume that the sum of the first k terms is:

$$S(k) = \frac{k}{2}[a_0 + a_n]$$

equivalent to

$$S(k) = \frac{k}{2}[2a_0 + (k-1)d]$$

To prove the inductive step for $S(k+1)$, consider:

$$S(k+1) = S(k) + a_0 + kd$$

Substituting the inductive hypothesis into the above equation yields:

$$S(k+1) = \frac{k}{2}[2a_0 + (k-1)d] + a_0 + kd$$

After simplifying, we aim to show that:

$$S(k+1) = \frac{k+1}{2}[2a_0 + kd]$$

This will complete the proof if we can establish that the simplified version of $S(k+1)$ matches the form of the inductive hypothesis. ■

Exercise 2.16 Prove the sum of geometric sequence is $S_n = \frac{a_0(1-r^n)}{1-r} =$ using mathematical induction. ■

Proof. We want to prove that the sum of the first n terms of a geometric sequence S_n with the first term a and common ratio r (where $r \neq 1$) is given by:

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

Base Case (n=1):
The sum of the first term is simply the term itself:

$$S_1 = a$$

which agrees with the formula.
Inductive Step:

Assume the formula holds for $n = k$, that is,

$$S_k = \frac{a(1-r^k)}{1-r}$$

We need to prove that it also holds for $n = k+1$:

$$S_{k+1} = \frac{a(1-r^{k+1})}{1-r}$$

Starting with the inductive hypothesis for S_k and adding the $(k+1)$-th term ar^k to both sides, we have:

$$S_k + ar^k = \frac{a(1-r^k)}{1-r} + ar^k$$

Simplifying, we obtain:

$$S_{k+1} = S_k + ar^k = \frac{a - ar^{k+1}}{1-r}$$

which is the same as the formula for S_{k+1}, thus completing the proof. ■

Exercise 2.17 Given a sequence $\{a_n\}$ and a series $S_n = an^2 + bn + c (a \neq 0)$.

1. Find the general term a_n;
2. Is the sequence $\{a_n\}$ an arithmetic sequence?

Hint: How can we get the value of a term from the sum of a sequence? **Solution:**

1. For $n \geq 2$, $a_n = S_n - S_{n-1} = (an^2 + bn + c) - [a(n-1)^2 + b(n-1) + c]$
   $= (b+a) + (n-1) \cdot 2a$,
   Therefore, for $n = 1$, $a_1 = (b+a) + (1-1) \cdot 2a = b + a + c - S_1$,
   and the general term is

$$a_n = \begin{cases} a+b+c & (n=1) \\ (b+a) + (n-1) \cdot 2a & (n \geq 2) \end{cases}$$

2. Since $c = 0$, a_n can be simplified to $a_n = a + b$, which is constant and equals $2a$ when $n \geq 2$. This implies $\{a_n\}$ is an arithmetic sequence with common difference $2a$, provided a, b are constants and $a \neq 0$.
   Note: From S_n we can deduce $a_n = S_n - S_{n-1}$ when $n \geq 2$. Since $a_1 = S_1$, the sequence $a_n = S_n - S_{n-1}$ (for $n \geq 2$) and a_1 is the first term. The sequence $\{a_n\}$ is an arithmetic sequence.
   Therefore, the general term a_n can be expressed as:

$$a_n = \begin{cases} S_1 & (n=1) \\ S_n - S_{n-1} & (n \geq 2) \end{cases}$$

   Given the series $\{a_n\}$ with $S_n = an^2 + bm + c (a \neq 0)$, the sequence $\{a_n\}$ is an arithmetic sequence with common difference $2a$ when $c = 0$.

Exercise 2.18 Given constants a, b, c, consider the sum $S_n = 1^2+2^2+3^2+\ldots+n(n+1)^2 = \frac{n(n+1)}{12}(an^2+bn+c)$, where $an^2+bn+c \neq 0$. ■

Proof. For $n=1$, we have $\frac{1}{6}(a+b+c)$, thus $a_1 = 4 = \frac{1}{6}(a+b+c)$. For $n=2$, we have $\frac{22}{2} = 11 = \frac{1}{2}(4a+b+c)$, thus $a_2 = 22 = 9a+3b+c$. For $n=3$, $a_3 = 70 = 9a+3b+c$.

From these equations, we find that:

$$\begin{aligned} a+b+c &= 24 \\ 4a+b+c &= 44 \\ 9a+3b+c &= 70 \end{aligned}$$

Solving the system, we get $a=3$, $b=11$, $c=10$. For $n=1,2,3$, the sum can be expressed as:

$$1\cdot 2^2+2\cdot 3^2+\ldots+n(n+1)^2 = \frac{n(n+1)}{12}(3n^2+11n+10),$$

thus, $S_n = 1\cdot 2^2+2\cdot 3^2+\ldots+n(n+1)^2$.

For a general term k, $S_k = \frac{k(k+1)}{12}(3k^2+11k+10)$. Therefore,

$$\begin{aligned} S_{k+1} &= S_k + (k+1)(k+2)^2 \\ &= \frac{k(k+1)}{12}(3k^2+11k+10)+(k+1)(k+2)^2 \\ &= \frac{k(k+1)}{12}((k+2)(3k+5)+(k+1)(k+2)^2) \\ &= \frac{(k+1)(k+2)}{12}(3k^2+5k+12k+24) \\ &= \frac{(k+1)(k+2)}{12}(3(k+1)^2+11(k+1)+10). \end{aligned}$$

Hence, by induction, we can show that for $n=k+1$ the sum is valid.

Finally, with $a=3$, $b=11$, $c=10$, we confirm that the given sequence is indeed a second-order arithmetic sequence. ■

Exercise 2.19 Evaluate:

1. $S = \sum_1^n \frac{N}{2^n}$
2. $S = \sum_1^n \frac{3n-2}{5^{n-1}}$

■

Solution:

(1) Given the series $S_n = \frac{1}{2}+\frac{2}{4}+\frac{3}{8}+\cdots+\frac{n}{2^n}$, we can write:

$$S_n - \frac{1}{2}S_n = \frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^n}-\frac{n}{2^{n+1}}$$

This simplifies to:

$$\frac{1}{2}S_n = \frac{1}{2}\left(1-\left(\frac{1}{2}\right)^n\right) = \frac{1}{2} - \frac{1}{2^{n+1}} = \frac{1}{2} - \frac{n}{2^{n+1}} + \frac{n}{2^{n+1}}$$

Hence, the series sum is:

$$S_n = 2 - \frac{1}{2^{n-1}} - \frac{n}{2^n} = 2 - \frac{n+2}{2^n}$$

(2) Considering the series $S_n = 1 + \frac{4}{5} + \frac{7}{25} + \cdots + \frac{3n-2}{5^{n-1}}$, we proceed similarly:

$$\left(1-\frac{1}{5}\right)S_n = 1 + \frac{3}{5} + \frac{3}{25} + \cdots + \frac{3}{5^{n-1}} - \frac{3n-2}{5^n}$$

The terms form a geometric series, so we get:

$$S_n = 1 + \frac{3}{5}\left(1 + \frac{1}{5} + \frac{1}{25} + \cdots + \frac{1}{5^{n-2}}\right) - \frac{3n-2}{5^n}$$

Applying the formula for the sum of a geometric series, we find:

$$S_n = 1 + \frac{3}{5}\left(\frac{1-\left(\frac{1}{5}\right)^{n-1}}{1-\frac{1}{5}}\right) - \frac{3n-2}{5^n}$$

Simplifying, we obtain:

$$S_n = 1 + \frac{3}{5}\left(\frac{5^n-5}{4\cdot 5^{n-1}}\right) - \frac{3n-2}{5^n}$$

Further simplification gives us:

$$S_n = \frac{35}{16} - \frac{12n+7}{16\cdot 5^{n-1}}$$

Computer Science is most commonly known as an engineering subject, while the unanimous pursuit of all engineering subjects are solving problems. This chapter delves in to methods to solve problems, which is also known as **algorithm**. In the world of Computer Science, everything is proceeded in a methodical manner, and the very dependency of this is algorithm. Meanwhile, we will introduce pseudocode, one of the most important tools for algorithm analysis, as well as the representation of number in Computer Science.

3.1 Numbers

Every reader could be quite surprised when seeing the title for this section. Yes, numbers, we have known what is number since the very beginning when we get to learn math as toddlers. In this section, we will explain the system of number, not only will we figure out how numbers and their operations are defined, but how they are categorized.

3.1.1 Typology of Numbers

This part recalls the type of numbers we've learned since primary school and their set notations.

1. **Natural Numbers**
   - Definition: Natural numbers are the set of positive integers used for counting and ordering, which do not include zero or negative numbers.
   - Set Notation:

$$\mathbb{N} = \{1, 2, 3, \ldots\}$$

2. **Integers**
   - Definition: Integers are all the whole numbers including positive natural numbers, their negatives, and zero.

- Set Notation:

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$$

3. **Rational Numbers**
   - Definition: Rational numbers are numbers that can be expressed as the quotient of two integers, a fraction $\frac{a}{b}$, where a and b are integers and $b \neq 0$. The set includes all integers and fractions.
   - Set Notation:

$$\mathbb{Q} = \left\{ \frac{a}{b} \mid a, b \in \mathbb{Z}, b \neq 0 \right\}$$

R For numbers that are written in finite decimal places, if there is a looping part in the decimal places, it is still recoganized as rational numbers. For instance, $\frac{1}{3}$ could be repesented as $0.33333\ldots$. Usually, we use upperline to mark the repeating part, in this case it is$0.\overline{3}$. But for those who have infinitely non-repeating decimal places, such as $\pi = 3.1415926535\ldots$, we categorize it as irrational number, as they cannot be written in the $\frac{a}{b}$ form.

4. **Irrational Numbers**
   - Definition: Irrational numbers are real numbers that cannot be expressed as a ratio of two integers. The decimal expansion of irrational numbers is non-terminating and non-repeating. Examples include π and $\sqrt{2}$.
   - Set Notation:

$$\mathbb{I} = \{x \in \mathbb{R} \mid x \notin \mathbb{Q}\}$$

   (Note: $\mathbb{I}$ is used here for illustrative purposes and is not a standard symbol.)

5. **Real Numbers**
   - Definition: The real numbers include both rational and irrational numbers, encompassing all points on an infinitely extended number line. The set of real numbers is continuous and is composed of all limits of sequences of rational numbers.
   - Set Notation:

$$\mathbb{R} = \{x \mid x \text{ is a limit of a sequence of rational numbers}\}$$

6. **Prime Numbers**
   - Definition: A prime number is a natural number greater than 1 that has no positive divisors other than 1 and itself. In other words, p is prime if $p > 1$ and if p is divisible only by 1 and p.
   - Set Notation:

$$\mathbb{P} = \{p \in \mathbb{N} \mid p > 1 \text{ and } p \text{ has no divisors other than 1 and } p\}$$

   - Examples: The first few prime numbers are:

$$2, 3, 5, 7, 11, 13, 17, \ldots$$

The following Venn diagram shows the relationship between different number sets.

Figure 3.1: Venn Diagram of Number Sets

R Complex number is currently not a something necessary, as our discussion so far only falls in the real number set.

3.1.2 The Real Number System

Considering the real numbers are the only system involved so far this book, here, we provide a new mathematical perspective that is different what we were taught, to understand the real number system. We will introduce three axioms that real number holds.

Definition 3.1 — Field Axioms. A set S with operations $+$ and $\cdot$ and distinguished elements 0 and 1 with $0 \neq 1$ is a *field* if the following properties hold for all $x, y, z \in S$:

A0:	$x+y \in S$	Closure
A1:	$(x+y)+z = x+(y+z)$	Associativity
A2:	$x+y = y+x$	Commutativity
A3:	$x+0 = x$	Identity
A4:	given x, there is a $w \in S$ such that $x+w = 0$	Inverse
M0:	$x \cdot y \in S$	Closure
M1:	$(x \cdot y) \cdot z = x \cdot (y \cdot z)$	Associativity
M2:	$x \cdot y = y \cdot x$	Commutativity
M3:	$x \cdot 1 = x$	Identity
M4:	for $x \neq 0$, there is a $w \in S$ such that $x \cdot w = 1$	Inverse
DL:	$x \cdot (y+z) = x \cdot y + x \cdot z$	Distributive Law

The operations $+$ and $\cdot$ are called addition and multiplication. The elements 0 and 1 are the additive identity element and the multiplicative identity element, respectively.

It follows from these axioms that the additive inverse and multiplicative inverse (of a nonzero x) are unique. The additive inverse of x is the **negative** of x, written as $-x$. To define subtraction of y from x, we let $x - y = x + (-y)$. The multiplicative inverse of x is the **reciprocal** of x, written as x^{-1}. **The element 0 has no reciprocal**. To define division of x by y when $y \neq 0$, we let $\frac{x}{y} = x \cdot (y^{-1})$. We write $x \cdot y$ as xy and $x \cdot x$ as x^2. We use parentheses where helpful to clarify the order of operations.

Definition 3.2 — Order Axioms. A positive set in a field F is a set $P \subset F$ such that for $x, y \in F$,

P1: $x, y \in P$ implies $x + y \in P$ — Closure under Addition

P2: $x, y \in P$ implies $xy \in P$ — Closure under Multiplication

P3: $x \in F$ implies exactly one of $x = 0$, $x \in P$, $-x \in P$ — Trichotomy

An ordered field is a field with a positive set P. In an ordered field, we define $x < y$ to mean $y - x \in P$. The relations $\leq, >$, and $\geq$ have analogous definitions in terms of P.

Note that $P = \{x \in F : x > 0\}$. Another phrasing of trichotomy is that each ordered pair (x, y) satisfies exactly one of $x < y$, $x = y$, $x > y$. If $S \subseteq F$, then $\beta \in F$ is an **upper bound** for S if $x \leq \beta$ for all $x \in S$.

Definition 3.3 — Completeness Theorem. An ordered field F is complete if every nonempty subset of F that has an upper bound in F has a least upper bound in F.

This theorem ensures the square roots of positive real numbers.

Axiom **??** and **??** imply many familiar property of arithmetic:

Proposition 3.1 — Arithmetic in $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$. Each of $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ is closed under addition and multiplication, $\mathbb{Z}$ and $\mathbb{Q}$ are closed under subtraction, and the set of nonzero numbers in $\mathbb{Q}$ is closed under division.

The next four propositions state properties of an ordered field F. All statements apply for each choice of $x, y, z, u, v \in F$.

Proposition 3.2 Elementary consequences of the field axioms.

a) $x + z = y + z$ implies $x = y$

b) $x \cdot 0 = 0$

c) $(-x)y = -(xy)$

d) $-x = (-1)x$

e) $(-x)(-y) = xy$

f) $xz = yz$ and $z \neq 0$ imply $x = y$

g) $xy = 0$ implies $x = 0$ or $y = 0$

Proposition 3.3 — Properties of an ordered field..

O1: $x \leq x$ — Reflexive Property

O2: $x < y$ and $y < x$ imply $x = y$ — Antisymmetric Property

O3: $x < y$ and $y < z$ imply $x < z$ — Transitive Property

O4: At least one of $x < y$ and $y < x$ holds — Total Ordering Property

Proposition 3.4 — More properties of an ordered field..

F1: $x \leq y$ implies $x + z \leq y + z$ — Additive Order Law

F2: $x < y$ and $0 < z$ imply $xz < yz$ — Multiplicative Order Law

F3: $x < y$ and $w < v$ imply $x + w < y + v$ — Addition of Inequalities

F4: $0 \leq x$ and $0 \leq w$ imply $xw \leq xw$ — Multiplication of Inequalities

Proposition 3.5 — Still more properties of an ordered field..

(a) $x < y$ implies $-y < -x$

(b) $x \leq y$ and $z \leq 0$ imply $yz \leq xz$

(c) $0 \leq x$ and $0 \leq y$ imply $0 \leq xy$

(d) $0 \leq x^2$

(e) $0 < 1$

(f) $0 < x$ implies $0 < x^{-1}$

(g) $0 < x < y$ implies $0 < y^{-1} < x^{-1}$

3.1.3 Floor, Ceiling, and Remainder

This Section discusses more form numbers that may not be as familiar as real numbers. We first introduce **Integer Function**:

Definition 3.4 — Floor and Ceiling. If x is any real number, we write

$$\lfloor x \rfloor = \text{the greatest integer less than or equal to } x \text{ (the floor of } x\text{)}$$

$$\lceil x \rceil = \text{the least integer greater than or equal to } x \text{ (the ceiling of } x\text{)}$$

Note that the x could be not just a variable, but a mathematical expression.

■ **Example 3.1**

$$\lfloor \sqrt{2} \rfloor = 1, \quad \lceil \sqrt{2} \rceil = 2, \quad \left\lceil \frac{1}{2} \right\rceil = 0, \quad \left\lfloor -\frac{1}{2} \right\rfloor = 0, \quad \left\lceil -\frac{1}{2} \right\rceil = -1 \text{ (not zero!)};$$

■

3.1.3.1 Properties of Integer Function

We will look into the properties of integer function with its graph.

Figure 3.2: Visualization of $\lceil x \rceil$ and $\lfloor x \rfloor$

Theorem 3.1 — Properties of Integer Function. Keep in mind the following important properties of integer function, which is often used in algorithm analysis or other mathematical proof.

1. $\lfloor x \rfloor \leq x \leq \lceil x \rceil$
2. $\forall x \in \mathbb{Z}, \lceil x \rceil = \lfloor x \rfloor$
3. $x \notin \mathbb{Z} \iff \lceil x \rceil = \lfloor x \rfloor + 1$
4. $\lfloor -x \rfloor = -\lceil x \rceil$; $\lceil -x \rceil = -\lfloor x \rfloor$
5. $x - 1 < \lfloor x \rfloor \leq x \leq \lceil x \rceil < x + 1$

The proofs to these conclusions are quite basic, and is therefore not provided here.

3.1.3.2 Remainder and Integer Function

We introduce a new operation that we have learned before in this section with its notation.

▪ **Notation 3.1 — modulo operator.** The modulo operation, denoted as $a \bmod n$, finds the remainder when one integer a is divided by another integer n. If a divided by n gives a quotient q with remainder r, then $a = nq + r$, and r would be the result of $a \bmod n$.

Moving back to the integer function, we have:

Theorem 3.2

$$\forall x, y \in \mathbb{R}, x \bmod y = x - y \left\lfloor \frac{x}{y} \right\rfloor, \quad \text{if } y \neq 0; \quad x \bmod 0 = x$$

Below is the proof to the first conclusion, where the properties of integer function are applied.

Proof. By the Division Algorithm, for any integer x and any positive integer y, there exist unique integers q and r such that $x = qy + r$ and $0 \leq r < |y|$, where q is the quotient and r is the remainder. The floor function $\left\lfloor \frac{x}{y} \right\rfloor$ yields the largest integer less than or equal to $\frac{x}{y}$, which by definition is the quotient q. Thus, we have:

$$\left\lfloor \frac{x}{y} \right\rfloor = q$$

and therefore:

$$y \left\lfloor \frac{x}{y} \right\rfloor = yq$$

Subtracting this from x gives:

$$x - y \left\lfloor \frac{x}{y} \right\rfloor = x - yq = r$$

The uniqueness of the quotient and remainder in the Division Algorithm ensures that this value of r is the remainder from the modulus operation. Therefore, we have:

$$x \quad \bmod y = x - y \left\lfloor \frac{x}{y} \right\rfloor$$

which is the remainder when x is divided by y, completing the proof.

■

Also, from definition, by dividing y on both sides of the equation:

$$0 \leq \frac{x}{y} - \lfloor \frac{x}{y} \rfloor = \frac{x \bmod y}{y} < 1$$

Among which, $x \bmod y < y$.

Proof. By the definition of the modulo operation, $x \mod y$ can be written as $x - y \left\lfloor \frac{x}{y} \right\rfloor$. Since $\left\lfloor \frac{x}{y} \right\rfloor$ is the greatest integer less than or equal to $\frac{x}{y}$, we have:

$$x - y \left\lfloor \frac{x}{y} \right\rfloor \geq x - y \cdot \frac{x}{y} = x - x = 0.$$

Furthermore, because $\left\lfloor \frac{x}{y} \right\rfloor$ is less than $\frac{x}{y}$, it follows that:

$$x - y \left\lfloor \frac{x}{y} \right\rfloor < x - y \cdot \left(\frac{x}{y} - 1 \right) = y.$$

Hence, $0 \leq x \mod y < y$.

■

And we have:

Corollary 3.1 By above-mentioned conclusions:
a) If $y > 0$, then $0 \leq x \bmod y < y$.
b) If $y < 0$, then$0 \geq x \bmod y > y$.
c) $x - (x \bmod y)$ is an integral multiple of y.

We call $x \bmod y$ the remainder when x is divided by y. We call $\lfloor \frac{x}{y} \rfloor$ the quotient. We have $x \bmod y = 0$ if and only if x is a multiple of y, that is, if and only if x is divisible by y. The notation $y \mid x$, read "y divides x", means that y is a positive integer and $x \bmod y = 0$.

3.1.4 exercises

Exercise 3.1 Let F be a field consisting of exactly three elements $0, 1, x$. Prove that $x + x = 1$ and that $x \cdot x = 1$. Obtain the addition and multiplication tables for F.

■

Hint: Think on the property of filed: inverse of addition and multiplication.

Proof. Since F is a field, it has the properties of both a group under addition and a group under multiplication (excluding 0 for the latter).

Part 1: Proof that $x + x = 1$

1. In a group, every element has an additive inverse. In F, the additive inverse of 0 is 0 itself, since $0 + 0 = 0$.
2. The additive inverse of 1 cannot be 1 itself because $1 + 1 = 1$ would imply $1 = 0$, which is a contradiction. Therefore, 1's additive inverse must be some other element of F, which can only be x. Hence, $1 + x = 0$.
3. The element x must also have an additive inverse in F, which cannot be 0 (as 0's inverse is 0) and cannot be 1 (as 1's inverse is x). The only option left is x itself. Thus, $x + x = 0$.

4. Given $x+x=0$ and $1+x=0$, by the cancellation law, it must be that $x=1$. However, this contradicts the assumption that x is distinct from 1. Therefore, our assumption that $x+x=0$ is incorrect.
5. The only remaining possibility is $x+x=1$.

Part 2: Proof that $x \cdot x = 1$

1. Similarly, in a multiplicative group (excluding 0), every non-zero element has a multiplicative inverse. For 1, the multiplicative inverse is 1 itself since $1 \cdot 1 = 1$.
2. The element x must have a multiplicative inverse. It cannot be 0 since 0 is not invertible, and it cannot be 1 since 1 is already serving as its own inverse.
3. The only remaining option for the multiplicative inverse of x is x itself. Hence, $x \cdot x = 1$.

Now we construct the addition and multiplication tables for F:

Addition Table:

$+$	0	1	x
0	0	1	x
1	1	x	0
x	x	0	1

Multiplication Table:

$\times$	0	1	x
0	0	0	0
1	0	1	x
x	0	x	1

■

Problem 3.1 Is there a field with exactly four elements? Is there a field with exactly six elements?

Exercise 3.2 Let n be an integer, and let x be a real number. Prove that:

a) $\lfloor x \rfloor < n$ if and only if $x < n$;
b) $n \leq \lfloor x \rfloor$ if and only if $n \leq x$;
c) $\lfloor x \rfloor \leq n$ if and only if $x \leq n$;
d) $n < \lfloor x \rfloor$ if and only if $n < x$;
e) $\lfloor x \rfloor = n$ if and only if $x-1 < n \leq x$ and if and only if $n \leq x < n+1$;
f) $\lfloor x \rfloor = n$ if and only if $x \leq n < x+1$ and if and only if $n-1 < x \leq n$.

■

Proof. **Part (a):** By definition, $\lfloor x \rfloor$ is the greatest integer less than or equal to x. Therefore, $\lfloor x \rfloor < n$ means that x is less than n but not equal to it since $\lfloor x \rfloor$ cannot be greater than x.

Part (b): If $n \leq \lfloor x \rfloor$, then n must also be less than or equal to x because $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

Part (c): The statement $\lfloor x \rfloor \leq n$ indicates that the largest integer less than or equal to x is also less than or equal to n, which directly implies $x \leq n$.

Part (d): If $n < \lfloor x \rfloor$, then n is strictly less than the integer part of x, which means that n is also strictly less than x.

Part (e): For $\lfloor x \rfloor = n$ to be true, x must be greater than $n-1$ but not reach $n+1$, hence the inequality $x-1 < n \leq x$ and $n \leq x < n+1$.

Part (f): Similarly, $\lfloor x \rfloor = n$ if x has not reached $n+1$ yet, which is the same as saying $x \leq n < x+1$, and also if x is greater than $n-1$ and less than or equal to n, hence $n-1 < x \leq n$. ■

Exercise 3.3 Using the previous exercise, prove that $\lfloor -x \rfloor = -\lceil x \rceil$. ■

Proof. From the previous exercise, we have the following properties of the floor function for a real number x and an integer n:

1. $x-1 < \lfloor x \rfloor \leq x$;
2. $n \leq x$ if and only if $n \leq \lfloor x \rfloor$.

Using these properties, we want to show that $\lfloor -x \rfloor = -\lceil x \rceil$.

Consider the number $-x$. Applying property 1, we get:

$$-x-1 < \lfloor -x \rfloor \leq -x$$

Adding 1 to all parts of the inequality, we obtain:

$$-x < \lfloor -x \rfloor + 1 \leq -x+1$$

Since $\lceil x \rceil$ is the smallest integer greater than or equal to x, $\lfloor -x \rfloor + 1$ is the smallest integer greater than $-x$, which is $-\lfloor x \rfloor$. Thus:

$$\lfloor -x \rfloor = -\lfloor x \rfloor - 1 = -\lceil x \rceil$$

This completes the proof. ■

Exercise 3.4 Prove that $\lceil (k-1)/2 \rceil = \lfloor k/2 \rfloor$ and $\lfloor (k-1)/2 \rfloor = \lceil k/2 \rceil \ \forall\, k \in \mathbb{Z}$. ▪

Proof. We will prove the statement by cases.

When k is even, let $k = 2n$. $\lceil (2n-1)/2 \rceil = \lceil n - \frac{1}{2} \rceil = n, and \lfloor 2n/2 \rfloor = n$.

When k is odd, let $k = 2n+1$. $\lceil 2n/2 \rceil = n$, and $\lfloor 2n+1/2 \rfloor = \lfloor n + \frac{1}{2} = n$.

The proof for $\lfloor (k-1)/2 \rfloor = \lceil k/2 \rceil$ is similar. ■

3.2 Algorithm and Algorithm Analysis

This Section discusses what is algorithm, and more importantly, how algorithms are assessed.

3.2.1 Algorithm

3.2.1.1 What is an Algorithm?

Definition 3.5 — Algorithm. An algorithm is a well-defined, step-by-step procedure or sequence of instructions designed to solve a specific class of problems:

- Each step in an algorithm must be clear and unambiguous.
- Algorithms must be solvable, meaning they should be able to produce a correct solution for any valid input within a finite amount of time.
- An algorithm must terminate, i.e., it should have a defined end, at which point the goal has been achieved and the final output is produced.

Here is an example of multiplication algorithm for better understanding of the concept.

▪ **Example 3.2 — Russian Peasant Multiplication.** To find the product of integers M and N, both larger than one:

1. Start two columns on a page, one labeled "A" and the other "B"; and put the value of M under A and the value of N under B.
2. Repeat

(a) calculate a new A-value by multiplying the old A-value by 2; and
(b) calculate a new B-value by dividing the old B-value by 2 and reducing the result by a half if necessary to obtain an integer;

Until the B-value equals one.

3. Go down the columns crossing out the A-value whenever the B-value is even.
4. Add up the remaining A-values and "return" the sum.

■

To show how it works, assume $A = 73$ and $B = 41$.

A	B
73	41
146	20 ($20\frac{1}{2}$ is reduced to 20)
292	10
584	5
1168	2 ($2\frac{1}{2}$ is reduced to 2)
2336	1

Table 3.1: Execution of RPM

Sum of the remaining A-values: $2336 + 584 + 73 = 2993$.

Let's review this algorithm referring to definition **??**.

1. Clarity and Accuracy: All the instructions are clear and manipulable, nothing is ambiguous.
2. Solvability: It is no doubt that for any two real numbers, we can find their product.
3. Termination: For which ever numbers, we can always solve the problem in limited steps, as the terminate condition is when $B = 1$, while B is divided by 2 (integer division) repetitively.

Hence, the RPM is a good example of algorithm, and with this, we could tell whether something else is an algorithm or not.

3.2.1.2 Pseudocode

Pseudocode is a simplified, half-code, half-natural language script used by software developers and algorithm designers to outline the structure of a program or algorithm. It's not executable code, but rather a high-level representation of the algorithm's logic. The purpose of pseudo-code is to express the design of an algorithm in a form that can be easily translated into actual programming languages. It is written in a way that is understandable to people who do not necessarily know the syntax of programming languages. Pseudo-code allows the designer to focus on the core logic of the algorithm without getting bogged down with the syntactic details of a particular programming language. It often uses control structures like if-then-else, while, for, and others that are common to many high-level languages. Understanding pseudocode is quite easy as it quite close to natural languages.

Here's the RPM transcribed in pseudocode:

Algorithm 1 Russian Peasant Multiplication

```
procedure RPM(A, B)
    product ← 0
    while B > 0 do
        if B is odd then
            product ← product + A
        end if
        A ← A × 2
        B ← B ÷ 2
    end while
    return product
end procedure
```

To explicit:

- **procedure RPM(A, B)**: Defines a procedure or function named ‘RPM‘ taking two parameters ‘A‘ and ‘B‘.
- **product ← 0**: The assignment operator ‘←‘ is used to assign the value on the right (0 here) to the variable on the left (‘product‘).
- **while B > 0 do**: Begins a ‘while‘ loop that continues as long as the condition ‘B > 0‘ is true. The ‘do‘ indicates that the following block of code will execute if the condition is met.
- **if B is odd then**: A conditional statement that checks if ‘B‘ is odd. If it is, the subsequent statement is executed.
- **product ← product + A**: An assignment operation that adds ‘A‘ to ‘product‘ and assigns the sum back to ‘product‘.
- **end if**: Marks the end of the ‘if‘ statement.
- **A ← A × 2**: Multiplies the value of ‘A‘ by 2 and then assigns the result back to ‘A‘.
- **B ← B ÷ 2**: Divides the value of ‘B‘ by 2 (integer division) and assigns the result back to ‘B‘.
- **end while**: Marks the end of the ‘while‘ loop.
- **return product**: The ‘return‘ statement indicates the output or result of the procedure, which here is the value of the variable ‘product‘.
- **end procedure**: Marks the end of the ‘RPM‘ procedure.

All later algorithms in this book will be presented using pseudocode.

3.2.2 Algorithm Analysis

Now let’s take a look at this algorithm from another aspects. Is this a good or a bad algorithm. People assess algorithms by examine its **complexity**, which could be either **space complexity** or **time complexity**. The former refers to the the relationship between the input and the space needed to execute the algorithm, the latter, similarly, refers to the time needed. Many tools are available to quantify complexity, for both space and time complexity, **the big O notation** is the most common measurement.

Definition 3.6 — The Big O Notation. Big O notation is used to classify algorithms according to how their running time or space requirements grow as the input size grows.

The notation describes an upper limit on the time an algorithm could possibly take to complete, given the size of the input. For a function $f(n)$, where n is the scale of input for the algorithm, the Big O notation is formally defined as follows with c as a positive constant:

$$O(f(n)) = \{g(n) : \text{Where } c \text{ and } n_0 \text{ such that } 0 \leq g(n) \leq c \cdot f(n) \text{ for all } n \geq n_0\}$$

The following table provides common time complexities using Big O notation:

$f(n)$	Description
1	Constant
$\log n$	Logarithmic
n	Linear
$n \log n$	Linearithmic
n^2	Quadratic
n^3	Cubic
2^n	Exponential
$n!$	Factorial

Table 3.2: Common time complexities in Big O notation

We can visualize it using function graph

Figure 3.3: Time Complexity Visualization

So which space and time complexity does this algorithm fall in?

3.2.2.1 Time Complexity

To analyze the time complexity, we usually focus on the termination condition or the number of iteration for the algorithm. For the real number B in RPM, each time it is divided by 2 until $B = 1$. Therefore, the total number of iteration will be $\log_2 B$, which is categorized in $O(\log n)$.

R Actually, the time complexity of an algorithm could be shown by strict proof using MI. Here is the proof on the time complexity of RPM.

Proof. We will use mathematical induction to prove that the number of steps in the algorithm is proportional to $\log_2(B)$.
Base Case:
When $B = 1$, the algorithm requires only one step. This is consistent with $\log_2(1) = 0$, which satisfies our complexity class $O(\log n)$.
Inductive Hypothesis:
Assume that for a positive integer k, when $B = k$, the algorithm operates within $\log_2(k)$ steps.
Inductive Step:
Consider $B = 2k$. In the first step of the algorithm, B is halved to k, and S is doubled. From this point, based on our inductive hypothesis, reaching $B = 1$ requires $\log_2(k)$ steps.

Hence, for $A = 2k$, the total number of steps is $\log_2(k) + 1$. Using the properties of logarithms, we have:

$$\begin{aligned}\log_2(k) + 1 &= \log_2(k) + \log_2(2)\\ &= \log_2(2k)\end{aligned}$$

Therefore, for any $A = 2k$, the total number of steps is also $\log_2(2k)$, proving that for any positive integer A, the time complexity of the Russian Peasant Multiplication is $O(\log A)$. ■

3.2.2.2 Space Complexity

The space complexity of the Russian Peasant Multiplication algorithm is determined by the amount of memory required to store the operands and the intermediate results. Initially, only two numbers need to be stored: the multiplicands. As the algorithm proceeds, we need additional space to keep track of the current product. Since the algorithm does not use any complex data structures and only requires a fixed number of variables, the space complexity is $O(1)$, indicating constant space usage. It does not depend on the size of the input operands, as the memory required does not increase with larger numbers.

3.2.3 Exercises

Exercise 3.5 A non-recursive Square and Multiply Algorithm to calculate b^n.
Precondition: n is a positive integer and b is of any type that can be multiplied.
Postcondition: the value returned is equal $(b)^n$.

1. Show that the algorithm terminates. Let a_k denote the value of a after the kth iteration of the while-loop, and let $s = \lfloor \lg(n) \rfloor$. Prove by Mathematical Induction on k that For any nonnegative integer k, after k iterations of the while-loop:

$$2^{s-k} \leq a_k < 2^{s-k+1}$$

2. Proof of correctness. Use Mathematical Induction on k to prove For any nonneg-

ative integer k, after k iterations of the while-loop:

$$(\text{square})^a \times \text{product} = (b)^n$$

■

Algorithm 2 Square and Multiply Algorithm

```
product ← 1
square ← b
a ← n
while a > 1 do
    if a mod 2 ≠ 0 then                    ▷ If a is odd
        product ← product × square
    end if
    square ← square × square
    a ← ⌊a/2⌋                              ▷ Integer division
end while
return product × square
```

1. Base Case $k = 0$: For $k = 0$, $a_0 = n$, which is the initial value of a. Since $s = \lfloor \lg(n) \rfloor$, it is the greatest integer less than or equal to $\lg(n)$, therefore $2^s \leq n < 2^{s+1}$. So, for $k = 0$, the predicate holds because:

$$\frac{2^s}{2^0} \leq a_0 = n < 2 \times \frac{2^s}{2^0}$$

Inductive Step: Assume $P(k)$ holds for some nonnegative integer k. That is:

$$\frac{2^s}{2^k} \leq a_k < 2 \times \frac{2^s}{2^k}$$

We need to show $P(k+1)$ holds. During each iteration, a is halved (integer division by 2), which gives us:

$$a_{k+1} = \left\lfloor \frac{a_k}{2} \right\rfloor$$

Since a_k is an integer, $\left\lfloor \frac{a_k}{2} \right\rfloor$ will either be $\frac{a_k}{2}$ or $\frac{a_k - 1}{2}$, depending on whether a_k is even or odd. Thus, we have:

$$\frac{2^s}{2^{k+1}} \leq \left\lfloor \frac{a_k}{2} \right\rfloor < 2 \times \frac{2^s}{2^{k+1}}$$

Since $a_k < 2 \times \frac{2^s}{2^k}$, dividing by 2 gives $\frac{a_k}{2} < \frac{2^s}{2^k}$, and therefore:

$$a_{k+1} < 2 \times \frac{2^s}{2^{k+1}}$$

Similarly, $\frac{2^s}{2^k} \leq a_k$ implies $\frac{2^s}{2^{k+1}} \leq \frac{a_k}{2}$, so we have:

$$\frac{2^s}{2^{k+1}} \leq a_{k+1}$$

This completes the inductive step and thus, by induction, $P(k)$ holds for all nonnegative integers k. ■

2. We need to prove that after k iterations of the while-loop, the invariant holds:

$$(\text{square})^a \times \text{product} = b^n$$

Base Case $k = 0$: Initially, product $= 1$, square $= b$, and $a = n$. So,

$$(\text{square})^a \times \text{product} = b^n$$

is trivially true.

Inductive Step: Assume the invariant holds after k iterations, i.e.,

$$(\text{square}_k)^{a_k} \times \text{product}_k = b^n$$

Now, consider the $k+1$th iteration. There are two cases:

Case 1 (a_k is odd): The product is updated by multiplying it with square_k, and we have:

$$\text{product}_{k+1} = \text{product}_k \times \text{square}_k$$

Since a_k is odd, we can write $a_k = 2m+1$ for some integer m, and after the iteration, a becomes $a_{k+1} = m$. The invariant becomes:

$$(\text{square}_k)^{2m+1} \times \text{product}_k = (\text{square}_k)^m \times \text{product}_{k+1} = b^n$$

Case 2 (a_k is even): The product remains the same, and a_k can be written as $2m$, so the invariant remains:

$$(\text{square}_k)^{2m} \times \text{product}_k = (\text{square}_k)^m \times \text{product}_k = b^n$$

since $\text{square}_{k+1} = (\text{square}_k)^2$ and $a_{k+1} = m$.

In both cases, after the iteration, square is squared, so we get:

$$\text{square}_{k+1} = (\text{square}_k)^2$$

and therefore, the invariant still holds as:

$$(\text{square}_{k+1})^{a_{k+1}} \times \text{product}_{k+1} = b^n$$

Thus, by mathematical induction, the invariant holds true for every iteration of the loop, proving the correctness of the algorithm. ■

throughout our mathematical journey, equalities are always the very basics of most conclusion, and that is also which we start to learn math for. However, inequalities are not such a know thing as equalities. Inequalities have many tricky characteristics so that we have to take with care, or things could go wrong. This chapter covers the fundamentals of inequality as a crucial tool for problem-solving in Computer Science.

4.1 Inequality basics

We all know about inequalities, and the first thing to clarify is the relationship between sizes. How to determine the size relationship between certain numbers? Since the basis for comparing "numbers" with each other corresponds to one, it is stipulated on the number line that the points increase from left to right, and therefore the numbers they represent increase in turn. Listed in ascending order, that is:

Let a, b be two real numbers, and the points on the number line are denoted as A, B respectively. If A is to the right of B, we say $a > b$; if A is to the left of B, we say $a < b$; if A coincides with B, we say $a = b$.

Thus, for any two real numbers, one and only one of the following three situations must hold:

$$a > b; \quad a = b; \quad a < b.$$

The above relationship is also known as the one-dimensional coordinate law.

$$a > b \Leftrightarrow a - b > 0$$
$$a < b \Leftrightarrow a - b < 0$$
$$a = b \Leftrightarrow a - b = 0$$

Where the symbol $\Leftrightarrow$ (double arrow), read as "if and only if," means that the truth of two propositions depends on each other. That is to say, if one proposition is true, then the other proposition is also true; conversely, if one proposition is false, then the other proposition is also false.

Based on the derivation of inequalities, in most cases, the above principles are sufficient. The following mathematical laws involve the geometric and algebraic meanings of the sizes of real numbers and the relationship between them. They are the basis for comparing the sizes of two real numbers and for proving inequalities by comparison. Let's review some basic properties of inequalities that are the foundation for our further study.

- Symmetry: $a > b$ if and only if $b < a$.
- Transitivity: If $a > b$ and $b > c$, then $a > c$.
- Addition (Subtraction): If $a > b$, then $a + c > b + c$.
- Multiplication (Division): If $a > b$ and $c > 0$, then $ac > bc$; if $a > b$ and $c < 0$, then $ac < bc$.
- Exponentiation: If $a > b$, then $a^n > b^n$, where n is a positive integer, and $n \geq 2$.
- Root Extraction (Power Root): If $a > b > 0$, then $\sqrt[n]{a} > \sqrt[n]{b}$, where n is a positive integer, and $n \geq 2$.
- If $a > b$ and $c > d$, then $a + c > b + d$.
- If $a > b > 0$ and $c > d > 0$, then $ac > bd$.

4.1.1 Exercises

Exercise 4.1 Explain the following statement.

1. If $a > b$, then $\frac{a}{c} > \frac{b}{c}$;
2. If $ac < bc$, then $a < b$;
3. If $a < b$, then $\frac{1}{a} > \frac{1}{b}$;
4. If $ac^2 > bc^2$, then $a > b$;
5. If $a > b$, then $a^n > b^n$.

Solution:

1. If $c > 0$, multiplying both sides of $a > b$ by the positive number $\frac{1}{c}$ preserves the inequality, hence $\frac{a}{c} > \frac{b}{c}$. If $c < 0$, the direction of the inequality would be reversed, which is not given in the condition, hence we assume $c > 0$.
2. Dividing both sides of $ac < bc$ by c (assuming $c \neq 0$), we get $a < b$ because division by a positive number preserves the inequality, and division by a negative number reverses it.
3. Taking the reciprocal of both sides of $a < b$ reverses the inequality because a and b are on opposite sides of the fraction line, hence $\frac{1}{a} > \frac{1}{b}$ (assuming $a, b > 0$ to avoid division by zero).
4. Dividing both sides of $ac^2 > bc^2$ by c^2 (assuming $c \neq 0$) preserves the inequality, hence $a > b$ because c^2 is positive regardless of whether c is positive or negative.
5. Raising both sides of $a > b$ to a power n (assuming n is a positive integer) preserves the inequality because both a and b are raised to the same power, hence $a^n > b^n$.

Exercise 4.2 Given the inequality $a > b > 0$, $c < d < 0$, $f < 0$, show that:

$$\frac{f}{a-c} > \frac{f}{b-d}.$$

Proof. Since $a > b > 0$ and $c < d < 0$, then $a - c > b - d$ because subtracting a smaller negative number is the same as adding a larger positive number. Given that $f < 0$, when

dividing by a larger positive number, the result is smaller because a negative number divided by a positive number yields a negative result, and the further away the divisor is from zero, the smaller the quotient.

Therefore:

$$\frac{f}{a-c} > \frac{f}{b-d}.$$

■

4.2 Solving Quadratic Inequality

Recall that, when we solve Quadratic equations, we use discriminant to solve equations, which is also applicable to inequalities. For a quadratic inequality of the form $ax^2 + bx + c > 0$ or $ax^2 + bx + c < 0$ (where $a > 0$), the solution set can be determined by the discriminant $\Delta = b^2 - 4ac$:

1. If $\Delta > 0$, the quadratic equation $ax^2 + bx + c = 0$ has two distinct real roots x_1 and x_2, and $x_1 < x_2$. The solution set for $y = ax^2 + bx + c$ being greater than zero (when $y = 0$) is for values of x either less than x_1 or greater than x_2, and the solution set for $ax^2 + bx + c < 0$ is $\{x \mid x_1 < x < x_2\}$.
2. If $\Delta = 0$, then $ax^2 + bx + c = 0$ has one real root, specifically $x_1 = x_2 = -\frac{b}{2a}$. The solution set for $y = ax^2 + bx + c$ being greater than zero is all x except $x \neq -\frac{b}{2a}$, and there is no solution set where $ax^2 + bx + c < 0$.
3. If $\Delta < 0$, then $ax^2 + bx + c = 0$ has no real roots, and the parabola $y = ax^2 + bx + c$ does not intersect the x-axis. The solution set for $ax^2 + bx + c > 0$ is all real numbers, and there is no solution set where $ax^2 + bx + c < 0$.

Figure 4.1: Quadratic function graphs based on the discriminant.

■ **Example 4.1** Solve the following quadratic inequalities:

1. $4x^2 + 6x + 2 < 0$;
2. $4x^2 + 4x + 1 < 0$;
3. $-3x^2 + x - 6 < 0$.

■

Solution:

1. The discriminant $\Delta = 6^2 - 4 \times 4 \times 2 = 4 > 0$, so the quadratic equation $4x^2 + 6x + 2 = 0$ has two real roots $x_1 = -1, x_2 = -\frac{1}{2}$. Hence, the solution set for the inequality is $x \in (-\infty, -1) \cup \left(-\frac{1}{2}, \infty\right)$.
2. The discriminant $\Delta = 4^2 - 4 \times 4 \times 1 = 0$, so the quadratic equation $4x^2 + 4x + 1 = 0$ has one real double root. Therefore, the inequality has no solution set.

3. The discriminant $\Delta = 1^2 - 4 \times (-3) \times (-6) = -71 < 0$, so the quadratic equation $-3x^2 + x - 6 = 0$ has no real roots. Therefore, the solution set for the inequality $3x^2 - x + 6 > 0$ is all real numbers, and thus the solution set for the given inequality is also all real numbers.

4.3 important Inequalities

In this part, we explore two fundamental inequalities in mathematics: the Triangle Inequality and the Arithmetic-Geometric Mean (AGM) Inequality. Each section provides a comprehensive overview, including detailed proofs and corollaries.

4.3.1 The Triangle Inequality

The Triangle Inequality is a fundamental relation in geometry and analysis, asserting that the sum of the lengths of any two sides of a triangle must be greater than or equal to the length of the remaining side.

Definition 4.1 — Triangular Inequality. For any real numbers a and b, the Triangle Inequality is given by:

$$|a+b| \leq |a| + |b|$$

Proof. The proof of the Triangle Inequality considers the sign of a and b:

- **Case 1:** If a and b have the same sign, the inequality follows directly.
- **Case 2:** If a and b have opposite signs, assume $a > 0$ and $b < 0$. Then, $|a+b| \leq a - b = |a| + |b|$.

Thus, the Triangle Inequality is proven. ■

R This important inequality will also be seen in other chapters.

We also have the following conclusion by preliminary algebra:

$$\begin{aligned}
|a+b| \leqslant |a| + |b| &\Leftrightarrow |a+b|^2 \leqslant (|a|+|b|)^2 \\
&\Leftrightarrow (a+b)^2 \leqslant |a|^2 + 2|a||b| + |b|^2 \\
&\Leftrightarrow a^2 + 2ab + b^2 \leqslant a^2 + 2|a|\,|b| + b^2 \\
&\Leftrightarrow ab \leqslant |a|\,|b| \\
&\Leftrightarrow ab \leqslant |ab|.
\end{aligned}$$

The quality holds only when $ab \geq 0$.

Geometric Explanation of Triangular Inequality

Consider a and b are random numbers on a number axis:

- If $ab \geq 0$, then they are both in the same half-axis (both positive or negative). In this case, the distance between a and $-b$ is the sum of the distance from both points to the origin of the number axis.
- Now consider $ab < 0$, either of them is positive, and the other is negative. In this case, the distance between a and $-b$ is shorter than the sum of the distance from both points to the origin of the number axis.

Figure 4.2: Triangular Inequality: $ab < 0$

We also have:

Theorem 4.1 For $a, b, c \in \mathbb{R}$, $|a-c| \leq |a-b| + |b-c|$.

Proof. Since $a-c$, $a-b$, $b-c$ follows the relationship of triangular inequality. We have:

$$(a-b)(b-c) \geq 0$$

meaning that b is between a and c, which could be proven by sketching. ■

Corollary 4.1 $||a|-|b|| \leq |a+b|$

Proof.

$$|a| = |(a+b)-b| \leq |a+b| + |-b| = |a+b| + |b|.$$

Therefore,

$$|a| - |b| \leq |a+b|,$$

and similarly it can be proven that

$$|b| - |a| \leq |a+b|.$$

Hence,

$$||a-b|| \leq |a+b|.$$

■

Corollary 4.2 $|a| - |b| \leq |a+b|$

Proof. By corollary **??**:

$$||a| - |b|| \leq |a + (-b)|,$$

therefore,

$$||a| - |-b|| \leq |a - b|.$$

■

4.3.2 The Arithmetic-Geometric Mean Inequality

The AGM Inequality(also known as AM-GM inequality) states that for any set of non-negative real numbers, the arithmetic mean is always greater than or equal to the geometric mean.

Definition 4.2 — AGM Inequality. For non-negative real numbers $a_1, a_2, \ldots, a_n$, the AGM Inequality is:

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 \cdot a_2 \cdots a_n}$$

Many methods are available to prove this important conclusion. Below is the proof by MI.

Proof. We prove that for any non-negative real numbers $x_1, \ldots, x_n$, the following inequality holds:

$$\alpha^n \geq x_1 x_2 \cdots x_n$$

where α is the arithmetic mean of the numbers, with equality if and only if all the numbers are equal.

R This is because, the inequality is equivalent to AGM:

$$\frac{a_1 + a_2 + \ldots + a_n}{n} \geq \sqrt[n]{a_1 \cdot a_2 \ldots a_n} \Longleftrightarrow \left(\frac{a_1 + a_2 + \ldots + a_n}{n}\right)^n \geq a_1 \cdot a_2 \ldots a_n$$

Induction Basis: For $n = 1$, the statement is trivially true, as the arithmetic mean of a single number is the number itself.
Induction Hypothesis: Assume that the AM-GM inequality holds for n non-negative real numbers.
Induction Step: Consider $n+1$ non-negative real numbers $x_1, \ldots, x_{n+1}$. Their arithmetic mean α satisfies:

$$\alpha = \frac{x_1 + \cdots + x_n + x_{n+1}}{n+1}$$

If all the x_i are equal to α, then we have equality in the AM-GM statement, and we are done. In the case where some are not equal to α, there must exist at least one number

greater and one smaller than α. Without loss of generality, we can reorder our x_i to ensure that $x_n > \alpha > x_{n+1}$, which gives us:

$$(x_n - \alpha)(\alpha - x_{n+1}) > 0 \tag{4.1}$$

Define a new number y with:

$$y = x_n + x_{n+1} - \alpha \geq x_n - \alpha > 0$$

Since all numbers from x_1 to y are non-negative.

$$\begin{aligned}(n+1)\alpha &= x_1 + \cdots + x_{n-1} + x_n + x_{n+1} \\ n\alpha &= x_1 + \cdots + x_{n-1} + \underbrace{x_n + x_{n+1} - \alpha}_{=y},\end{aligned}$$

This shows that α is also a geometric mean of the n-sequence $x_1, \ldots, x_{n-1}, y$.
By the induction hypothesis:

$$\alpha^{n+1} = \alpha^n \cdot \alpha \geq x_1 x_2 \cdots x_{n-1} y \cdot \alpha.$$

From equation **??**, we have:

$$(\underbrace{x_n + x_{n+1} - \alpha}_{=y})\alpha - x_n x_{n+1} = (x_n - \alpha)(\alpha - x_{n+1}) > 0,$$

Hence:

$$y \cdot \alpha > x_n x_{n+1}$$

By substituting we have:

$$\alpha^{n+1} > x_1 x_2 \cdots x_{n-1} x_n x_{n+1}$$

This completes the proof. ■

Geometric representation

We can use function graphs to visualize the relationship implied in the AGM inequality. On the left and right-hand-side of the inequality sign, there are two functions with n variables. For simplicity, we use $n = 2$ to show its visualization. We obtain this interactive graph by Mathematica. This simple model involves three variables only. x_1 and x_2 are the values to be used to calculate AM and GM. We take one of the input value as preimage on the x-axis, and the AM or GM as the image on y-axis. The scroll on the top indicates the value of x_2. When $x_2 = 0$, the graph is shown below. It can be seen that the AGM inequality holds, and only when $x_2 = x_1 = 0$, there is an intersection $(0,0)$.

Figure 4.3: AGM when $x_2 = 0$

As we increase the x_2 value, the GM curve rise up, and the intersection moves on the right direction, remaining on the AM curve. Below is the visualization for $x_2 = 5$. There is no any overlap between graphs except the intersection $(5,5)$.

Figure 4.4: AGM when $x_2 = 5$

Problem 4.1 Where will the intersection go if we have $x_2 = 10$ for this model?

Now, I think everything is clear, but only things so far. From now on, we will extend the understanding of function graph to a new dimension, in the 3 dimension space. Functions in the 3D space is extremely important for multi-variable calculus, and learning linear algebra also requires us to abstract numbers in more than 3 dimension, but goes up to n dimension space. But don't worry, the discussion here only involve simple concepts.

We have three variables: x_1, x_2, y, where y is the value of AM or GM from x_1 and x_2. Coincidently, we have three coordinate axis in a 3D space. In the three-dimensional coordinate system, we assign two preimages x_1, x_2 to x and y axis, and the mean to the z axis. In this way we can get two function graphs, or surface, in the real sense, in the coordinate, which look like in the figure below.

Figure 4.5: AGM 3D Visualization

The plane in blue represents the set of all possible arithmetic mean derived from all combinations of x_1 and x_2 with $x_1, x_2 \in [0, 10]$, and the other surface, almost below the AM plane, is exactly the geometric mean curve that has a similar definition to the former. This is quite different from the functions we have known, since they are only a line or a curve in the Cartesian coordinate, yet in the 3D coordinate, function could be surface. We will explore more about multi-variable functions in the future.

It is noticeable that the two surfaces are actually independent. They intersect in a line that go across the space. That specific line is actually a set of all the ordered pairs (x_1, x_2) where $x_1 = x_2$. We can actually relate it to the graph in the Cartesian coordinate. If we change the view point to the front of the curves shaped in the 3D space, we see something like this picture.

Figure 4.6: AGM 3D Visualization (Front)

Isn't this quite similar to the 2D graph? Actually, the first image is exactly a horizontal cross-section of the 3D function graph, and the intersection in the 2D graph changes as the position of intersection line changes. These explain AGM inequality's geometric meaning.

Conclusions of AGM inequality

AGM inequality has brought us to these conclusions:

Corollary 4.3 — Mean Inequality Corollary. If $x, y > 0$, then

$$\frac{2xy}{x+y} \leq \sqrt{xy} \leq \frac{x+y}{2}.$$

Equality holds in each inequality only when $x = y$.

Proof. Proposition **??** yields $\sqrt{xy} \leq \frac{x+y}{2}$. We obtain the other inequality from this by multiplying both sides by the positive number $\frac{2\sqrt{xy}}{x+y}$, leading to:

$$\sqrt{xy} \cdot \frac{2\sqrt{xy}}{x+y} \leq \frac{x+y}{2} \cdot \frac{2\sqrt{xy}}{x+y},$$

which simplifies to:

$$\frac{2xy}{x+y} \leq \sqrt{xy}.$$

Thus, we have shown that $\frac{2xy}{x+y} \leq \sqrt{xy} \leq \frac{x+y}{2}$, with equality if and only if $x = y$. ■

The expression $\frac{2xy}{x+y}$ is the harmonic mean of x and y. It arises in the study of average rates. For example, consider traveling a distance d at rate r_1 in time t_1 and making the return trip at rate r_2 in time t_2. The harmonic mean gives us the average rate r for the full trip.

Assuming we travel the same distance d for both trips, we have:

$$r_1 t_1 = d$$
$$r_2 t_2 = d$$

The average rate r for the full trip is computed as follows:

$$r = \frac{2d}{t_1 + t_2} = \frac{2d}{\frac{d}{r_1} + \frac{d}{r_2}} = \frac{2d}{\frac{r_1 r_2 (t_1 + t_2)}{r_1 r_2}} = \frac{2r_1 r_2}{r_1 + r_2}$$

The other important corollary of AGM inequality is:

Corollary 4.4 $a^2 + b^2 \geq 2ab$

Proof. $a^2 + b^2 \geq 2ab \iff a^2 + b^2 - 2ab \geq 0 \iff (a-b)^2 \geq 0.$ ■

4.3.3 Exercises

Exercise 4.3 Prove that for any triangle with sides a, b, and c, the following inequality holds:

$$\frac{a}{b+c} + \frac{b}{a+c} + \frac{c}{a+b} > 1$$

■

Proof. Given a triangle with sides a, b, and c, we know that for any triangle, the sum of the lengths of any two sides is greater than the length of the third side. Therefore, $a < b+c$, $b < a+c$, and $c < a+b$.

Now, consider the inequality $\frac{a}{b+c} > \frac{a}{a+b+c}$. This is true because $b+c > a$ implies that the denominator of the left-hand side is smaller than that of the right-hand side while keeping the numerator constant.

Similarly, we can show that $\frac{b}{a+c} > \frac{b}{a+b+c}$ and $\frac{c}{a+b} > \frac{c}{a+b+c}$.

Adding these three inequalities, we get:

$$\frac{a}{b+c} + \frac{b}{a+c} + \frac{c}{a+b} > \frac{a+b+c}{a+b+c}$$

Simplifying the right-hand side, we obtain:

$$\frac{a}{b+c} + \frac{b}{a+c} + \frac{c}{a+b} > 1$$

Thus, the inequality is proven.

■

Exercise 4.4 Given $\varepsilon > 0$, if $|x-a| < \frac{\varepsilon}{4}$ and $|y-b| < \frac{\varepsilon}{6}$, prove that:

$$|2x+3y-2a-3b| < \varepsilon.$$

Hint: Rearrange it in a form that is good for using triangle inequality.

Proof.

$$\begin{aligned}
|2x+3y-2a-3b| &= |2(x-a)+3(y-b)| \\
&\leq |2(x-a)| + |3(y-b)| \\
&= 2|x-a| + 3|y-b| \\
&< 2 \times \frac{\varepsilon}{4} + 3 \times \frac{\varepsilon}{6} \\
&= \varepsilon.
\end{aligned}$$

■

Exercise 4.5 Prove for any real numbers a, b, c, d that:

$$|a-b| + |b-c| + |c-d| + |d-a| \geq |a-c| + |b-d|$$

Proof.

1. **Application of the Triangle Inequality:** The triangle inequality states that for any real numbers x, y, z:

   $$|x-y| \leq |x-z| + |z-y|$$

   We can apply this inequality to certain terms in our original problem. For example, consider $|a-c|$ and $|b-d|$.

2. **Separate Applications of the Triangle Inequality:** For $|a-c|$, we have:

$$|a-c| \leq |a-b| + |b-c|$$

For $|b-d|$, we have:

$$|b-d| \leq |b-c| + |c-d|$$

3. **Combining Inequalities:** Adding the above two inequalities, we get:

$$|a-c| + |b-d| \leq |a-b| + 2|b-c| + |c-d|$$

4. **Simplification and Rearrangement:** Notice that $2|b-c|$ appears on the right side of the inequality. However, since $|b-c|$ is non-negative, we can remove one $|b-c|$ and the inequality still holds. Hence, we have:

$$|a-c| + |b-d| \leq |a-b| + |b-c| + |c-d|$$

This is the reverse of the inequality in our original problem, so we can conclude:

$$|a-b| + |b-c| + |c-d| + |d-a| \geq |a-c| + |b-d|$$

5. **Conclusion:** Therefore, the original inequality is proved.

■

Exercise 4.6 Prove that for any real numbers x, y, z, the following inequality holds:

$$|x+y+z| \leq |x| + |y| + |z|$$

Proof. We will use the triangle inequality which states that for any real numbers a and b:

$$|a+b| \leq |a| + |b|$$

First, apply the triangle inequality to x and y:

$$|x+y| \leq |x| + |y|$$

Now, let $a = x+y$ and $b = z$, and apply the triangle inequality again:

$$|(x+y)+z| \leq |x+y| + |z|$$

Substitute the first inequality into the second one:

$$|x+y+z| \leq |x| + |y| + |z|$$

This completes the proof. ■

Exercise 4.7 Prove that for any sequence of real numbers $x_1, x_2, \ldots, x_n$, the following inequality holds:

$$|x_1 + x_2 + \ldots + x_n| \leq |x_1| + |x_2| + \ldots + |x_n|$$

Proof. We will prove this by induction on the number of terms n.

Base case ($n = 1$): For a single real number x_1, the inequality trivially holds as:

$$|x_1| = |x_1|$$

Inductive step: Assume the inequality holds for some $n = k$, i.e.,

$$|x_1 + x_2 + \ldots + x_k| \leq |x_1| + |x_2| + \ldots + |x_k|$$

Now, consider the case when $n = k + 1$. By the triangle inequality, we have:

$$|x_1 + x_2 + \ldots + x_k + x_{k+1}| \leq |x_1 + x_2 + \ldots + x_k| + |x_{k+1}|$$

Using the induction hypothesis, we can then write:

$$|x_1 + x_2 + \ldots + x_k + x_{k+1}| \leq (|x_1| + |x_2| + \ldots + |x_k|) + |x_{k+1}|$$

$$|x_1 + x_2 + \ldots + x_k + x_{k+1}| \leq |x_1| + |x_2| + \ldots + |x_k| + |x_{k+1}|$$

This completes the inductive step and thus, by the principle of mathematical induction, the inequality holds for all positive integers n. ■

Exercise 4.8 Let a, b, c be positive real numbers. Prove the inequality:

$$(a + b + c)(a^2 + b^2 + c^2) > 9abc.$$

Proof. By the Arithmetic Mean-Geometric Mean Inequality (AM-GM Inequality), we have:

$$\frac{a^2 + b^2 + c^2}{3} \geq \sqrt[3]{a^2b^2c^2},$$

$$\frac{a + b + c}{3} \geq \sqrt[3]{abc}.$$

Cubing both sides of the inequalities, we get:

$$(a^2 + b^2 + c^2)^3 \geq 27a^2b^2c^2,$$

$$(a + b + c)^3 \geq 27abc.$$

Multiplying the resulting inequalities, we obtain:

$$(a^2 + b^2 + c^2)^3(a + b + c)^3 \geq (27a^2b^2c^2)(27abc).$$

Taking the cube root of both sides, we arrive at:

$$(a^2 + b^2 + c^2)(a + b + c) \geq 9abc.$$

Note that the inequality is strict when a, b, c are positive real numbers, thus:

$$(a^2 + b^2 + c^2)(a + b + c) > 9abc.$$

■

Exercise 4.9 Given non-negative real numbers a, b, c such that $a+b+c=1$, we want to prove that:

$$a^2+b^2+c^2 \geq \frac{1}{3}.$$

■

Proof. We will use the Arithmetic Mean-Geometric Mean Inequality (AM-GM Inequality) which states that for any non-negative real numbers x, y, z, the following holds:

$$\frac{x+y+z}{3} \geq \sqrt[3]{xyz}.$$

Applying this to a, b, c, we have:

$$\frac{a+b+c}{3} \geq \sqrt[3]{abc} \Rightarrow \frac{1}{3} \geq \sqrt[3]{abc}.$$

Cubing both sides of the inequality yields:

$$\frac{1}{27} \geq abc.$$

Now, by AM-GM applied to a^2, b^2, c^2, we get:

$$\frac{a^2+b^2+c^2}{3} \geq \sqrt[3]{a^2b^2c^2}.$$

Since $a^2b^2c^2$ is the square of abc, it follows that:

$$\frac{a^2+b^2+c^2}{3} \geq (abc)^{\frac{2}{3}}.$$

Given $\frac{1}{27} \geq abc$, we have:

$$(abc)^{\frac{2}{3}} \leq \left(\frac{1}{27}\right)^{\frac{2}{3}} = \frac{1}{9}.$$

Therefore, we can conclude that:

$$\frac{a^2+b^2+c^2}{3} \geq \frac{1}{9}.$$

Multiplying through by 3, we obtain the desired inequality:

$$a^2+b^2+c^2 \geq \frac{1}{3}.$$

Hence, we have proved that $a^2+b^2+c^2 \geq \frac{1}{3}$ as required. ■

Exercise 4.10 Consider the function

$$f(x,y,z) = \frac{x}{y} + \frac{y}{z} + \frac{z}{x}$$

for all positive real numbers x, y, and z. Find the minimal value of the function. ■

Proof. For (x,y,z) positive real numbers, we have

$$f(x,y,z) = \frac{x}{y} + \frac{y}{z} + \frac{z}{x}$$

can be rewritten as

$$f(x,y,z) = 6 \cdot \left(\frac{1}{6} \cdot \frac{x}{y} + \frac{1}{6} \cdot \frac{y}{z} + \frac{1}{6} \cdot \frac{z}{x} + \frac{1}{6} \cdot \frac{x}{y} + \frac{1}{6} \cdot \frac{y}{z} + \frac{1}{6} \cdot \frac{z}{x} \right)$$

Setting $x_1 = \frac{x}{y}, x_2 = x_3 = \frac{1}{2} \cdot \frac{y}{z}, x_4 = x_5 = x_6 = \frac{1}{3} \cdot \frac{z}{x}$, and applying the AM-GM inequality for $n = 6$, we get

$$f(x,y,z) \geq 6 \cdot \sqrt[6]{x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5 \cdot x_6} = 6 \cdot \sqrt[6]{\frac{1}{2 \cdot 2 \cdot 3 \cdot 3 \cdot 3} \cdot \frac{x}{y} \cdot \frac{y}{z} \cdot \frac{z}{x}}$$

which simplifies to

$$f(x,y,z) \geq 6 \cdot \sqrt[6]{\frac{1}{2^2 \cdot 3^3}} = 2^{2/3} \cdot 3^{1/2}$$

Further, we know that the two sides are equal exactly when all the terms of the mean are equal. Thus,

$$f(x,y,z) = 2^{2/3} \cdot 3^{1/2}$$

when

$$\frac{x}{y} = \frac{1}{2} \cdot \frac{y}{z} = \frac{1}{3} \cdot \frac{z}{x}$$

All the points (x,y,z), satisfying these conditions lie on a half-line starting at the origin and are given by,

$$(x,y,z) = \left(t, \frac{3}{2}\sqrt{3}t, \frac{3}{2}\sqrt{3}t \right)$$

with $t > 0$. ■

4.3.4 Cauchy-Schwarz Inequality

Cauchy-Schwarz inequality is another important inequality that is used for mathematical proofs.

Theorem 4.2 — Cauchy-Schwarz Inequality. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be real numbers. Then

$$(a_1b_1 + a_2b_2)^2 \leq (a_1^2 + a_2^2)(b_1^2 + b_2^2)$$

Proof.

$$\begin{aligned}
&(a_1 + a_2)(b_1^2 + b_2^2) - (a_1b_1 + a_2b_2)^2 \geq 0 \\
&\Leftrightarrow a_1^2b_1^2 + a_2^2b_2^2 + a_1^2b_2^2 + a_2^2b_1^2 - a_1^2b_1^2 - a_2^2b_2^2 - 2a_1a_2b_1b_2 \geq 0 \\
&\Leftrightarrow a_1^2b_2^2 - 2a_1a_2b_1b_2 + a_2^2b_1^2 \geq 0 \\
&\Leftrightarrow (a_1b_2 - a_2b_1)^2 \geq 0
\end{aligned}$$

Hence, the equality holds if an only if when $a_1b_2 = a_2b_1$. ■

This is not a difficult proof, but it does not show all the implications of the inequality, especially in geometry. Let's take a look at how we can derive its vector form.

In the inequality of the inner product in vector space, let α and β be the directed line segments determined by vectors **a** and **b** respectively, with terminal coordinates a_1, a_2 and b_1, b_2, and let

$$\alpha = (a_1, a_2), \quad \beta = (b_1, b_2),$$

Then, α and β are not both zero, and the angle between them is denoted as $\langle \alpha, \beta \rangle$, with the convention that

$$0 \leq \langle \alpha, \beta \rangle \leq \pi.$$

The $\cos\langle \alpha, \beta \rangle$ is called the cosine of the angle (inner product) between vectors α and β, denoted as $\alpha \cdot \beta$, and

$$\begin{aligned}
&\alpha \cdot \beta = a_1b_1 + a_2b_2, \\
&|\alpha| = \sqrt{\alpha \cdot \alpha} = \sqrt{a_1^2 + a_2^2}, \\
&|\beta| = \sqrt{\beta \cdot \beta} = \sqrt{b_1^2 + b_2^2},
\end{aligned}$$

Therefore,

$$\cos\langle \alpha, \beta \rangle = \frac{a_1b_1 + a_2b_2}{\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2}},$$

$$\cos^2\langle \alpha, \beta \rangle = \left(\frac{a_1b_1 + a_2b_2}{\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2}} \right)^2 \leq 1,$$

which implies

$$(a_1^2 + a_2^2)(b_1^2 + b_2^2) \geq (a_1b_1 + a_2b_2)^2,$$

and since

$$\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2} \geq |a_1b_1 + a_2b_2|. \tag{4.2}$$

It is evident that $\cos^2\langle \alpha, \beta \rangle = 1$ implies $\langle \alpha, \beta \rangle = 0$ or π, which corresponds to vectors α and β being parallel or anti-parallel. When the angle is 0, vectors A and B are in the same direction, which is the condition for equality.

Theorem 4.3 — Cauchy-Schwarz Inequality(vector). Assume that $\alpha = (a_1, a_2)$ and $\beta = (b_1, b_2)$ are two planar vectors, then

$$|\alpha||\beta| \geq |\alpha \cdot \beta|$$

and

$$\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2} \geq |a_1 b_1 + a_2 b_2|$$

It is noticeable that when we take $a_2 = b_2 = 0$, we have

$$|a_1| + |b_1| \geq |a_1 + b_1|$$

which is exactly the triangle inequality for vector that we have discussed earlier.

4.3.5 Rearrangement Inequality

This section introduces **Rearrangement Inequality**, whose direct conclusions are the QM-AM-GM-HM inequality.

We start with an example. Suppose there are four boxes containing \$10, \$20, \$50 and \$100 bills, respectively. You may take 2 bills from one box, 3 bills from another, 4 bills from another, and 5 bills from the remaining box. What is the maximum amount of money you can get?

Clearly, youd want to take as many bills as possible from the box with largest-value bills! So you would take 5 \$100 bills, 4 \$50 bills, 3 \$20 bills, and 2 \$10 bills, for a grand total of

$$5 \cdot \$100 + 4 \cdot \$50 + 3 \cdot \$20 + 2 \cdot \$10 = \$780. \tag{4.3}$$

Suppose instead that your arch-nemesis (who isn't very good at math) is picking the bills instead, and he asks you how many bills he should take from each box. In this case, to minimize the amount of money he gets, youd want him to take as many bills as possible from the box with lowest-value bills. So you tell him to take 5 \$10 bills, 4 \$20 bills, 3 \$50 bills, and 2 \$100 bills, for a grand total of

$$5 \cdot \$10 + 4 \cdot \$20 + 3 \cdot \$50 + 2 \cdot \$100 = \$480. \tag{4.4}$$

The maximum is attained when the number of bills taken and the denominations are similarly sorted as in Eq. **??** and the minimum is attained when they are oppositely sorted as in **??**. The Rearrangement Inequality formalizes this observation.

Theorem 4.4 Rearrangement Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be real numbers (not necessarily positive) with

$$x_1 \leq x_2 \leq \ldots \leq x_n, \quad \text{and} \quad y_1 \leq y_2 \leq \ldots \leq y_n,$$

and let σ be a permutation of $\{1, 2, \ldots, n\}$. (That is, σ sends each of $1, 2, \ldots, n$ to a different value in $\{1, 2, \ldots, n\}$.) Then the following inequality holds:

$$x_1 y_n + x_2 y_{n-1} + \ldots + x_n y_1 \leq x_1 y_{\sigma(1)} + x_2 y_{\sigma(2)} + \ldots + x_n y_{\sigma(n)} \leq x_1 y_1 + x_2 y_2 + \ldots + x_n y_n.$$

Proof. We prove the inequality on the right by induction on n. The statement is obvious for $n = 1$. Suppose it true for $n-1$. Let m be an integer such that $\sigma(m) = n$. Since

$$x_n \geq x_m \quad \text{and} \quad y_n \geq y_{\sigma(n)},$$

we have

$$0 \leq (x_n - x_m)(y_n - y_{\sigma(n)})$$

$$\implies x_m y_n + x_n y_{\sigma(n)} \leq x_m y_{\sigma(n)} + x_n y_n.$$

Hence

$$x_1 y_{\sigma(1)} + \ldots + x_m y_{\sigma(m)} + \ldots + x_n y_{\sigma(n)} \leq x_1 y_{\sigma(1)} + \ldots + x_m y_n + \ldots + x_n y_n.$$

By the induction hypothesis,

$$x_1 y_{\sigma(1)} + \ldots + x_m y_n + \ldots + x_n y_{\sigma(n-1)} \leq x_1 y_1 + \ldots + x_m y_m + \ldots + x_n y_{n-1}.$$

Thus the RHS is at most $x_1 y_1 + \ldots + x_{n-1} y_{n-1} + x_n y_n$, as needed. To prove the LHS, apply the above with $-y_i$ instead of y_i (noting that negating an inequality reverses the sign). ■

R The equality holds if and only if $a_1 = a_2 = \cdots a_n$ or $b_1 = b_2 = \cdots b_n$

4.3.6 Exercises

Exercise 4.11 Prove that for any real numbers a, b, c, and d, the following inequality holds:

$$a^2 + b^2 + c^2 + d^2 \geq ab + bc + cd + da.$$

Proof. Consider the sequences (a, b, c, d) and (b, c, d, a). By the Cauchy-Schwarz inequality, we have:

$$(a^2 + b^2 + c^2 + d^2)(b^2 + c^2 + d^2 + a^2) \geq (ab + bc + cd + da)^2.$$

$$\left(a^2 + b^2 + c^2 + d^2\right)^2 \geq (ab + bc + cd + da)^2$$

$$a^2 + b^2 + c^2 + d^2 \geq ab + bc + cd + da.$$

This completes the proof. ■

Exercise 4.12 In mathematics, the QM-AM-GM-HM inequalities, also known as the mean inequality chain, state the relationship between the harmonic mean, geometric mean, arithmetic mean, and quadratic mean (also known as root mean square). It follows that:

$$0 < \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \leq \sqrt[n]{x_1 x_2 \cdots x_n} \leq \frac{x_1 + x_2 + \cdots + x_n}{n} \leq \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} \quad (4.5)$$

The right part of the inequality is a direct conclusion from Cauchy-Schwarz inequality. Prove it by substituting special values to fit the relation. ■

Proof. Cauchy-Schwarz inequality states that:

$$\left(a_1^2+a_2^2+\cdots+a_n^2\right)\left(b_1^2+b_2^2+\cdots+b_n^2\right) \geq (a_1b_1+a_2b_2+\cdots+a_nb_n)^2$$

where all terms are real numbers. Let $b_1 = b_2 = \cdots = b_n = 1$, we have:

$$\left(x_1^2+x_2^2+\cdots+x_n^2\right)(1+1+\cdots+1) \geq (x_1+x_2+\cdots+x_n)^2$$

$$\frac{x_1^2+x_2^2+\cdots+x_n^2}{n} \geq \left(\frac{x_1+x_2+\cdots+x_n}{n}\right)^2$$

$$\sqrt{\frac{x_1^2+\cdots+x_n^2}{n}} \geq \frac{x_1+\cdots+x_n}{n}$$

This completes the proof. ■

If you are interested in the complete proof, check this link.

Exercise 4.13 Prove that, let a,b,c be positive real numbers such that $a+b+c=1$. Then the following inequality holds:

$$\frac{1}{a}+\frac{1}{b}+\frac{1}{c} \geq 9.$$

■

Proof. By applying the Cauchy-Schwarz inequality to the sequences $(\sqrt{a},\sqrt{b},\sqrt{c})$ and $(\frac{1}{\sqrt{a}},\frac{1}{\sqrt{b}},\frac{1}{\sqrt{c}})$, we have:

$$\left((\sqrt{a})^2+(\sqrt{b})^2+(\sqrt{c})^2\right)\left(\left(\frac{1}{\sqrt{a}}\right)^2+\left(\frac{1}{\sqrt{b}}\right)^2+\left(\frac{1}{\sqrt{c}}\right)^2\right)$$

$$\geq \left(\sqrt{a}\cdot\frac{1}{\sqrt{a}}+\sqrt{b}\cdot\frac{1}{\sqrt{b}}+\sqrt{c}\cdot\frac{1}{\sqrt{c}}\right)^2$$

Simplifying both sides of the inequality gives us:

$$(a+b+c)\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}\right) \geq (1+1+1)^2$$

Given that $a+b+c=1$, substituting this into the inequality yields:

$$1\cdot\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}\right) \geq 3^2$$

Therefore, we have:

$$\frac{1}{a}+\frac{1}{b}+\frac{1}{c} \geq 9.$$

This completes the proof. ■

Exercise 4.14 We have proven the Cauchy-Schwarz inequality for planar vectors, can you expand theorem **??** to 3D vectors? Hint: Express your solution in the form of Eq. **??**. No extra thinking is needed. ■

Solution:

In the 3D space, we have

Corollary 4.5 Assume that $\alpha = (a_1, a_2, a_3)$ and $\beta = (b_1, b_2, b_3)$ are two 3D vectors:

$$\sqrt{a_1^2 + a_2^2 + a_3^2}\sqrt{b_1^2 + b_2^2 + b_3^2} \geq |a_1 b_1 + a_2 b_2 + a_3 b_3|.$$

Proof. In the inequality of the inner product in vector space, let α and β be the directed line segments determined by vectors **a** and **b** respectively, with terminal coordinates a_1, a_2, a_3 and b_1, b_2, b_3, and let

$$\alpha = (a_1, a_2, a_3), \quad \beta = (b_1, b_2, b_3),$$

Then, α and β are not both zero, and the angle between them is denoted as $\langle \alpha, \beta \rangle$, with the convention that

$$0 \leq \langle \alpha, \beta \rangle \leq \pi.$$

The $\cos\langle \alpha, \beta \rangle$ is called the cosine of the angle (inner product) between vectors α and β, denoted as $\alpha \cdot \beta$, and

$$\alpha \cdot \beta = a_1 b_1 + a_2 b_2 + a_3 b_3,$$

$$|\alpha| = \sqrt{\alpha \cdot \alpha} = \sqrt{a_1^2 + a_2^2 + a_3^2},$$

$$|\beta| = \sqrt{\beta \cdot \beta} = \sqrt{b_1^2 + b_2^2 + b_3^2},$$

Therefore,

$$\cos\langle \alpha, \beta \rangle = \frac{a_1 b_1 + a_2 b_2 + a_3 b_3}{\sqrt{a_1^2 + a_2^2 + a_3^2}\sqrt{b_1^2 + b_2^2 + b_3^2}},$$

$$\cos^2\langle \alpha, \beta \rangle = \left(\frac{a_1 b_1 + a_2 b_2 + a_3 b_3}{\sqrt{a_1^2 + a_2^2 + a_3^2}\sqrt{b_1^2 + b_2^2 + b_3^2}} \right)^2 \leq 1,$$

which implies

$$(a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) \geq (a_1 b_1 + a_2 b_2 + a_3 b_3)^2,$$

and since

$$\sqrt{a_1^2 + a_2^2 + a_3^2}\sqrt{b_1^2 + b_2^2 + b_3^2} \geq |a_1 b_1 + a_2 b_2 + a_3 b_3|.$$

which completes the proof. ■

Exercise 4.15 We have shown that the Cauchy-Schwarz inequality still works in the 3-dimensional space, why not try to show that it holds for all vectors in the n-dimension space? Think about the strategy to be used and show your meticulous proof for nD vector. Hint: We are actually trying to find a generalized form of the inequality. Try to express it in an open form (summation). ■

Solutions:

Theorem 4.5 — Generalized Cauchy-Schwarz Inequality(vector). We consider vectors in n-dimentional space. Let $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ be two n-dimensional vectors. Then the following inequality holds:

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

Proof. We will prove the Cauchy-Schwarz inequality for n-dimensional vectors using mathematical induction.

Base Case: For $n = 1$, we have:

$$(a_1 b_1)^2 \leq (a_1^2)(b_1^2)$$

which is obviously true, as both sides are equal.

Inductive Step: Assume the inequality holds for $n = k$, that is:

$$\left(\sum_{i=1}^{k} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{k} a_i^2\right)\left(\sum_{i=1}^{k} b_i^2\right)$$

We need to show that it also holds for $n = k+1$:

$$\left(\sum_{i=1}^{k+1} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{k+1} a_i^2\right)\left(\sum_{i=1}^{k+1} b_i^2\right)$$

Expanding both sides, we get:

LHS:

$$\left(\sum_{i=1}^{k} a_i b_i + a_{k+1} b_{k+1}\right)^2 = \left(\sum_{i=1}^{k} a_i b_i\right)^2 + 2a_{k+1} b_{k+1} \sum_{i=1}^{k} a_i b_i + a_{k+1}^2 b_{k+1}^2$$

RHS:

$$\left(\sum_{i=1}^{k} a_i^2 + a_{k+1}^2\right)\left(\sum_{i=1}^{k} b_i^2 + b_{k+1}^2\right)$$

$$= \left(\sum_{i=1}^{k} a_i^2\right)\left(\sum_{i=1}^{k} b_i^2\right) + b_{k+1}^2 \sum_{i=1}^{k} a_i^2 + a_{k+1}^2 \sum_{i=1}^{k} b_i^2 + a_{k+1}^2 b_{k+1}^2$$

By the inductive hypothesis, we have:

$$\left(\sum_{i=1}^{k} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{k} a_i^2\right)\left(\sum_{i=1}^{k} b_i^2\right)$$

It remains to show that the additional terms also satisfy the inequality:

$$2a_{k+1}b_{k+1}\sum_{i=1}^{k}a_ib_i+a_{k+1}^2b_{k+1}^2\leq b_{k+1}^2\sum_{i=1}^{k}a_i^2+a_{k+1}^2\sum_{i=1}^{k}b_i^2+a_{k+1}^2b_{k+1}^2$$

This is equivalent to showing:

$$\sum_{i=1}^{k}(a_{k+1}b_i-a_ib_{k+1})^2\geq 0$$

Which is true since it is a sum of squares.
Hence, by mathematical induction, the inequality holds for all n-dimensional vectors.

R If this is not quite clear for you, here's the explanation:
Cancel $a_{k+1}^2b_{k+1}^2$ on both sides, we have:

$$2a_{k+1}b_{k+1}\sum_{i=1}^{k}a_ib_i\leq a_{k+1}^2\sum_{i=1}^{k}b_i^2+b_{k+1}^2\sum_{i=1}^{k}a_i^2$$

Moving the terms on the same side of the inequality sign:

$$a_{k+1}^2\sum_{i=1}^{k}b_i^2+b_{k+1}^2\sum_{i=1}^{k}a_i^2-2a_{k+1}b_{k+1}\sum_{i=1}^{k}a_ib_i\geq 0$$

This is an interesting expression since the coefficients and the terms of the summation can all fit the terms of sum of square: $(a\pm b)^2=a^2+b^2\pm 2ab$. we can expand LHS as:

$$LHS=a_{k+1}^2\left(b_1^2+b_2^2+\cdots+b_k^2\right)+b_{k+1}^2\left(a_1^2+a_2^2+\cdots+a_k^2\right)-2a_{k+1}b_{k+1}(a_1b_1+a_2b_2+\cdots+a_kb_k)$$

Fitting it into sum of square formula $(a\pm b)^2=a^2+b^2\pm 2ab$ with $a=a_{k+1}b_i$ and $b=a_ib_{k+1}$.
We have:

$$\sum_{i=1}^{k}(a_{k+1}b_i-a_ib_{k+1})^2=LHS$$

■

Exercise 4.16 Having proven the generalized Cauchy-Schwarz Inequality by using vector properties, you must have built up some confidence. These problems that seem tricky are actually easy to prove. Now we have one last proof for Cauchy-Schwarz Inequality (it's true I promise, no more proof for this inequality.). Since we have proved the generalized form by using vector, recall that we also prove it by pure algebra analysis in theorem **??**. Now, prove the generalized form without using vector, and think about the similarity and difference of the two proofs. ■

solution:

Proof. The equality holds if and only if $\frac{a_1}{b_1}=\frac{a_2}{b_2}=\ldots=\frac{a_n}{b_n}$ (assuming $b_i\neq 0$ for all i; if $b_j=0$, then $a_j=0$, for $j=1,2,\ldots,n$).

Case 1: If all $a_i=0$, the inequality obviously holds.

Case 2: If not all $a_i = 0$, consider the following quadratic in x:

$$f(x) = (a_1x + b_1)^2 + (a_2x + b_2)^2 + \ldots + (a_nx + b_n)^2$$

It is clear that for all x, $f(x) \geq 0$ since it is a sum of squares.

Expanding $f(x)$, we get:

$$f(x) = (a_1^2 + a_2^2 + \ldots + a_n^2)x^2 + 2(a_1b_1 + a_2b_2 + \ldots + a_nb_n)x + (b_1^2 + b_2^2 + \ldots + b_n^2)$$

The discriminant Δ of this quadratic is:

$$\Delta = 4(a_1b_1 + a_2b_2 + \ldots + a_nb_n)^2 - 4(a_1^2 + a_2^2 + \ldots + a_n^2)(b_1^2 + b_2^2 + \ldots + b_n^2)$$

For $f(x) \geq 0$ to hold for all x, the discriminant Δ must be less than or equal to zero. This leads to:

$$(a_1b_1 + a_2b_2 + \ldots + a_nb_n)^2 - (a_1^2 + a_2^2 + \ldots + a_n^2)(b_1^2 + b_2^2 + \ldots + b_n^2) \leq 0$$

Hence, the inequality is proven. Equality holds if and only if the numbers are proportional. ■

Exercise 4.17 If one did not know the Cauchy Schwarz inequality, but knew Lagrange's identity, then how could one derive the Cauchy-Schwarz inequality.

Theorem 4.6 — Lagrange's Identity. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be real numbers. Then

$$\left(\sum_{k=1}^{n} a_k^2\right)\left(\sum_{k=1}^{n} b_k^2\right) - \left(\sum_{k=1}^{n} a_kb_k\right)^2 = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(a_ib_j - a_jb_i)^2$$

Prove this identity with summation form of generalized Cauchy-Schwarz inequality. Hint: Loosely speaking, Lagrange's identity says that the left-hand side in the Cauchy- Schwarz inequality is off from the right-hand side of the Cauchy-Schwarz inequality by the error term. ■

Proof. Cauchy Inequality states that:

$$\left(\sum_{i=1}^{n} a_ib_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

Assuming the Cauchy-Schwarz inequality is true, we aim to demonstrate Lagrange's Identity. The inequality gives us the left-hand side of the identity directly. To arrive at the right-hand side, we need to show the existence of an error term which is the sum of squares of the differences between the products of the components of **a** and **b**.

Consider the expression $\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(a_ib_j - a_jb_i)^2$, which expands to include all combinations of the product differences. Each term is of the form $a_i^2b_j^2 - 2a_ib_ia_jb_j + a_j^2b_i^2$, and when summed over all i and j, the middle terms $-2a_ib_ia_jb_j$ combine to give us $-2\left(\sum_{k=1}^{n} a_kb_k\right)^2$, which is the term we subtract on the left-hand side of the identity to obtain the equality.

Thus, by rearranging the terms and recognizing that the right-hand side sum is always non-negative, we establish Lagrange's Identity as an equality.

■

Exercise 4.18 Consider the function $f(x) = \frac{(x+k)^2}{x^2+1}$ where k is a positive whole number. Show that $f(x) \leq k^2 + 1$. ■

Hint: Try to fit Cauchy-Schwarz inequality. There are more than one method.

Proof. We will prove that $\frac{(x+k)^2}{x^2+1} \leq k^2+1$, which is equivalent to $(x+k)^2 \leq (k^2+1)(x^2+1)$.

Notice that $(k^2+1)(x^2+1)$ has (by Cauchy-Schwarz Inequality):

$$(k^2+1)(x^2+1) \geq (kx+1)^2 = k^2x^2 + 2kx + 1$$

where $x \in \mathbb{R}$ and $k \in \mathbb{Z}^+$.

We also have $(x+k)^2 = x^2 + 2kx + k^2$, now we need to compare k^2x^2+1 and x^2+k^2. Cancel the $2kx$ term, we need to show that

$$x^2 + k^2 \leq k^2x^2 + 1$$

Simplifying, we see:

$$x^2(1-k^2) \leq 1 - k^2$$

Since $1-k^2$ is non-positive for $k \in \mathbb{Z}^+$, dividing by $1-k^2$ (which reverses the inequality due to the negative sign) gives us:

$$x^2 \geq 1$$

Which holds true for all real x. Therefore, the original inequality is proven.

Below is how you can construct a new inequality to prove it.

Firstly, we utilize the vector dot product form of the Cauchy-Schwarz Inequality over the real numbers. To apply the vector dot product form, we identify corresponding elements for two vectors. Let's choose vectors **a** and **b**:

$$\mathbf{a} = \begin{bmatrix} x \\ k \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ \frac{1}{x} \end{bmatrix}$$

The Cauchy-Schwarz Inequality is stated as:

$$(\mathbf{a} \cdot \mathbf{b})^2 \leq (\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b})$$

Applying this to our case:

$$\left(x \cdot 1 + k \cdot \frac{1}{x}\right)^2 \leq (x^2+k^2)\left(1^2 + \left(\frac{1}{x}\right)^2\right)$$

$$\left(\frac{x+k}{x}\right)^2 \leq (x^2+k^2)\left(1 + \frac{1}{x^2}\right)$$

$$\left(\frac{(x+k)^2}{x^2}\right) \leq (x^2+k^2)\left(\frac{x^2+1}{x^2}\right)$$

$$\frac{(x+k)^2}{x^2+1} \leq k^2+1$$

Therefore, we have proved the given inequality $f(x) = \frac{(x+k)^2}{x^2+1} \leq k^2+1$ using the Cauchy-Schwarz Inequality.

■

Exercise 4.19 For $a,b,c > 0$ prove that
(a) $a^a b^b c^c \geq a^b b^c c^a$.
(b) $a^a b^b c^c \geq (abc)^{\frac{a+b+c}{3}}$.

■

Proof. Since ln is an increasing function, we take the ln of both sides to find that the inequalities are equivalent to

$$a\ln a + b\ln b + c\ln c \geq b\ln a + c\ln b + a\ln c$$
$$a\ln a + b\ln b + c\ln c \geq \frac{a+b+c}{3}(\ln a + \ln b + \ln c).$$

■

Exercise 4.20 Chebyshev inequality is an important conclusion about random variables. It states that: Let $a_1, a_2, \ldots, a_n; b_1, b_2, \ldots, b_n$ be real numbers arranged in ascending order for a_i and in descending order for b_i, such that:
(1) If $a_1 \leq a_2 \leq \ldots \leq a_n$ and $b_1 \geq b_2 \geq \ldots \geq b_n$, then

$$\frac{a_1b_1 + a_2b_2 + \ldots + a_nb_n}{n} \geq \left(\frac{a_1+a_2+\ldots+a_n}{n}\right)\left(\frac{b_1+b_2+\ldots+b_n}{n}\right)$$

(2) If $a_1 \leq a_2 \leq \ldots \leq a_n$ but $b_1 \leq b_2 \leq \ldots \leq b_n$, then

$$\frac{a_1b_1 + a_2b_2 + \ldots + a_nb_n}{n} \leq \left(\frac{a_1+a_2+\ldots+a_n}{n}\right)\left(\frac{b_1+b_2+\ldots+b_n}{n}\right)$$

■

Proof. For (1) where $a_1 \leq a_2 \leq \ldots \leq a_n$, and $b_1 \geq b_2 \geq \ldots \geq b_n$, we have the following:

$$\begin{aligned}
a_1b_1 + a_2b_2 + \ldots + a_nb_n &= a_1b_1 + a_2b_2 + \ldots + a_nb_n \\
a_1b_1 + a_2b_2 + \ldots + a_nb_n &\geq a_1b_2 + a_2b_3 + \ldots + a_nb_1 \\
a_1b_1 + a_2b_2 + \ldots + a_nb_n &\geq a_1b_3 + a_2b_4 + \ldots + a_{n-1}b_1 + a_nb_2 \\
&\vdots \\
a_1b_1 + a_2b_2 + \ldots + a_nb_n &\geq a_1b_n + a_2b_1 + \ldots + a_nb_{n-1}.
\end{aligned}$$

Adding these n inequalities together, we obtain:

$$n(a_1b_1 + a_2b_2 + \ldots + a_nb_n) \geq (a_1 + a_2 + \ldots + a_n)(b_1 + b_2 + \ldots + b_n).$$

After dividing by n^2, it follows that:

$$\frac{a_1b_1 + a_2b_2 + \ldots + a_nb_n}{n} \geq \left(\frac{a_1 + a_2 + \ldots + a_n}{n}\right)\left(\frac{b_1 + b_2 + \ldots + b_n}{n}\right).$$

This holds especially if $a_1 = a_2 = \ldots = a_n$ and $b_1 = b_2 = \ldots = b_n$, which is trivial. For (2) where the sequences are oppositely sorted, the proof is similar. ■

Exercise 4.21 Suppose $a_1, a_2, \ldots, a_n > 0$ and let $s = a_1 + \ldots + a_n$. Prove that

$$\frac{a_1}{s - a_1} + \ldots + \frac{a_n}{s - a_n} \geq \frac{n}{n-1}$$

■

Proof. The left-hand side of the inequality can be seen as the average of n fractions where the numerators are the a_i's and the denominators are $s - a_i$'s.

Since the sum of the numerators equals the sum of the a_i's and the sum of the denominators is $n(s - \frac{s}{n}) = (n-1)s$, we can apply Chebyshev's Inequality because it states that if $a_1 \leq \ldots \leq a_n$ and $b_1 \leq \ldots \leq b_n$, then the arithmetic mean of the products a_ib_i is greater than or equal to the product of the arithmetic means of a_i and b_i.

By rearranging the terms, we can match the a_i's with the inverses of denominators in a way that corresponds to the conditions of Chebyshev's Inequality:

$$\sum_{i=1}^{n} a_i \cdot \frac{1}{s - a_i} \geq \frac{1}{n}\sum_{i=1}^{n} a_i \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{1}{s - a_i}.$$

This simplifies to:

$$\frac{a_1}{s - a_1} + \ldots + \frac{a_n}{s - a_n} \geq \frac{s}{(n-1)s} = \frac{n}{n-1}.$$

Therefore, we have shown that:

$$\frac{a_1}{s - a_1} + \ldots + \frac{a_n}{s - a_n} \geq \frac{n}{n-1}.$$

This completes the proof.

■

Exercise 4.22 Prove the following for $x, y, z > 0$:

(a) $\frac{x^2}{y} + \frac{y^2}{x} \geq x + y$.

(b) $\frac{x^2}{y^2} + \frac{y^2}{z^2} + \frac{z^2}{x^2} \geq \frac{x}{z} + \frac{y}{x} + \frac{z}{y}$.

(c) $\frac{xy}{z^2} + \frac{yz}{x^2} + \frac{zx}{y^2} \geq \frac{x}{y} + \frac{y}{z} + \frac{z}{x}$.

■

Proof. (a) Without loss of generality, $x \geq y$. Then $x^2 \geq y^2$ and $\frac{1}{y} \geq \frac{1}{x}$, i.e., (x^2, y^2) and $(\frac{1}{y}, \frac{1}{x})$ are similarly sorted. Thus

$$\frac{x^2}{y} + \frac{y^2}{x} \geq \frac{x^2}{x} + \frac{y^2}{y} = x + y.$$

(b) Letting $a = \frac{x}{y}$, $b = \frac{y}{z}$, $c = \frac{z}{x}$, the inequality is equivalent to

$$a^2 + b^2 + c^2 \geq ab + bc + ca.$$

This is true by the rearrangement inequality applied to the similarly sorted sequences (a, b, c) and (a, b, c).

(c) Let $a = \frac{1}{x}\frac{1}{y^{1/3}}\frac{1}{z^{1/3}}$, $b = \frac{1}{y}\frac{1}{z^{1/3}}\frac{1}{x^{1/3}}$, and $c = \frac{1}{z}\frac{1}{x^{1/3}}\frac{1}{y^{1/3}}$. Then the inequality to prove becomes

$$a^3 + b^3 + c^3 \geq a^2b + b^2c + c^2a$$

which was proved in problem 1.

■

In **??**, we introduced the categorization of numbers. This chapter will unveil the most special subset of number that we have known, which is complex number. Complex number is a powerful mathematical tool for Computer Science, especially for some machine learning algorithms and computer graphics.

In the 18th century, some mathematicians are confused by the roots of equations, especially high-ordered equation. In some cases, they have to get the square root of a negative number, which is non-existent in the real number filed. In this context, mathematicians extend their knowledge to complex number. For equations such as $x^2 = -1$, we can easily find out that the root is $\sqrt{-1}$, which is not defined for all real numbers. Mathematicians define $\sqrt{-1}$ as **Imaginary Number**, denoted by i.

Definition 5.1 — imaginary Number. The equation $x^2 = -1$ is used to define imaginary number i, which gives $i^2 = -1$. The two roots are i and $-i$ respectively.

Formally, a compex number is defined as:

Definition 5.2 A complex number is an expression of the form $a + bi$, where a and b are real numbers. The set of all complex numbers is denoted by $\mathbb{C}$. That is,

$$\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}$$

The letter z is often used to denote a complex number.

- If $a = 0$, then $z = bi$ is said to be an imaginary number.
- If $b = 0$, then $z = a$ is a real number.

The real numbers and the imaginary numbers are subsets of $\mathbb{C}$

Definition 5.3 — Real and Imaginary Part. For a complex number $z = a + bi$, we define

$$\mathrm{Re}(z) = a \quad \text{and} \quad \mathrm{Im}(z) = b$$

where $\mathrm{Re}(z)$ is called the *real part* of z, and $\mathrm{Im}(z)$ is called the *imaginary part* of z.

Note: Both $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ are real numbers. That is, $\mathrm{Re} : \mathbb{C} \to \mathbb{R}$ and $\mathrm{Im} : \mathbb{C} \to \mathbb{R}$.

Sometimes we need to represent or simplify a given number to complex number. refer to the following examples

■ **Example 5.1** **a** Represent $\sqrt{-5}$ as an imaginary number. **b** Simplify $2\sqrt{-9}+4i$. ■

Solution

a $\sqrt{-5} = \sqrt{5} \times (-1)$
$= \sqrt{5} \times \sqrt{-1}$
$= i\sqrt{5}$

b $2\sqrt{-9}+4i = 2\sqrt{9} \times (-1)+4i$
$= 2 \times 3 \times i+4i$
$= 6i+4i$
$= 10i$

5.1 Algebra of Complex Number

This section discusses operations of complex number and some of its algebraic properties. For rationals, we have

- **Commutative Law of Addition**

$$a+b = b+a$$

- **Commutative Law of Multiplication**

$$ab = ba$$

- **Associative Law of Addition**

$$a+(b+c) = (a+b)+c$$

- **Associative Law of Multiplication**

$$a(bc) = (ab)c$$

- **Distributive Law**

$$(a+b)c = ac+bc,$$

for any rationals a, b, and c.

These basic rules are still available for complex number.

Definition 5.4 — Addition and Subtraction of Complex Number. The operations of addition and subtraction of complex numbers are given by

$$(a+bi) \pm (c+di) = (a \pm c)+(b \pm d)i,$$

Definition 5.5 — Multiplication of Complex Number. The multiplication of two complex numbers is defined by

$$(a+bi)(c+di) = (ac-bd)+(bc+ad)i.$$

where $i^2 = -1$.

Now let's consider division of complex number. If we write it as a fraction, we have two complex number with real and imaginary parts, which is quite tricky to handle. In this case we can use the technique to deal with fraction with root denominator, rationalization, to cancel the imaginary part in the denominator, as $i^2 = -1$.

Definition 5.6 — Division of Complex Number. The division of complex numbers is given by

$$\frac{a+bi}{c+di} := \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}i \quad (\text{if } c^2+d^2 \neq 0).$$

■ **Example 5.2** Find the quotient

$$\frac{(6+2i)-(1+3i)}{(-1+i)-2}.$$

■

Solution.

$$\begin{aligned}\frac{(6+2i)-(1+3i)}{(-1+i)-2} &= \frac{5-i}{-3+i} = \frac{5-i}{-3+i}\cdot\frac{(-3-i)}{(-3-i)} \\ &= \frac{-15-1-5i+3i}{9+1} \\ &= \frac{-16-2i}{10} \\ &= \frac{-8}{5} - \frac{1}{5}i.\end{aligned}$$

5.1.1 Exercises

Exercise 5.1 Verify the commutative, associative, and distributive laws for complex numbers. ■

Exercise 5.2 Notice that 0 and 1 retain their identity properties as complex numbers; that is, $0+z=z$ and $1\cdot z = z$ when z is complex.

(a) Verify that complex subtraction is the inverse of complex addition (that is, $z_3 = z_2 - z_1$ if and only if $z_3 + z_1 = z_2$).

(b) Verify that complex division, as given in the text, is the inverse of complex multiplication (that is, if $z_2 \neq 0$, then $z_3 = z_1/z_2$ if and only if $z_3 z_2 = z_1$).

■

Exercise 5.3 Prove that if $z_1 z_2 = 0$, then $z_1 = 0$ or $z_2 = 0$. ■

Exercise 5.4 Show that $\Re(iz) = -\Im z$ for every complex number z. ■

Hint: Prove using $z = a+bi$ directly.

Exercise 5.5 Let k be an integer. show that

$$i^{4k} = 1,\ i^{4k+1} = i,\ i^{4k+2} = -1,\ i^{4k+3} = -i$$

and thus evalueate

$$3i^{11}+6i^3+\frac{8}{i^{20}}+i^{-1}$$

Proof. We know that $i^2=-1$. Therefore, we can express powers of i in terms of powers of -1:

$$\begin{aligned}
i^{4k} &= (i^2)^{2k} = (-1)^{2k} = 1,\\
i^{4k+1} &= i^{4k}\cdot i = 1\cdot i = i,\\
i^{4k+2} &= i^{4k}\cdot i^2 = 1\cdot(-1) = -1,\\
i^{4k+3} &= i^{4k+2}\cdot i = (-1)\cdot i = -i.
\end{aligned}$$

Now we can evaluate the given expression:

$$\begin{aligned}
&3i^{11}+6i^3+\frac{8}{i^{20}}+i^{-1} = 3i^{4(2)+3}+6i^{4(0)+3}+\frac{8}{i^{4(5)}}+i^{-1}\\
&= 3(-i)+6(-i)+\frac{8}{1}+\frac{1}{i}\\
&= -3i-6i+8-i\cdot\left(\frac{1}{i}\cdot\frac{i}{i}\right) = -3i-6i+8-\frac{i}{i^2} = -3i-6i+8-\frac{i}{-1} = -3i-6i+8+i\\
&= 8-10i
\end{aligned}$$

Therefore, the evaluated expression is $8-10i$. ■

Exercise 5.6 Solve each of the following equations for z.

(a) $iz = 4-zi$
(b) $\frac{z}{1-z} = 1-5i$
(c) $(2-i)z+8z^2 = 0$
(d) $z^2+16 = 0$

Exercise 5.7 The complex numbers z_1, z_2 satisfy the system of equations

$$\begin{aligned}
(1-i)z_1+3z_2 &= 2-3i,\\
iz_1+(1+2i)z_2 &= 1.
\end{aligned}$$

Find z_1, z_2.

Hint: To find z_1 and z_2, we solve the system of equations and find that

$$z_1 = 1+i$$

$$z_2 = -i$$

Exercise 5.8 The straightforward method of computing the product $(a+bi)(c+di) = (ac-bd)+i(bc+ad)$ requires four (real) multiplications (and two signed additions). On most computers multiplication is far more time-consuming than addition. Devise an algorithm for computing $(a+bi)(c+di)$ with only three multiplications (at the cost of extra additions).

- Write this algorithm in pseudocode

Hint: Start with $(a+b)(c+d)$, this is calleded "Karatsuba's algorithm"
Solution:

Algorithm 3 Karatsuba's algorithm for multiplying two complex numbers

```
procedure KARATSUBAMULTIPLY(a, b, c, d)
    ac ← a · c
    bd ← b · d
    abcd ← (a + b) · (c + d)
    real ← ac − bd
    imag ← abcd − ac − bd
    return real + imag · i
end procedure
```

5.2 Point representation of Complex Number

This section delves into the representation of Complex Number in the Cartesian coordinate, with which we are already quite familiar. In this system, we use ordered pairs like (a,b) to show the position of a given point. In the study of complex number, we can draw all complex number on a Cartesian coordinate which we call **Complex Plane**.

Figure 5.1: Complex Plane

Definition 5.7 — Complex Plane. The complex plane is a two-dimensional space where each point represents a complex number. The horizontal axis represents the real part of the number, and the vertical axis represents the imaginary part. This allows for a geometric interpretation of complex numbers.

Now that we have point representation on the complex plane, we can gauge the length of a complex number, which we call **absolute value** or **modulus**, as what we do to vectors.

Definition 5.8 — Modulus of Complex Number. The absolute value or modulus of the number $z = a + bi$ is denoted by $|z|$ and is given by

$$|z| := \sqrt{a^2 + b^2}.$$

In particular,

$$|0| = 0, \quad \left|\frac{i}{2}\right| = \frac{1}{2}, \quad |3-4i| = \sqrt{9+16} = 5.$$

Similarly, we can also gain the distance between two complex numbers in the plane by taking them as two points.

Definition 5.9 — Distance between Complex Number. Let $z_1 = a_1 + b_1 i$ and $z_2 = a_2 + b_2 i$.

$$|z_1 - z_2| = |(a_1 - a_2) + (b_1 - b_2)i| = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2} \tag{5.1}$$

We can use this property to describe curves in the plane.

■ **Example 5.3** Draw the area that satisfies $|z - z_0| = 1$, where $z_0 = 2 + 2i$. ■

This set consists of all points z whose distance from z_0 is r.

■ **Example 5.4** Describe the set of points z that satisfy the equations

(a) $|z+2| = |z-1|$,

(b) $|z-1| = \operatorname{Re} z + 1$.

■

Solution.

(a) A point z satisfies Eq. (a) if and only if it is equidistant from the points -2 and 1. Hence, Eq. (a) is the equation of the perpendicular bisector of the line segment joining -2 and 1; that is, Eq. (a) describes the line $x = -\frac{1}{2}$.
A more routine method for solving Eq. (a) is to set $z = x + iy$ in the equation and perform the algebra:

$$\begin{aligned} |z+2| &= |z-1|, \\ |x+iy+2| &= |x+iy-1|, \\ (x+2)^2 + y^2 &= (x-1)^2 + y^2, \\ 4x+4 &= -2x+1, \\ x &= -\frac{1}{2}. \end{aligned}$$

(b) The geometric interpretation of Eq. (b) is less obvious, so we proceed directly with the mechanical approach and derive

$$\sqrt{(x-1)^2 + y^2} = x+1 \iff y^2 = 4x$$

which is a parabola.

Another important concept is **complex conjugate**.

Geometrically, the conjugate of a complex number is its reflection on the x-axis. The complex conjugate of the number $z = a + bi$ is denoted by $\bar{z} = a - bi$. It follows that $z = \bar{z}$ if and only if z is a real number. Also, it is clear that the conjugate of the sum (difference) of two complex numbers is equal to the sum (difference) of their conjugates; that is,

$$z_1 + z_2 = \bar{z}_1 + \bar{z}_2$$

$$z_1 - z_2 = \bar{z}_1 - \bar{z}_2$$

We also have:

$$\overline{(z_1 z_2)} = \bar{z}_1 \bar{z}_2$$

and

$$\overline{\left(\frac{z_1}{z_2}\right)} = \frac{\bar{z}_1}{\bar{z}_2} \quad (z_2 \neq 0);$$

$$\operatorname{Re} z = \frac{z + \bar{z}}{2};$$

$$\operatorname{Im} z = \frac{z - \bar{z}}{2i};$$

The proof is not very complex, and is therefore exercises for this section.

Figure 5.2: Complex Conjugate

5.2.1 Exercises

Exercise 5.9 Describe the set of points z in the complex plane that satisfies each of the following.

(a) $\operatorname{Im} z = -2$
(b) $|z-1+i| = 3$
(c) $|2z-i| = 4$
(d) $|z-1| = |z+i|$
(e) $|z| = \operatorname{Re} z + 2$
(f) $|z-1| + |z+1| = 7$
(g) $|z| = 3|z-1|$
(h) $\operatorname{Re} z \geq 4$
(i) $|z-i| < 2$
(j) $|z| > 6$

Proof. Here we describe the set of points z in the complex plane that satisfies each of the following conditions:

(a) For $\operatorname{Im} z = -2$, the Cartesian equation is $y = -2$.
(b) For $|z-1+i| = 3$, the Cartesian equation is $(x-1)^2 + (y+1)^2 = 9$.
(c) For $|2z-i| = 4$, the Cartesian equation is $(2x)^2 + (2y-1)^2 = 16$.
(d) For $|z-1| = |z+i|$, the Cartesian equation is $(x-1)^2 + y^2 = x^2 + (y+1)^2$.
(e) For $|z| = \operatorname{Re} z + 2$, the Cartesian equation is $x^2 + y^2 = x + 2$.
(f) For $|z-1| + |z+1| = 7$, this is the equation of an ellipse with foci at (1,0) and (-1,0).
(g) For $|z| = 3|z-1|$, the Cartesian equation is $x^2 + y^2 = 3(\sqrt{(x-1)^2 + y^2})$.
(h) For $\operatorname{Re} z \geq 4$, the Cartesian equation is $x \geq 4$.
(i) For $|z-i| < 2$, the Cartesian equation is $(x)^2 + (y-1)^2 < 4$.
(j) For $|z| > 6$, the Cartesian equation is $x^2 + y^2 > 36$.

■

Exercise 5.10 Prove that if $\bar{z}^2 = z^2$, then z is either real or pure imaginary.

Proof. Let $z = a + bi$ where a and b are real numbers and i is the imaginary unit. Then $\bar{z} = a - bi$.

According to the premise:

$$\bar{z}^2 = (a-bi)^2 = a^2 - 2abi + b^2 i^2 = a^2 - 2abi - b^2$$

$$z^2 = (a+bi)^2 = a^2 + 2abi + b^2 i^2 = a^2 + 2abi - b^2$$

For $\bar{z}^2$ to equal z^2, the real parts and the imaginary parts of $\bar{z}^2$ and z^2 must be equal, thus:

$$a^2 - b^2 = a^2 - b^2$$

$$-2ab = 2ab$$

The real parts are already equal. For the imaginary parts to be equal, $-2ab$ must equal $2ab$, which is only possible if $ab = 0$. This means that either $a = 0$ or $b = 0$ (or both).

If $a = 0$, then z is pure imaginary. If $b = 0$, then z is real. Hence, if $\bar{z}^2 = z^2$, z must be either real or pure imaginary. ■

Exercise 5.11 Show that:

- $|z_1 z_2| = |z_1||z_2|$ (the modulus of a product is the product of the moduli)
- $\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|}$ (the modulus of a quotient is the quotient of the moduli)
- $|z_1| + |z_2| \leq |z_1 + z_2|$ (triangle inequality)

■

Exercise 5.12 Show that:

- $z_1 + z_2 = \overline{z_1} + \overline{z_2}$
- $z_1 z_2 = \overline{z_1}\,\overline{z_2}$
- $\bar{z}z = |z|^2$
- $z + \bar{z} = 2Re(z)$
- $kz = \overline{kz}$, for $k \in \mathbb{R}$

■

Exercise 5.13 Prove that $\bar{z}^k = (\overline{z^k})$ for every integer k (provided $z \neq 0$ when k is negative). ■

Proof. Consider a complex number $z = a + bi$, where a and b are real numbers and i is the imaginary unit. The complex conjugate of z is $\bar{z} = a - bi$.

The statement is trivially true for $k = 0$ and $k = 1$. Now let k be any positive integer. We proceed by induction on k:

Base case ($k = 1$):

$$\bar{z}^1 = \bar{z} = a - bi = \overline{z^1}$$

The base case holds.

Inductive step: Assume the statement is true for k, i.e., $\bar{z}^k = \overline{z^k}$. We need to show that $\bar{z}^{k+1} = \overline{z^{k+1}}$.

$$\bar{z}^{k+1} = \bar{z}^k \cdot \bar{z} = \overline{z^k} \cdot \bar{z}$$

Using the property of complex conjugates that $\overline{zw} = \bar{z} \cdot \bar{w}$, we have:

$$\overline{z^k} \cdot \bar{z} = \overline{z^k \cdot z} = \overline{z^{k+1}}$$

This completes the inductive step.

For negative integers k, we have $\bar{z}^k = \overline{z^{-k}} = (\overline{z^{-1}})^k = (\bar{z})^k$, since the complex conjugate of a reciprocal is the reciprocal of the complex conjugate, and the induction hypothesis applies.

Therefore, $\bar{z}^k = (\overline{z^k})$ for every integer k. ■

Exercise 5.14 Let $a_1, a_2, \ldots, a_n$ be real constants. Show that if z_0 is a root of the polynomial equation $z^n + a_1 z^{n-1} + a_2 z^{n-2} + \ldots + a_n = 0$, then so is $\overline{z_0}$. ■

Proof. Suppose z_0 is a root of the polynomial $P(z) = z^n + a_1 z^{n-1} + a_2 z^{n-2} + \ldots + a_n$. This means that:

$$P(z_0) = z_0^n + a_1 z_0^{n-1} + a_2 z_0^{n-2} + \ldots + a_n = 0$$

Taking the complex conjugate of the entire equation, we have:

$$\overline{P(z_0)} = \overline{z_0^n + a_1 z_0^{n-1} + a_2 z_0^{n-2} + \ldots + a_n} = \overline{0}$$

Since the a_i are real, $\overline{a_i} = a_i$, and using the property that $\overline{z+w} = \overline{z} + \overline{w}$ and $\overline{zw} = \overline{z} \cdot \overline{w}$, we get:

$$\overline{P(z_0)} = \overline{z_0}^n + a_1 \overline{z_0}^{n-1} + a_2 \overline{z_0}^{n-2} + \ldots + a_n$$

Since $\overline{0} = 0$, the equation simplifies to:

$$P(\overline{z_0}) = \overline{z_0}^n + a_1 \overline{z_0}^{n-1} + a_2 \overline{z_0}^{n-2} + \ldots + a_n = 0$$

Therefore, $\overline{z_0}$ is also a root of the polynomial $P(z)$. ■

R This statement is actually **Conjugate Root Theorem**, which we will discuss further details about in finding complex roots of equations.

Exercise 5.15 We have noted that the conjugate $\overline{z}$ is the reflection of the point z in the real axis (the horizontal line $y = 0$). Show that the reflection of z in the line $ax + by = c$ (where a, b, c are real) is given by

$$\frac{2ic + (b - ai)\overline{z}}{b + ai}$$

■

Hint: In general, the reflection of the point (x_1, y_1) in the line $ax + by + c = 0$ given by this formula:

$$x_2 = \frac{x_1(b^2 - a^2) - 2aby_1 - 2ac}{a^2 + b^2}$$

$$y_2 = \frac{y_1(a^2 - b^2) - 2abx_1 - 2bc}{a^2 + b^2}$$

Proof. Now the reflection of $z = (x, y)$ in the line $ax + by - c = 0$ is $w = u + iv$, $u, v \in \mathbb{R}$, where:

$$u = \frac{x(b^2 - a^2) - 2aby + 2ac}{a^2 + b^2},$$
$$v = \frac{y(a^2 - b^2) - 2abx + 2bc}{a^2 + b^2}.$$

Then:

$$w = u + vi = \frac{x(b^2 - a^2) - 2aby + 2ac}{a^2 + b^2} + \frac{y(a^2 - b^2) - 2abx + 2bc}{a^2 + b^2} i.$$

Next, we'll prove now:

$$w = \frac{2ic + (b - ai)\overline{z}}{b + ai}.$$

We'll factorize and simplify w in $(*)$:

$$\begin{aligned} w &= \frac{x(b^2-a^2)-2aby+2ac}{a^2+b^2} + \frac{y(a^2-b^2)-2abx+2bc}{a^2+b^2}i \\ &= \frac{x(b^2-a^2)-2aby+2ac+y(a^2-b^2)i-2abxi+2bci}{a^2+b^2}. \end{aligned}$$

Simplify more:

$$\begin{aligned} (b^2-a^2-2abi) &= (b-ai)^2, \\ (a^2i-b^2i-2ab) &= (-i)(b^2-a^2+2ab) \\ &= (-i)(b+ai)^2 \quad \left(\text{since } \frac{1}{i} = -i\right), \\ &= (-i)(b-ai)(b+ai)^2 \\ &= i(b-ai) \\ &= i(b-ai)(b+ai). \end{aligned}$$

Now we'll put all of them in the (1),

$$\begin{aligned} w &= \frac{x(b-ai)^2-yi(b-ai)^2+2ci(b-ai)}{b+ai} \\ &= \frac{(b-ai)x(b-ai)-yi(b-ai)+2ci}{b+ai} \\ &= \frac{2ci+(b-ai)\bar{z}}{b+ai}. \end{aligned}$$

Therefore, the reflection of z in the line $ax+by=c$ is:

$$w = \frac{2ic+(b-ai)\bar{z}}{b+ai}.$$

■

5.3 Vector and Polar Form

5.3.1 Vector Form of Complex Number

Now that we can express complex numbers as points scattered in the complex plane, we can take them as direction vectors in the plane as shown in figure **??**.

Figure 5.3: Complex Number as Vector

With this, we can apply all possible operations of vectors to complex numbers, such as the **parallelogram law** as in figure **??**.

Figure 5.4: Parallelogram Law

Examing this figure, it actually reminds us of a important conclusion mentioned in earlier chapter. If we focus on the lower triangle of the parallelogram, and denote these two complex numbers as z_1 and z_2 respectively. They hold that:

$$|z_1 + z_2| \leq |z_1| + |z_2|$$

This is exactly the geometrical meaning of triangular inequality as in definition **??**. And these are all we need to know about complex number's vector form, as it is not used very common.

5.3.2 Polar Form of Complex Number

This section discusses one of the most commonly used form of complex number. We start with introducing a new coordinate syste.

5.3.2.1 The Polar Coordinate

Polar coordinates provide an alternative to Cartesian coordinates for describing the location of points in a two-dimensional plane. While Cartesian coordinates use a grid of horizontal and vertical lines to specify a point by its horizontal (x) and vertical (y) distances from an origin, polar coordinates specify a point by its distance from a reference point (called the pole, analogous to the origin) and an angle relative to a reference direction.

Figure 5.5: Complex Number in Polar Coordinate

A point's location in polar coordinates is given as (r, θ), where r is the radial distance from the pole, and θ is the angular coordinate, typically measured in radians from the positive x-axis (the reference direction). Polar coordinates are particularly useful in situations where the geometry of a problem has rotational symmetry, making it more natural and simpler to work with angles and radii than with rectangular coordinates.

One of the most common applications of polar coordinates is in the field of trigonometry, complex numbers, and vector calculus, where they provide a more straightforward approach to solving problems involving circular motion, periodic functions, and fields.

5.3.2.2 Polar Expression of Complex Number

Now that we know the existence of polar form, how can we get it from Cartesian form? In the Cartesian coordinate, the positions are expressed, straightforwardly, in the horizontal and vertical distance to y and x-axis, while in polar coordinate, we express them using the angle between the line from the origin to the specific position and the horizontal axis, as well as the length of this line. Is there a mapping or relation between the points in these two types of coordinates? Some clever students may have noticed that trigonometric functions are the keys here. we readily derive the equations expressing the rectangular (or Cartesian) coordinates (x, y) in terms of the polar coordinates (r, θ):

$$x = r\cos\theta, \quad y = r\sin\theta. \tag{5.2}$$

On the other hand, the expressions for (r,θ) in terms of (x,y) contain some minor but troublesome complications. Indeed the coordinate r is given, unambiguously, by

$$r = \sqrt{x^2+y^2} = |z|. \tag{5.3}$$

However, observe that although it is certainly true that $\tan\theta = \frac{y}{x}$, the natural conclusion

$$\theta = \tan^{-1}\left(\frac{y}{x}\right) \tag{5.4}$$

is *invalid* for points z in the second and third quadrants (since the standard interpretation of the arctangent function places its range in the first and fourth quadrants). Since an angle is fixed by its sine *and* cosine, θ is uniquely determined by the pair of equations

$$\cos\theta = \frac{x}{|z|}, \quad \sin\theta = \frac{y}{|z|}, \tag{5.5}$$

And all these lead us to the polar form of a given complex number:

$$z = x+iy = r(\cos\theta + i\sin\theta) = r\operatorname{cis}\theta \tag{5.6}$$

Considering the circularity of radiant, we introduce the **argument** of complex number.

Definition 5.10 — Arguement of Complex Number. In the study of complex numbers, the *argument* of a complex number z, denoted as $\arg(z)$, is a fundamental concept representing the angle between the positive real axis and the line segment that joins the origin with the point z in the complex plane. Specifically, if $z = x+iy$, where x and y are real numbers, then $\arg(z)$ is defined as the angle θ in polar coordinates that satisfies $x = r\cos(\theta)$ and $y = r\sin(\theta)$, where r is the magnitude of z. The value of $\arg(z)$ is usually given in radians and, by convention, is restricted to the interval $(-\pi,\pi]$, known as the principal value. The argument provides a complete description of the direction in which the point z lies from the origin, serving as a crucial tool in the fields of complex analysis, phasor calculus in electrical engineering, and in the representation of waves and oscillations.

■ **Example 5.5** Find $\arg(1+\sqrt{3}i)$ and write $1+\sqrt{3}i$ in polar form. ■

Solutions:

Note that $r = |1+\sqrt{3}i| = 2$ and that the equations $\cos\theta = \frac{1}{2}$, $\sin\theta = \frac{\sqrt{3}}{2}$ are satisfied by $\theta = \frac{\pi}{3}$. Hence $\arg(1+\sqrt{3}i) = \frac{\pi}{3} + 2k\pi, k = 0,\pm1,\pm2,\ldots$ [in particular, $\operatorname{Arg}(1+\sqrt{3}i) = \frac{\pi}{3}$]. The polar form of $1+\sqrt{3}i$ is

$$2(\cos\frac{\pi}{3} + i\sin\frac{\pi}{3}) = 2\operatorname{cis}\frac{\pi}{3}.$$

More properties of polar form complex number could be derived with properties of trigonometric identities. Now suppose we have

$$z_1 = r_1(\cos\theta_1 + i\sin\theta_1), \quad z_2 = r_2(\cos\theta_2 + i\sin\theta_2)$$

then we have

$$z_1z_2 = r_1r_2[(\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + i(\sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2)]$$

which follows that

$$z_1 z_2 = r_1 r_2 \left[\cos\left(\theta_1 + \theta_2\right) + i \sin\left(\theta_1 + \theta_2\right)\right] \tag{5.7}$$

R The compound angle formula is applied in the proof, which should already been covered in high school syllabus. You may refer to this link for further information. A concise proof is provided below.

Proof of the Composite Angle Formulas. To prove the sine of the sum of two angles α and β, we can use the unit circle and the definition of sine and cosine.

Consider a unit circle where point A corresponds to angle α, and point B corresponds to angle $\alpha+\beta$. Point A has coordinates $(\cos(\alpha), \sin(\alpha))$, and point B has coordinates $(\cos(\alpha+\beta), \sin(\alpha+\beta))$.

By rotating point A by angle β, we can form a right triangle where the new point, C, has coordinates $(\cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta), \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta))$.

The coordinates of point C represent the cosine and sine of the sum of angles α and β due to the rotation. Therefore, we have:

$$\sin(\alpha+\beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$$

$$\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$$

Similarly, by considering the rotation in the opposite direction (subtracting angle β from α), we can derive the formulas for the sine and cosine of the difference of two angles:

$$\sin(\alpha-\beta) = \sin(\alpha)\cos(\beta) - \cos(\alpha)\sin(\beta)$$

$$\cos(\alpha-\beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)$$

Thus, we have proven the composite angle formulas for sine and cosine. ■

The abbreviated version of Eq. **??** reads as follows:

$$\overline{z_1 z_2} = (r_1 \operatorname{cis} \theta_1)(r_2 \operatorname{cis} \theta_2) = (r_1 r_2) \operatorname{cis}(\theta_1 + \theta_2)$$

and we see that

The modulus of the product is the product of the moduli:

$$|\overline{z_1 z_2}| = |z_1||z_2| \quad (= r_1 r_2);$$

The argument of the product is the sum of the arguments:

$$\arg \overline{z_1 z_2} = \arg z_1 + \arg z_2 \quad (= \theta_1 + \theta_2).$$

As division is the inverse of multiplication, we can get the following statements by similar method:

$$\frac{z_1}{z_2} = \frac{r_1}{r_2}\left[\cos\left(\theta_1 - \theta_2\right) + i \sin\left(\theta_1 - \theta_2\right)\right] = \frac{r_1}{r_2} \operatorname{cis}\left(\theta_1 - \theta_2\right) \tag{5.8}$$

$$\arg\left(\frac{z_1}{z_2}\right) = \arg z_1 - \arg z_2 \tag{5.9}$$

$$\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|} \tag{5.10}$$

■ **Example 5.6** Write the quotient $\frac{(1+i)}{(\sqrt{3}-i)}$ in polar form. ■

Proof. The polar forms for $1+i$ and $\sqrt{3}-i$ are

$$1+i = |1+i|\operatorname{cis}(\arg(1+i)) = \sqrt{2}\operatorname{cis}\left(\frac{\pi}{4}\right),$$

$$\sqrt{3}-i = 2\operatorname{cis}\left(-\frac{\pi}{6}\right).$$

Hence, from Eq. **??**, we have

$$\frac{1+i}{\sqrt{3}-i} = \frac{\sqrt{2}}{2}\operatorname{cis}\left(\frac{\pi}{4}-\left(-\frac{\pi}{6}\right)\right) = \frac{\sqrt{2}}{2}\operatorname{cis}\left(\frac{5\pi}{12}\right).$$

■

5.3.3 Exercises

Exercise 5.16 Find the following:

(a) $\left|\frac{1+2i}{-2-i}\right|$
(b) $|(1+i)(2-3i)(4i-3)|$
(c) $\left|\frac{i(2+i)^3}{(1-i)^2}\right|$
(d) $\left|\frac{(\pi+i)^{100}}{(\pi-i)^{100}}\right|$

■

Exercise 5.17 Draw the following vectors:

(a) $7\operatorname{cis}\left(\frac{3\pi}{4}\right)$
(b) $4\operatorname{cis}\left(-\frac{\pi}{6}\right)$
(c) $\operatorname{cis}\left(\frac{3\pi}{4}\right)$
(d) $3\operatorname{cis}\left(\frac{27\pi}{4}\right)$

■

Exercise 5.18 Find the argument of each of the following complex numbers and write each in polar form:

(a) $-\frac{1}{2}$
(b) $-3+3i$
(c) $-\pi i$
(d) $-2\sqrt{3}-2i$
(e) $(1-i)(-\sqrt{3}+i)$
(f) $(\sqrt{3}-i)^2$
(g) $\frac{-1+\sqrt{3}i}{2+2i}$
(h) $\frac{-\sqrt{7}(1+i)}{\sqrt{3}+i}$

■

Exercise 5.19 Show geometrically that the nonzero complex numbers z_1 and z_2 satisfy $z_1+z_2 = |z_1|+|z_2|$ if and only if they have the same argument. ■

Proof. If z_1 and z_2 have the same argument, it follows that:

We see here if $z_1 = r_1(\cos\theta + i\sin\theta)$, $z_2 = r_2(\cos\theta + i\sin\theta)$, then

$$\begin{aligned}|z_1 + z_2| &= |r_1(\cos\theta + i\sin\theta) + r_2(\cos\theta + i\sin\theta)| \\ &= |(r_1 + r_2)(\cos\theta + i\sin\theta)| \\ &= r_1 + r_2 = |z_1| + |z_2|\end{aligned}$$

■

Exercise 5.20 Given the vector z, interpret geometrically the vector $(\cos\phi + i\sin\phi)z$.

■

Solution:

Let $z \in \mathbb{C}$, $z = r(\cos\theta + i\sin\theta)$

$$(\cos\phi + i\sin\phi)z = (\cos\phi + i\sin\phi)r(\cos\theta + i\sin\theta)$$

$$= r[\cos(\theta + \phi) + i\sin(\theta + \phi)]$$

$$= r\,\mathrm{cis}(\theta + \phi)$$

Geometrically this: $r\,\mathrm{cis}(\theta + \phi)$ means a vector that its length r and argument is $\theta + \phi$, we can get it by rotating vector z about origin through an angle ϕ in the counterclockwise direction.

Exercise 5.21 Show that $|z_1 z_2 z_3| = |z_1||z_2||z_3|$ and thus prove

$$|\prod_{i=1}^{n} z_i| = \prod_{i=1}^{n} |z_i|$$

R $\prod$ is called pi (upper case) notation. Take it as the sigma notation for multiplication.

■

Proof. Consider the complex numbers $z_1 = r_1(\cos\theta_1 + i\sin\theta_1)$, $z_2 = r_2(\cos\theta_2 + i\sin\theta_2)$, and $z_3 = r_3(\cos\theta_3 + i\sin\theta_3)$, where r_1, r_2, and r_3 are the moduli of z_1, z_2, and z_3 respectively, and θ_1, θ_2, and θ_3 are the arguments. The product $z_1 z_2 z_3$ is then given by:

$$z_1 z_2 z_3 = r_1 r_2 r_3(\cos(\theta_1 + \theta_2 + \theta_3) + i\sin(\theta_1 + \theta_2 + \theta_3))$$

The modulus of this product is:

$$|z_1 z_2 z_3| = |r_1 r_2 r_3(\cos(\theta_1 + \theta_2 + \theta_3) + i\sin(\theta_1 + \theta_2 + \theta_3))| = r_1 r_2 r_3$$

Since the moduli of z_1, z_2, and z_3 are r_1, r_2, and r_3 respectively, it follows that:

$$|z_1 z_2 z_3| = |z_1||z_2||z_3|$$

Now, extend this property to a product of n complex numbers:

$$\prod_{i=1}^{n} z_i = \prod_{i=1}^{n} r_i(\cos\theta_i + i\sin\theta_i)$$

Taking the modulus of both sides:

$$\left|\prod_{i=1}^{n} z_i\right| = \left|\prod_{i=1}^{n} r_i(\cos\theta_i + i\sin\theta_i)\right| = \prod_{i=1}^{n} r_i$$

Thus:

$$\prod_{i=1}^{n} |\overline{z_i}| = \prod_{i=1}^{n} |z_i|$$

as required.

MI is also a option:

Base case ($n = 1$): For a single complex number z_1, the statement is trivially true since $|\overline{z_1}| = |z_1|$.

Inductive step: Assume the statement holds for $n = k$, i.e.,

$$\prod_{i=1}^{k} |\overline{z_i}| = \prod_{i=1}^{k} |z_i|$$

Now consider $n = k+1$ complex numbers. By the induction hypothesis and the property that the modulus of a product is equal to the product of the moduli for any two complex numbers a and b, namely $|ab| = |a||b|$, we have:

$$\prod_{i=1}^{k+1} |\overline{z_i}| = |\overline{z_{k+1}}| \prod_{i=1}^{k} |\overline{z_i}| = |z_{k+1}| \prod_{i=1}^{k} |z_i| = \prod_{i=1}^{k+1} |z_i|$$

Thus, by mathematical induction, the statement is proved for all $n \in \mathbb{N}$. ■

Exercise 5.22 Show that

(a) $\arg z_1 z_2 z_3 = \arg z_1 + \arg z_2 + \arg z_3$

(b) $\arg z_1 \overline{z_2} = \arg z_1 - \arg z_2$.

Solution: Let $z_1, z_2, z_3 \in \mathbb{C}$,

(a) we'll prove $\arg(z_1 z_2 z_3) = \arg z_1 + \arg z_2 + \arg z_3$

Let $z_1 = r_1(\cos\theta_1 + i\sin\theta_1)$, where $\arg z_1 = \theta_1$

$z_2 = r_2(\cos\theta_2 + i\sin\theta_2)$, where $\arg z_2 = \theta_2$

$z_3 = r_3(\cos\theta_3 + i\sin\theta_3)$, where $\arg z_3 = \theta_3$

We'll prove $\arg z_1 z_2 z_3 = \arg z_1 + \arg z_2 + \arg z_3$

At first we'll find $z_1 z_2 z_3$:

$$\begin{aligned}
z_1 z_2 z_3 &= (z_1 z_2) z_3 = r_1 r_2 [\cos(\theta_1+\theta_2) + i\sin(\theta_1+\theta_2)] z_3 \\
&= r_1 r_2 r_3 [\cos(\theta_1+\theta_2) + i\sin(\theta_1+\theta_2)](\cos\theta_3 + i\sin\theta_3) \\
&= r_1 r_2 r_3 [\cos(\theta_1+\theta_2)\cos\theta_3 - \sin(\theta_1+\theta_2)\sin\theta_3 \\
&\quad + i(\sin(\theta_1+\theta_2)\cos\theta_3 + \cos(\theta_1+\theta_2)\sin\theta_3)] \\
&= r_1 r_2 r_3 [\cos(\theta_1+\theta_2+\theta_3) + i\sin(\theta_1+\theta_2+\theta_3)]
\end{aligned}$$

Therefore, $\arg z_1 z_2 z_3 = \theta_1 + \theta_2 + \theta_3 = \arg z_1 + \arg z_2 + \arg z_3$.

(b) we'll prove $\arg z_1\overline{z_2} = \arg z_1 - \arg z_2$
Therefore, $z_2 = r_2(\cos\theta_2 + i\sin\theta_2)$ and hence

$$\overline{z_2} = r_2(\cos(-\theta_2) + i\sin(-\theta_2)) = r_2(\cos\theta_2 - i\sin\theta_2)$$

Now we find the argument of $z_1\overline{z_2}$:

$$\begin{aligned}
\arg z_1\overline{z_2} &= \arg\,[(r_1(\cos\theta_1 + i\sin\theta_1))(r_2(\cos\theta_2 - i\sin\theta_2))] \\
&= \arg\,[r_1r_2((\cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2) \\
&\quad + i(\cos\theta_1\sin\theta_2 - \cos\theta_2\sin\theta_1))] \\
&= \arg\,[r_1r_2(\cos(\theta_1 - \theta_2) + i\sin(\theta_1 - \theta_2))] \\
&= \theta_1 - \theta_2 \\
&= \arg z_1 - \arg z_2
\end{aligned}$$

Thus, $\arg z_1\overline{z_2} = \arg z_1 - \arg z_2$.

Exercise 5.23 Recall that the dot (scalar) product of two planar vectors $v_1 = (x_1, y_1)$ and $v_2 = (x_2, y_2)$ is given by:

$$v_1 \cdot v_2 = x_1x_2 + y_1y_2.$$

Show that the dot product of the vectors represented by the complex numbers z_1 and z_2 is given by

$$z_1 \cdot z_2 = Re(\overline{z_1}z_2).$$

■

Proof. Let $z_1, z_2 \in \mathbb{C}$.
We can represent z_1, z_2 as vectors as follows:

$$z_1 = (x_1, y_1) = x_1 + y_1 j, \quad z_2 = (x_2, y_2) = x_2 + y_2 j$$

$$\begin{aligned}
&z_1 \cdot z_2 = (x_1 + y_1 j) \cdot (x_2 + y_2 j) \\
&= x_1x_2 + y_1y_2 + x_1y_2(j \cdot j) + x_2y_1(i \cdot j) \\
&= x_1x_2 + y_1y_2 + 0 + 0 \quad (\because i^2 = j^2 = -1, i \cdot j = 0) \\
&\therefore z_1 \cdot z_2 = x_1x_2 + y_1y_2 \quad (1)
\end{aligned}$$

Now we'll find:

$$\overline{z_1}z_2 = (x_1 - y_1 j)(x_2 + y_2 j) = x_1x_2 + y_1y_2 + x_1y_2(-j) - x_2y_1 j$$

$$\therefore Re(\overline{z_1}z_2) = x_1x_2 + y_1y_2 \quad (2)$$

From (1) and (2) we get:

$$z_1 \cdot z_2 = Re(\overline{z_1}z_2) = x_1x_2 + y_1y_2$$

■

Exercise 5.24 We have proven the **Generalized Triangle Inequality** in last chapter that:

$$\left|\sum_{k=1}^{n} z_k\right| \leq \sum_{k=1}^{n} |z_k|$$

For complex numberz_1, z_2, and z_3, prove that:

$$\left|\frac{m_1z_1+m_2z_2+m_3z_3}{m_1+m_2+m_3}\right| \leq 1$$

Proof. We'll take at first:

$$\left|\frac{m_1z_1+m_2z_2+m_3z_3}{m_1+m_2+m_3}\right| = \frac{|m_1z_1+m_2z_2+m_3z_3|}{|m_1+m_2+m_3|}$$

Now we'll find:

$$|m_1z_1+m_2z_2+m_3z_3| \leq |m_1z_1|+|m_2z_2|+|m_3z_3| \quad \text{(Triangle inequality)}$$

$$= m_1|z_1|+m_2|z_2|+m_3|z_3|$$

$$\leq m_1+m_2+m_3 \quad (\text{since } m_1,m_2,m_3>0 \text{ and } |z_1|,|z_2|,|z_3| \leq 1)$$

$$= |m_1+m_2+m_3|$$

Therefore:

$$|m_1z_1+m_2z_2+m_3z_3| \leq |m_1+m_2+m_3|$$

$$\therefore \left|\frac{m_1z_1+m_2z_2+m_3z_3}{m_1+m_2+m_3}\right| = \frac{|m_1z_1+m_2z_2+m_3z_3|}{|m_1+m_2+m_3|} \leq 1$$

■

5.4 Exponential Form

The other form for complex number is its exponential form with base e. Euler's formula relate the exponential expression to the polar expression. However, the proof of Euler's formula require further knowledge of calculus, such as Taylor series. You may access this link for a proof without calculus. You can also just skip the proof for now, since it does not affect the problem-solving in this chapter.

Definition 5.11 — Euler's Formula.

$$e^{iy} = \cos y + i\sin y \tag{5.11}$$

You can relate this formula to the polar form very easily, since the right-hand side is equivalent to$cisy$, which means y is exactly the principal argument of the complex number, θ.

Euler's Formula enables us to write the polar form of a complex number as

$$z = r\text{cis}\theta = r(\cos\theta + i\sin\theta) = re^{i\theta}.$$

Thus, we can (and do) drop the awkward "cis" artifice and use, as the standard polar representation,

$$z = re^{i\theta} = |z|e^{\arg z}.$$

In particular, notice the following identities:

$$e^{i0} = e^{2\pi i} = e^{-2\pi i} = e^{4\pi i} = e^{-4\pi i} = \cdots = 1,$$

$$e^{(\pi/2)i} = i, \quad e^{(-\pi/2)i} = -i, \quad e^{\pi i} = -1.$$

Observe also that $\left|e^{i\arg z}\right| = 1$ and that Euler's equation leads to the following representations of the customary trigonometric functions:

$$\begin{aligned} \cos\theta &= \operatorname{Re} e^{i\theta} = \tfrac{e^{i\theta}+e^{-i\theta}}{2}, \\ \sin\theta &= \operatorname{Im} e^{i\theta} = \tfrac{e^{i\theta}-e^{-i\theta}}{2i}. \end{aligned}$$

Proof. We start with Euler's formula which states that for any real number θ,

$$e^{i\theta} = \cos(\theta) + i\sin(\theta). \tag{5.12}$$

The complex conjugate of $e^{i\theta}$ is $e^{-i\theta}$, which gives us

$$e^{-i\theta} = \cos(-\theta) + i\sin(-\theta) = \cos(\theta) - i\sin(\theta), \tag{5.13}$$

since cosine is an even function, $\cos(-\theta) = \cos(\theta)$, and sine is an odd function, $\sin(-\theta) = -\sin(\theta)$.

To derive the expression for the cosine function, we take the sum of $e^{i\theta}$ and $e^{-i\theta}$, and divide by 2:

$$\cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2}. \tag{5.14}$$

This is because the imaginary parts $i\sin(\theta)$ and $-i\sin(\theta)$ cancel each other out, leaving the sum of $\cos(\theta) + \cos(\theta)$, which is then divided by 2.

To derive the expression for the sine function, we take the difference of $e^{i\theta}$ and $e^{-i\theta}$, and divide by $2i$:

$$\sin(\theta) = \frac{e^{i\theta} - e^{-i\theta}}{2i}. \tag{5.15}$$

Here, the real parts $\cos(\theta) - \cos(\theta)$ cancel out, leaving the difference of $i\sin(\theta) - (-i\sin(\theta))$ which is $2i\sin(\theta)$, and then dividing by $2i$ yields $\sin(\theta)$.

This completes the derivation of the exponential forms of the sine and cosine functions. ■

The rules derived in definition **??** for multiplying and dividing complex numbers in polar form now find very natural expressions:

$$z_1 z_2 = (r_1 e^{i\theta_1})(r_2 e^{i\theta_2}) = (r_1 r_2) e^{i(\theta_1+\theta_2)}, \tag{5.16}$$

$$\frac{z_1}{z_2} = \frac{r_1 e^{i\theta_1}}{r_2 e^{i\theta_2}} = \left(\frac{r_1}{r_2}\right) e^{i(\theta_1-\theta_2)}, \tag{5.17}$$

and complex conjugation of $z = re^{i\theta}$ is accomplished by changing the sign of i in the exponent:

$$\bar{z} = re^{-i\theta}. \tag{5.18}$$

■ **Example 5.7** Compute (a) $\frac{1+i}{\sqrt{3}-i}$ and (b) $(1+i)^{24}$. ■

Solution:

(a) This quotient was evaluated using the cis operator in Example 1.11 of Sec. 1.3; using the exponential the calculations take the form

$$1+i = \sqrt{2}\text{cis}\left(\frac{\pi}{4}\right) = \sqrt{2}e^{i\pi/4},$$

$$\sqrt{3}-i = 2\text{cis}\left(-\frac{\pi}{6}\right) = 2e^{-i\pi/6},$$

and, therefore,

$$\frac{1+i}{\sqrt{3}-i} = \frac{\sqrt{2}e^{i\pi/4}}{2e^{-i\pi/6}} = \frac{\sqrt{2}}{2}e^{i5\pi/12}.$$

(b) The exponential forms become

$$(1+i)^{24} = \left(\sqrt{2}e^{i\pi/4}\right)^{24} = \left(\sqrt{2}^{24}\right)e^{i24\pi/4} = 2^{12}e^{i6\pi} = 2^{12}.$$

5.4.1 De Moivre's Theorem

De Moivre's Theorem is a fundamental result in complex analysis that connects complex numbers and trigonometry. Named after the French mathematician Abraham de Moivre, the theorem provides a formula for raising complex numbers to any power using polar coordinates.

Given a complex number expressed in polar form as $z = r(\cos\theta + i\sin\theta)$, where r is the modulus and θ is the argument of the complex number, De Moivre's Theorem states that:

Theorem 5.1 — De Moivre's Theorem.

$$z^n = r^n(\cos n\theta + i\sin n\theta) \tag{5.19}$$

or in exponential form:

$$\left(e^{i\theta}\right)^n = \underbrace{e^{i\theta}e^{i\theta}\cdots e^{i\theta}}_{(n \text{ times })} = e^{i\theta+i\theta+\cdots+i\theta} = e^{in\theta} \tag{5.20}$$

for any integer n. This elegant relationship not only simplifies the computation of powers of complex numbers but also lays the foundation for finding roots of complex numbers.

The theorem is particularly useful because it transforms a potentially difficult multiplication problem into a much simpler form by taking advantage of the properties of exponential functions and Euler's formula, which expresses complex exponentiation in terms of sine and cosine:

$$e^{i\theta} = \cos\theta + i\sin\theta \tag{5.21}$$

Through this lens, De Moivre's Theorem is often used to derive results in trigonometry, such as trigonometric identities for sine and cosine of multiple angles, and it plays a crucial role in the field of complex analysis. This theorem could be proven by simple MI.

Proof. We proceed by induction on n.

Base case: For $n = 1$, the statement holds trivially:

$$(\cos(\theta) + i\sin(\theta))^1 = \cos(\theta) + i\sin(\theta).$$

Inductive step: Assume that the theorem holds for some integer k, that is

$$(\cos(\theta) + i\sin(\theta))^k = \cos(k\theta) + i\sin(k\theta).$$

Now consider the case for $k+1$:

$$\begin{aligned}
(\cos(\theta) + i\sin(\theta))^{k+1} &= (\cos(\theta) + i\sin(\theta))^k(\cos(\theta) + i\sin(\theta)) \\
&= (\cos(k\theta) + i\sin(k\theta))(\cos(\theta) + i\sin(\theta)) \quad \text{(by inductive hypothesis)} \\
&= \cos(k\theta)\cos(\theta) - \sin(k\theta)\sin(\theta) \\
&\quad + i(\sin(k\theta)\cos(\theta) + \cos(k\theta)\sin(\theta)) \\
&= \cos((k+1)\theta) + i\sin((k+1)\theta) \quad \text{(using angle addition formulas).}
\end{aligned}$$

Thus, the theorem holds for $k+1$.

By induction, the theorem is true for all integers n. ■

5.4.2 Exercises

Exercise 5.25 Write each of the given numbers in Cartesian form.

(a) $e^{-i\pi/4}$
(b) $\frac{e^{1+i3\pi}}{e^{-1+i\pi/2}}$
(c) e^{ei}
(d) $\frac{e^{3i}-e^{-3i}}{2i}$
(e) $2e^{3+i\pi/6}$
(f) e^z, where $z = 4e^{i\pi/3}$

■

Exercise 5.26 Write each of the given numbers in exponential form.

(a) $\frac{1-i}{3}$
(b) $-8\left(1+\sqrt{3}i\right)$
(c) $(1+i)^6$
(a) $\cos\left(\frac{2\pi}{9}\right) + i\sin\left(\frac{2\pi}{9}\right)$
(b) $\frac{2+2i}{-\sqrt{3}+i}$

(c) $\frac{2i}{3e^i}$

Exercise 5.27 Consider a complex number sequence $\{x_n\}$, where $x_n = (1+i)^n$ and n is a non-negative integer. Calculate the sum of the first N terms of the sequence, $S_N = \sum_{n=0}^{N-1} x_n$.

Hint: Utilize the properties of powers of complex numbers and summation formulas to solve this problem. Consider converting $1+i$ into its exponential form to simplify the calculation. **Solution:**

Let $x_n = (1+i)^n$, then the sum of the first N terms of the sequence S_N is:

$$S_N = \sum_{n=0}^{N-1} x_n$$

Using the exponential form of $1+i$, which is $\sqrt{2}e^{i\frac{\pi}{4}}$, we have:

$$S_N = \sum_{n=0}^{N-1} \left(\sqrt{2}e^{i\frac{\pi}{4}}\right)^n$$

The sum of a geometric series with the common ratio r is:

$$S_N = \frac{1-r^N}{1-r}$$

Substituting $r = \sqrt{2}e^{i\frac{\pi}{4}}$ into the formula gives us:

$$S_N = \frac{1-\left(\sqrt{2}e^{i\frac{\pi}{4}}\right)^N}{1-\sqrt{2}e^{i\frac{\pi}{4}}}$$

This can be further simplified depending on the value of N.

Exercise 5.28 Show that for $z = e^{x+iy}$, the modulus $|z|$ is e^x and the argument $\arg(z)$ is $y+2k\pi$ for $k = 0, \pm 1, \pm 2, \ldots$.

Solution: For $z = x+iy$,

$$\begin{aligned} e^z &= e^x(\cos(y) + i\sin(y)) \\ &= e^x\cos(y) + e^x i\sin(y) \end{aligned}$$

Let this be a complex number that is, $z_1 = e^z = e^x\cos(y) + e^x i\sin(y)$. Now, modulus of z_1,

$$
\begin{aligned}
|z_1| &= \sqrt{Re(z_1)^2 + Im(z_1)^2} \\
&= \sqrt{(e^x\cos(y))^2 + (e^x\sin(y))^2} \\
&= \sqrt{e^{2x}(\cos^2(y) + \sin^2(y))} \\
&= \sqrt{e^{2x}} \\
&= e^x
\end{aligned}
$$

Therefore $|z_1| = |e^{x+iy}| = e^x$.
Similarly, argument of z_1,

$$
\begin{aligned}
\arg(z_1) &= \tan^{-1}\left(\frac{Im(z_1)}{Re(z_1)}\right) \\
&= \tan^{-1}\left(\frac{e^x\sin(y)}{e^x\cos(y)}\right) \\
&= \tan^{-1}(\tan(y))
\end{aligned}
$$

Since $\tan(y)$ is a periodic function with period 2π,

$$\arg(z_1) = \arg(e^{x+iy}) = y + 2k\pi, \forall k \in \mathbb{Z}$$

Exercise 5.29 Show that, for all z,
(a) $e^{z+\pi i} = -e^z$
(b) $e^{\bar{z}} = \overline{e^z}$

Solution:
$e^{z+\pi i} = -e^z$
We know for $z = x+iy$, $e^z = e^x(\cos y + i\sin y)$.
Consider the left-hand side for $z = x+iy$,

$$
\begin{aligned}
e^{z+\pi i} &= e^{x+i(y+\pi)} \\
&= e^x(\cos(y+\pi) + i\sin(y+\pi)) \\
&= e^x(-\cos y - i\sin y) \\
&= -e^x(\cos y + i\sin y) \\
&= -e^z
\end{aligned}
$$

$e^{\bar{z}} = \overline{e^z}$ We know, for $z = x+iy$, $e^z = e^x(\cos y + i\sin y)$. Therefore,

$$e^{\bar{z}} = e^x(\cos y - i\sin y) \quad (1)$$

For $\bar{z} = x - iy$,

$$
\begin{aligned}
e^{\bar{z}} &= e^{x-iy} \\
&= e^x(\cos(-y) + i\sin(-y)) \\
&= e^x(\cos y - i\sin y) \quad (2)
\end{aligned}
$$

From equations (1) and (2), we can say that,

$$e^{\bar{z}} = \overline{e^{z}}$$

Exercise 5.30 Show that $e^z = e^{z+2\pi i}$ for all z. (The exponential function is periodic with period $2\pi i$.) ■

textbfSolution:

We have to prove that $e^z = e^{z+2\pi i}$ for all z.

Start with the right-hand side. Consider for $z = x + iy$,

$$\begin{aligned} e^{z+2\pi i} &= e^{x+iy+2\pi i} \\ &= e^{x+i(y+2\pi)} \end{aligned}$$

We know that

$$e^z = e^{x+iy} = e^x(\cos y + i \sin y)$$

Applying this to the above,

$$e^{x+i(y+2\pi)} = e^x(\cos(y+2\pi) + i\sin(y+2\pi))$$

We know that sin and cos are both periodic functions with period being 2π. This means that

$$\begin{aligned} \cos(y+2\pi) &= \cos y \\ \sin(y+2\pi) &= \sin y \end{aligned}$$

Thus,

$$\begin{aligned} e^{z+2\pi i} &= e^x(\cos y + i \sin y) \\ &= e^{x+iy} \\ &= e^z \end{aligned}$$

Therefore, e^z is a periodic function with period $2\pi i$.

5.5 Finding Complex Roots

In this section, we will not be too obsessed with the form of complex number, but focus on the practical reason why we need complex number: finding roots. Because of this, this section could be a little tedious, as it is more like middle school algebra, nothing really new but algebra techniques, yet they are important for further mathematical learning.

5.5.1 Solving Quadratic Equations Over the Complex Numbers

In the middle school, we have been introduced many ways to find Solutions for quadratic equations, like factorization or using the determinant δ.

Let's make q quick recap on the determinant of quadratic equation. A quadratic equation is a second-order polynomial equation in a single variable x with the form $ax^2 + bx + c = 0$, where a, b, and c are constants, and $a \neq 0$. The expression under the square root, $b^2 - 4ac$, is known as the discriminant. The discriminant can tell us about the nature of the roots:

- If the discriminant is positive, there are two distinct real roots.
- If the discriminant is zero, there is one real root (also known as a repeated or double root).
- If the discriminant is negative, there are no real roots, but two complex roots.

The solutions to the quadratic equation are known as the roots of the equation, which can be found using the quadratic formula:

Theorem 5.2 — Root(s) of Quadratic Equation.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Proof. Starting with the standard form of the quadratic equation:

$$ax^2 + bx + c = 0$$

Divide through by a (where $a \neq 0$) to normalize the quadratic term:

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0$$

Subtract $\frac{c}{a}$ from both sides:

$$x^2 + \frac{b}{a}x = -\frac{c}{a}$$

To complete the square, add $\left(\frac{b}{2a}\right)^2$ to both sides:

$$x^2 + \frac{b}{a}x + \left(\frac{b}{2a}\right)^2 = \left(\frac{b}{2a}\right)^2 - \frac{c}{a}$$

Write the left side as a square and simplify the right side:

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}$$

Take the square root of both sides:

$$x + \frac{b}{2a} = \pm\sqrt{\frac{b^2 - 4ac}{4a^2}}$$

Solve for x:

$$x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a}$$

Combining the terms gives us the quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

This concludes the introduction and proof of the determinant and roots of quadratic equations. ■

Also, since $i^2 = -1$, we can rewrite a sum of two squares as a difference of two squares:

Corollary 5.1 — Sum of Two Squares.

$$\begin{aligned} z^2 + a^2 &= z^2 - (ai)^2 \\ &= (z+ai)(z-ai) \end{aligned}$$

■ **Example 5.8** Factorize:

1. $z^2 + 16$
2. $2z^2 + 6$

■

Solution:

$$\begin{aligned} z^2 + 16 &= z^2 - 16i^2 \\ &= (z+4i)(z-4i) \end{aligned}$$

$$\begin{aligned} 2z^2 + 6 &= 2(z^2+3) \\ &= 2(z^2 - 3i^2) \\ &= 2(z+\sqrt{3}i)(z-\sqrt{3}i) \end{aligned}$$

For more complex equations, we tend to use the root formula.

■ **Example 5.9** Solve each of the following equations for z:

1. $z^2 + z + 3 = 0$
2. $2z^2 - z + 1 = 0$
3. $z^2 = 2z - 5$

■

Solution:

$$z^2 + z + 3 = \left(z - \left(-\frac{1}{2} - \frac{\sqrt{11}}{2}i\right)\right)\left(z - \left(-\frac{1}{2} + \frac{\sqrt{11}}{2}i\right)\right)$$

Hence $z^2 + z + 3 = 0$ has solutions

$$z = -\frac{1}{2} - \frac{\sqrt{11}}{2}i \quad \text{and} \quad z = -\frac{1}{2} + \frac{\sqrt{11}}{2}i$$

$$2z^2 - z + 1 = 2\left(z - \left(-\frac{1}{4} - \frac{\sqrt{7}}{4}i\right)\right)\left(z - \left(-\frac{1}{4} + \frac{\sqrt{7}}{4}i\right)\right)$$

Hence $2z^2 - z + 1 = 0$ has solutions

$$z = \frac{1}{4} - \frac{\sqrt{7}}{4}i \quad \text{and} \quad z = \frac{1}{4} + \frac{\sqrt{7}}{4}i$$

$$z^2 - 2z + 5 = 0$$

Now apply the quadratic formula:

$$\begin{aligned} z &= \frac{2 \pm \sqrt{-16}}{2} \\ &= \frac{2 \pm 4i}{2} \\ &= 1 \pm 2i \end{aligned}$$

The solutions are $1 + 2i$ and $1 - 2i$.

5.5.2 Solving Polynomial Equations on Complex Number

Now let's try to solve some more tricky and general problems. Quadratic equation is actually a special case of polynomial equations, which is defined as:

Definition 5.12 — Polynomial Equation. a polynomial of degree n is an expression of the form

$$P(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

where the coefficients a_i are complex numbers and $a_n \neq 0$.

When we divide the polynomial $P(z)$ by the polynomial $D(z)$ we obtain two polynomials, $Q(z)$ the quotient and $R(z)$ the remainder, such that

$$P(z) = D(z)Q(z) + R(z)$$

and either $R(z) = 0$ or $R(z)$ has degree less than $D(z)$.

If $R(z) = 0$, then $D(z)$ is a factor of $P(z)$. Here we introduce two important theorems about polynomial.

Theorem 5.3 — Remainder theorem. Let $\alpha \in \mathbb{C}$. When a polynomial $P(z)$ is divided by $z - \alpha$, the remainder is $P(\alpha)$.

Proof. Consider the polynomial division of $P(z)$ by $z - \alpha$, expressed as

$$P(z) = (z - \alpha)Q(z) + R \tag{5.22}$$

where $Q(z)$ is the quotient and R is the remainder. Since R is a polynomial with degree less than that of $z - \alpha$, R is a constant. Evaluating $P(z)$ at $z = \alpha$ gives:

$$P(\alpha) = (\alpha - \alpha)Q(\alpha) + R = R. \tag{5.23}$$

Therefore, the remainder of the division of $P(z)$ by $z - \alpha$ is $R = P(\alpha)$. ■

Theorem 5.4 — Factor theorem. Let $\alpha \in \mathbb{C}$. Then $z - \alpha$ is a factor of a polynomial $P(z)$ if and only if $P(\alpha) = 0$.

Proof. Suppose that $z - \alpha$ is a factor of $P(z)$. Then $P(z)$ can be written as

$$P(z) = (z - \alpha)Q(z) \tag{5.24}$$

for some polynomial $Q(z)$. Evaluating at $z = \alpha$ yields:

$$P(\alpha) = (\alpha - \alpha)Q(\alpha) = 0. \tag{5.25}$$

This shows that if $z - \alpha$ is a factor of $P(z)$, then $P(\alpha) = 0$.

Conversely, if $P(\alpha) = 0$, by the Remainder Theorem we have

$$\begin{aligned} P(z) &= (z - \alpha)Q(z) + P(\alpha) \\ &= (z - \alpha)Q(z) \end{aligned} \tag{5.26}$$

as $P(\alpha) = 0$. Therefore, $P(z)$ is divisible by $z - \alpha$, and $z - \alpha$ is a factor of $P(z)$. ■

We start with a simple example.

■ **Example 5.10** Factorize $P(z) = z^3 + z^2 + 4$. ■

Solution: Use the factor theorem to find the first factor:

$$P(-1) = -1 + 1 + 4 \neq 0$$

$$P(-2) = -8 + 4 + 4 = 0$$

Therefore $z+2$ is a factor. We obtain $P(z) = (z+2)(z^2 - z + 2)$ by division.
We can factorize $z^2 - z + 2$ by completing the square:

$$\begin{aligned} z^2 - z + 2 &= \left(z^2 - z + \frac{1}{4}\right) + 2 - \frac{1}{4} \\ &= \left(z - \frac{1}{2}\right)^2 - \frac{7}{4}i^2 \\ &= \left(z - \frac{1}{2} + \frac{\sqrt{7}}{2}i\right)\left(z - \frac{1}{2} - \frac{\sqrt{7}}{2}i\right) \end{aligned}$$

Hence

$$P(z) = (z+2)\left(z - \frac{1}{2} + \frac{\sqrt{7}}{2}i\right)\left(z - \frac{1}{2} - \frac{\sqrt{7}}{2}i\right)$$

It is noticeable that two of the roots in the example are conjugate to each other. Let's recall that, in exercise **??**, we have proven this as "Conjugate Root Theorem".

Theorem 5.5 Let $P(z)$ be a polynomial with real coefficients. If $a + bi$ is a solution of the equation $P(z) = 0$, with a and b real numbers, then the complex conjugate $a - bi$ is also a solution.

With this theorem, we can solve equations of higher order more quickly.

■ **Example 5.11** Let $P(z) = z^3 - 3z^2 + 5z - 3$.

a. Use the factor theorem to show that $z - 1 + \sqrt{2}i$ is a factor of $P(z)$.

b. Find the other linear factors of $P(z)$.

■

Solution: a. To show that $z - (1 - \sqrt{2}i)$ is a factor, we must check that $P(1 - \sqrt{2}i) = 0$. We have

$$P(1 - \sqrt{2}i) = (1 - \sqrt{2}i)^3 - 3(1 - \sqrt{2}i)^2 + 5(1 - \sqrt{2}i) - 3 = 0$$

Therefore $z - (1 - \sqrt{2}i)$ is a factor of $P(z)$.

b. Since the coefficients of $P(z)$ are real, the complex linear factors occur in conjugate pairs, so $z - (1 + \sqrt{2}i)$ is also a factor.

To find the third linear factor, first multiply the two complex factors together:

$$(z-(1-\sqrt{2}i))(z-(1+\sqrt{2}i)) = z^2-(1-\sqrt{2}i)z-(1+\sqrt{2}i)z+(1-\sqrt{2}i)(1+\sqrt{2}i)$$

$$= z^2-(1-\sqrt{2}i+1+\sqrt{2}i)z+1+2 = z^2-2z+3$$

Therefore, by inspection, the linear factors of $P(z) = z^3-3z^2+5z-3$ are

$$z-1+\sqrt{2}i, \quad z-1-\sqrt{2}i, \quad \text{and} \quad z-1$$

■ **Example 5.12** Factorise: $P(z) = z^6-1$ ■

Solution: $P(z) = z^6-1$ is factored as:

$$P(z) = (z^3+1)(z^3-1)$$

To factor z^3+1 and z^3-1, we recognize these as a sum and difference of cubes, respectively.

We have

$$z^3+1 = (z+1)\left(z^2-z+1\right)$$

$$= (z+1)\left(\left(z-\frac{1}{2}\right)^2-\left(\frac{\sqrt{3}}{2}i\right)^2\right)$$

$$= (z+1)\left(z-\frac{1}{2}+\frac{\sqrt{3}}{2}i\right)\left(z-\frac{1}{2}-\frac{\sqrt{3}}{2}i\right)$$

By a similar method, we have

$$z^3-1 = (z-1)\left(z^2+z+1\right)$$

$$= (z-1)\left(\left(z+\frac{1}{2}\right)^2-\left(\frac{\sqrt{3}}{2}i\right)^2\right)$$

$$= (z-1)\left(z+\frac{1}{2}+\frac{\sqrt{3}}{2}i\right)\left(z+\frac{1}{2}-\frac{\sqrt{3}}{2}i\right)$$

Therefore

$$z^6-1 = (z+1)(z-1)\left(z-\frac{1}{2}+\frac{\sqrt{3}}{2}i\right)\left(z-\frac{1}{2}-\frac{\sqrt{3}}{2}i\right)\left(z+\frac{1}{2}+\frac{\sqrt{3}}{2}i\right)\left(z+\frac{1}{2}-\frac{\sqrt{3}}{2}i\right)$$

For this kind of simple expression that we can find some solutions easily, we can also use a more simplified method. All you need to do is to apply long division to the original

expression and the product of known factors. Here we know that $z = \pm 1$ are two solutions, and $(z+1)\cdot(z-1) = z^2 - 1$. Thus, we apply long division.

$$
\begin{array}{r|l}
 & z^4 + z^2 + 1 \\
\hline
z^2 - 1 & z^6 - 0z^5 - 0z^4 - 0z^3 + 0z^2 - 0z - 1 \\
 & -(z^6 - 0z^4) \\
\hline
 & 0z^5 + z^4 \\
 & -(0z^5 + z^4 - z^2) \\
\hline
 & 0z^4 + z^2 \\
 & -(0z^4 + z^2 - 1) \\
\hline
 & 1
\end{array}
$$

The expression $z^4 + z^2 + 1$ can be factored over the complex numbers. It can be written as a quadratic in terms of z^2:

$$u^2 + u + 1 = 0, \quad \text{where} \quad u = z^2$$

Using the quadratic formula to solve for u:

$$u = \frac{-1 \pm \sqrt{-3}}{2} = \frac{-1 \pm i\sqrt{3}}{2}$$

Thus, the roots for z^2 are:

$$z^2 = \frac{-1 + i\sqrt{3}}{2}, \quad z^2 = \frac{-1 - i\sqrt{3}}{2}$$

The factorization of the original polynomial is therefore:

$$z^4 + z^2 + 1 = \left(z^2 + \frac{1}{2} - \frac{i\sqrt{3}}{2}\right)\left(z^2 + \frac{1}{2} + \frac{i\sqrt{3}}{2}\right)$$

This can be further factored to obtain the linear factors in terms of z:

$$z^4 + z^2 + 1 = \left(z - \frac{1}{2} + \frac{\sqrt{3}}{2}i\right)\left(z - \frac{1}{2} - \frac{\sqrt{3}}{2}i\right)\left(z + \frac{1}{2} + \frac{\sqrt{3}}{2}i\right)\left(z + \frac{1}{2} - \frac{\sqrt{3}}{2}i\right)$$

So far, we have solved problems on polynomials of different orders. It seems so far that a n order polynomial has exactly n solutions. Is this a generalized law? This is actually a corollary under Fundamental Theorem of Algebra.

Theorem 5.6 — Theorem of Algebra. Every polynomial $P(z) = a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0$ of degree n, where $n \geq 1$ and the coefficients a_i are complex numbers, has at least one linear factor in the complex number system.

Given any polynomial $P(z)$ of degree $n \geq 1$, the theorem tells us that we can factorize $P(z)$ as

$$P(z) = (z - \alpha_1)Q(z)$$

for some $\alpha_1 \in \mathbb{C}$ and some polynomial $Q(z)$ of degree $n-1$. By applying the fundamental theorem of algebra repeatedly, it can be shown that:

Corollary 5.2 A polynomial of degree n can be factorized into n linear factors in $\mathbb{C}$:

$$i.e., P(z) = a_n(z - \alpha_1)(z - \alpha_2)\dots(z - \alpha_n), \text{ where } \alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{C}$$

A polynomial equation can be solved by first rearranging it into the form $P(z) = 0$, where $P(z)$ is a polynomial, and then factorizing $P(z)$ and extracting a solution from each factor. If $P(z) = (z - \alpha_1)(z - \alpha_2)\dots(z - \alpha_n)$, then the solutions of $P(z) = 0$ are $\alpha_1, \alpha_2, \dots, \alpha_n$. The solutions of the equation $P(z) = 0$ are also referred to as the zeroes or the roots of the polynomial $P(z)$.

■ **Example 5.13** Solve each of the following equations over $\mathbb{C}$:
(a) $P(z) = z^2 + 64 = 0$
(b) $P(z) = z^3 + 3z^2 + 7z + 5 = 0$
(c) $P(z) = z^3 - iz^2 - 4z + 4i = 0$

■

Solution:
(a)

$$(z + 8i)(z - 8i) = 0$$

(b)

$$= (z+1)\left(z + \frac{1}{2} - \frac{\sqrt{7}}{2}i\right)\left(z + \frac{1}{2} + \frac{\sqrt{7}}{2}i\right)$$

(c)

$$(z - i)(z - 2)(z + 2) = 0$$

5.5.3 Solving Equation with De Moivre's Theorem

With the help of complex number, we have "conquered" polynomial equations. Now we will focus on solving equations with de Moivre's theorem.

Equations of the form $z^n = a$, where $a \in \mathbb{C}$, are often solved by using De Moivre's theorem.

Write both z and a in polar form, as $z = r\text{cis}\theta$ and $a = q\text{cis}\varphi$.

Then $z^n = a$ becomes

$$(r\text{cis}\theta)^n = q\text{cis}\varphi$$

$$\therefore \quad r^n\text{cis}(n\theta) = q\text{cis}\varphi \quad \text{(using De Moivre's theorem)}$$

Compare modulus and argument:

$$r^n = q \quad \text{and} \quad \text{cis}(n\theta) = \text{cis}\varphi$$

$$r = \sqrt[n]{q}$$

$$n\theta = \varphi + 2k\pi \quad \text{where} \quad k \in \mathbb{Z}$$

$$\theta = \frac{1}{n}(\varphi + 2k\pi) \quad \text{where} \quad k \in \mathbb{Z}$$

This will provide all the solutions of the equation.

▪ **Example 5.14** Solve $z^3 = 1$. ▪

Solution Let $z = r\text{cis}\theta$. Then

$$(r\text{cis}\theta)^3 = 1\text{cis}0$$

$$\therefore \quad r^3\text{cis}(3\theta) = 1\text{cis}0$$

$$\therefore \quad r^3 = 1 \quad \text{and} \quad 3\theta = 0 + 2k\pi \quad \text{where} \quad k \in \mathbb{Z}$$

$$\therefore \quad r = 1 \quad \text{and} \quad \theta = \frac{2k\pi}{3} \quad \text{where} \quad k \in \mathbb{Z}$$

Hence the solutions are of the form $z = \text{cis}\left(\frac{2k\pi}{3}\right)$, where $k \in \mathbb{Z}$.

We start finding solutions.

For $k = 0$: $z = \text{cis}0 = 1$

For $k = 1$: $z = \text{cis}\left(\frac{2\pi}{3}\right)$

For $k = 2$: $z = \text{cis}\left(\frac{4\pi}{3}\right) = \text{cis}\left(-\frac{2\pi}{3}\right)$

For $k = 3$: $z = \text{cis}(2\pi) = 1$

The solutions begin to repeat.

The three solutions are 1, $\text{cis}\left(\frac{2\pi}{3}\right)$, and $\text{cis}\left(-\frac{2\pi}{3}\right)$. We can plot these solutions on the complex plane.

Figure 5.6: Solutions for $z^3 = 1$

The solutions are shown to lie on the unit circle at intervals of $\frac{2\pi}{3}$ around the circle. We find that the number of roots, of course, follows that the number of roots equal to the order of expression.

To generalize, we can see that there are exactly m distinct mth roots of unity, denoted by

$$\boxed{1^{1/m} = e^{i2k\pi/m} = \cos\frac{2k\pi}{m} + i\sin\frac{2k\pi}{m} \quad (k = 0,1,2,\ldots,m-1).} \tag{5.27}$$

Theorem 5.7 — Solution of $z^n = a$. For $n \in \mathbb{N}$ and $a \in \mathbb{C}$, the solutions of the equation $z^n = a$ are called the nth roots of a.

- The solutions of $z^n = a$ lie on a circle with center the origin and radius $|a|^{1/n}$.
- There are n solutions and they are equally spaced around the circle at intervals of $\frac{2\pi}{n}$. This observation can be used to find all solutions if one is known.

More generally, consider any natural number $n \geq 2$. Using De Moivre's theorem, we can show that the nth roots of unity are

$$1, \operatorname{cis}\left(\frac{2\pi}{n}\right), \operatorname{cis}\left(\frac{4\pi}{n}\right), \ldots, \operatorname{cis}\left(\frac{2(n-1)\pi}{n}\right)$$

So the nth roots of unity form a geometric sequence with common ratio $\omega = \operatorname{cis}\left(\frac{2\pi}{n}\right)$. We can list the terms of this sequence as $1, \omega, \omega^2, \ldots, \omega^{n-1}$. The sum of the terms is

$$1 + \omega + \omega^2 + \ldots + \omega^{n-1} = \frac{\omega^n - 1}{\omega - 1}$$

since $\omega^n = 1$.

Proof. The nth roots of unity are solutions to the equation $z^n = 1$. By expressing 1 in polar form as 1cis0, and applying De Moivre's theorem, $(r\operatorname{cis}\theta)^n = r^n\operatorname{cis}(n\theta)$, we find that for $z^n = 1$, we must have $z = \operatorname{cis}\left(\frac{2k\pi}{n}\right)$, where k is an integer from 0 to $n-1$.

These solutions can be written as a sequence where each term after the first is obtained by multiplying the previous term by $\operatorname{cis}\left(\frac{2\pi}{n}\right)$, making it a geometric sequence with a common ratio of $\omega = \operatorname{cis}\left(\frac{2\pi}{n}\right)$.

The sum of a geometric sequence with n terms and a common ratio $r \neq 1$ is given by $S_n = a_1\frac{1-r^n}{1-r}$. For our sequence, $a_1 = 1$ and $r = \omega$, so the sum is $S_n = \frac{1-\omega^n}{1-\omega}$. Since $\omega^n = \operatorname{cis}(2\pi) = 1$, the numerator becomes $1 - 1 = 0$, hence the sum S_n is zero. ■

To obtain the mth roots of an *arbitrary* (nonzero) complex number $z = re^{i\theta}$, we generalize the idea and, reasoning similarly, conclude that the m distinct mth roots of z are given by

$$\boxed{z^{1/m} = \sqrt[m]{|z|}e^{i(\theta+2k\pi)/m} \quad (k = 0,1,2,\ldots,m-1)} \tag{5.28}$$

Equivalently, we can form these roots by taking any single one such as given in (3) and multiplying by the mth roots of unity.

▪ **Example 5.15** Find all the cube roots of $\sqrt{2} + i\sqrt{2}$. ▪

Solution: The polar form for $\sqrt{2} + i\sqrt{2}$ is

$$\sqrt{2} + i\sqrt{2} = 2e^{i\pi/4}.$$

Putting $|z| = 2$, $\theta = \pi/4$, and $m = 3$ into Eq. **??**, we obtain

$$(\sqrt{2} + i\sqrt{2})^{1/3} = \sqrt[3]{2}e^{i(\pi/12+2k\pi/3)} \quad (k = 0,1,2).$$

Therefore, the three cube roots of $\sqrt{2} + i\sqrt{2}$ are $\sqrt[3]{2}(cos\pi/12 + i\sin\pi/12)$, $\sqrt[3]{2}(cos3\pi/4 + i\sin 3\pi/4)$, and $\sqrt[3]{2}(cos17\pi/12 + i\sin 17\pi/12)$.

5.5.4 Exercises

Exercise 5.31 Let $P(z) = 2z^3 + 9z^2 + 14z + 5$.

a. Use the factor theorem to show that $z+2-i$ is a linear factor of $P(z)$.

b. Write down another complex linear factor of $P(z)$.

c. Hence find all the linear factors of $P(z)$ over $\mathbb{C}$.

Solution

a. By the factor theorem, if $z+2-i$ is a factor of $P(z)$, then $P(-2+i) = 0$. Calculating this we get:

$$\begin{aligned} P(-2+i) &= 2(-2+i)^3 + 9(-2+i)^2 + 14(-2+i) + 5 \\ &= 2(-8-12i+6i^2) + 9(4-8i+2i^2) + 14(-2+i) + 5 \\ &= 2(-8-12i-6) + 9(4-8i-2) + (-28+14i) + 5 \\ &= -28-24i-12+36-72i-18-28+14i+5 \\ &= 0. \end{aligned}$$

Thus, $z+2-i$ is a linear factor of $P(z)$.

b. By the conjugate root theorem, if $z+2-i$ is a root, then its conjugate $z+2+i$ is also a root.

c. Having found two complex roots, we can divide $P(z)$ by the product of the corresponding factors to find the remaining factor. Let's perform the division (this is a mock example; the actual division should be computed):

$$P(z) = (z+2-i)(z+2+i)(z-\alpha)$$

After performing the division, we find that the remaining factor is $z - \frac{1}{2}$. Thus, the linear factors of $P(z)$ are:

$$P(z) = (z+2-i)(z+2+i)\left(z - \frac{1}{2}\right)$$

Exercise 5.32 For a cubic polynomial $P(x)$ with real coefficients, it is given that $P(2+i) = 0$, $P(1) = 0$ and $P(0) = 10$. Express $P(x)$ in the form $P(x) = ax^3 + bx^2 + cx + d$ and solve the equation $P(x) = 0$.

solution:

Given that $P(x)$ is a cubic polynomial with real coefficients and $P(2+i) = 0$, we know that its conjugate $P(2-i) = 0$ as well, due to the complex conjugate root theorem. Moreover, since $P(1) = 0$, we can write $P(x)$ as:

$$P(x) = a(x-(2+i))(x-(2-i))(x-1)$$

Expanding this and simplifying, we get:

$$P(x) = a((x-2)^2+1)(x-1)$$

Since $P(0) = 10$, we can find a by substituting $x = 0$:

$$10 = a((0-2)^2+1)(0-1)$$

$$10 = a(4+1)(-1)$$

$$10 = -5a$$

$$a = -2$$

Thus, the polynomial is:

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

$$P(x) = -2(x^2-4x+4+1)(x-1)$$

$$P(x) = -2(x^2-4x+5)(x-1)$$

$$P(x) = -2(x^3-x^2-4x^2+4x+5x-5)$$

$$P(x) = -2x^3+10x^2-18x+10$$

To solve for $P(x) = 0$, we now have the equation:

$$-2x^3+10x^2-18x+10 = 0$$

Since $P(2+i) = 0$, by conjugate root theorem, $p(2-i) = 0$, and we also have $p(1) = 0$. The order of this expression is 3, by the fundamental algebra theorem, we have found all solutions.

Exercise 5.33 If $z = 1+i$ is a zero of the polynomial $z^3+az^2+bz+10-6i$, find the constants a and b, given that they are real. ■

solution:

Since $z = 1+i$ is a zero, by the Complex Conjugate Root Theorem, $z = 1-i$ is also a zero. Substituting $z = 1+i$ into the polynomial gives:

$$\begin{aligned}(1+i)^3+a(1+i)^2+b(1+i)+10-6i &= 0\\ (1+3i-3-i)+a(1+2i-1)+b+10-6i &= 0\\ (-2+2i)+a(2i)+b+10-6i &= 0\\ (-2+(2a-6)i)+b+10 &= 0\end{aligned}$$

For the polynomial to be zero, both the real and imaginary parts must be zero. Therefore, we have two equations:

$$-2+b+10=0 \quad \text{(Real part)}$$
$$2a-6=0 \quad \text{(Imaginary part)}$$

Solving these equations gives us a and b:

$$b=-8$$
$$a=3$$

Thus, the constants are $a=3$ and $b=-8$.

Exercise 5.34 Let n be a positive integer. Prove that

$$\arg(z^n) = n\text{Arg}(z) + 2k\pi, \quad k = 0, \pm 1, \pm 2, \ldots,$$

for $z \neq 0$ ■

Proof. Let's take $z \in \mathbb{C}$. We can write z in its polar form as

$$z = |z|(\cos\theta + i\sin\theta)$$

By definition of arg$z$ and Arg$z$ we have: $\arg(z) = \theta + 2k\pi, k \in \mathbb{Z}, \theta \in [-\pi, \pi]$ and $\text{Arg}(z) = \arg_(z) = \theta, \theta \in [-\pi, \pi]$.

Now if we take the polar form of z^n, where $n \in \mathbb{N}, n \neq 0$, we have

$$z^n = |z|^n(\cos(n\theta) + i\sin(n\theta))$$

$$\Rightarrow \arg(z^n) = n\theta + 2k\pi$$

$$= n\text{Arg}(z) + 2k\pi, \quad k \in \mathbb{Z}$$

This concludes the proof. ■

Exercise 5.35 Find all the values of the following.

(a) $(-16)^{1/4}$
(b) $1^{1/5}$
(c) $i^{1/4}$
(d) $(1-\sqrt{3}i)^{1/3}$
(e) $(i-1)^{1/2}$
(f) $\left(\frac{2i}{1+i}\right)^{1/6}$

■

Exercise 5.36 Find all four roots of the equation $z^4+1=0$ and use them to deduce the factorization $z^4+1 = (z^2-\sqrt{2}z+1)(z^2+\sqrt{2}z+1)$. ■

Solution:

First, we solve the equation:

$$\begin{aligned} z^4+1&=0\\ z^4&=-1\\ z&=\sqrt[4]{-1} \end{aligned}$$

Since $-1=\cos\pi+i\sin\pi$, we can find the roots using De Moivre's theorem:

$$z=\sqrt[4]{\cos\pi+i\sin\pi}=\sqrt[4]{1}(\cos(\pi+2k\pi)/4+i\sin(\pi+2k\pi)/4),\quad k=0,1,2,3$$

Thus, the roots are:

$$\begin{aligned} z_0&=\cos\frac{\pi}{4}+i\sin\frac{\pi}{4}=\frac{1}{\sqrt{2}}+i\frac{1}{\sqrt{2}}\\ z_1&=\cos\frac{3\pi}{4}+i\sin\frac{3\pi}{4}=-\frac{1}{\sqrt{2}}+i\frac{1}{\sqrt{2}}\\ z_2&=\cos\frac{5\pi}{4}+i\sin\frac{5\pi}{4}=-\frac{1}{\sqrt{2}}-i\frac{1}{\sqrt{2}}\\ z_3&=\cos\frac{7\pi}{4}+i\sin\frac{7\pi}{4}=\frac{1}{\sqrt{2}}-i\frac{1}{\sqrt{2}} \end{aligned}$$

To show the factorization, we pair the roots:

$$\begin{aligned} z^4+1&=(z-z_0)(z-z_1)(z-z_2)(z-z_3)\\ &=\left(z-\frac{1}{\sqrt{2}}-i\frac{1}{\sqrt{2}}\right)\left(z+\frac{1}{\sqrt{2}}-i\frac{1}{\sqrt{2}}\right)\\ &\quad\times\left(z+\frac{1}{\sqrt{2}}+i\frac{1}{\sqrt{2}}\right)\left(z-\frac{1}{\sqrt{2}}+i\frac{1}{\sqrt{2}}\right)\\ &=\left(z^2-\sqrt{2}z+1\right)\left(z^2+\sqrt{2}z+1\right) \end{aligned}$$

And we have shown the factorization.

Exercise 5.37 Let α and β be the mth and nth roots of unity, respectively. We want to show that the product $\alpha\beta$ is a kth root of unity for some integer k. ■

Solution:

Let's consider α as the m-th root of unity and β the n-th root of unity. Then we have

$$\alpha^m=1$$

$$\beta^n=1$$

Let's show that $\alpha\beta$ is a root of unity. If $\alpha\beta$ is a root of unity then there exists some $l\in\mathbb{N}$ such that $(\alpha\beta)^l=1$.

For $l=mn$ (since m and n are relatively prime), we have

$$(\alpha\beta)^l=(\alpha\beta)^{mn}$$

$$=\alpha^{mn}\beta^{mn}$$

$$= (\alpha^m)^n(\beta^n)^m$$

$$= 1^n 1^m$$

$$= 1$$

Thus, $\alpha\beta$ is indeed a root of unity, specifically an mn-th root of unity.

Exercise 5.38 Let m and n be positive integers that have no common factor. Prove that the set of numbers $(z^{1/n})^m$ is the same as the set of numbers $(z^m)^{1/n}$. We denote this common set of numbers by $z^{m/n}$. Show that

$$z^{m/n} = \sqrt[n]{|z|^m}\left[\cos\left(\frac{m}{n}(\theta+2k\pi)\right) + i\sin\left(\frac{m}{n}(\theta+2k\pi)\right)\right]$$

for $k = 0, 1, \ldots, n-1$. ■

Proof. Let's take $m, n \in \mathbb{N}$ so that they are relatively prime and $z \in \mathbb{C}$ so that $\operatorname{Arg}(z) = \theta$. We have:

$$\begin{aligned}(z^{1/n})^m = (\sqrt[n]{z})^m &= \left(\sqrt[n]{|z|e^{i\theta}}\right)^m \\ &= \left(\sqrt[n]{|z|}e^{i\theta/n}\right)^m \\ &= |z|^{m/n}e^{im\theta/n}\end{aligned}$$

Now as m and n are relatively prime we have

$$\frac{2k\pi}{n}\frac{m}{n} \qquad k = 0, \ldots, n-1$$

from where

$$\begin{aligned}|z|^{m/n}e^{im(\theta+2k\pi)/n} &= |z|^{m/n}e^{im\theta/n+im2k\pi/n} \\ &= (|z|^m)^{1/n}(e^{im\theta})^{1/n} \\ &= ((|z|^m e^{im\theta})^{1/n} \\ &= (z^m)^{1/n}\end{aligned}$$

To show the expression we take the previous result and expand it:

$$\begin{aligned}z^{m/n} &= |z|^{m/n}e^{im\theta/n} \\ &= \sqrt[n]{|z|^m}\left[\cos\left(\frac{m\theta}{n}+\frac{2km\pi}{n}\right) + i\sin\left(\frac{m\theta}{n}+\frac{2km\pi}{n}\right)\right] \\ &= \sqrt[n]{|z|^m}\left[\cos\left(\frac{m}{n}(\theta+2k\pi)\right) + i\sin\left(\frac{m}{n}(\theta+2k\pi)\right)\right]\end{aligned}$$

■

Exercise 5.39 Use the conclusion in last problem to evaluate $(i-i)^{\frac{3}{2}}$. ■

Solution: It has been shown that

$$\boxed{z^{m/n} = \sqrt[n]{|z|^m}\left[\cos\left(\frac{m}{n}(\theta+2k\pi)\right)+i\sin\left(\frac{m}{n}(\theta+2k\pi)\right)\right]} \tag{5.29}$$

We have:

$$z = (1-i)^{3/2} \Rightarrow m = 3, n = 2$$

$$|z| = \sqrt{1^2+(-1)^2} = \sqrt{2}$$

$$\theta = \text{Arg}z = 2\pi - \cot^{-1}\left(\frac{-1}{1}\right) = \frac{7\pi}{4}$$

When we put everything in the formula we get

$$z = \sqrt[2]{\sqrt{2}^3}\left[\cos\left(\frac{3}{2}\left(\frac{7\pi}{4}+2k\pi\right)\right)+i\sin\left(\frac{3}{2}\left(\frac{7\pi}{4}+2k\pi\right)\right)\right], k = 0,1$$

$$z = \sqrt[2]{2^{3/2}}\left[\cos\left(\frac{21\pi}{8}+\frac{3k\pi}{2}\right)+i\sin\left(\frac{21\pi}{8}+\frac{3k\pi}{2}\right)\right], k = 0,1$$

$$z_0 = 2^{3/4}\left[\cos\left(\frac{21\pi}{8}\right)+i\sin\left(\frac{21\pi}{8}\right)\right]$$

$$z_1 = 2^{3/4}\left[\cos\left(\frac{21\pi}{8}+\frac{3\pi}{2}\right)+i\sin\left(\frac{21\pi}{8}+\frac{3\pi}{2}\right)\right]$$

6.1 Boolean Expression and Truth Table 145
6.2 Predicates and Quantifiers 155
6.3 Digital Circuits of Logic 155
6.4 Logic of Deduction and Induction 155

7 Preliminary Number Theory and Cryptography 157
7.1 Divisibility and Modular Arithmetic 157
7.2 Number Representations and Algorithms 164
7.3 Primes and Greatest Common Divisors .. 171
7.4 Solving Congruence 177

8 Search and Sorting with Further Algorithm Analysis 179
8.1 Searching algorithms 179

9 Graph Theory 193

In the first chapter of this book, we discussed the very basic of mathematics, proof and propositions. This chapter aims to excavate mathematical logics in further details. With Boolean Algebra, we can explain more thoroughly on the mechanism of logics and mathematical proof, and on top of that, we can figure out how computer functions, as well as how integrated circuits are constructed.

6.1 Boolean Expression and Truth Table

We call Boolean Algebra "Algebra" of course, because it possesses property of Algebra. We will start with basic algebra operation and thus, proceed to get the rule of Boolean operation, which we call the "Truth Table".

6.1.1 Property of Algebra Operation

For normal algebra operation of numbers, we have the following general law by representing the number in x, y, and z. In this case, we call x, y, and z variables as in programming.

Table 6.1: Common Algebraic Laws

Law	**Addition Expression**	**Multiplication Expression**
Identity	$x+0=x$	$x\cdot 1=x$
Property of Zero		$x\cdot 0=0$
Inverse	$x+(-x)=0$	$x\cdot x^{-1}=1$, for $x\neq 0$
Commutative	$x+y=y+x$	$x\cdot y=y\cdot x$
Associative	$(x+y)+z=x+(y+z)$	$(x\cdot y)\cdot z=x\cdot(y\cdot z)$
Distributive		$x\cdot(y+z)=x\cdot y+x\cdot z$

for all possible values of these variables, these laws hold. Why these laws are important is that all algebra expressions is derived from these laws, which means we can prove any equality that is correct. We take -1 as an example. One may say without a second thought that the answer is 1. But why? Can we prove it? Now we try to prove it using the common algebraic laws.

■ **Example 6.1** Prove that $(-1)\times(-1)=1$ ■

Proof.

$$
\begin{aligned}
&(-1)\times(-1)\\
=&((-1)\times(-1))+0\ (Addition\ Identity\ )\\
=&((-1)\times(-1))+\ ((-1)+1)\ (Addition\ Inverse)\\
=&(((-1)\times(-1)+(-1))+1\ (Associative\ Law\ of\ Addition)\\
=&(((-1)\times(-1)+((-1)\times 1))+1\ (Multiplication\ Identity)\\
=&((-1)\times((-1)+1))+1\ (Distributive\ Law)\\
=&((-1)\times 0)\ +\ 1\ (Addition\ Inverse)\\
=&0+1\ (Multiplication\ Property\ of\ Zero)\\
=&1+0\ (Commutative\ Law\ of\ Addition)\\
=&1\ (Addition\ Identity)
\end{aligned}
$$

R Some people may not understand why we have to write $0+1$ into $1+0$. This is because Addition Identity is only defined in the table as $x+0=x$, so we need to apply Commutative law of addition to fit it into the known conclusion.

■

After seeing this, you may think, do we really have to know this to make sure that $-1\times-1=1$? Of course not, but keep in mind that **Every mathematical conclusion cannot be simply referred as rules such as "a negative times negative give you a positive number", but by meticulous, reasonable, and replicable proof.**

6.1.2 Boolean Expression and Truth Table

Hold on a second, isn't this chapter on Boolean Algebra? Why we still need to go over these old knowledge from primary school? Well, this is because we will prove Boolean expression in the same way. Before that, we introduce Boolean value and operations.

Definition 6.1 — Boolean Values. In mathematics and computer science, a Boolean value is defined as an element of the Boolean domain B, which can be mathematically represented as:

$$B=\{0,1\}$$

where:

- 0 typically represents *false*
- 1 typically represents *true*

Some may be confused by true and false here. Just recall what we have done to propositions in the first chapter in this book. When a statement holds for given condition, we take it as correct, while incorrect when it does not hold. Boolean value is the basis of Boolean Expression, and **all valid boolean expression could be reduced or simplified to a Boolean value**.

Definition 6.2 — Boolean Operations. Boolean algebra involves operations such as AND, OR, NOT, XOR, which operate on these Boolean values. These operations are defined as follows:

- **AND** ($\wedge$): An operation on two Boolean values that returns *true* if both operands are *true*, otherwise returns *false*.
- **OR** ($\vee$): An operation on two Boolean values that returns *true* if at least one of the operands is *true*, otherwise returns *false*.
- **NOT** ($\neg$): A unary operation that returns *true* if the operand is *false* and vice versa.
- **XOR** ($\oplus$): An operation on two Boolean values that returns *true* if the operands are different, otherwise returns *false*.

R There are more Boolean operators to be discussed later.

In computer science, Boolean values are fundamental in conditional statements and loops, where they determine the flow of control in algorithms and programs. They are also essential in the design of electronic circuits and digital computing. The following table shows the truth table of these basic Boolean operations. the first column shows the combination of inputs for the specific operator, and the second column shows the result of operation.

A	**$\neg$A**
F	T
T	F

A	**B**	**A $\wedge$ B**
F	F	F
F	T	F
T	F	F
T	T	T

A	**B**	**A $\vee$ B**
F	F	F
F	T	T
T	F	T
T	T	T

A	**B**	**A $\oplus$ B**
F	F	F
F	T	T
T	F	T
T	T	F

Table 6.2: Common Boolean Operators Truth Tables

It's noticeable that different Boolean operators may differ in the number of input. $\neg$ operator takes only one Boolean variable (input) to get an output by negating the input, however the rest take two inputs and produce one output.

At the beginning of the section, we have defined Boolean space $B = \{0, 1\}$, but why George Boole(The British mathematician who initiate this idea) define Boolean value with a set containing only zero and one? This is actually because, **The foundation of Boolean Algebra is set upon operations involving 0 and 1**. To explicit, I need to clarify that among the four operators introduced above, $\neg$, $\wedge$, and $\vee$ are the very basic of all Boolean expressions, which means other Boolean operators could be written in an equivalent form with these basic operators. For example, we can write XOR($\oplus$) in the other three operators:

$$A \oplus B = (A \wedge \neg B) \vee (\neg A \wedge B) \tag{6.1}$$

It doesn't really matter if you find the RHS strange,as all we need now is just know the fact that we can write it in this form. We will discuss the expression in details in the next

section. Now let's refocus on Boolean space $\{0,1\}$. We all know that there are four basic operations in algebra which are addition, subtraction, multiplication, and division. But generally, we say there are only two basic operations in algebra, which are $+$ and $\times$, as subtraction and division are just inverse operation of addition and multiplication. Let's look in to the table of addition and multiplication of 0 and 1 listed below.

A	B	$A \times B$
0	0	0
0	1	0
1	0	0
1	1	1

A	B	$A+B$
0	0	0
0	1	1
1	0	1
1	1	1

Table 6.3: Addition and Multiplication Rule of 0 and 1

Now you may have realized that this is exactly the truth table for $\wedge$ and $\vee$, you may check table **??**.

6.1.3 Boolean Identities

Now recall that $A \oplus B = (A \wedge \neg B) \vee (\neg A \wedge B)$. In this expression, we cannot get a direct answer from the RHS by using the truth table of $\vee$. So, we need to try harder to get the truth table for the expression as shown in the table below.

Table 6.4: Truth table for $(A \wedge \neg B) \vee (\neg A \wedge B)$

A	B	$\neg B$	$A \wedge \neg B$	$\neg A$	$\neg A \wedge B$	$(A \wedge \neg B) \vee (\neg A \wedge B)$
F	F	T	F	T	F	F
F	T	F	F	T	T	T
T	F	T	T	F	F	T
T	T	F	F	F	F	F

With this method, we can find truth table for more complex expressions, and some of them are categorized as **Basic Boolean Identities**. The LHS and RHS are could be proven qaual by listing truth table respectively.

Table 6.5: Basic Boolean Identities

Identity	**AND Form**	**OR Form**
Idempotent Law	$x \cdot x = x$	$x + x = x$
Identity Law	$x \cdot 1 = x$	$x + 0 = x$
Domination Law	$x \cdot 0 = 0$	$x + 1 = 1$
Complement Law	$x \cdot \neg x = 0$	$x + \neg x = 1$
Double Negation Law	$\neg(\neg x) = x$	$\neg(\neg x) = x$
Commutative Law	$x \cdot y = y \cdot x$	$x + y = y + x$
Associative Law	$x \cdot (y \cdot z) = (x \cdot y) \cdot z$	$x + (y + z) = (x + y) + z$
Distributive Law	$x \cdot (y + z) = (x \cdot y) + (x \cdot z)$	$x + (y \cdot z) = (x + y) \cdot (x + z)$
De Morgan's Law	$\neg(x + y) = \neg x \cdot \neg y$	$\neg(x \cdot y) = \neg x + \neg y$
Absorption Law	$x \cdot (x + y) = x$	$x + (x \cdot y) = x$

R If you are confused by the algebra form of Boolean identities, just take x, y, z as A, B, and C; take 0/1 as F/T; take + as $\vee$, and $\times$ as $\wedge$.

It is quite important to be familiar with these rules, as they are just as essential as the normal algebra laws we've learned before, since we may simplify complex Boolean expressions using these rules. Also, the proof of some of these identities using truth table will be added to the problem set for this section. If you remain any doubt or confusion about any laws given, just try to get the truth table for the LHS and RHS of the identity, simple as that.

This is actually not yet a complete table of Boolean Identities, two laws are still missing from the table, because they are relevant to secondary operator, one of which is already mentioned($\oplus$).

With these basic Boolean laws, we can derive more interesting and useful theorems. A commonly used theorem when simplifying Boolean expression is concensus (or also called redundancy) theorem.

Theorem 6.1 — Concensus Theorem. The Consensus Theorem helps simplifying Boolean expressions by eliminating a redundant term, and like other laws, it has both OR and AND form. It states that for any Boolean variables x, y, and z:

$$xy \vee \bar{x}z \vee yz = xy \vee \bar{x}z$$

Which is equivalent to

$$(x \vee y)(\bar{x} \vee z)(y \vee z) = (x \vee y)(\bar{x} \vee z)$$

Proof.

$$\begin{aligned} xy \vee \bar{x}z \vee yz &= xy \vee \bar{x}z \vee (x \vee \bar{x})yz \\ &= xy \vee \bar{x}z \vee xyz \vee \bar{x}yz \\ &= (xy \vee xyz) \vee (\bar{x}z \vee \bar{x}yz) \\ &= xy(1 \vee z) \vee \bar{x}z(1 \vee y) \\ &= xy \vee \bar{x}z \end{aligned}$$

Or in another notation.

$$\begin{aligned} xy + \bar{x}z + yz &= xy + \bar{x}z + (x + \bar{x})yz \\ &= xy + \bar{x}z + xyz + \bar{x}yz \\ &= (xy + xyz) + (\bar{x}z + \bar{x}yz) \\ &= xy(1 + z) + \bar{x}z(1 + y) \\ &= xy + \bar{x}z \end{aligned}$$

■

Secondary Boolean Operators

Definition 6.3 — Secondary Boolean Operators. Secondary operators are derived from basic operators:

- Material conditional: $x \to y = \neg x \vee y$
- Material biconditional: $x \leftrightarrow y = (x \wedge y) \vee (\neg x \wedge \neg y)$
- Exclusive OR (XOR): $x \oplus y = \neg(x \leftrightarrow y) = (x \vee y) \wedge (\neg x \vee \neg y) = (x \wedge \neg y) \vee (\neg x \wedge$

$y)$

Their truth tables for the secondary operation are below:

x	y	$x \to y$	$x \leftrightarrow y$	$x \oplus y$
0	0	1	1	0
1	0	0	0	1
0	1	1	0	1
1	1	1	1	0

Table 6.6: Truth values of material conditional, biconditional, and XOR for all possible inputs.

The $\to$ operation holds that $x \to x = 1$. This expression means that when $x = y$, then $x \to y$ must be true. The complete form of secondary operators and the implication identity will be the last Boolean identity to be covered in this chapter.

1. Material Conditional ($\to$): The material conditional $x \to y$ is read as "if x then y" or "x implies y". It represents the logical implication. The truth value of $x \to y$ is false only when x is true and y is false; in all other cases, it is true. This is a bit counter-intuitive when x is false because the implication will be true regardless of the value of y. It can be expressed using basic operations as $\neg x \vee y$.
2. Material Biconditional ($\leftrightarrow$): The material biconditional $x \leftrightarrow y$, also known as logical equivalence, is the operation that is true when x and y have the same truth values, and false otherwise. It's often read as " x if and only if y". It can be formulated as $(x \wedge y) \vee (\neg x \wedge \neg y)$, meaning both x and y are true, or both are false.
3. Exclusive OR (XOR): The exclusive OR $x \oplus y$ is true when x and y have different truth values that is, one is true and the other is false. It differs from the regular OR operation in that $x \oplus y$ is false when both x and y are true. It can be represented as $(x \wedge \neg y) \vee (\neg x \wedge y)$.

With these basic identities and operators, we can prove more interesting Boolean theorems. Below are more theorems to be proved as example. Maybe you are ready to draw a truth table to prove them. However, with these basic Boolean Laws, we actually don't have to do that, but prove it by using Algebra techniques.

■ **Example 6.2** Prove the following identities using the table given earlier. ■

Expression	**Name**
$(x \to \text{False}) = (\neg x)$	$\neg$ as $\to$
$(x \to y) = (\neg y \to \neg x)$	contrapositive
$((x \to y) \wedge (x \to z)) = (x \to (y \wedge z))$	implication
$((x \to y) \wedge (\neg x \to y)) = y$	absurdity
$(x \to (\neg x)) = (\neg x)$	contradiction

Proof. To prove $\neg$ as $\to$, we just need to use the definition of $\to$.

$$x \to \text{ False} = \neg x \vee False = \neg x + 0 = \neg x \text{ (OR Identity Law)}$$

The contrapositive identity states that $(x \to y)$ is logically equivalent to $(\neg y \to \neg x)$.

$$\begin{aligned} x \to y &\equiv \neg x \vee y \\ &\equiv y \vee \neg x \text{ (Commutative Law)} \\ &\equiv \neg y \to \neg x \end{aligned}$$

R This proof shows why proof by contrapositive (mentioned in chapter1) is correct.

The implication identity can be proved by showing that $(x \to y) \wedge (x \to z)$ is equivalent to $x \to (y \wedge z)$.

$$\begin{aligned} (x \to y) \wedge (x \to z) &\equiv (\neg x \vee y) \wedge (\neg x \vee z) \\ &\equiv \neg x \vee (y \wedge z) \\ &\equiv x \to (y \wedge z) \end{aligned}$$

The absurdity identity states that $(x \to y) \wedge (\neg x \to y)$ is equivalent to y.

$$\begin{aligned} (x \to y) \wedge (\neg x \to y) &\equiv (\neg x \vee y) \wedge (x \vee y) \\ &\equiv y \vee (x \wedge \neg x) \\ &\equiv y \vee \text{False} \\ &\equiv y \end{aligned}$$

The contradiction identity states that $x \to (\neg x)$ is equivalent to $\neg x$.

$$\begin{aligned} x \to (\neg x) &\equiv \neg x \vee (\neg x) \\ &\equiv \neg x \end{aligned}$$

■

R Some may feel confused by the notation of Boolean expression we use in this chapter. Sometimes we use logic notation, while sometimes use algebra operation. Both are ok, it's really up to your preference.

When you look through the whole process, you'll find that we are actually doing the same thing as simplifying Boolean expressions, where all the laws could be used so that we don't need to draw truth table again and again.

6.1.4 Boolean Function

Now we have learned most of the rules of Boolean operations; we use the rest of this section to bring you some further ideas of Boolean operation. Just like we can take a normal algebra expression, such as $2x+3$, as a function $f(x) = 2x+3$; we can define **Boolean Function** with Boolean operation.

Definition 6.4 — Boolean Function. A Boolean function is a function that returns a Boolean value (true or false) for each possible combination of Boolean inputs. $f : B^n \to B$ is a function from the set of n-tuples of Boolean values to the set of Boolean values, where $B = \{0, 1\}$.

6.1.5 Exercises

Exercise 6.1 Use table **??** and $1+1=2$ to show that $(x+x)=(2\times x)$.

Proof.

$$\begin{aligned}(2\times x) &= (1+1)\times x\ (by\ 1+1=2)\\ &= 1\times x+1\times x\ (Distributive\ Law)\\ &= x+x\ (Multiplication\ Identity)\end{aligned}$$

■

Exercise 6.2 Show that $((-1)\times x)+x=0$ using the table.

Proof.

$$\begin{aligned}((-1)\times x)+x &= x(-1+1)\ (Distributive\ Law)\\ &= x\times 0\ (Addition\ Inverse)\\ &= 0\ (Multiplication\ Property\ of\ Zero)\end{aligned}$$

■

Exercise 6.3 Show that $(x+(((-1)\times(x+y))+z))+y=z$, you may use the conclusion of previous exercise.

Proof. To make it clear, we use square and curly bracket for different layers of parenthesis.

$$\begin{aligned}&\{x+[((-1)\times(x+y))+z]\}+y\\ &=\{x+[((-1)\times x)+((-1)\times y)]+z\}+y\ (Distributive\ Law\ )\\ &=\{x+((-1)\times x)+[((-1)\times y)+z]\}+y\ (Associative\ Law\ of\ Addition)\\ &=\{((-1)\times x)+x+\ [((-1)\times y)+z]\}+y(Commutative\ Law\ of\ Addition)\\ &=0+\{[((-1)\times y)+z]+y\}\ (Previous\ Conclusion)\\ &=\{[((-1)\times y)+z]+y\}+0\ (Commutative\ Law\ of\ Addition)\\ &=\{[((-1)\times y)+z]+y\}\ (Addition\ Identity)\\ &=((-1)\times y)+(z+y)\ (Associative\ Law\ of\ Addition)\\ &=(z+y)+((-1)\times y)\ (Commutative\ Law\ of\ Addition)\\ &=z+[y+((-1)\times y)]\ (Associative\ Law\ of\ Addition)\\ &=z+[((-1)\times y)+y]\ (Commutative\ Law\ of\ Addition)\\ &=z+0\ (Previous\ Conclusion)\\ &=z\ (Addition\ Identity)\end{aligned}$$

■

Exercise 6.4 Use truth table to show that De Morgan's Law is correct. ■

Proof. De Morgan's First Law: $\neg(A \wedge B) = \neg A \vee \neg B$

A	B	$A \wedge B$	$\neg(A \wedge B)$	$\neg A \vee \neg B$
0	0	0	1	1
0	1	0	1	1
1	0	0	1	1
1	1	1	0	0

De Morgan's Second Law: $\neg(A \vee B) = \neg A \wedge \neg B$

A	B	$A \vee B$	$\neg(A \vee B)$	$\neg A \wedge \neg B$
0	0	0	1	1
0	1	1	0	0
1	0	1	0	0
1	1	1	0	0

■

Exercise 6.5 Use truth table to show that Absorption Law is correct. ■

Proof.

x	y	$x \vee y$	$x \wedge (x \vee y)$
T	T	T	T
T	F	T	T
F	T	F	F
F	F	F	F

x	y	$x \wedge y$	$x \vee (x \wedge y)$
T	T	T	T
T	F	F	T
F	T	F	F
F	F	F	F

■

Exercise 6.6 Show the truth table of $x \vee ((\neg y) \wedge (\neg z))$. ■

Hint: For Boolean expression of 3 variable, there will be more combinations.

Solution:

Table 6.7: Truth table for the expression $x \vee ((\neg y) \wedge (\neg z))$

x	y	z	$\neg y$	$\neg z$	$(\neg y) \wedge (\neg z)$	$x \vee ((\neg y) \wedge (\neg z))$
0	0	0	1	1	1	1
0	0	1	1	0	0	0
0	1	0	0	1	0	0
0	1	1	0	0	0	0
1	0	0	1	1	0	1
1	0	1	1	0	0	1
1	1	0	0	1	0	1
1	1	1	0	0	0	1

Exercise 6.7 Given the expressions:

1. $(x \to y) \wedge (\neg x \to \neg y)$
2. $((\neg x) \vee y) \wedge (x \vee (\neg y))$
3. $\neg((x \wedge (\neg y)) \vee ((\neg x) \wedge y))$
4. $\neg((x \vee y) \wedge (\neg x \vee \neg y))$

Show that they are equivalent and find the common (simplified) form of the four expression ■

Proof.

$$\begin{aligned}
(x \to y) \wedge (\neg x \to \neg y) &\equiv (\neg x \vee y) \wedge (x \vee \neg y) && \text{(Implication equivalence)}\\
&\equiv (x \vee \neg y) \wedge (\neg x \vee y) && \text{(Commutativity of OR)}\\
&\equiv (x \vee \neg y) \wedge (y \vee \neg x) && \text{(Commutativity of AND)}
\end{aligned}$$

$$\begin{aligned}
((\neg x) \vee y) \wedge (x \vee (\neg y)) &\equiv (y \vee \neg x) \wedge (x \vee \neg y) && \text{(Commutativity of OR)}\\
&\equiv (x \vee \neg y) \wedge (y \vee \neg x) && \text{(Commutativity of AND)}
\end{aligned}$$

$$\begin{aligned}
\neg((x \wedge (\neg y)) \vee ((\neg x) \wedge y)) &\equiv \neg(x \wedge (\neg y)) \wedge \neg((\neg x) \wedge y) && \text{(De Morgan's laws)}\\
&\equiv (\neg x \vee y) \wedge (x \vee \neg y) && \text{(De Morgan's laws)}\\
&\equiv (x \vee \neg y) \wedge (y \vee \neg x) && \text{(Commutativity of OR and AND)}
\end{aligned}$$

$$\begin{aligned}
\neg((x \vee y) \wedge (\neg x \vee \neg y)) &\equiv \neg(x \vee y) \vee \neg(\neg x \vee \neg y) && \text{(De Morgan's laws)}\\
&\equiv (\neg x \wedge \neg y) \vee (x \wedge y) && \text{(De Morgan's laws)}\\
&\equiv (x \wedge y) \vee (\neg x \wedge \neg y) && \text{(Commutativity of OR)}\\
&\equiv (x \vee \neg y) \wedge (y \vee \neg x) && \text{(Distribution)}
\end{aligned}$$

These expressions could be simplified to $(\neg y \vee x) \wedge (\neg x \vee y)$, or $(y \to x) \wedge (x \to y)$, which can also be understood as x and y cannot be true of false in the same time. ■

Exercise 6.8 Show that the concensus theorem's two forms are equivalent.

$$xy \vee \bar{x}z \vee yz = xy \vee \bar{x}z$$

and

$$(x \vee y)(\bar{x} \vee z)(y \vee z) = (x \vee y)(\bar{x} \vee z)$$

are equivalent. ■

Solution: We can simplify the product form to the standard form easily.

$$\begin{aligned}(x \vee y)(\bar{x} \vee z)(y \vee z) &= x\bar{x} \vee xz \vee y\bar{x} \vee yz \vee xy \vee xz \vee y^2 \vee yz \\ &= 0 \vee xz \vee y\bar{x} \vee yz \vee xy \vee xz \vee y \vee yz \\ &= xz \vee y\bar{x} \vee yz \vee xy \\ &= xy \vee xz \vee y\bar{x}\end{aligned}$$

6.2 Predicates and Quantifiers

6.3 Digital Circuits of Logic

6.4 Logic of Deduction and Induction

In this chapter, we will look into one of the most exciting part of mathematics, and it is also a crucial cornerstone for computer science. The inception of number theory can be traced back to ancient civilizations, but it began to emerge as a distinct mathematical discipline with the work of the Greek mathematician Euclid. His monumental work, "Elements," laid the groundwork for the study of prime numbers and proved fundamental theorems, such as the infinitude of primes, which are still central to number theory today. The nature of whole numbers, especially the intriguing properties of prime numbersthe building blocks of all natural numbershas fascinated mathematicians for centuries. Over time, the field has expanded to include a rich array of topics such as the distribution of primes, the solutions of Diophantine equations, modular arithmetic, and the exploration of number-theoretic functions like the Riemann zeta function. Number theory's initial focus on prime numbers and divisibility has blossomed into a diverse and vibrant branch of mathematics, with applications ranging from cryptography to the theory of chaos.

7.1 Divisibility and Modular Arithmetic

We start with the divisibility of number, as it is the basis of many further topics. Divisibility serves as the cornerstone of number theory due to several pivotal reasons. It is the bedrock upon which the Fundamental Theorem of Arithmetic stands, declaring that each integer greater than 1 is uniquely decomposable into a product of prime numbers. These primes are the elemental units, akin to atoms in chemistry, from which all other numbers are constructed. The concepts of greatest common divisors and least common multiples emerge from divisibility, forming the basis of crucial algebraic structures such as rings and fields, and thereby extending number theory into algebraic realms. Furthermore, the pursuit of integer solutions to equations, known as Diophantine analysis, is deeply reliant on principles of divisibility. In the modern context, divisibility underpins cryptographic systems, leveraging the challenging task of prime factorization to secure digital communication. Moreover, divisibility naturally extends to modular arithmetic, enriching number theory with the study of congruences, which has profound implications across mathematics and its myriad applications. Therefore, the study of divisibility is not just foundational but also a connective thread that weaves through the entire fabric of number theory.

7.1.1 Division and Divisibility

To figure out problems related to **Divisibility**, we must figure out the definition and properties of **Division**. We were taught in the primary school that a division is an operation that divides a number to a certain part. For instance, $6 \div 2 = 3$, and $6 \div 4 = 1.5$. Sometimes, the division produce an integer as result, while sometimes not. We define division as follows:

Definition 7.1 — Division. The division of a number a by a non-zero number b is denoted by $a \div b$ or $\frac{a}{b}$, and it gives the quotient q and possibly a remainder r. The operation can be written as:

$$a = b \times q + r$$

where $0 \leq r < b$ and $q \in \mathbb{Z}$.

In $6 \div 2 = 3$, clearly we have $6 = 2 \times 3 + 0$, with $r = 0$, $q = 3$. But how could we explain $6 \div 4 = 1.5$? The quotient is a decimal instead of an integer. Fundamentally, we cannot separate the number 6 in to 4 parts to be represented as integer. In this case the division can be represented as $6 = 4 \times 0 + 6$, meaning $q = 0$ and $r = 6$. Actually, $6 \div 4 = 1.5$ is not defined as only division, but **True Division**.

Definition 7.2 — True Division. True division is concerned with the quotient including the remainder as a fractional part. In true division, when a is divided by b, the quotient q is a real number that can be represented as:

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

You may find that true division doesn't have a remainder. This is just because the purpose of these two operations are different, as true division only focus on the scale of each part of a, but division involves the divisibility and remainder, which will be discussed further in this chapter.

Now that we have defined division and true division, you may understand the following definition of divisibility easier.

Definition 7.3 — Divisibility. If a and b are integers with $a \neq 0$, we say that a divides b if there is an integer c such that $b = a \times c$ (or equivalently, if b a is an integer). When a divides b we say that a is a factor or divisor of b, and that b is a multiple of a. The notation a | b denotes that a divides b. We write $a \nmid b$ when a does not divide b

For example, we have $4 \mid 20$, because $20 \div 4$ gives an integer, however, $4 \nmid 21$, as the answer cannot be represented as integer. Besides, we have:

$$a \mid b \iff \exists c(ac = b). \tag{7.1}$$

Problem 7.1 Let n and d be positive integers. How many positive integers not exceeding n are divisible by d?

Solution: The positive integers divisible by d are all the integers of the form dk, where k is a positive integer. Hence, the number of positive integers divisible by d that do not exceed n equals the number of integers k with $0 < dk \leq n$, or with $0 < k \leq \frac{n}{d}$. Therefore, there are $\left\lfloor \frac{n}{d} \right\rfloor$ positive integers not exceeding n that are divisible by d.

Divisibility holds the following properties, which could be proven directly.

Theorem 7.1 — Properties of Divisibility. Let a, b, and c be integers, where $a \neq 0$, then:

1. if $a \mid b$ and $a \mid c$, then $a \mid (b+c)$
2. if $a \mid b$ then $a \mid bc$ for all integers c
3. if $a \mid b$ and $b \mid c$, then $a \mid c$

Proof. Since $a \mid b$ and $a \mid c$, there exist integers m and n such that $b = am$ and $c = an$. Therefore, $b + c = am + an = a(m+n)$, and since $m+n$ is an integer, it follows that $a \mid (b+c)$.

Given $a \mid b$, there exists an integer k such that $b = ak$. For any integer c, $bc = a(kc)$. Since kc is an integer (because the product of two integers is an integer), $a \mid bc$.

If $a \mid b$ and $b \mid c$, then there exist integers p and q such that $b = ap$ and $c = bq$. Substituting the expression for b into the equation for c gives $c = (ap)q = a(pq)$. Since pq is an integer, $a \mid c$. ■

With this we have the following conclusion.

Corollary 7.1 if a, b, and c are integers, where $a \neq 0$, such that $a \mid b$ and $a \mid c$, then $a \mid mb + nc$ whenever m and n are integers.

This could be proven in direct proof, which you will finish in the exercise.

7.1.2 Modular Arithmetic

In this section, we will focus on the remainder of division, as well as modular arithmetic. In the definition of division, we involved q and r, which denotes the quotient and the remainder of the operation. We use the following notations to denote each of both.

▪ **Notation 7.1 — div and mod.** For $a, b \in \mathbb{R}$. $a = b \times q + r$

$$q = a \textbf{ div } b$$

$$r = a \textbf{ mod } b$$

Besides, when a is an integer and b is a positive integer, we have a **div** $b = \lfloor a/b \rfloor$.

▪ **Example 7.1** What are the quotient and remainder when 93 is divided by 9? ▪

Solution: $93 = 9 \times 10 + 3$. $q = 10$, $r = 3$.
The quotient is 93 **div** $9 = 10 = \lfloor 93/9 \rfloor = \lfloor 10.3333\ldots \rfloor = 10$
The remainder is 93 **mod** $9 = 3 = 93 - 90$

▪ **Example 7.2** What are the quotient and remainder when -93 is divided by 9? ▪

Solution: $-93 = 9 \times (-11) + 6$. $q = -11$, $r = 6$.

Remember that we must make sure $r \geq 0$ as we defined earlier, even though $-93 = 9 \times (-10) - 3$. But remainder could be positive in other division algorithm, which we will discuss in the exercise.

We have already introduced the notation a **mod** m to represent the remainder when an integer a is divided by the positive integer m. We now introduce a different, but related, notation that indicates that two integers have the same remainder when they are divided by the positive integer m. But why we need to find and study the numbers with the same remainder by dividing the same positive integer? Studying numbers that yield the

same remainder when divided by a given positive integer is a fundamental part of number theory; later you will understand why say so.

Definition 7.4 If a and b are integers and m is a positive integer, then a is **congruent** to b **modulo** m if m divides $a-b$. We use the notation $a \equiv b(\bmod m)$ to indicate that a is congruent to b modulo m. We say that $a \equiv b(\bmod m)$ is a congruence and that m is its modulus (plural moduli). If a and b are not congruent modulo m, we write $a \not\equiv b(\bmod m)$.

Do note that mod and **mod** are different notations. The first represents a relation on the set of integers, whereas the second represents a function. However, they are still related.

Theorem 7.2 Let a and b be integers, and let m be a positive integer. $a \equiv b(\bmod\ m)$ if and only if a **mod** m = b **mod** m.

Proof. First, suppose $a \equiv b \pmod{m}$. By definition of congruence modulo m, m divides $a-b$, which means there exists some integer k such that $a-b=km$.

Dividing a and b by m, they both leave the same remainder r, since $a = q_1m+r$ and $b = q_2m+r$ for some integers q_1 and q_2. The remainder r in both cases is the same because the difference $a-b$ is a multiple of m, which does not affect the remainder.

Conversely, if $a \mod m = b \mod m$, then both a and b leave the same remainder when divided by m. Denote this common remainder as r.

We can write $a = q_1m+r$ and $b = q_2m+r$ for some integers q_1 and q_2. Subtracting these two equations, we get $a-b = (q_1-q_2)m$, which shows that $a-b$ is a multiple of m.

Therefore, m divides $a-b$, and by definition of congruence modulo m, we have $a \equiv b \pmod{m}$.

This completes the proof. ■

R Remember, when we say if and only if, we need proof from each of the both statements to the other.

Theorem 7.3 Let m be a positive integer. The integers a and b are congruent modulo m if and only if there is an integer k such that $a = b+km$.

The proof is similar and not complex, try to prove it in the exercise.

Theorem 7.4 Let m be a positive integer. If $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then

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

and

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

Proof. We use a direct proof. Because $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, by Theorem 4 there are integers s and t with $b = a+sm$ and $d = c+tm$. Hence,

$$b+d = (a+sm)+(c+tm) = (a+c)+m(s+t)$$

and

$$bd = (a+sm)(c+tm) = ac+m(at+cs+stm).$$

Hence,

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

and

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

■

▪ **Example 7.3** Because $7 \equiv 2 \pmod{5}$ and $11 \equiv 1 \pmod{5}$, it follows that

$$18 = 7+11 \equiv 2+1 = 3 \pmod{5}$$

and that

$$77 = 7 \cdot 11 \equiv 2 \cdot 1 = 2 \pmod{5}.$$

▪

Corollary 7.2 Let m be a positive integer and let a and b be integers. Then

$$(a+b) \bmod m = ((a \bmod m)+(b \bmod m)) \bmod m$$

and

$$ab \bmod m = ((a \bmod m)(b \bmod m)) \bmod m.$$

Proof. By the definitions of mod and of congruence modulo m, we know that $a \equiv (a \bmod m) \pmod{m}$ and $b \equiv (b \bmod m) \pmod{m}$. Hence,

$$a+b \equiv (a \bmod m)+(b \bmod m) \pmod{m}$$

and

$$ab \equiv (a \bmod m)(b \bmod m) \pmod{m}.$$

The equalities in this corollary follow from these last two congruence. ■

We can define arithmetic operations on $\mathbb{Z}_m$, the set of nonnegative integers less than m, that is, the set $\{0,1,\ldots,m-1\}$. In particular, we define addition of these integers, denoted by $\oplus_m$, by

$$a \oplus_m b = (a+b) \bmod m,$$

where the addition on the right-hand side of this equation is the ordinary addition of integers, and we define multiplication of these integers, denoted by $\odot_m$, by

$$a \odot_m b = (a \cdot b) \bmod m,$$

where the multiplication on the right-hand side of this equation is the ordinary multiplication of integers. The operations $\oplus_m$ and $\odot_m$ are called addition and multiplication modulo m and when we use these operations, we are said to be doing arithmetic modulo m.

■ **Example 7.4** Use the definition of addition and multiplication in $\mathbb{Z}_m$ to find $7 \oplus_{11} 9$ and $7 \odot_{11} 9$.

Solution: Using the definition of addition modulo 11, we find that

$$7 \oplus_{11} 9 = (7+9) \bmod 11 = 16 \bmod 11 = 5,$$

and

$$7 \odot_{11} 9 = (7 \cdot 9) \bmod 11 = 63 \bmod 11 = 8.$$

Hence, $7 \oplus_{11} 9 = 5$ and $7 \odot_{11} 9 = 8$. ■

R sometimes, normal notations are also used to express this calculation with subscript.

7.1.3 Exercises

Exercise 7.1 Prove corollary **??** that if a, b, and c are integers, where $a \neq 0$, such that $a \mid b$ and $a \mid c$, then $a \mid mb + nc$ whenever m and n are integers. ■

Proof. By theorem **??**, $a \mid b$ and $a \mid c$ gives us $a \mid mb$ and $a \mid nc$ (property 2). Hence, $a \mid (mb + nc)$ (property 1). This completes the proof. ■

Exercise 7.2 Prove theorem **??**, you may use other theorems or corollary in this chapter. ■

Proof. If $a \equiv b (\bmod\ m)$, that means we have $m \mid (a-b)$. There must be a integer k, such that $a = b + km$ (theorem **??**). Conversely, if we have a integer k, such that $a = b + km$, we have $km = a - b$. Hence, $m \mid (a-b)$, $a \equiv b (\bmod\ m)$. ■

Exercise 7.3 show that for integer $a,\ b,\ c\ \in \mathbb{Z}^+$, $a, c \neq 0$ and $ac \mid bc$,then $a \mid b$. ■

Proof. $ac \mid bc$ means that $bc = k(ac)$ for some integer k, so we have $b = ka$, this means $a \mid b$. ■

Exercise 7.4 Prove that if a and b are integers and a divides b, then a is odd or b is even. ■

Proof. We could try proof by contrapositive. Suppose for a is even and b is odd, $a, b \in \mathbb{Z}$, $a \mid b$. Then $b = ka$ for some integer k. Make $a = 2n$, $b = 2n+1$, $n \in \mathbb{Z}$, then there should be $2n+1 = 2kn$. However, in this case we have $k = \dfrac{2n+1}{2n}$, meaning that k is not an integer, which contradict the assumption. This completes the proof. ■

Exercise 7.5 Prove that if a is a positive integer, then $4 \nmid (a^2+2)$. ■

Proof. Consider any positive integer a. We know that a can be expressed in one of the following forms where n is an integer:

- $a = 4n$
- $a = 4n+1$
- $a = 4n+2$

- $a = 4n+3$

Now, we examine a^2 modulo 4 for each case:

$$\begin{aligned}
(4n)^2 &= 16n^2 = 4\cdot 4n^2 \equiv 0 \pmod 4,\\
(4n+1)^2 &= 16n^2+8n+1 = 4(4n^2+2n)+1 \equiv 1 \pmod 4,\\
(4n+2)^2 &= 16n^2+16n+4 = 4(4n^2+4n+1) \equiv 0 \pmod 4,\\
(4n+3)^2 &= 16n^2+24n+9 = 4(4n^2+6n+2)+1 \equiv 1 \pmod 4.
\end{aligned}$$

Adding 2 to a^2 in each case yields:

$$\begin{aligned}
a^2+2 &\equiv 2 \pmod 4 \quad \text{if } a^2 \equiv 0 \pmod 4,\\
a^2+2 &\equiv 3 \pmod 4 \quad \text{if } a^2 \equiv 1 \pmod 4.
\end{aligned}$$

In both cases, a^2+2 does not yield a remainder of 0 modulo 4. Thus, it cannot be divisible by 4.

Therefore, we have proven that for any positive integer a, the expression a^2+2 is not divisible by 4. ■

Exercise 7.6 Suppose that a and b are integers, $a \equiv 11 \mod 19$, and $b \equiv 3 \mod 19$. Find the integer c with $0 \le c \le 18$ such that

1. $c \equiv 13a \mod 19$.
2. $c \equiv 8b \mod 19$.
3. $c \equiv a-b \mod 19$.
4. $c \equiv 7a+3b \mod 19$.
5. $c \equiv 2a^2+3b^2 \mod 19$.
6. $c \equiv a^3+4b^3 \mod 19$.

■

Solution: This problem is equivalent to asking for the right-hand side mod 19. So we just do the arithmetic and compute the remainder upon division by 19.

1. $13\cdot 11 = 143 \equiv 10 \mod 19$
2. $8\cdot 3 = 24 \equiv 5 \mod 19$
3. $11-3 = 8 \mod 19$
4. $7\cdot 11+3\cdot 3 = 86 \equiv 10 \mod 19$
5. $2\cdot 11^2+3\cdot 3^2 = 269 \equiv 3 \mod 19$
6. $11^3+4\cdot 3^3 = 1439 \equiv 14 \mod 19$

Exercise 7.7 Let m be a positive integer. Show that $a \mod m = b \mod m$ if $a \equiv b \pmod m$. ■

Proof. Assume that $a \equiv b \pmod m$. This means that $m \mid a-b$, say $a-b = mc$, so that $a = b+mc$. Now let us compute $a \mod m$. We know that $b = qm+r$ for some nonnegative r less than m (namely, $r = b \mod m$). Therefore, we can write $a = qm+r+mc = (q+c)m+r$. By definition, this means that r must also equal $a \mod m$. That is what we wanted to prove. ■

Exercise 7.8 Show that if $a \equiv b(\bmod n)$ and $d \mid a, d \mid b$, $d \mid n$, then

$$\frac{a}{d} \equiv \frac{b}{d}(\bmod \frac{n}{d})$$

■

Proof. Since $a \equiv b \mod n$, by definition there exists an integer k such that $a = b + kn$. Now, since $d \mid a$ and $d \mid b$, we have $a = dm_1$ and $b = dm_2$ for some integers m_1 and m_2. Also, $d \mid n$ implies $n = dl$ for some integer l.

Substituting these into our first equation we get

$$dm_1 = dm_2 + k(dl)$$

Dividing through by d we obtain

$$m_1 = m_2 + kl$$

This can be rewritten as

$$\frac{a}{d} = \frac{b}{d} + k\left(\frac{n}{d}\right)$$

Since $k\left(\frac{n}{d}\right)$ is an integer, this shows that $\frac{a}{d}$ is congruent to $\frac{b}{d}$ modulo $\frac{n}{d}$, which is precisely

$$\frac{a}{d} \equiv \frac{b}{d} \mod \frac{n}{d}$$

Hence, the statement is proven. ■

7.2 Number Representations and Algorithms

In everyday life, we use decimal notation to express integers. Though this is not for all cases, as we are using a 60 base system for time. However, in computer science, binary, octal, and hexadecimal systems are widely used for number representation. The reason is that binary system consist of only 0 and 1, as we have mentioned in Boolean algebra, this system is perfect for logic operations in side the computer. What's more, 0 and 1 could be easily represented by the on and off of tiny switches in the integrated circuits. This chapter looks into the representation of number in different bases and their relationship, with a tress on binary representation that computer system uses. Meanwhile, we will go through algorithm of some operations between number, and analyze them accordingly.

7.2.1 Representations of Numbers and Base Conversion

In whichever base, a number could be expressed using the exponent of the base.

Theorem 7.5 — Representation of Number. Let b be an integer greater than 1. Then if n is a positive integer, it can be expressed uniquely in the form:
$n = a_k b^k + a_{k-1} b^{k-1} + \cdots + a_1 b + a_0$
where k is a nonnegative integer, $a_0, a_1, \cdots, a_k$ are negative integers less than b, and $a_k \neq 0$

For example, 1024 could be interpreted as $1024 = 1 \times 10^3 + \times 10^2 + 2 \times 10^1 + 4 \times 10^0$. Besides, the base of number constraints the possible numbers to be used on the bits. For example, under decimal system, we cannot take 10 as one digit, instead it is a number with two digits 1 and 0, because for bits under decimal system, the maximum only goes to 9. Therefore, you could take it as a fact that for a system of base b, the biggest value for one bit is $b-1$. Similarly, we see only 0 and 1 in binary, 1-7 in ocatl, 1-15 in hexadecimal expression.

To indicate the base of a number, we use the following subscript.

▪ **Notation 7.2 —** $xxxx_b$. We use $_b$ to show the base of a number, where b is the base number.

The other important fact is that numbers are just numbers, base is only the way we gauge them. However you change 121_{10}, to whatever base, it is still 121 in decimal, but just expressed in a different way.

Binary, Octal, and Hexadecimal Expression

In **binary expansion**, the base on any integer is 2, that is to say, the number in each digit is either 0 or 1. We can expand binary integer in the same ways as mentioned in last section.

▪ **Example 7.5** What is the decimal expansion of the integer that has $(101011111)_2$ as its binary expansion?

$(101011111)_2$ has nine digits, so we have:
$(101011111)_2 = 1 \times 2^9 + 1 \times 2^7 + 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 351$

▪

Decimal expansion of octal expansions could be calculated in the same way as the binary expansion. However, Sixteen different digits are required for hexadecimal expansions. Usually, the hexadecimal digits used are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, and F, where the letters A through F represent the digits corresponding to the numbers 10 through 15 (in decimal notation).

▪ **Example 7.6** What is the decimal expansion of the number with hexadecimal expansion $(2AE0B)_{16}$? ▪

In hexadecimal cases, the only difference is that capital latters are used to represent two-digit number in one digit. We still have:
$(2AE0B)_{16} = 2 \times 16^4 + 10 \times 16^3 + 14 \times 16^2 + 0 \times 16^1 + 11 \times 16^0 = 175627$

NOTICE: Each hexadecimal digit can be represented using four bits. For instance, we see that $(11100101)_2 = (E5)_{16}$ because $(1110)_2 = (E)_{16}$ and $(0101)_2 = (5)_{16}$. Bytes, which are bit strings of length eight, can be represented by two hexadecimal digits.

7.2.2 Base Conversion

Base conversion of Integer

Now we have learned how to express numbers in different base systems, but is there any way to convert them from one base to the other? You may have realized that a general solution is to convert the number back to decimal form, and then to the target base. Yes, it works, but can we make it better? One way to this is to use the general base conversion

algorithm. The most common algorithm to constructing the base b expansion of an integer n is as follows:
First, divide n by b to obtain a quotient and remainder, that is:

$$n = bq_0 + a_0 \qquad (0 \leq a_0 < b)$$

The remainder, a_0, is the rightmost digit in the base b expansion of n. Next, divide q_0 by b toobtain:

$$q_0 = bq_1 + a_1 \qquad (0 \leq a_1 < b)$$

We see that a_1 is the second digit from the right in the base b expansion of n. Then, we just continue this process until we obtain a quotient equal to zero. This algorithm produces the base b digits of n from the right to the left.

▪ **Example 7.7** Find the hexadecimal expansion of $(177130)_{10}$.

$$\begin{aligned} 177130 &= 16 \cdot 11070 + 10 \\ 11070 &= 16 \cdot 691 + 14 \\ 691 &= 16 \cdot 43 + 3 \\ 43 &= 16 \cdot 2 + 11 \\ 2 &= 16 \cdot 0 + 2 \end{aligned}$$

Therefore, $(177130)_{10} = (2B3EA)_{16}$ ▪

This method is equivalent to the following pseudocode.

Algorithm 4 Base conversion of an integer n to base b

Require: An integer n and a base b to convert to
Ensure: The base b expansion of n

```
function ConvertToBase(n, b)
    digits ← empty list
    while n > 0 do
        remainder ← n mod b
        Append remainder to digits
        n ← ⌊n/b⌋
    end while
    digits ← reverse digits
    return digits
end function
```

Base Conversion of Float

The previous algorithm only deals with integers, so we need a different approach for decimals. The algorithm for converting a floating-point number from base 10 to base b is given as follows:

▪ **Example 7.8** Convert the floating-point number 12.375 from base 10 to base 2.

- Integer part: 12 in base 10 is 1100 in base 2.

Algorithm 5 Convert a floating-point number to a different base

```
function CONVERTFLOATTOBASE(number, targetBase)
    integerPart ← ⌊number⌋
    fractionalPart ← number − integerPart
    baseInteger ← CONVERTINTEGERTOBASE(integerPart, targetBase)
    baseFraction ← ""
    while fractionalPart > 0 and length of baseFraction is less than limit do
        fractionalPart ← fractionalPart × targetBase
        baseFraction ← baseFraction + ⌊fractionalPart⌋
        fractionalPart ← fractionalPart − ⌊fractionalPart⌋
    end while
    return baseInteger + "." + baseFraction
end function
```

- Fractional part: 0.375 in base 10 to base 2.

$$0.375 \times 2 = 0.75 \rightarrow \text{integer part is } 0$$
$$0.75 \times 2 = 1.5 \rightarrow \text{integer part is } 1$$
$$0.5 \times 2 = 1 \rightarrow \text{integer part is } 1$$

- The fractional part in base 2 is .011.

Therefore, 12.375_{10} is 1100.011_2. ■

Also, when a number consist of both integer and decimal parts, we just deal with them separately and them put them back together later. You will see these exercises in the problem set.

Base Conversion between Binary, Octal, and Hexadecimal number

The base conversion between Binary, Octal and Hexadecimal numbers are much easier.

TABLE 1 **Hexadecimal, Octal, and Binary Representation of the Integers 0 through 15.**

Decimal	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Hexadecimal	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
Octal	0	1	2	3	4	5	6	7	10	11	12	13	14	15	16	17
Binary	0	1	10	11	100	101	110	111	1000	1001	1010	1011	1100	1101	1110	1111

Figure 7.1: Binary, Octal and Hexadecimal Representation

The conversion between binary, octal, and hexadecimal systems is straightforward because the bases of these systems (2 for binary, 8 for octal, and 16 for hexadecimal) are related by powers of 2. Specifically:

- Octal digits correspond to three binary digits since $8 = 2^3$. Each octal digit can be directly mapped to a unique combination of three binary bits.
- Hexadecimal digits correspond to four binary digits since $16 = 2^4$. Each hexadecimal digit can be directly mapped to a unique combination of four binary bits.

This relationship allows for simple grouping of binary bits into sets of three or four to convert to octal or hexadecimal, respectively, without any complex calculation or division.

■ **Example 7.9** Find the octal and hexadecimal expansions of $(11\ 1110\ 1011\ 1100)_2$ and the binary expansions of $(765)_8$ and $(A8D)_{16}$. ■

Solution: To convert $(11\ 1110\ 1011\ 1100)_2$ into octal notation we group the binary digits into blocks of three, adding initial zeros at the start of the leftmost block if necessary. These blocks, from left to right, are 011, 111, 010, 111, and 100, corresponding to 3, 7, 2, 7, and 4 respectively. Consequently, we get $(11\ 1110\ 1011\ 1100)_2 = (37274)_8$

We do the hexadecimal convention in the same way. These blocks, from left to right, are 0011, 1110, 1011, and 1100, corresponding to the hexadecimal digits 3, E, B, and C, respectively. Consequently, $(11\ 1110\ 1011\ 1100)_2 = (3EBC)_{16}$

NOTE: The conversion could be done inversely in the same way. If the number of bit is not the multiple of 4 when you are converting between 2 and 16 base number, fill those missing digits/bits with 0.

Whether you noticed or that the base conversion, essentially, is about division and modular arithmetic that we just learned? The general base conversion algorithm could be simplified to the following algorithm, where n denotes the number to be converted and b is the target base.

Algorithm 6 Constructing Base b Expansions

```
procedure BASE b EXPANSION(n, b: positive integers with b > 1)
    q ← n
    k ← 0
    while q ≠ 0 do
        a_k ← q mod b
        q ← ⌊q/b⌋
        k ← k + 1
    end while
    return (a_{k-1}, ..., a_1, a_0)          ▷ The base b expansion of n
end procedure
```

7.2.3 Operation Algorithms of Number

This section discusses algorithm of basic operation between numbers, of base2. In section, we express the binary number in the form of

$$a = (a_{n-1}a_{n-2}\ldots a_1a_0)_2, b = (b_{n-1}b_{n-2}\ldots b_1b_0)_2$$

where a and b each have n bits.

Addition Algorithm

Consider the problem of adding two integers in binary notation. A procedure to perform addition can be based on the usual method for adding numbers with pencil and paper. This method proceeds by adding pairs of binary digits together with carries, when they occur, to compute the sum of two integers. This procedure will now be specified in detail.

To add a and b, first add their rightmost bits. This gives

$$a_0 + b_0 = c_0 \cdot 2 + s_0,$$

where s_0 is the rightmost bit in the binary expansion of $a+b$ and c_0 is the carry, which is either 0 or 1. Then add the next pair of bits and the carry,

$$a_1 + b_1 + c_0 = c_1 \cdot 2 + s_1,$$

where s_1 is the next bit (from the right) in the binary expansion of $a+b$, and c_1 is the carry. Continue this process, adding the corresponding bits in the two binary expansions and the carry, to determine the next bit from the right in the binary expansion of $a+b$. At the last stage, add a_{n-1}, b_{n-1}, and c_{n-2} to obtain $c_{n-1} \cdot 2 + s_{n-1}$. The leading bit of the sum is $s_n = c_{n-1}$. This procedure produces the binary expansion of the sum, namely, $a+b = (s_n s_{n-1} \ldots s_1 s_0)_2$. Consider the binary addition of two numbers, 1011_2 and 1101_2.

The addition process is similar to that used in decimal addition, but it is performed in base 2. Below is the columnar addition process:

$$\begin{array}{rr} & 1011_2 \\ + & 1101_2 \\ \hline & 11000_2 \end{array}$$

Let's perform the addition step by step:

1. Start with the rightmost bits (least significant bits). Add $1+1$. Since this is base 2, $1+1 = 10_2$. Write down the 0 and carry the 1 over to the next column.
2. Move to the next column. Add $1+0+1$ (including the carry). This equals 10_2. Write down the 0 and carry the 1.
3. In the next column, add $0+1+1$. This equals 10_2. Write down the 0 and carry the 1.
4. For the leftmost bits (most significant bits), add $1+1+1$. This equals 11_2. Write down the 1 and carry the 1 to a new column to the left.
5. Write down the carry.

The final result is 11000_2.

This algorithm's pseudocode is as listed below.

Algorithm 7 Addition of Integers

```
procedure ADD(a, b: positive integers)
    the binary expansions of a and b are (a_{n-1} a_{n-2} ... a_1 a_0)_2
    and (b_{n-1} b_{n-2} ... b_1 b_0)_2, respectively
    c ← 0
    for j ← 0 to n − 1 do
        d ← ⌊(a_j + b_j + c)/2⌋
        s_j ← a_j + b_j + c − 2d
        c ← d
    end for
    s_n ← c
    return (s_n s_{n-1} ... s_1 s_0)_2          ▷ the binary expansion of the sum
end procedure
```

Time Complexity: The time complexity of the binary addition algorithm is $O(n)$, where n is the number of bits in the binary representation of the inputs.

Proof. Consider the binary addition algorithm which consists of a single loop that iterates n times, where n is the number of bits in the binary representations of the two integers being added.

Within the loop, the algorithm performs a constant number of operations for each bit:

- An addition of the j-th bits of the two numbers $a_j + b_j$.
- An addition of the carry from the previous step c.
- A division by 2 to compute the new carry d.
- A subtraction to determine the j-th bit of the sum s_j.
- An assignment of the new carry $c \leftarrow d$.

Since each of these operations has a constant time complexity, and they are all executed once for each bit, the overall time complexity of the loop is linear with respect to the number of bits. Therefore, the time complexity of the entire algorithm is $O(n)$.

Note that this analysis assumes that basic arithmetic operations (addition, division by 2, subtraction) can be performed in constant time. ■

Multiplication Algorithm

Using the same representation of binary number a and b, we can define binary multiplication as

$$\begin{aligned} ab &= a\left(b_0 2^0 + b_1 2^1 + \cdots + b_{n-1} 2^{n-1}\right) \\ &= a\left(b_0 2^0\right) + a\left(b_1 2^1\right) + \cdots + a\left(b_{n-1} 2^{n-1}\right) \end{aligned}$$

. The binary multiplication algorithm works similarly to traditional pencil-and-paper multiplication, but with binary digits. Given two binary numbers, we multiply each bit of the second number by the first number, shifting the result left for each subsequent bit.

Algorithm 8 Addition of Integers

```
procedure ADD(a, b: positive integers)
    the binary expansions of a and b are (a_{n-1} a_{n-2} ... a_1 a_0)_2
    and (b_{n-1} b_{n-2} ... b_1 b_0)_2, respectively
    c ← 0
    for j ← 0 to n − 1 do
        d ← ⌊(a_j + b_j + c)/2⌋
        s_j ← a_j + b_j + c − 2d
        c ← d
    end for
    s_n ← c
    return (s_n s_{n-1} ... s_1 s_0)_2          ▷ the binary expansion of the sum
end procedure
```

■ **Example 7.10 Find the product of $a = (110)_2$ and $b = (101)_2$.** First note that

$$a \cdot b_0 \cdot 2^0 = (110)_2 \cdot 1 \cdot 2^0 = (110)_2,$$

$$a \cdot b_1 \cdot 2^1 = (110)_2 \cdot 0 \cdot 2^1 = (0000)_2,$$

and

$$a \cdot b_2 \cdot 2^2 = (110)_2 \cdot 1 \cdot 2^2 = (11000)_2.$$

To find the product, add $(110)_2, (0000)_2$, and $(11000)_2$. Carrying out these additions (using Algorithm 2, including initial zero bits when necessary) shows that $ab = (11110)_2$.

Time Complexity: The time complexity of binary multiplication is $O(n^2)$, where n is the number of bits in the binary numbers. This is because each bit of one number is multiplied by each bit of the other number, resulting in n multiplications for each of the n bits. ∎

Algorithm for Div and Mod

The following algorithm is used to find the quotient and remainder of $a \div b$. This is a more general algorithm as it could handle the cases where a is negative

Algorithm 9 Computing div and mod.

```
procedure DIVISION_ALGORITHM(a: integer, d: positive integer)
    q ← 0
    r ← |a|
    while r ≥ d do
        r ← r − d
        q ← q + 1
    end while
    if a < 0 and r > 0 then
        r ← d − r
        q ← −(q + 1)
    end if
    return (q, r)          ▷ q = a ÷ d is the quotient, r = a mod d is the remainder
end procedure
```

7.2.4 Exercises

7.3 Primes and Greatest Common Divisors

In previous chapter, we learned the properties of division and divisibility, as well as modular arithmetic. These are the most essential parts of the whole number theory. In this section , we will discuss primes and its property. Prime is not something unfamiliar to us, since some may even have learned that from the kindergarten or primary school. We will look into the Algorithms that could help is find primes and delve into the algebraic properties of prime later. All these are foundations of cryptography.

7.3.1 Primes and Related Algorithms

To learn prime, of course we need to recap on its definition.

Definition 7.5 — Prime Number. An integer p greater than 1 is called prime if the only positive factors of p are 1 and p. A positive integer that is greater than 1 and is not prime is called **composite**.

Note that, by this definition, 1 is not a prime, as it is only 1 as the only factor.

The reason why prime numbers are so fascinating to mathematicians is that they have so many interesting and unique property, even except for what we have seen in its definition. Below is what we call the fundamental theorem of arithmetic.

Theorem 7.6 — The Fundamental Theorem of Arithmetic. For every integer $n > 1$, there exists a unique factorization into prime numbers, up to the order of the factors. Specifically, n can be expressed as

$$n = p_1^{a_1} \cdot p_2^{a_2} \cdot \ldots \cdot p_k^{a_k}$$

where $p_1 < p_2 < \ldots < p_k$ are prime numbers and $a_1, a_2, \ldots, a_k$ are positive integers. This factorization is unique, apart from the order of the prime factors.

Proof. We prove the theorem in two parts: existence and uniqueness.

Existence: We prove by mathematical induction that every integer greater than 1 can be written as a product of primes.

Base Case: For $n = 2$, the statement holds true since 2 is itself a prime number.

Inductive Step: Assume the statement holds for all integers greater than 1 and less than n. Now consider the integer n.

- If n is prime, then it is trivially a product of primes (itself).
- If n is not prime, it can be written as $n = a \cdot b$ where $1 < a, b < n$. By the inductive hypothesis, both a and b can be factored into a product of primes. Therefore, n can also be expressed as a product of primes by combining the prime factorization of a and b.

This completes the proof of existence.

Uniqueness: Assume, for the sake of contradiction, that there are two distinct prime factorization of n:

$$n = p_1^{a_1} \cdot p_2^{a_2} \cdot \ldots \cdot p_k^{a_k} = q_1^{b_1} \cdot q_2^{b_2} \cdot \ldots \cdot q_m^{b_m}$$

where p_i and q_j are prime numbers, and a_i, b_j are positive integers. To proceed to the rest of the proof, we need to use Euclid's lemma.

lemma:

Let p be a prime number. If p divides the product ab, where a and b are integers, then p divides a or p divides b.

Proof. Assume p is a prime that divides ab but does not divide a. We need to show that p must divide b.

Since p does not divide a, the greatest common divisor (gcd) of a and p is 1, i.e., $\gcd(a, p) = 1$. According to Bezout's identity, there exist integers x and y such that:

$$ax + py = 1$$

Multiplying both sides of the equation by b, we get:

$$abx + pby = b$$

R If this proof is not yet understandable for you, skip it, and check it back after we go over the rest of this chapter about linear combination and this theorem.

Since p divides ab (by assumption), p divides abx. Also, p obviously divides pby. Hence, p divides the sum $abx + pby$, which means p divides b.

This completes the proof, showing that if p divides ab and does not divide a, then p must divide b, in accordance with Euclid's lemma. ■

By Euclid's lemma, if a prime divides the product of two numbers, it must divide at least one of those numbers. Thus, p_1 must divide some q_j on the right-hand side. Since q_j is prime, we conclude that $p_1 = q_j$. Applying this argument symmetrically and repeatedly, we find that each set of prime factors must be identical to the other, contradicting the assumption of two distinct factorization.

Therefore, the prime factorization of any integer greater than 1 is unique, up to the order of the factors, which completes the proof of the Fundamental Theorem of Arithmetic. ■

▪ **Example 7.11** 100 can be taken as $100 = 2 \cdot 2 \cdot 5 \cdot 5 = 2^2 \cdot 5^2$. Both 2 and 5 are primes. ▪

Now that we know these interesting properties of prime and its significance, how do we find primes, not just correctly, but also efficiently? The most basic algorithm that anyone could find is that we can actually check numbers one by one. To find whether an integer n is prime, what we need to do is to check every number from 2 to $n-1$, and use them to divide n one by one. If anyone of them divides n, then n is not prime, and vice versa.

Trial Division

We actually can apply a more efficient way to determine whether an integer is prime or not. Considering the following theorem.

Theorem 7.7 If n is a composite integer, then n has a prime divisor less than or equal to $\sqrt{n}$.

Proof. If n is composite, by the definition of a composite integer, we know that it has a factor a with $1 < a < n$. Hence, by the definition of a factor of a positive integer, we have $n = ab$, where b is a positive integer greater than 1. We will show that $a \leq \sqrt{n}$ or $b \leq \sqrt{n}$. If $a > \sqrt{n}$ and $b > \sqrt{n}$, then $ab > \sqrt{n} \cdot \sqrt{n} = n$, which is a contradiction. Consequently, $a \leq \sqrt{n}$ or $b \leq \sqrt{n}$. Because both a and b are divisors of n, we see that n has a positive divisor not exceeding $\sqrt{n}$. This divisor is either prime or, by the fundamental theorem of arithmetic, has a prime divisor less than itself. In either case, n has a prime divisor less than or equal to $\sqrt{n}$. ■

Below is the pseudocode for this procedure.

Algorithm 10 Primality Test Using Trial Division

```
function IsPrime(n)
    if n ≤ 1 then
        return false
    end if
    for i ← 2 to n − 1 do
        if n mod i = 0 then
            return false
        end if
    end for
    return true
end function
```

■ **Example 7.12** Show that $\sqrt{105}$ is prime. ■

textbfSolution: Primes less than $\sqrt{105}$ are 2, 3, 5, 7. 101 is not divisible by any of them, so it is not composite, and thus it is prime.

We also need an efficient algorithm to factorize a give number to primes. Here is how is works. It works as follows.

Algorithm 11 Prime Factorization

```
function PRIMEFACTORIZATION(n)
    factors ← an empty list
    for p ← 2 to ∞ do
        while n mod p == 0 do
            Add p to factors
            n ← n/p
        end while
        if p > n/p then
            break
        end if
    end for
    if n > 1 then
        Add n to factors
    end if
    return factors
end function
```

■ **Example 7.13** Find the prime factorization of 8964. ■

Solution:

1. 8964 is even, so start with 2: $8964 \div 2 = 4482$.
2. 4482 is also even, divide by 2 again: $4482 \div 2 = 2241$.
3. 2241 is not divisible by 2. The next primes to try are 3, 5, 7, 11, and so on. We find that 2241 is not divisible by any of these primes until we reach 31.
4. 2241 divided by 31 gives us 71, which is a prime number.

Thus, the prime factorization of 8964 is $2 \times 2 \times 31 \times 71$, or in exponent form, $2^2 \times 31 \times 71$.

There are also other method to finding prime numbers or tell whether a number is prime or not. If you are willing to delve into it, you may check Sieve of Eratosthenes, Sieve of Ktkin. These are only small part of all possible method, which we will learn later, such as Fermat's Little Theorem in solving congruence, as well as Monte Carlo method in probability.

7.3.2 Greatest Common Divisors and Least Common Multiples

Now we will discuss yet another primary school topic, GCD and LCM. Briefly recap their definitions:

Definition 7.6 — Greatest Common Divisior. Let a and b be integers, not both zero. The largest integer d such that $d \mid a$ and $d \mid b$ is called the greatest common divisor of a and b. The greatest common divisor of a and b is denoted by $\gcd(a,b)$.

■ **Example 7.14** Find $\gcd(12,24)$ and $\gcd(17,22)$.

These are easy-to-solve problems. The largest number that divides 12 and 24 is 6, so $\gcd(12,24) = 6$,while 17 and 22 do not have common divisor other than 1, so $\gcd(17,22) = 1$. ■

For 17 and 22, we have $\gcd(17,22) = 1$. In this case, we call that they are **Relatively Prime**, because their greatest common divisor is 1.

We can generalize this idea to pairwise Relatively prime.

Definition 7.7 — Pairwise Relatively Prime. The integers $a_1, a_2, ..., a_n$ are pairwise relatively prime if $\gcd(a_i, a_j) = 1$ whenever $1 \leq i < j \leq n$.

For example, 9, 16, 23 are pairwise relatively prime, becuase $\gcd(9,16) = 0$, $\gcd(9,23) = 0$, $\gcd(16,23) = 0$. So we could say that a sequence of number is pairwise relatively prime if and only if any combination of 2 of the numbers in The sequence a, b, has gcd $(a,b) = 1$.

The other way to find gcd of two number is using their prime factorization. Suppose that the prime factorizations of the positive integers a and b are

$$a = p_1^{a_1} p_2^{a_2} \cdots p_n^{a_n}, b = p_1^{b_1} p_2^{b_2} \cdots p_n^{b_n},$$

. where each exponent is a nonnegative integer, and where all primes occurring in the prime factorization of either a or b are included in both factorizations, with zero exponents if necessary. Then $\gcd(a,b)$ is given by

$$\gcd(a,b) = p_1^{\min(a_1,b_1)} p_2^{\min(a_2,b_2)} \cdots p_n^{\min(a_n,b_n)},$$

where $\min(x,y)$ represents the minimum of the two numbers x and y. To show that this formula for $\gcd(a,b)$ is valid, we must show that the integer on the right-hand side divides both a and b, and that no larger integer also does. This integer does divide both α and b, because the power of each prime in the factorization does not exceed the power of this prime in either the factorization of a or that of b. Further, no larger integer can divide both α and b, because the exponents of the primes in this factorization cannot be increased, and no other primes can be included.

■ **Example 7.15** Find $\gcd(100,250)$.

$$\gcd(120,500) = 2^{\min(3,2)} 3^{\min(1,0)} 5^{\min(1,3)} = 2^2 3^0 5^1 = 20.$$

■

This method could be finetuned to find the least common multiple of two integers. Still according to the fundamental theorem of arithmetic, every integer could find a unique factorization as product of prime numbers.

Definition 7.8 The least common multiple of the positive integers a and b is the smallest positive integer that is divisible by both a and b. The least common multiple of a and b is denoted by $\operatorname{lcm}(a,b)$.

Suppose the prime factorizations of a and b are given by

$$a = p_1^{a_1} p_2^{a_2} \ldots p_n^{a_n}, \quad b = p_1^{b_1} p_2^{b_2} \ldots p_n^{b_n},$$

where a_i and b_i are the exponents of the prime factors of a and b, respectively.

The product ab can be expressed as a prime factorization:

$$ab = p_1^{a_1+b_1} p_2^{a_2+b_2} \dots p_n^{a_n+b_n}.$$

From this, we can deduce that any common multiple of a and b must be a product of their prime factors raised to at least the maximum exponent found in either a or b. Therefore, the $\text{lcm}(a,b)$ can be expressed as

$$\text{lcm}(a,b) = p_1^{\max(a_1,b_1)} p_2^{\max(a_2,b_2)} \dots p_n^{\max(a_n,b_n)}.$$

This ensures that $\text{lcm}(a,b)$ is divisible by both a and b, and it is the smallest such number with this property.

■ **Example 7.16** Find $\text{lcm}(120,500)$.

$$\text{lcm}(120,500) = 2^{\max(2,3)} 3^{\max(1,0)} 5^{\max(1,3)} = 8 \times 3 \times 125 = 3000$$

■

Have you noticed, that ab is actually the product of $\text{lcm}(a,b)$ and $\gcd(a,b)$? Becuase the order of each term in the prime factorization of ab is a sum of two number, what ever which number is the maximum of the minimum, we always have the order of $a_n + b_n$. For 120 and 500 we have

$$ab = 120 \times 500 = 60000 = \text{lcm}(a,b) \times \gcd(a,b) = 3000 \times 20$$

Theorem 7.8 Let a and b be positive integers. Then $ab = \gcd(a,b) \cdot \text{lcm}(a,b)$.

The Euclidean Algorithm

Greatest Common Divisors as Linear Combinations

The greatest common divisor of two integers a and b can be expressed in the form

$$sa + tb$$

where s and t are integers. In other words, $\gcd(a,b)$ can be expressed as a linear combination with integer coefficients of a and b.

This property of GCD is exactlly the Bézout's Theorem, which states that:

Theorem 7.9 — Bézout's Theorem. Let a and b be integers, not both zero. There exist integers x and y such that

$$ax + by = \gcd(a,b),$$

where $\gcd(a,b)$ denotes the greatest common divisor of a and b. The integers x and y are known as Bézout coefficients. the equation$\gcd(a,b) = sa + tb$ is called Bézout's identity.

Proof. The proof of Bézout's identity uses the property that for nonzero integers a and b, dividing a by b leaves a remainder of r_1 strictly less than $|b|$ and $\gcd(a,b) =. \gcd(r_1,b)$.Then

by repeated applications of the Euclidean division algorithm, we have $a = bx_1 + r_1$, $0 < r_1 < |b|$ $b = r_1x_2 + r_2$, $0 < r_2 < r_1$

$$r_{n-1} = r_nx_{n+1} + r_{n+1}, \quad 0 < r_{n+1} < r_n$$

$$r_n = r_{n+1}x_{n+2}$$

where the r_{n+1} is the last nonzero remainder in the division process. Now, as illustrated in the example above, we can use the second to last equation to solve for r_{n+1} as a combination of r_n and r_{n-1}. Unfolding this, we can solve for r_n as a combination of r_{n-1} and r_{n-2} etc. until we eventually write r_{n+1} as alinear combination of a and b. Since r_{n+1} is the last nonzero remainder in the division process, it is the greatest common divisor of a and b,which proves Bézout's identity. □ ■

7.4 Solving Congruence

7.4.1 Linear Congruences

Definition 7.9 — Linear Congruence. Congruence of the form as follows

$$ax \equiv b(\bmod m)$$

is called **Linear Congruence**, where $a, b, m \in \mathbb{Z}^+$, and x is the variable to be solved. To solve this congruence, we need to find $\bar{a}$, which we call **inverse of a modulo m** if a, such that $a\bar{a} \equiv 1(\bmod m)$.

7.4.2 The Chinese Remainder Theorem

7.4.3 Fermat's Little Theorem

In this chapter, we will look into common search and sorting algorithms in computer science. This could be treated as an extension to prior knowledge in this book, as almost every chapter are involved in the analysis and proof of these algorithms. Stay calm if you find it a little difficult, since it takes some time for everyone to fully understand the notions. Some of you may think it's quite hilarious to put sorting and searching algorithm in a chapter about discrete math, but they are exactly the algorithms that requires to be analyzed using the most mathematical tools.

8.1 Searching algorithms

The very first thing we need to do is figure out what is searching, before we do any math.

Definition 8.1 — Searching algorithm. Given an array $A[1], A[2], A[3], \ldots, A[n]$ and a target value T, find an index j where $T = A[j]$ or determine that no such index exists because T is not in the array A.

8.1.1 Searching in Unordered Array List

We proceed to analyze the most common case, where we need to find the element T from an unordered array. In this case, what we need to do is just like counting numbers. We will check each element in the array A, from $A[1]$ to $A[n]$, until we find T. This way of reaching in pure sequential order is called **Linear Search**. The pseudocode is shown below. This is the easiest algorithm in this book, and it is clear that for an array of size n, this algorithm has a time complexity of $O(n)$ in the worst cases. Now consider that, what if we are lucky enough to find our target T in the very first element in the array? Actually, computer scientists have the other notation to show the best-case performance of a algorithm, which is ω notation.

▪ **Notation 8.1 — Omega Notation.** The Omega notation describes the lower bound of an algorithm's running time. It measures the best-case scenario, at least this fast. For a given function g(n), we denote $\omega(g(n))$ the set of functions:

$$\Omega(g(n)) = \{f(n) : \exists c > 0, n_0 > 0 \text{ such that } 0 \leq cg(n) \leq f(n) \text{ for all } n \geq n_0\}$$

Algorithm 12 Linear search for a target value in an array

```
procedure LINEARSEARCH(A, T)
    for j = 1 to j = n do
        if A[j] == T then
            return j
        end if
    end for
    return "not found"
end procedure
```

Yet there is another notation to show the overall, or average performance of an algorithm.

▪ **Notation 8.2 — Theta Notation.** The Theta notation tightly bounds a function from above and below, meaning it defines an algorithm's running time both in the best and worst cases. For a given function $g(n)$, we denote $\Theta(g(n))$ the set of functions:

$$\Theta(g(n)) = \{f(n) : \exists c_1 > 0, c_2 > 0, n_0 > 0 \text{ such that } 0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n) \text{ for all } n \geq n_0\}$$

Theta notation is used to evaluate the average time complexity of executing an algorithm. For instance, we use $\bar{p}$ to denote the average of accessed elements. We have:

$$\bar{p} = \frac{1}{n}\sum_{i=i}^{n} i = \frac{1}{n} \times \frac{n(n+1)}{2} = \frac{n+1}{2}$$

Which means this algorithm has $\Theta(\frac{n+1}{2})$ time complexity.

8.1.2 Searching in Ordered Array List

It seems that we do not have any better way to design a more efficient algorithm for finding target in unordered array. Now we will discuss the case for an ordered array.

Definition 8.2 An ordered array refers to an array A that satisfies $A[1] \leq \cdots \leq A[n]$ or $A[1] \geq \cdots, \geq A[n]$.

In this case, the target T and any random element $A[n]$ in the array has the following properties

- If $T < A[i]$, then T cannot occur among $A[i], A[i+1], \ldots, A[n]$, and we need only search the list from $A[1]$ to $A[i-1]$
- If $A[i] < T$, then T cannot occur among $A[1], A[2], \ldots, A[i]$, and we need only search the list from $A[i+1]$ to $A[n]$

If we adopt a policy of making the worst case as favorable as we can, we should probe in the middle (or as near to the middle as possible) of the list (or sublist) we're searching. Then, if we don't find the target there, we will have at most half the list to search.

Suppose the partition of the original list we are searching in is from $A[p]$ to $A[q]$. We can get the median by $\frac{p+q}{2}$, however, when $p+q$ is odd, we cannot get a valid integer as position of element. As a result, we probe $A[i]$ where $i = \lfloor \frac{p+q}{2} \rfloor$ instead. The resulting search algorithm where we probe in the middle of the current sublist is **Binary Search**.

Algorithm 13 Binary Search for a target value in a sorted array

```
procedure BINARYSEARCH(A, T)
    p ← 1
    q ← n
    while p ≤ q do
        j ← ⌊(p+q)/2⌋
        if A[j] == T then
            return j
        else if A[j] < T then
            p ← j+1
        else
            q ← j−1
        end if
    end while
    return "not found"
end procedure
```

R If this code is a little confusing for you, you may check here for complete visualization of binary search.

Walkthrough with $n = 12$ and $A = (3,5,8,8,9,16,29,41,50,63,64,67)$.
If $T = 99$, then, we can show the variables involved in a table. Suppose the sublist

p	j	q	$A[j]$	relation	output
1	6	12	16	$A[j] < T$	-
7	9	12	50	$A[j] < T$	-
10	11	12	64	$A[j] < T$	-
12	12	12	67	$A[j] < T$	-
13	-	12	-	-	T is not in A

Table 8.1: Binary Search Iterations

we're searching is from $A[p]$ up to $A[q]$ of length $k = q - p + 1$ and we probe unsuccessfully at $A[j]$:

$$\underbrace{A[p]\ldots A[j-1]}_{k1} \quad \text{A}[j] \quad \underbrace{A[j+1]\ldots A[q]}_{k2}$$

If $A[j] \geq T$, then we will search from $A[p]$ up to $A[j-1]$ of length $k_1 = j - p$.

If $A[j] < T$, then we will search from $A[j+1]$ up to $A[q]$ of length $k_2 = q - j$.

We would like k_1 and k_2 to be equal, but if that's not possible (because $q - p$ is odd), let's consistently make k_1 the smaller value. How should we choose j so that $k_1 \leq k_2$ and k_2 is as small as possible? To make $k_2 = q - j$ as small as possible, we need to make j as large as possible. We want

$$k_1 + k_2 \leq k_2 + k_2,$$

that is,

$$q - p \leq 2(q - j) = 2q - 2j$$

or

$$2j \leq 2q - (q - p) = q + p.$$

The largest integer j such that $j \leq (q+p)/2$ is $\lfloor (q+p)/2 \rfloor$, and this is the j-value used in the algorithm.

Theorem 8.1 — Binary Search Sublist Lengths. On each iteration of Binary Search with $j = \lfloor \frac{p+q}{2} \rfloor$, the lengths of the sublists will be $k_1 = \lfloor \frac{k-1}{2} \rfloor$ such that $k_1 \leq k_2 = \lceil \frac{k-1}{2} \rceil \leq \frac{k}{2}$.

Proof. Since $j \leq \frac{q+p}{2} < j+1$, subtracting p from each of these three expressions gives

$$j - p \leq \frac{q+p}{2} - p < j - p + 1$$
$$k_1 \leq \frac{q-p}{2} < k_1 + 1.$$

Thus, $k_1 = \lfloor \frac{q-p}{2} \rfloor = \lfloor \frac{k-1}{2} \rfloor$.

Because $k_1 + k_2 = k - 1$, we know that

- if k is odd, say $k = 2r+1$, then $k - 1 = 2r$ and $k_1 = r = k_2 < \frac{k}{2}$,
- if k is even, say $k = 2r$, then $k - 1 = 2r - 1$ and $k_1 = r - 1 < r = k_2 = \frac{k}{2}$.

Thus, $k_1 = \lfloor \frac{k-1}{2} \rfloor \leq k_2 = \lceil \frac{k-1}{2} \rceil \leq \frac{k}{2}$. ■

Theorem 8.2 — Binary Search Termination. Binary Search terminates after at most $\lceil \lg(n) \rceil + 1$ probes.

Proof. Let $w = \lfloor \lg(n) \rfloor$. If, in some instance, Binary Search has not terminated after w (unsuccessful) probes, then the current value of p must be $\leq$ the current value of q, and the length of the current sublist, k, must be $\leq n/2^w$.

//Prove this by MI.

Since $w \leq \lg(n) < w + 1$,

$$2^w \leq n < 2 \times 2^w$$

//then dividing each by 2^w

$$1 \leq n/2^w < 2.$$

//which implies that $k = 1$

For the next iteration, $p = q$ and so $j = p$.

//And we probe the one remaining entry in A.

If $A[j] < T$, then $p \leftarrow j + 1 > q$, so Binary Search terminates after this probe; if $A[j] > T$, then $q \leftarrow j - 1 < p$, so Binary Search terminates after this probe; and if $A[j] = T$, then Binary Search terminates after this one last probe.

Therefore, Binary Search terminates after at most $\lceil \lg(n) \rceil + 1$ probes. ■

To show that the binary search algorithm is correct, we also need to show its **loop invariant**.

Definition 8.3 In computer science, a loop invariant is a logical assertion that is true before and after each iteration of a program loop. It is a condition that is true at the beginning and end of every iteration of a loop.

Assume that T is in A and $T = A[i]$, $i \in [p,q]$.

Theorem 8.3 — Loop Invariance of Binary Search. After k iterations of the loop, if $T = A[i]$, then $p \leq i \leq q$.

Proof. By Mathematical Induction on k where $k \in \{0..\}$

Step 1. After $k = 0$ iterations of the loop, that is, before the loop is done once

$$p = 1 \quad \text{and} \quad q = n,$$

and therefore,

$$\text{if} \quad T = A[i] \quad \text{then} \quad p \leq i \leq q. \quad (\text{P}(0) \text{ is True})$$

Step 2. Assume $\exists w \in \mathbb{N}$ such that after w iterations of the loop, if $T = A[i]$, then $p \leq i \leq q$. (This is P(w).)

Step 3. Suppose there is another iteration, the $w+1^{st}$. That is, T has not been found in w unsuccessful probes, and now, $p \leq q$. On the next iteration, we calculate a new j-value,

$$j_{\text{new}} \leftarrow \left\lfloor \frac{(p+q)}{2} \right\rfloor$$

Since $p \leq q$,

$$p + p \leq p + q \leq q + q$$

and

$$p = \frac{p+p}{2} \leq \frac{p+q}{2} \leq \frac{q+q}{2} = q$$

and so $p \leq j_{\text{new}} \leq q$. In fact, if $p = q$, then $j_{\text{new}} = p$, and if $p < q$, then $p \leq j_{\text{new}} < q$.

In the remainder of the proof, we will let p^* and q^* denote the values of p and q at the end of the iteration. There are three cases to consider:

Case 1. If $A[j_{\text{new}}] < T$, then T cannot occur at or before position j_{new} hence

$$\text{if} \quad T = A[i] \quad \text{then} \quad p < j_{\text{new}} + 1 \leq i \leq q. \quad \text{Here} \quad p^* = j_{\text{new}} + 1 \quad \text{and} \quad q^* = q.$$

Case 2. If $A[j_{\text{new}}] > T$, then T cannot occur at or after position j_{new} hence

$$\text{if} \quad T = A[i] \quad \text{then} \quad p \leq i < j_{\text{new}} - 1 < q. \quad \text{Here} \quad p^* = p \quad \text{and} \quad q^* = j_{\text{new}} - 1.$$

Case 3. If $A[j_{\text{new}}] = T$, then neither p nor q are changed hence, from Step 2

$$\text{if} \quad T = A[i] \quad \text{then} \quad p \leq i \leq q. \quad \text{In this case,} \quad p^* = p \quad \text{and} \quad q^* = q.$$

Therefore, after this next iteration,

$$\text{if} \quad T = A[i] \quad \text{then} \quad p^* \leq i \leq q^*. \quad \text{in every case}$$

■

R When the repeat-loop in Binary Search terminates,

if $T = A[i]$, then $p \leq i \leq q$. (That is, this conditional statement is True.)

But if $p > q$, there is no index i where $p \leq i \leq q$.

(The consequent must be False. The antecedent must be False.)

If $p > q$, there is no index i where $T = A[i]$.

(If $p > q$, then T cannot be in A. Therefore, when Binary Search terminates,

the target has been found at position j or

($p > q$ and so T is not in A.)

Binary Search is correct and very efficient.

(compared to Linear Search)

Branching Diagrams

Definition 8.4 — Branching Diagram. A General Branching Diagram for an algorithm is a tree that shows all possible sequences of operations the algorithm might do.

Branching tree is an important tool for algorithm analysis; when n is not very large, we can draw the branching diagram for algorithm analysis.

Suppose that we have an array with $n = 12$. In this branching tree, each cell represent a result of iterating element in the array. Certainly, we have $A[6]$ as the first element to be accessed. With 6 visited, there are two possible consequence of comparisons which will lead us to either $A[3]$ or $A[9]$, and so forth.

Figure 8.1: Branching Tree of Binary Search

This kind of diagram is known as a **Binary Tree**: it's rooted at the vertex placed at the top of the diagram, and is binary in the sense that from any vertex, there are at most two edges downward in the diagram. The vertices with no downward edges are called **leaves**. The other vertices are called **internal vertices**. Every vertex, except the root, is the end of exactly one edge to it from a vertex above it in the diagram.

8.1.3 Binary Search: Version 2

Now we look into the other version of binary search.

Algorithm 14 Binary Search Version 2 for a target value in a sorted array

```
procedure BINARYSEARCH(A, T)
    p ← 1
    q ← n
    while p < q do
        j ← ⌊(p+q)/2⌋
        if A[j] < T then
            p ← j+1
        else
            q ← j
        end if
    end while
    if A[p] = T then
        return p
    else
        return not found
    end if
end procedure
```

There is only one major difference between these 2 versions. For p and q, the way they are updated in the second version is different, as q is assigned by j instead of $j-1$. It seems that this is only a trivial change in the algorithm, mathematically, however, it changes the way how it works completely. We refer to the same example where $T = 99$.

Table 8.2: Execution steps of Binary Search when $T = 99$

p	j	q	$p < q$	$A[j]$	$A[j] < T$	Output
1	6	12	t	16	t	-
7	9	12	t	50	t	-
10	11	12	t	64	t	-
12	-	12	f	-	-	T is not in A

Comparing with the table **??**, the number of iteration decreases by 1.

Is this algorithm really correct? To figure it out, recall that to show an algorithm correct, we need to secure its loop invariance and whether it is terminable. Actually, this algorithm operates on two partitions of the array only, but not 3, as in the previous version. Suppose the sublist we're searching is from $A[p]$ up to $A[q]$ of length $k = q - p + 1$, and we make the comparison is $A[j] < T$?:

$$\underbrace{A[p]\ldots A[j]}_{k1} \quad \underbrace{A[j+1]\ldots A[q]}_{k2}$$

In this case, $A[j]$ is not considered as an independent element, but the last element of sublist p to j. If $A[j] < T$, the Procedure will try to find it in $k2$ by assign the index of the

smallest item in it to p. And if $A[j] \geq T$, it is just the other way around. Recall that in theorem **??**, when $j = \lfloor \frac{p+q}{2} \rfloor$ the length of $k1$ will be $\lfloor \frac{k-1}{2} \rfloor$, while length of $k2$ is $\lceil \frac{k-1}{2} \rceil$. However, $\forall k \in \mathbb{Z}$, $\lceil \frac{k-1}{2} \rceil = \lfloor \frac{k}{2} \rfloor$ (proven in previous chapter).

Since $k_1 + k_2 = k$, we have

$$\text{if } k \text{ is even, say } k = 2r, \text{ then } k_1 = r = k_2 = \frac{k}{2} \text{ and}$$

$$\text{if } k \text{ is odd, say } k = 2r+1, \text{ then } k_2 = r < \frac{k}{2} < r+1 = k_1.$$

In general,

$$k_2 = \left\lfloor \frac{k}{2} \right\rfloor \leq \frac{k}{2} \leq k_1 = \left\lceil \frac{k}{2} \right\rceil.$$

Therefore, at some iterations of binary search #2, the length of the next sublist is more than half the length of the previous sublist.

Let $L(w)$ denote the length of the sublist still to be searched after w iterations of the while loop.

Theorem 8.4 After w iterations of the loop

$$\left\lfloor \frac{n}{2^w} \right\rfloor \leq L(w) \leq \left\lceil \frac{n}{2^w} \right\rceil.$$

Proof. Denote the statement as $P(w)$

By Mathematical Induction on w where $w \in \{0, \ldots\}$.

Step 1. After $w = 0$ iterations of the loop, before the loop is done once the length of the current sublist,

$$L(0) = n = \left\lceil \frac{n}{2^0} \right\rceil = \left\lfloor \frac{n}{2^0} \right\rfloor.$$

$P(0)$ is True.

Suppose that for the first few iterations of the loop (even though p and q may have been changed), these bounds on the length of the current sublist are maintained.

Step 2. Assume there exists $m \in \mathbb{N}$ such that after m iterations of the loop

$$\left\lfloor \frac{n}{2^m} \right\rfloor \leq L(m) \leq \left\lceil \frac{n}{2^m} \right\rceil.$$

This is $P(m)$.

Step 3. Suppose there is another iteration, the $m+1$. That is, now $p < q$. We know that

$$\left\lfloor \frac{L(m)}{2} \right\rfloor \leq L(m+1) \leq \left\lceil \frac{L(m)}{2} \right\rceil.$$

If we can show that

$$\lceil L(m)/2 \rceil \leq \lceil \lceil n/2^m \rceil /2 \rceil \text{ and } \lceil \lceil n/2^m \rceil /2 \rceil \leq \lceil n/2^{m+1} \rceil$$

we will have the upper bound we want: $L(m+1) \leq \left\lceil \frac{n}{2^{m+1}} \right\rceil$.

We deal with these (and the corresponding inequalities with the floor function) in a more general setting in the next two lemmas.

Remember that $\lceil r \rceil$ is (defined to be) the smallest integer greater than or equal to the (real) number r and $\lfloor r \rfloor$ is the largest integer less than or equal to r. Also, recall that whenever y is not an integer, $\lfloor y \rfloor < y < \lfloor y \rfloor + 1$.

It takes some extra step to show this. We will prove the inductive hypothesis by 2 lemmas.

Lemma 1: If x and y are real numbers and $x < y$, then $\lceil x \rceil \leq \lceil y \rceil$ and $\lfloor x \rfloor \leq \lfloor y \rfloor$.

Proof. The floor and ceiling functions are nondecreasing. Since $x < y \leq \lceil y \rceil \in \mathbb{Z}$ and $\lceil x \rceil$ is the smallest integer $\geq x$, $\lceil x \rceil \leq \lceil y \rceil$. Since $\lfloor x \rfloor \leq x < y$ and $\lfloor y \rfloor$ is the largest integer $\leq y$, $\lfloor y \rfloor \geq \lfloor x \rfloor$. ■

Lemma 2: For any real number x, $\left\lfloor \frac{\lfloor x \rfloor}{2} \right\rfloor = \left\lfloor \frac{x}{2} \right\rfloor$ and $\left\lceil \frac{\lceil x \rceil}{2} \right\rceil = \left\lceil \frac{x}{2} \right\rceil$.

Proof. If x is an integer, then $\lfloor x \rfloor = x = \lceil x \rceil$, and therefore,

$$\left\lfloor \frac{\lfloor x \rfloor}{2} \right\rfloor = \left\lfloor \frac{x}{2} \right\rfloor$$

and

$$\left\lceil \frac{\lceil x \rceil}{2} \right\rceil = \left\lceil \frac{x}{2} \right\rceil$$

. Suppose now that x is not an integer, and let Q denote $\left\lfloor \frac{x}{2} \right\rfloor$. Then, we have

$$Q < \frac{x}{2} < Q + 1 \quad // \frac{x}{2} \text{ cannot be an integer.}$$

$$\Leftrightarrow 2Q < x < 2Q + 2.$$

In fact,

$$2Q \leq |x| < 2Q + 2,$$

// because $2Q \in \mathbb{Z}$ and

$$2Q < \lceil x \rceil \leq 2Q + 2.$$

Then,

$$Q \leq \left\lfloor \frac{x}{2} \right\rfloor < \frac{x}{2} < \frac{x}{2} < \lceil x \rceil \leq Q + 1,$$

and so

$$\left\lfloor \frac{\lfloor x \rfloor}{2} \right\rfloor = \left\lfloor \frac{x}{2} \right\rfloor = Q$$

and

$$\left\lceil \frac{\lceil x \rceil}{2} \right\rceil = \left\lceil \frac{x}{2} \right\rceil = Q + 1.$$

■

Returning to the proof of Theorem 4.2.1, and taking $x = n/2^m$, from Lemma 2, we have

$$\left\lfloor \frac{n/2^m}{2} \right\rfloor = \left\lfloor \frac{n/2^m+1}{2} \right\rfloor \quad \text{and} \quad \left\lfloor \frac{n/2^m}{2} \right\rfloor = \left\lfloor \frac{n/2^m+1}{2} \right\rfloor.$$

Since

$$\left\lfloor \frac{n/2^m}{2} \right\rfloor \leq L(m) \leq \left\lceil \frac{n/2^m}{2} \right\rceil, \quad //\text{from Step 2}$$

$$\left\lfloor \frac{n/2^m}{2} \right\rfloor \leq \frac{L(m)}{2} \leq \left\lceil \frac{n/2^m}{2} \right\rceil \quad //\text{and applying Lemma A}$$

$$\left\lfloor \frac{n/2^m}{2} \right\rfloor \leq \frac{L(m)}{2} \quad \text{and} \quad \left\lceil \frac{L(m)}{2} \right\rceil \leq \left\lceil \frac{n/2^m}{2} \right\rceil.$$

Thus,

$$L(m+1) \leq \left\lfloor \frac{L(m)}{2} \right\rfloor \quad \text{and} \quad L(m+1) \geq \left\lceil \frac{L(m)}{2} \right\rceil$$

that is,

$$\left\lfloor \frac{n/2^m+1}{2} \right\rfloor \leq L(m+1) \leq \left\lceil \frac{n/2^m+1}{2} \right\rceil.$$

The completes the proof. ■

Theorem 8.5 After at most $\lceil \lg(n) \rceil$ comparisons of the form "Is $A[j] < T$?" and when $p = q$ in the last comparison, the length of the current sublist is 1.

Proof. Let $Q = \lceil \lg(n) \rceil$. Then

$$Q - 1 < \lg(n) \leq Q \text{ so } 2^{Q-1} < n \leq 2^Q.$$

For any integer k

$$\frac{2^{Q-1}}{2^k} = \frac{2^Q}{2^{k+1}} < \frac{n}{2^k} < \frac{2^Q}{2^k} = 2^{Q-k}.$$

By theorem **??** and lemma 1

$$L(k) \leq \left\lfloor \frac{n}{2^k} \right\rfloor \leq \frac{2^Q}{2^k} = 2^{Q-k} \quad //Q-k \in \mathbb{P} \text{ so } 2^{Q-k} \in \mathbb{P}.$$

and

$$L(k) \geq \left\lceil \frac{n}{2^k} \right\rceil \geq \frac{2^{Q-1}}{2^k} = 2^{Q-1-k} \quad //\text{And } 2^{Q-1-k} \in \mathbb{P}.$$

In particular,

$$2 = 2^1 \leq L(Q-2) \leq 2^2$$

and

$$1 = 2^0 \leq L(Q-1) \leq 2^1 = 2.$$

Thus, the while loop cannot stop before $Q-1$ iterations are done, though it might stop after exactly $Q-1$ iterations are done, but if $L(Q-1) = 2$, it must stop after (one more iteration) exactly Q iterations. ■

We can also visualize the version2 of binary search in a Branching Diagram. Below is a search in a sorted list with length $n = 12$. This diagram is also a Binary Tree. But in it,

Figure 8.2: Binary Search V2 with $n = 12$

from every internal vertex, there are exactly two edges downward in the diagram (never just one). This kind of Binary Tree is said to be a full Binary Tree. We will look into trees in detail in the next section.

Theorem 8.6 If T is a full binary tree with m internal vertices, then T has $m+1$ leaves.

Proof. We proceed by induction on the number of internal vertices, m.

Base Case: When $m = 1$, there is only one internal vertex, which is the root of the tree. By the definition of a full binary tree, the root must have two children, and since these children cannot be internal vertices (as there is only one internal vertex, the root itself), they must be leaves. Hence, the tree has $1+1 = 2$ leaves, which satisfies our theorem.

Inductive Step: Assume that the theorem holds for a full binary tree with m internal vertices, such that the tree has $m+1$ leaves. Now consider a full binary tree with $m+1$

internal vertices. By the definition of a full binary tree, adding an internal vertex means we are adding two children to an existing leaf (which is now converted into an internal vertex). This addition results in one less leaf (the converted internal vertex) and two more leaves (the new children). Therefore, the number of leaves increases by $2-1=1$ compared to the number of leaves in the tree with m internal vertices.

So, a tree with $m+1$ internal vertices will have $(m+1)+1$ leaves. Hence, by the principle of mathematical induction, the theorem is true for all $m \geq 1$. ■

8.1.4 Exercises

Exercise 8.1 Suppose T is not found by Binary Search. Where will T fit into the array? Where should it be inserted?

1. Is the final q-value always one less than the final p-value?
2. When is $A[q] < T < A[p]$?
3. If $p = 1$ (that is, p was never changed), is $T < A[1]$?
4. If $q = n$ (that is, q was never changed), is $A[n] < T$?

■

Solution:

1. The final value of q is not necessarily always one less than the final p value. It depends on the search process and the position where T would fit if it were in the array.
2. The condition $A[q] < T < A[p]$ holds when T would be inserted between $A[q]$ and $A[p]$ in a sorted array to maintain the sorted order.
3. If $p = 1$, it means $T < A[1]$, and T should be inserted at the beginning of the array.
4. If $q = n$, it means $A[n] < T$, and T should be inserted at the end of the array.

Exercise 8.2 Consider the best performance of binary search in Θ notation and worst performance in big O notation. Is it possible to get the average performance in Θ notation? If yes, state the precise time complexity; if not, explain why. ■

Solution:

- Best time complexity: The best case scenario occurs when the target value is at the midpoint of the array, which happens with a constant time complexity, thus $\Omega(1)$. However, best case scenarios are generally not described using Θ notation as it implies a tight bound that is both the upper and lower limit, which is not applicable for the best case.
- Worst time complexity: Binary search has a worst-case time complexity of $O(\log n)$ when the target value is not within the initial bounds of the midpoint checks, thus requiring the maximum number of iterations to converge on the target.
- Average time complexity: The average time complexity of binary search is generally represented as $O(\log n)$. This is due to the fact that, on average, the number of steps required to find an element in a sorted array is logarithmic relative to the array size. Precise average case analysis of binary search is complex as it depends on the probability distribution of the target element's position. It is generally not represented in Θ notation because the lower and upper bounds for the average case cannot be tightly bounded as they can be for the best and worst cases.

Exercise 8.3 Prove by MI that $\forall\ k \in \mathbb{N}$, if Binary Search is applied to an array of length n and has not terminated after k (unsuccessful) probes, then the length of the current sublist must be less or equal to $\frac{n}{2^k}$ ■

Proof. We use $L(w)$ to denote the length of the current sublist after w probes.

Base Case: When $k = 0$, the sorting is not yet started, so $L(0) = n$, and we have $L(0) = \frac{n}{2^0} = n$.

Inductive hypothesis: Now assume that when $k = w$, it holds that:

$$L(w) \leq \frac{n}{2^w}$$

Inductive Step: Thus, we need to prove that for $k = w+1$, $L(w+1) \leq \frac{n}{2^{w+1}}$. We have proven earlier in theorem **??** that after w loops, $\lfloor n/2^w \rfloor \leq L(w) \leq \lceil n/2^w \rceil$, and the length of the sublist to be probed after $k+1$ unsuccessful probes is $\frac{n}{2^{w+1}}$. While $\frac{n}{2^{w+1}} \leq \lceil n/2^{w+1} \rceil$. Refer to the definition of $w(k)$ in the statement of problem, $w \in \mathbb{N}$, so $\frac{n}{2^{w+1}} = \lceil n/2^{w+1} \rceil$. Hence, $L(w+1) \leq \frac{n}{2^{w+1}}$, this completes the proof. ■

13.2 Basic Derivative Rules 204
13.3 Higher-order Derivatives 204
13.4 Derivatives of Abnormal Function 204
13.5 Differentiation 204
13.6 Related Rates 204
13.7 Taylor Series 204
13.8 Applications of Differential Calculus 204

14 integral calculus 205
14.1 Fundamentals of Integration 206
14.2 Techniques of Integration 206
14.3 Applications of Integration 206
14.4 Improper Integrals 206
14.5 Numerical Integration Methods 206

15 Differential Equation 207

16 Infinite Series 209

10.1 Monotonicity of Function

10.2 Parity of Function

10.3 Periodicity of Function

11.1 Abstract Function

11.2 Piece-wise Function

12.1 Limit of Sequence

12.2 limit of Function

12.3 Continuity

12.4 Application of Limit

13.1 Derivative Basics

13.1.1 Definition of Derivative

13.1.2 Geometric Meaning of Derivative

13.1.3 Physical Meaning of Derivative

13.2 Basic Derivative Rules

13.2.1 Derivatives of Elementary Functions

13.2.2 Product Rule, Quotient Rule, Chain Rule

13.3 Higher-order Derivatives

13.3.1 Second-order Derivatives and Applications

13.3.2 Calculation and Significance of Higher-order Derivatives

13.4 Derivatives of Abnormal Function

13.4.1 Derivatives of Implicit Functions

13.4.2 Derivatives of Parametric Equations

13.5 Differentiation

13.6 Related Rates

13.6.1 Relationships between Rates of Change of Different Quantities

13.7 Taylor Series

13.7.1 Taylor Expansion of Functions

13.7.2 Application of Taylor Series

13.8 Applications of Differential Calculus

13.8.1 Tangents and Normals of Curves

14.1 Fundamentals of Integration

14.1.1 Definition of the Integral

14.1.2 Properties of Integrals

14.1.3 The Fundamental Theorem of Calculus

14.2 Techniques of Integration

14.2.1 Basic Integration Formulas

14.2.2 Integration by Substitution

14.2.3 Integration by Parts

14.2.4 Trigonometric Integrals

14.2.5 Partial Fractions

14.3 Applications of Integration

14.3.1 Area Under Curves

14.3.2 Volumes of Solids of Revolution

14.3.3 Arc Length and Surface Area

14.3.4 Center of Mass and Moments

14.4 Improper Integrals

14.4.1 Convergence and Divergence of Improper Integrals

14.4.2 Applications of Improper Integrals

14.5 Numerical Integration Methods

14.5.1 The Trapezoidal Rule

14.5.2 Simpson's Rule

17 integral calculus 213

18 Introduction to Multivariable Functions 215

19 Partial Derivatives 217

20 Multiple Integrals 219

21 Vector Calculus 221
21.1 Vector Fields 221
21.2 Gradient, Divergence, and Curl 221
21.3 Line and Surface Integrals 221

18.0.1 Concepts of Multivariable Functions

18.0.2 Graphs and Contour Plots

19.0.1 Definition and Interpretation
19.0.2 Higher-Order Partial Derivatives
19.0.3 Chain Rule in Multiple Variables

20.0.1 Double Integrals

20.0.1.1 Iterated Integrals

20.0.1.2 Double Integrals over General Regions

20.0.2 Triple Integrals

20.0.2.1 Cylindrical and Spherical Coordinates

21.1 Vector Fields
21.2 Gradient, Divergence, and Curl
21.3 Line and Surface Integrals

23 Matrices and Systems of Equations 227
24 Determinant of Matrix 229
25 Orthogonality 231
26 Linear Transformations 233
27 Eigenvalues and Eigenvector .. 235
28 Singular Value Decomposition . 237
29 Complex Vector and Matrices 239
30 Matrix Differential Calculus 241

31 Counting Method and Basic Combinatorics 245

32 Probability Theorems 247

33 Discrete Distribution 249

34 Continuous Distribution 251

35 Joint Cumulative Distribution ... 253

36 Joint Cumulative Distribution ... 255

37 Limit Theory in Probability 257

38 Stochastic Process 259

41 Graphical Statistics 267

42 Statistical Inference 269
42.1 Parameter Estimation 269
42.2 Interval Estimation 269
42.3 Hypothesis Testing 269
42.4 Variance Inference 269
42.5 Bayesian Inference 269

43 Hypothesis Testing 271

44 Regression and Regressive Analysis 273

45 Basic Multi-variable Statistical analysis 275

42.1 Parameter Estimation
42.2 Interval Estimation
42.3 Hypothesis Testing
42.4 Variance Inference
42.5 Bayesian Inference

46 Measuring of Information 279
47 Information Entropy 281
48 Joined Entropy and Conditional Entropy 283
49 Cross Entropy and Relative Entropy 285
50 Mutual Information 287
51 Differential Entropy 289