Mathematics And Axiomatization

Jinshuo Li

Shanghai Jiao Tong University

January 12, 2026

Contents

Preface

At the heart of mathematics lies a beautiful and fundamental tension: the tension between our innate, intuitive grasp of the world and the uncompromising demand for absolute certainty.

We all begin as naive mathematicians. We perceive patterns, sense relationships, and manipulate numbers and shapes with an instinctive confidence. This ”naive understanding” is the soil from which all mathematical curiosity grows. It is natural, powerful, and profoundly human.

Yet, as history has shown, intuition alone can be a treacherous guide, leading to contradictions and uncertainties when pushed beyond its limits. The great edifice of modern mathematics, therefore, could not be built upon this soil alone. It required foundations dug deep into the bedrock of logical rigor.

For hundreds of generations, starting from the simplest details, mathematicians have continuously abstracted mathematical concepts and built an edifice of logic. Geometers began with the most basic geometric structures—points, lines, and planes—establishing the system of Euclidean plane geometry through careful postulates and proofs. This foundational framework, with its emphasis on congruence, similarity, and the properties of space, eventually developed and expanded into modern advanced geometry, where non-Euclidean alternatives challenged long-held assumptions about parallel lines and curvature. From there, it blossomed into topology, the study of shapes and spaces that remain invariant under continuous deformations, and even gave rise to concepts such as manifolds, which provide the mathematical scaffolding for understanding higher-dimensional realities in physics and beyond.

Algebraists, meanwhile, started from the most fundamental concept—quantity itself—and built simple, elementary algebra around operations like addition, subtraction, and solving equations for unknowns. Over time, they further abstracted algebraic structures, recognizing patterns in groups, rings, and fields, leading to disciplines like linear algebra, which models vector spaces and transformations essential to everything from computer graphics to quantum mechanics, and abstract algebra as we know it today, a realm of pure structure that underpins cryptography, coding theory, and the symmetries of the universe.

This process of abstraction extended to other branches as well. In analysis, scholars began with the intuitive notions of limits and continuity, formalizing them into the rigorous calculus of Newton and Leibniz, which evolved into real and complex analysis, measure theory, and functional analysis—tools that allow us to grapple with infinities, probabilities, and the behavior of functions in infinite-dimensional spaces. Number theorists, drawing from the primal fascination with integers and primes, constructed arithmetic systems that branched into analytic number theory, algebraic number theory, and even the profound mysteries of the Riemann Hypothesis, connecting primes to the zeros of complex functions.

This book is about the construction of that edifice. It is the story of the journey from the fertile but fuzzy landscape of intuition to the crystalline structure of formal axiomatic systems. We will witness how mathematicians, through centuries of intellectual toil, learned to distill their intuitive ideas into precise definitions and unequivocal rules—axioms.

These axioms are the cornerstone of the mathematical skyscraper. Chosen with care, they are both simple enough to be self-evident and powerful enough to support an ever-ascending tower of ideas. Each new floor—a theorem, a theory, a whole new discipline like calculus or algebra—is constructed securely upon the layers beneath it, its integrity guaranteed by the logical connections that bind it to the foundation.

This architectural principle is what makes mathematics the most unifying of all languages. It connects the geometric world of shapes with the algebraic world of equations, and the discrete world of integers with the continuous world of analysis, weaving them into a single, grand, and coherent narrative. It bridges the probabilistic uncertainties of statistics with the deterministic certainties of logic, and even links the abstract realms of set theory—where infinities are tamed and paradoxes resolved—with the applied worlds of computer science and engineering.

The skyscraper stands as a testament to the power of human reason. Yet, Gödel's incompleteness theorems cast a fascinating and necessary shadow. They remind us that even the most perfectly constructed skyscraper cannot contain the tools to verify the stability of its own deepest foundations from within. This is not a flaw that collapses the structure; rather, it is a profound insight into its very nature.

It tells us that mathematics is not a static, completed temple of absolute truth, but a living, growing, and endlessly fascinating human adventure. The inability to achieve a final, self-verifying system is not a weakness, but a source of strength—it guarantees that the adventure will never end, that there will always be new horizons to explore and new questions to ask. It invites us to embrace the unknown, to push the boundaries of what we can formalize, and to find beauty in the interplay between certainty and mystery.

This manuscript is an invitation to join this great adventure. It is designed for those who are not satisfied with merely being told a result; it is for those who wish to stand at the drawing board and understand, step by logical step, how the skyscraper was designed and built. Along the way, we will encounter the triumphs of discovery, the pitfalls of early misconceptions, and the elegant resolutions that have shaped the field into what it is today.

With each new concept mastered, the view from the skyscraper grows more magnificent.

Welcome.

Jinshuo Li
Shanghai Jiao Tong University
2025.12.2, in Shanghai

Chapter 0: Basic Knowledge and Notations

All mathematical insights undergo a process from "naïve understanding" to "rigorous formulation". Mathematics itself is the same; all axiomatized languages cannot be separated from the empiricist's naïve descriptive language.

To introduce certain special mathematical symbols that will be frequently used in the future, I have specially added a Chapter Zero before Chapter One, intended to present parts of the knowledge concerning mathematical language and mathematical logic. In this chapter, I will also introduce knowledge related to axioms; however, in the future, we will not necessarily define axioms so rigorously every time.

0.1 Propositional Logic

0.1.1 Naïve introduction to propositional logic

Before we start the main part of the propositional logic, let's start with a question: what is a proposition? Take a look at the sentences below:

1. The equation $x^2 + 1 = 0$ has a real-number solution;
2. There are infinitely many prime numbers;
3. Every even integer $n \geq 4$ is a sum of two prime numbers (The famous Goldbach Conjecture);
4. What time is it?
5. This statement is false;

In the sentences above, the first 3 sentences share the same properties: 1) they are sentences or assertions that declare facts. 2) They are either true or false, but not both. However, the 4th sentence is not a declarative sentence, and we can't judge the correctness of the 5th sentence.

So, to study mathematics, we need the sentences that assert one or several facts. Sometimes, whether it is true or not is not that important once it can be determined. It doesn't mean that other types of expressions are not vital. We value this kind of expression because all the axioms, definitions, and theorems are written in this form.

Definition 0.1.1: Proposition

Propositions are sentences or assertions that declare facts and are either true or false, but not both.

To make propositions more actionable, we use **propositional variables** to represent different propositions. Commonly used propositional variables are letters like $p, q, r, \ldots$

Sometimes we need to connect different propositional variables to form a new proposition. And we need to use words like "and", "or", "imply" to explain their relationships.

Definition 0.1.2: Logical operators/connectives

Logical operators/connectives are marks that are used to connect propositional variables, forming compound propositions.

Here are some common operators:

$\neg$ negation ("not")

$\wedge$ conjunction ("and")

$\vee$ disjunction ("or")

$\rightarrow$ **(or $\Rightarrow$)** conditional ("imply")

$\leftrightarrow$ biconditional ("if and only if" ("iff"))

$\oplus$ exclusive-or

Definition 0.1.3: Compound proposition

A compound proposition is built from propositional variables or constants through logical operators.

Definition 0.1.4: Atomic proposition

An atomic proposition is a proposition that can't be divided any further. It is the most basic building block of logical expressions and is considered an indivisible semantic unit.

If we want to determine whether an atomic proposition is true or not, we need to use correlated knowledge of different fields in mathematics. But how can we determine the truth of a compound proposition? We can use truth tables!

Definition 0.1.5: Truth table

A truth table is a mathematical table used in logic to compute the functional values of logical expressions based on their inputs.

Negation

p	$\neg p$
T	F
F	T

Conjunction

p	q	$p \wedge q$
T	T	T
T	F	F
F	T	F
F	F	F

Disjunction

p	q	$p \lor q$
T	T	T
T	F	T
F	T	T
F	F	F

Conditional (Imply)

p	q	$p \to q$
T	T	T
T	F	F
F	T	T
F	F	T

Biconditional (iff)

p	q	$p \leftrightarrow q$
T	T	T
T	F	F
F	T	F
F	F	T

Exclusive-or

p	q	$p \oplus q$
T	T	F
T	F	T
F	T	T
F	F	F

Just like numerical computation, logical operators also have precedence. Here is the order of precedence for logical operators (from highest to lowest):

$$(), [] \quad \neg \quad \wedge \quad \vee \quad \oplus \quad \rightarrow \quad \leftrightarrow$$

Apart from this, operators that appear first are of higher precedence. In short, we compute higher-precedence logical operators first, then lower ones.

Using these laws, we can define the truth value of different propositions. But there exist some sub-classes of compound propositions:

Tautology A tautology is a compound proposition that is always true, no matter what the truth values of the propositional variables are.

Contradiction A compound proposition that is always false, regardless of the truth values of propositional variables.

Contingency A compound proposition that is neither a tautology nor a contradiction. It can be true or false, depending on the value of the variables.

Based on the definitions above, now we can rigorously define what logical equivalence is.

Definition 0.1.6: Logical Equivalence

We can say compound propositions p, q are **logically equivalent** if $p \leftrightarrow q$ is a tautology. If p and q are logically equivalent, we denote it as $p \equiv q$.

This definition means that for different truth values of every atomic proposition, p and q come out with the same truth value. In further study, we will know that several of the operators above will be enough to express all the propositions.

Here are a few frequently used examples. These can be used directly.

$$\begin{aligned}
\text{Identity Laws:}\quad & p \wedge T \equiv p \\
& p \vee F \equiv p \\
\text{Domination Laws:}\quad & p \vee T \equiv T \\
& p \wedge F \equiv F \\
\text{Absorption Laws:}\quad & p \vee (p \wedge q) \equiv p \\
& p \wedge (p \vee q) \equiv p \\
\text{Negation Laws:}\quad & p \vee \neg p \equiv T \\
& p \wedge \neg p \equiv F \\
\text{Commutative Laws:}\quad & p \vee q \equiv q \vee p \\
& p \wedge q \equiv q \wedge p \\
\text{Associative Laws:}\quad & (p \vee q) \vee r \equiv p \vee (q \vee r) \\
& (p \wedge q) \wedge r \equiv p \wedge (q \wedge r) \\
\text{Distributive Laws:}\quad & p \vee (q \wedge r) \equiv (p \vee q) \wedge (p \vee r) \\
& p \wedge (q \vee r) \equiv (p \wedge q) \vee (p \wedge r) \\
\text{De Morgan's Laws:}\quad & \neg(p \wedge q) \equiv \neg p \vee \neg q \\
& \neg(p \vee q) \equiv \neg p \wedge \neg q
\end{aligned}$$

Definition 0.1.7: Satisfiability

A compound proposition is **satisfiable** if it is true under some truth assignment for propositional variables. A truth assignment that makes a compound proposition true is called a **solution**. If a compound proposition is not satisfiable, we say it is **unsatisfiable**.

For now, we can use a truth table and logical equivalence to check whether a compound proposition is satisfiable. For readers majoring in computer science, they can use special algorithms to solve the so-called "Boolean Satisfiability Problem", like heuristic searching, etc.

0.1.2 Functional Completeness

According to previous study, we know that there are connectives for different variables to form compound propositions. However, do we really need that many connectives to represent all the relationships and compound propositions? The answer is absolutely NO! In some cases, we only need two connectives to reach the so-called state of functional completeness!

Definition 0.1.8: Disjunctive Normal Form

A compound proposition is in **disjunctive normal form (DNF)** if it is a disjunction of conjunctions of propositional variables or their negations.

Theorem 0.1.1

Every compound proposition is logically equivalent to some compound proposition in disjunctive normal form.

Proof. It's not difficult to prove it. According to previous content, we know that we can make a truth table for every compound proposition. According to the truth table, we can rewrite it in the form of disjunctive normal form. For example, we have the truth table of the exclusive-or connective:

p	q	$p \oplus q$
T	T	F
T	F	T
F	T	T
F	F	F

The content of the middle two lines (where the result is T) represents the solution to this compound proposition. Then we can write a proposition like this:

$$p \oplus q \equiv (p \land \neg q) \lor (\neg p \land q)$$

Every part of the compound proposition connected with the disjunction represents a solution of the satisfiable problem of the proposition. That means if we can make a truth table for all the compound propositions, we can rewrite them in the form of disjunctive normal form. From this perspective, the theorem is quite clear. □

Let's move further: can we rewrite the compound proposition in other forms? The answer is obviously YES!

Definition 0.1.9: Conjunctive Normal Form

A compound proposition is in **conjunctive normal form (CNF)** if it is a conjunction of disjunctions of propositional variables or their negations.

Theorem 0.1.2

Every compound proposition is logically equivalent to some compound proposition in conjunctive normal form.

Proof. We will use the truth table again to prove this theorem. Assume we have a proposition A. Firstly, we use Theorem 0.1.1 to rewrite $\neg A$ into the form of disjunctive normal form. Next, we take the negation of $\neg A$, and using De Morgan's Law, we can now rewrite A in the form of conjunctive normal form. □

Theorem 0.1.1 and Theorem 0.1.2 are quite a marvel. In circuit design, since we cannot design a circuit component for every logical connector, we must use as few circuit components as possible to express all logical expressions. And according to these theorems, we no longer need to use that many connectives. Only conjunction, disjunction, and negation are necessary to construct logically equivalent propositions to any propositions in the world.

Definition 0.1.10: Functionally Complete

A collection of logical operators is **functionally complete** if any compound proposition is logically equivalent to a compound proposition that involves only the logical operators in the collection.

Thus, we can declare that the collection $\{\neg, \vee, \wedge\}$ is functionally complete. Also, according to De Morgan's Law:

$$p \wedge q \equiv \neg(\neg p \vee \neg q)$$
$$p \vee q \equiv \neg(\neg p \wedge \neg q)$$

We found that conjunction can be expressed using disjunction and negation, and disjunction can be expressed using conjunction and negation. Which means that the collections $\{\neg, \vee\}$ and $\{\neg, \wedge\}$ are functionally complete.

Remark 0.1.1

The most striking fact is that there exists a connective that is functionally complete in itself (e.g., NAND or NOR)! But we won't use it in the future, so you can search by yourselves.

0.2 Predicate Logic

In many cases, using propositional logic is enough to cover the usage. But there is a kind of statement that can't be expressed using the tool of propositional logic. In such cases, we need to use predicate logic to formulate logical expressions.

Definition 0.2.1: Predicate

A **predicate** is in the form of a statement with variables. A predicate is also known as a **propositional function**. A predicate can be denoted as $P(x, y, z, \dots)$, and x, y, z are known as the variables.

Definition 0.2.2: Quantifiers

We use **quantifiers** to indicate the quantity of elements being referred to. We have two types of quantifiers, $\forall$ and $\exists$.
$\forall$ The **universal quantifier** $\forall$ means "for all".
$\exists$ The **existential quantifier** $\exists$ means "there exists".

For example, the statement "Every SJTUer is a talent" can be written in the form of predicate logic: $\forall x(\text{SJTUer}(x) \to \text{Talent}(x))$.

Variables in predicates all have restricted domains; they are the **domains** of predicates. For all the values in the domain, we can rewrite the predicates in the form of propositions. If the domain is finite, $\{x_1, x_2, \dots, x_n\}$, then:

$$\forall x P(x) \equiv P(x_1) \wedge P(x_2) \wedge \cdots \wedge P(x_n)$$
$$\exists x P(x) \equiv P(x_1) \vee P(x_2) \vee \cdots \vee P(x_n)$$

But if the domain is infinite, we can only express the logic in the form of predicate logic.

- $\forall x P(x)$ stands for "For every x in domain, $P(x)$". If there is a value a in the domain such that $P(a)$ is false, the statement is false. a is called a **counterexample**.
- $\exists x P(x)$ stands for "There exists an x, $P(x)$". If for all values a in the domain, $P(a)$ is false, the statement is false.

However, not all sentences with predicates and quantifiers can form a predicate logic. For example, $\forall x(P(x) \wedge Q(y))$ is not a statement with predicate logic. Because as the value of y varies, we can't decide whether the sentence is true or false, which makes its truth value unknown.

So, how can we make sure a sentence is a predicate logic? In the case of $\forall x(P(x) \wedge Q(y))$, it is y that causes trouble. And x, which is connected with a quantifier, did not cause any inconvenience. So, we call variables that are bounded only if they follow a quantifier, and they are called the **bounded variables**. Otherwise, they are the **free variables**. If every variable in a sentence is bounded, it's a proposition.

Just like propositional logic, predicate logic can also be connected using logical connectives. And there are also a few arithmetic laws.

De Morgan's Laws for Quantifiers

$$\neg \forall x P(x) \equiv \exists x \neg P(x)$$
$$\neg \exists x P(x) \equiv \forall x \neg P(x)$$

Asymmetric Associative Laws

$$\forall x \forall y P(x, y) \equiv \forall y \forall x P(x, y)$$
$$\exists x \exists y P(x, y) \equiv \exists y \exists x P(x, y)$$

Note

The associative laws are asymmetric. The following two assertions are not necessarily valid. In some cases, they can be false.

$$\forall x \exists y P(x, y) \not\equiv \exists y \forall x P(x, y)$$

Definition 0.2.3: Logical Equivalence in Predicate Logic

If ϕ and ψ are statements with predicates and quantifiers, but without free variables, then we say that ϕ and ψ are **logically equivalent**, written as $\phi \equiv \psi$, if, no matter which concrete predicates (for the predicate symbols) and domains (for variables) are given, the truth values of ϕ, ψ coincide.

Also, we can define universal validity (just like tautology in propositional logic), and predicate logic also has satisfiability. As this chapter only provides a brief introduction, it will not be elaborated further here.

In this chapter, we came to know what logic is in the form of mathematics. This overview of propositional and predicate logic establishes the formal basis for clear reasoning. Logic is the indispensable scaffolding of rational thought—the *a priori framework* that precedes all specific knowledge. Its importance doesn't need much explanation.

Keywords: propositional logic, predicate logic, logical equivalence
Reference: Discrete Mathematics and Its Applications (Eighth Edition), Kenneth H. Rosen, McGraw-Hill Education.

Chapter 1

The Axioms of ALL

In modern mathematics, is there any single axiom that can be called "The Axioms of all Branches"? It's quite a tricky question because researchers in different fields have different answers for it. But if we take the intersection of the answers from most of them, we would find only one element in this set: The Set Theory.

1.1 The Naïve Set Theory

Actually, if you are a student majoring in engineering or applied mathematics, learning naïve set theory is enough for you to cover the math you need during your work and study. But we need to know that the naïve theory is basically an intuitive definition, which is not included as a part of the modern axiomatic set theory. Although we say that the naïve set theory is a really clear and useful definition of the concept "set", we have to declare that the naïve set theory is incomplete because naïve set theory itself cannot resolve Russell's paradox. We will introduce the paradox that causes the third mathematical crisis to readers later.

Definition 1.1.1: Set

A **set** is an unordered collection of distinct objects, called **elements** or **members** of the set. A set is said to *contain* its elements. We write $a \in A$ to denote that a is an element of the set A. The notation $a \notin A$ means a is not an element of the set.

The concept of set is quite simple. A set can be the container of anything, like numbers, points, other sets, even an apple can be a set. The only requirement is that the elements in a set must be distinct. For example, the set $\{1, 2, 2, 3\}$ is actually equal to the set $\{1, 2, 3\}$ because the element 2 is repeated.

Note

An object can only be in one of the two states: 'belonging to A' or 'not belonging to A'; it cannot be in both states at the same time, nor can it be in neither state.

Definition 1.1.2: Subset

We say that the set A is a **subset** of B, written $A \subseteq B$, if every element of A is an element of B. Additionally, A is a **proper subset** of B, written $A \subset B$, if $A \subseteq B$ and $A \neq B$.

The concept of subset is quite straightforward. If all the elements in set A can be found in set B, then A is a subset of B. For example, $\{1, 2\} \subseteq \{1, 2, 3\}$, and $\{1, 2\} \subset \{1, 2, 3\}$. However, in later study, we will find that the concept of subset creates contradcitions in naïve set theory, which is the root cause of Russell's paradox.

Definition 1.1.3: Equality

Two sets A and B are **equal** (i.e., they are the same set), written $A = B$, if they contain the same members.

Set equality can be determined using the definition of subset. If $A \subseteq B$ and $B \subseteq A$, then $A = B$. This is the formal definition of set equality.

Here are several common set operations, shown below:

Definition 1.1.4: Set Union

The **union** of sets A and B is denoted as $A \cup B$, which contains all the elements of A together with the elements of B.

Definition 1.1.5: Set Intersection

The **intersection** of sets A and B is denoted as $A \cap B$, which contains the elements that are the members of both A and B.

Definition 1.1.6: Set difference

The **difference** of the set A w.r.t B, written as $A - B$ (or $A \setminus B$), is the set consisting of those elements of A that are not in B.

Definition 1.1.7: Empty set

The **empty set**, denoted $\emptyset$, is the set that contains no elements.

Note that the empty set is a subset of every set. And there is only one empty set in the whole universe of sets. Which means that if two sets have no elements, they are equal.

Definition 1.1.8: Power set

The **power set** $\mathcal{P}(x)$ (aka 2^x): the set consisting of all subsets of x.

According to the definition above, we can easily find several conclusions below:

Corollary 1.1.1

$\emptyset \subseteq A$ for any set A.

Corollary 1.1.2

Assume that the number of elements in the finite set x is $\mathrm{card}(x)$, then $\mathrm{card}(\mathcal{P}(x)) = 2^{\mathrm{card}(x)}$.

Corollary 1.1.3

If the set x is empty, then $\mathrm{card}(\mathcal{P}(\emptyset)) = 1$. (This inference can be seen as an extension of the inference 1.1.2).

Russell's Paradox

Russell's Paradox makes the naïve set theory incomplete. Russell assumed there exists a set X that contains sets that don't include themselves.

$$X = \{x \mid x \notin x\}$$

And then he found X doesn't belong to set X nor does it belong to set X.

$$X \in X \implies X \notin X$$
$$X \notin X \implies X \in X$$

So, does X belong to X or not? Thus, this creates a contradiction, because X either belongs to X or does not belong to X, which is determined by the nature of naive set theory itself. However, the properties of the naïve set theory allow the existence of set X, which creates a contradiction.

In order to mend such a flaw present in naive set theory, countless mathematicians devoted themselves tirelessly and proposed the ZFC axiomatic system, which ultimately became the core of modern set theory.

1.2 The Axiomatic Set Theory

1.2.1 The ZFC Axioms System

ZFC stands for **Zermelo-Fraenkel Set theory with the axiom of choice**, which includes 9 different axioms (numbered from ZF1 to ZF8 and AC). We will now introduce them to the readers one by one.

Principles:

- Either $a \in A$ or $a \notin A$ but not both.
- A formal language is required for constructing meaningful statements.
- Every object is a set, and every set is an object.

ZF1 (Axiom of Extensionality) If X and Y have the same elements, then $X = Y$.

$$\forall X \forall Y (\forall u (u \in X \leftrightarrow u \in Y) \rightarrow X = Y)$$

(ZF1 defines the "=" in set theory)

ZF2 (Axiom of the Unordered Pair) For any a and b, there exists a set $\{a, b\}$ that contains exactly a and b. (Also called Axiom of Pairing)

$$\forall a \forall b \exists Z \forall u (u \in Z \leftrightarrow (u = a \vee u = b))$$

(ZF2 constructs the unordered pairs and allows the existence of ordered pairs.)

ZF3 (Axiom of Subsets) Assume ϕ is a property with parameter p, then for any X and p, there exists a Set $Y = \{u \in X \mid \phi(u, p)\}$ that contains all those $u \in X$ that have the property ϕ. (also called Axiom of Separation or Axiom of Comprehension)

$$\forall X \forall p \exists Y \forall u (u \in Y \leftrightarrow (u \in X \wedge \phi(u, p)))$$

(ZF3 is the key axiom that prevents the situation Russell's paradox described. Assume $X = \{x \in C \mid x \notin x\}$. According to the axiom of subsets, X must be constructed from an existing set C. This form of definition automatically excluded X from being a "set of all sets".)

ZF4 (Axiom of the Sum Set) For any X, there exists a set $Y = \bigcup X$, the union of all elements of X. (Also called Axiom of Union)

$$\forall X \exists Y \forall u (u \in Y \leftrightarrow \exists Z (Z \in X \land u \in Z))$$

(ZF4 allows the construction of a union set, allowing mathematicians to construct bigger sets and form more complex mathematical structures.)

ZF5 (Axiom of the Power Set) For any X, there exists a set $Y = \mathcal{P}(X)$, the set of all subsets of X.

$$\forall X \exists Y \forall u (u \in Y \leftrightarrow u \subseteq X)$$

(ZF5 defines the concept of a power set.)

ZF6 (Axiom of Infinity) There exists an infinite set.

$$\exists S (\emptyset \in S \land \forall x (x \in S \to x \cup \{x\} \in S))$$

(ZF6 allows the existence of infinite sets, which is the basis of the set of $\mathbb{N}$.)

ZF7 (Axiom of Replacement) If F is a function, then for any X, there exists a set $Y = F[X] = \{F(x) \mid x \in X\}$.

$$\forall X [(\forall x \in X \exists! y \phi(x, y)) \to \exists Y \forall y (y \in Y \leftrightarrow \exists x \in X \phi(x, y))]$$

(ZF7 allows mathematicians to construct new sets using existing sets and a function. It can imply ZF3.)

ZF8 (Axiom of Foundation) Every non-empty set A contains a member that is disjoint from A.

$$\forall A (A \neq \emptyset \to \exists x (x \in A \land x \cap A = \emptyset))$$

(ZF8 avoids infinite nesting of sets ($A \in A$) and is a powerful aid to ZF3 when refuting Russell's paradox. If $A \in A$, construct $B = \{A\}$. Then A is the only element in B. By ZF8, $A \cap B = \emptyset$. But $A \in B$ and $A \in A$, so $A \in A \cap B$, which means $A \cap B \neq \emptyset$. This is a contradiction.)

AC (Axiom of Choice) Every family of nonempty sets has a choice function.

$$\forall \mathcal{F} [(\emptyset \notin \mathcal{F}) \to \exists f : \mathcal{F} \to \bigcup \mathcal{F} (\forall A \in \mathcal{F} (f(A) \in A))]$$

(AC is fundamental as it guarantees the ability to make infinitely many simultaneous, non-constructive choices, which is indispensable for proving a vast number of crucial theorems across diverse fields of mathematics.)

Although Gödel's incompleteness theorems tell us that the ZFC axiom system cannot be proven to be consistent, most mathematicians believe that ZFC is consistent because it has not yet produced any fundamental contradiction that threatens the integrity of mathematics. Therefore, it can be regarded as a reliable foundation for modern mathematics.

Of course, other axiomatic systems exist, like the NBG (von Neumann–Bernays–Gödel) axiomatic system. But for practical purposes, ZFC is enough. From the ZFC axiomatic system, we know that no set includes everything.

1.3 Extensions of Axiomatic Set Theory

1.3.1 Ordered Pairs and Cartesian Product

Ordered Pairs

Now, with the help of ZFC axiomatic set theory, we can define ordered pairs:

Definition 1.3.1: Ordered Pair

We define the ordered pair (a, b) as:

$$(a, b) := \{\{a\}, \{a, b\}\}$$

$\{a\} \in \{\{a\}, \{a, b\}\}$ guarantees the order of the pair (a, b).

Corollary 1.3.1

$(a, b) = (c, d)$ if and only if $a = c$ and $b = d$.

Corollary 1.3.2

$(a, a) = \{\{a\}, \{a, a\}\} = \{\{a\}\}$.

Cartesian Product

Definition 1.3.2: Cartesian Product

The Cartesian Product of two sets A and B is defined as:

$$A \times B := \{(a, b) \in 2^{2^{x \cup y}} \mid a \in A \wedge b \in B\}$$

Furthermore, we can define the n-ary Cartesian product as below:

$$X_1 \times X_2 \times \cdots \times X_n := \{(x_1, x_2, \ldots, x_n) \mid x_i \in X_i \text{ for } i = 1, \ldots, n\}$$

Cartesian Product is widely used in Mathematical Analysis, Analytic Geometry, and Group Theory, etc.

1.3.2 Relations and Their Special Types

Relations

Now, based on the concept of Cartesian Product, we can define Relations.

Definition 1.3.3: Relation

Suppose we are given two sets X and Y. A **relation** R from X to Y is a subset of the Cartesian Product $X \times Y$.

$$R \subseteq X \times Y$$

If $X = Y$, we can say that R is a relation on A. We write xRy for $(x, y) \in R$.

For all relations, there are descriptions unique to themselves, which we denote as $P(x, y)$. Then a relation can be described as $R = \{(x, y) \in X \times Y \mid P(x, y)\}$.

Remark 1.3.1

Mark that the Y is the range of the relation, and X is the domain of the relation. This fact may contradict our common sense of the relation.

$$\mathrm{dom}R := \{x \in \cup \cup R | \exists y (x, y) \in R\}$$

$$\mathrm{ran}R := \{y \in \cup \cup R | \exists x (x, y) \in R\}$$

Some special types of relations

Here are a few properties that can be used to classify different relations on a set A.

Reflexive R is reflexive if and only if $\forall a \in A((a,a) \in R)$.

Antireflexive (or Irreflexive) R is antireflexive if and only if $\forall a \in A((a,a) \notin R)$.

Symmetric R is symmetric if and only if $\forall a,b \in A((a,b) \in R \to (b,a) \in R)$.

Antisymmetric R is antisymmetric if and only if $\forall a,b \in A[((a,b) \in R \land (b,a) \in R) \to a = b]$.

Transitive R is transitive if and only if $\forall a,b,c \in A[((a,b) \in R \land (b,c) \in R) \to (a,c) \in R]$.

Using these properties, we can define a very important concept in mathematics: the equivalence relation.

Definition 1.3.4: Equivalence Relation

Suppose R is a relation on A. If R simultaneously possesses reflexivity, symmetry, and transitivity, then R is an **equivalence relation** on A.

We can use an equivalence relation to classify a set.

Definition 1.3.5: Equivalence Class

We define the **equivalence class** $[a]$ of an element $a \in A$ by:

$$[a] = \{x \in A \mid xRa\}$$

We can completely classify a set using equivalence classes. For example, in SJTU, we can classify all students by their nationality. This relation satisfies reflexivity, symmetry, and transitivity, so it's an equivalence relation.

Similarly, we can define order relations.

Definition 1.3.6: Partial Order

Suppose R is a relation on A. If R simultaneously possesses reflexivity, antisymmetry, and transitivity, then R is a **partial order relation** on A. A set A with a partial order R is a **partially ordered set (poset)**, denoted (A,R) or $(A,\preceq)$.

Definition 1.3.7: Total/Linear Order

Suppose R is a partial order relation on A. If for all $a,b \in A$, either aRb or bRa is true, then R is a **total order relation** on A.

Definition 1.3.8: Strict Partial Order

Suppose R is a relation on A. If R simultaneously possesses antireflexivity and transitivity, then R is a **strict partial order relation** on A. (Note: antireflexivity and transitivity imply asymmetry).

We can also define inverse and composite relations.

Mind that the domain and range of the inverse relation are swapped compared to the original relation.

Definition 1.3.9: Inverse Relation

We define the **inverse relation** $R^{-1} \subseteq Y \times X$:

$$R^{-1} = \{(y, x) \mid (x, y) \in R\}$$

In more formal language:

$$R^{-1} = \{(x, y) \in \mathrm{ran}R \times \mathrm{dom}R | yRx\}$$

Definition 1.3.10: Composite Relation

Let $R \subseteq X \times Y$ and $S \subseteq Y \times Z$. We define the **composite relation** $S \circ R \subseteq X \times Z$ by:

$$S \circ R = \{(x, z) \mid \exists y \in Y((x, y) \in R \wedge (y, z) \in S)\}$$

Theorem 1.3.1

1. $(R^{-1})^{-1} = R$
2. $(R \circ S)^{-1} = S^{-1} \circ R^{-1}$
3. $(R \circ (S \cup T)) = (R \circ S) \cup (R \circ T)$

However, mind that$R \circ (S \cap T) \neq (R \circ S) \cap (R \circ T)$. Here is an counterexample: We define $U = \{a, b, c, d\}$ $R = \{(a, b), (a, c)\}$, $S = \{(b, d)\}$, $T = \{(c, d)\}$. We know that $R \circ (S \cap T) = \emptyset$ because $S \cap T = \emptyset$. But $(R \circ S) \cap (R \circ T) = \{(a, d)\}$, which is an counterexample.

Using the concept of equivalence relation, we can "divide" a set precisely.

Definition 1.3.11: Partition

Assume A is a nonempty set. A collection P of non-empty subsets of A ($P \subseteq \mathcal{P}(A)$) is a **partition** on A if:
1. No set in P is empty: $\forall S \in P(S \neq \emptyset)$.
2. The union of sets in P is A: $\bigcup_{S \in P} S = A$.
3. The sets in P are pairwise disjoint: $\forall S_1, S_2 \in P(S_1 \neq S_2 \to S_1 \cap S_2 = \emptyset)$.

Theorem 1.3.2

Every partition of a set A corresponds to an equivalence relation on A, and vice versa.

1.3.3 A Brief Example: Measure

Here comes a question: can we assign a non-negative value representing the "length" of a subset of $\mathbb{R}$?

Definition 1.3.12: Measure

A **measure** μ on a collection $\mathcal{M}$ (of subsets of a set X) is a function $\mu : \mathcal{M} \to [0, \infty]$ such that for any pairwise-disjoint countable-infinite sequence of sets $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{M}$ (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), if their union is also in $\mathcal{M}$, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

This key property is called **countable additivity**.

Does there exist a measure m on $\mathcal{P}(\mathbb{R})$ (i.e., all subsets of the real numbers) that satisfies the following conditions (our naïve understanding of "length"):

1. $m([0,1]) = 1$.
2. Translation invariance: $m(A + x) = m(A)$ for any $A \subseteq \mathbb{R}$ and $x \in \mathbb{R}$, where $A + x = \{a + x \mid a \in A\}$. Which means that the same measure apply to all the set with the same structure on the axis.
3. Countable additivity.

Unfortunately, the answer is no. We construct the **Vitali Set**, which can't be measured following these conditions.

Proof of the existence of a non-measurable set (Vitali Set). Define an equivalence relation $\sim$ on $[0,1]$ by $x \sim y \iff x - y \in \mathbb{Q}$. This is an equivalence relation. Let V be the set of its equivalence classes. According to the Axiom of Choice, we can choose exactly one element from each equivalence class, altogether to form a set of representatives M. We can assume that $M \subseteq [0,1]$. Let $\mathbb{Q}^* = \mathbb{Q} \cap [-1,1]$. For each $q \in \mathbb{Q}^*$, define $M_q = M + q = \{m + q \mid m \in M\}$. We have the following conclusions:

1. All M_q (for $q \in \mathbb{Q}^*$) are pairwise disjoint. (If $y \in M_q \cap M_r$, then $y = m_1 + q = m_2 + r$. This means $m_1 - m_2 = r - q \in \mathbb{Q}$, so $m_1 \sim m_2$. By construction of M, this implies $m_1 = m_2$, so $q = r$.)
2. $[0,1] \subseteq \bigcup_{q \in \mathbb{Q}^*} M_q$. (For any $x \in [0,1]$, let $m \in M$ be the representative x is equivalent to, so $x - m = q \in \mathbb{Q}$. Since $x, m \in [0,1]$, $q \in [-1,1]$. Thus $x = m + q \in M_q$.)
3. $\bigcup_{q \in \mathbb{Q}^*} M_q \subseteq [-1,2]$. (Since $M \subseteq [0,1]$ and $\mathbb{Q}^* \subseteq [-1,1]$.)

Now, let's try to measure M. Assume $m(M) = c$. By translation invariance, $m(M_q) = m(M) = c$ for all $q \in \mathbb{Q}^*$. By countable additivity (since $\mathbb{Q}^*$ is countable):

$$m\left(\bigcup_{q \in \mathbb{Q}^*} M_q\right) = \sum_{q \in \mathbb{Q}^*} m(M_q) = \sum_{q \in \mathbb{Q}^*} c$$

From our inclusions:

$$m([0,1]) \le m\left(\bigcup_{q \in \mathbb{Q}^*} M_q\right) \le m([-1,2])$$

$$1 \le \sum_{q \in \mathbb{Q}^*} c \le 3$$

If $c = 0$, then $1 \le 0$, a contradiction. If $c > 0$, then $1 \le \infty$, which is not a contradiction, but $\sum c \le 3$ implies $\infty \le 3$, a contradiction. Thus, M cannot be assigned a measure $m(M)$, and is **non-measurable**. ☐

Additional Knowledge: The Definition of Lebesgue Measure

Remark 1.3.2

Note: This section provides supplementary material on the definition of the Lebesgue measure. It presents both the original constructive approach and the modern, abstract definition. This content is provided for a deeper historical and theoretical context and can be considered optional for the first reading.

1. Lebesgue's Original Idea & Construction

The original idea, as developed by Henri Lebesgue, is a constructive process for defining the measure of a set $E \subset \mathbb{R}^n$. It starts with simple sets (intervals) and then approximates more complex sets from the outside.

Definition 1.3.13: Lebesgue Outer Measure and Measurability

The **Lebesgue outer measure** $m^*(E)$ of any set $E \subset \mathbb{R}^n$ is defined by covering E with a **countable** collection of n-dimensional intervals (or cubes) and taking the infimum of the total volume of such coverings.

$$m^*(E) = \inf \left\{ \sum_{k=1}^{\infty} \ell(I_k) : E \subset \bigcup_{k=1}^{\infty} I_k \right\}$$

where $\{I_k\}$ is a countable collection of n-dimensional intervals, and $\ell(I_k)$ is the product of the lengths of its sides (its volume).
A set E is called **Lebesgue measurable** if for every $\epsilon > 0$, there exists an open set $O \supset E$ such that the outer measure of the difference $m^*(O \setminus E) < \epsilon$. This is Carathéodory's criterion, a later improvement that perfectly captures the idea that a measurable set can be "approximated closely" by open sets.
For a measurable set E, its Lebesgue measure $m(E)$ is simply defined as its outer measure: $m(E) = m^*(E)$.

2. Modern (Improved) Definition via σ-Algebras

The modern approach, which is more abstract and powerful, defines the Lebesgue measure as the completion of a measure defined on a specific σ-algebra. This is the standard definition found in most modern textbooks on measure theory.

Definition 1.3.14: Lebesgue Measure via σ-Algebra

Let $\mathcal{B}(\mathbb{R}^n)$ be the Borel σ-algebra on $\mathbb{R}^n$. The **Lebesgue σ-algebra**, denoted $\mathcal{L}$, is the completion of $\mathcal{B}(\mathbb{R}^n)$ with respect to the Lebesgue measure.
The **Lebesgue measure** is the unique measure

$$m : \mathcal{L} \to [0, \infty]$$

satisfying the following properties:

1. **Translation Invariance:** For any $A \in \mathcal{L}$ and $x \in \mathbb{R}^n$, $m(A + x) = m(A)$.
2. **Normalization:** The measure of the unit cube is 1: $m([0, 1]^n) = 1$.
3. **Countable Additivity:** For any countable collection $\{E_i\}_{i=1}^{\infty}$ of pairwise disjoint Lebesgue measurable sets, $m\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} m(E_i)$.

Equivalently, it is the unique extension of the pre-measure defined on the algebra of elementary sets (finite unions of intervals) to the full Lebesgue σ-algebra, via Carathéodory's extension theorem.

3. The Vitali Set: An Example of a Non-Measurable Set

An important consequence of the properties of the Lebesgue measure is the existence of sets that are not Lebesgue measurable. The most famous example is the **Vitali set**, constructed by Giuseppe Vitali in 1905.

Definition 1.3.15: Construction of the Vitali Set

Consider the interval $[0, 1] \subset \mathbb{R}$. Define an equivalence relation $\sim$ on $[0, 1]$ by:

$$x \sim y \iff x - y \in \mathbb{Q}.$$

This partitions $[0, 1]$ into equivalence classes. Using the Axiom of Choice, we select exactly one element from each equivalence class to form a set $V \subset [0, 1]$. This set V is called a **Vitali set**.

Theorem 1.3.3: The Vitali Set is Not Lebesgue Measurable

The Vitali set V is not Lebesgue measurable.

Proof. The proof proceeds by contradiction. Assume V is Lebesgue measurable.

Consider the rational numbers in $[-1,1]$, denoted $\mathbb{Q}\cap[-1,1]$. For each $q \in \mathbb{Q}\cap[-1,1]$, define the translation:

$$V_q = V + q = \{v + q : v \in V\}.$$

These sets V_q are pairwise disjoint. If $v_1+q_1 = v_2+q_2$ for $v_1, v_2 \in V$ and $q_1, q_2 \in \mathbb{Q}$, then $v_1-v_2 = q_2-q_1 \in \mathbb{Q}$, which implies $v_1 \sim v_2$. Since V contains exactly one element from each equivalence class, $v_1 = v_2$ and thus $q_1 = q_2$.

By translation invariance of the Lebesgue measure, if V is measurable, then each V_q is measurable and $m(V_q) = m(V)$.

Now, observe that:

$$[0,1] \subset \bigcup_{q\in\mathbb{Q}\cap[-1,1]} V_q \subset [-1,2].$$

If V is measurable with $m(V) = 0$, then by countable additivity:

$$m\left(\bigcup_q V_q\right) = \sum_q m(V_q) = 0,$$

which contradicts $m([0,1]) = 1$.

If $m(V) > 0$, then:

$$m\left(\bigcup_q V_q\right) = \sum_q m(V_q) = \infty,$$

which contradicts $m([-1,2]) = 3$.

Therefore, our assumption that V is measurable must be false. The Vitali set V is not Lebesgue measurable. We can't find any countable intervals to cover the Vitali set. □

Remark 1.3.3

The existence of non-measurable sets like the Vitali set demonstrates that the Lebesgue measure cannot be extended to all subsets of $\mathbb{R}^n$ while preserving translation invariance and countable additivity. This result relies on the Axiom of Choice, and indeed, it can be shown that in models of set theory without the Axiom of Choice, all subsets of $\mathbb{R}$ can be Lebesgue measurable.

1.3.4 Another Example: Closure of Relation

Having defined the fundamental properties of relations, we now address a natural question: if a relation R on a set A lacks a certain property, what is the *smallest* relation containing R that *does* possess that property? This leads to the concept of the **closure** of a relation.

Definition 1.3.16: Closure

Let P be a property of relations (such as reflexivity, symmetry, or transitivity). The P**-closure** of a relation R on a set A is the smallest relation S on A such that:

1. $R \subseteq S$

2. S has the property P

There are three important types of closure:

Reflexive Closure $R' = R \cup I_A$, where $I_A = \{(a, a) \mid a \in A\}$ is the identity relation on A.

Symmetric Closure $R' = R \cup R^{-1}$.

Transitive Closure $R^* = \bigcup_{n=1}^{\infty} R^n$, where $R^1 = R$ and $R^{n+1} = R^n \circ R$.

Proof that R^* is the Transitive Closure. We must show R^* is transitive and is the smallest such relation containing R.

1. **(Transitivity)** Let $(x, y) \in R^*$ and $(y, z) \in R^*$. By definition, $(x, y) \in R^m$ for some $m \geq 1$ and $(y, z) \in R^n$ for some $n \geq 1$. By definition of composition, this implies $(x, z) \in R^{m+n}$. Since $m + n \geq 1$, $(x, z) \in \bigcup_{k=1}^{\infty} R^k = R^*$. Thus R^* is transitive.
2. **(Minimality)** Let T be any transitive relation such that $R \subseteq T$. We must show $R^* \subseteq T$. We know $R^1 = R \subseteq T$. Assume $R^k \subseteq T$ (inductive hypothesis). Let $(a, c) \in R^{k+1} = R^k \circ R$. Then $\exists b$ such that $(a, b) \in R^k$ and $(b, c) \in R$. By the inductive hypothesis, $(a, b) \in T$. Since $R \subseteq T$, $(b, c) \in T$. Because T is transitive, $(a, b) \in T \wedge (b, c) \in T \implies (a, c) \in T$. Thus $R^{k+1} \subseteq T$. By induction, $R^n \subseteq T$ for all $n \geq 1$. Therefore, $R^* = \bigcup_{n=1}^{\infty} R^n \subseteq T$. This shows R^* is the smallest transitive relation containing R.

□

1.3.5 Mapping and Function

In mathematical expressions, we use the language of set theory to define what a map is.

Definition 1.3.17: Map / Function

A **map** (or **function**) f from A to B, denoted $f : A \to B$, is a relation $f \subseteq A \times B$ such that for every $x \in A$, there is a *unique* object $y \in B$ such that $(x, y) \in f$. We write this unique y as $f(x)$.

- The set A is called the **domain** of f.
- The set B is called the **codomain** of f.
- The set $\{f(x) \mid x \in A\} \subseteq B$ is called the **range** or **image** of f.

The uniqueness requirement is: $\forall x \in A \forall y_1, y_2 \in B[((x, y_1) \in f \wedge (x, y_2) \in f) \to y_1 = y_2]$.

Definition 1.3.18: Injection

A function $f : A \to B$ is an **injection** (or one-to-one) if and only if:

$$\forall x_1, x_2 \in A(f(x_1) = f(x_2) \to x_1 = x_2)$$

This means that each element in the domain corresponds to a unique element in the codomain.

Definition 1.3.19: Surjection

A function $f : A \to B$ is a **surjection** (or onto) if and only if:

$$\forall y \in B \exists x \in A(f(x) = y)$$

This means that the codomain is equal to the range.

Definition 1.3.20: Bijection

A function $f : A \to B$ is a **bijection** if and only if it is both an injection and a surjection.

Definition 1.3.21: Inverse Function

Let $f : A \to B$ be a function. A function $g : B \to A$ is called the **inverse function** (or inverse map) of f if and only if it satisfies the following two conditions:

1. $g \circ f = id_A$ (where id_A is the identity map on A)
2. $f \circ g = id_B$ (where id_B is the identity map on B)

The inverse map of f, if it exists, is unique and denoted by f^{-1}. A function has an inverse if and only if it is a bijection.

Definition 1.3.22: Function Composition

Let $f : A \to B$ and $g : B \to C$ be two functions. The **composition** of g and f is a new function denoted $g \circ f : A \to C$ defined as:

$$(g \circ f)(x) = g(f(x)) \quad \text{for all } x \in A$$

Function (Special Types)

Definition 1.3.23: Real Function

If the domain $X \subseteq \mathbb{R}$ and the codomain $Y \subseteq \mathbb{R}$, the mapping is called a **real function of one variable**, denoted $y = f(x)$.

Piecewise Function A piecewise function is a function that is defined by multiple sub-functions, each of which applies to a certain interval or region of the main function's domain.

$$f(x) = \begin{cases} f_1(x) & \text{if condition 1} \\ f_2(x) & \text{if condition 2} \\ \vdots & \vdots \end{cases}$$

Implicit Function An implicit function is a function that is defined by an equation relating its variables, e.g., $F(x, y) = 0$, rather than by an explicit formula $y = f(x)$.

Parametric Function A parametric function describes a curve by expressing the coordinates of the points on the curve as functions of a third variable, called a parameter, t.

$$\begin{cases} x = f(t) \\ y = g(t) \end{cases} \qquad \text{for } t \text{ in some interval } I$$

Definition 1.3.24: Basic Elementary Functions

Basic elementary functions are a finite set of fundamental functions:

1. Constant Functions: $f(x) = c$.
2. Power Functions: $f(x) = x^\alpha$.
3. Exponential Functions: $f(x) = a^x$ (where $a > 0, a \neq 1$).
4. Logarithmic Functions: $f(x) = \log_a x$.

5. Trigonometric Functions: $\sin x, \cos x, \tan x$, etc.
6. Inverse Trigonometric Functions: $\arcsin x, \arccos x$, etc.

Definition 1.3.25: Elementary Function

An **elementary function** is any function that can be obtained from the basic elementary functions by performing a finite number of the following operations:

- Arithmetic Operations: Addition, subtraction, multiplication, division.
- Composition: The operation of function composition.

Definition 1.3.26: Operation

An **operation** is a function from a set to itself. More specifically, an **n-ary operation** ω on a set X is a function $\omega : X^n \to X$.

1.3.6 Ordered Structure

Well-order

In previous parts, we introduced several ordered relations. Now we will define the concept and property of well-order. This is the last part of chapter 1, and well-order can be seen as a bridge to the next chapter.

Definition 1.3.27: Well-ordered Set

A totally ordered set $(W, \leq)$ is called a **well-ordered set** if it satisfies that every non-empty subset has a least element.

$$\forall S \subseteq W(S \neq \emptyset \rightarrow \exists s \in S \forall x \in S(s \leq x))$$

Theorem 1.3.4

The set of natural numbers $\mathbb{N}$ under the usual order $\leq$ is well-ordered. (This is the **Well-Ordering Principle**).

Proof. Consider an arbitrary non-empty subset S of $\mathbb{N}$. Start checking each natural number from 0 onwards. The first number that belongs to S is the least element. This process must terminate in a finite number of steps, because if no such element were ever found, it would imply that S is empty, contradicting the assumption. Thus, $\mathbb{N}$ under the usual order is a well-ordered set. $\square$

It seems that there is a very close connection between the well-ordering principle and the set of natural numbers. And this principle of "starting from the least element" is seemingly very symmetrical to Mathematical Induction (MI). We will reveal their relationships in the next chapter.

Keywords: Set, Axiom, ZFC Axiomatic Set theory, Ordered Pairs, Cartesian Product, Relations, Well-ordered Sets.
Reference: Discrete Mathematics and Its Applications (Eighth Edition), Kenneth H. Rosen, McGraw-Hill Education.

Chapter 2

Mathematical Analysis: Part I

We now embark on a great undertaking: building a rigorous framework for calculus. Our journey through logic and set theory has provided the tools; the ordered structures of Chapter 1 have revealed a critical insight—the rational numbers, though dense, are incomplete. They possess gaps, like the legendary irrational $\sqrt{2}$, which defy representation as a ratio.

This chapter confronts that insufficiency. We will construct the real number system, a complete, continuous tapestry woven to fill these voids. Its cornerstone is the Completeness Axiom, which guarantees that bounded sets have precise bounds, a property the rationals fatally lack.

Upon this unshakable foundation, we will erect the central pillars of analysis: the precise theory of limits, the formal definition of continuity, and the powerful machinery of the derivative and integral in the next chapter. This is the transition from intuitive calculation to profound understanding—from calculus to Analysis.

2.1 Extension of the Number System

If asked about the foundation of mathematical analysis, the answer is clear: the axiom of real numbers. In this part, we will start from Peano Axioms to form the axioms of natural numbers. And for practical purposes, we extend it to integers and rational numbers. Then, to support theorems of calculus, we construct the axioms of real numbers and complex numbers.

Before we officially start this part, we need to clarify three simple principles:

Motivation Principle (Solving a Limitation) Each expansion is driven by the need to perform an operation that is not always possible within the smaller number system. The primary goal is to achieve **closure** under this new operation.

Embedding Principle (Preserving the Original Structure) The smaller, original number system must be isomorphic to a subsystem of the new, larger number system. This is achieved by constructing an **injective embedding** that preserves all the essential operations (like addition and multiplication) and properties of the original system.

Minimality Principle (The "Smallest" Extension) The new number system should be the "smallest" or "most economical" extension that satisfies the first two principles. It should introduce *only* the elements necessary to solve the limitation, without any superfluous structure.

2.1.1 Peano Axioms and Natural Numbers

Peano axioms define natural numbers using 5 axioms:

Definition 2.1.1: Peano's Axioms

A set $\mathbb{N}$ is called the **natural number set** if it satisfies the following properties (with a "successor" function $S : \mathbb{N} \to \mathbb{N}$):

1. $0 \in \mathbb{N}$. (Zero is a natural number.)
2. $\forall a \in \mathbb{N}(S(a) \in \mathbb{N})$. (If a is a natural number, the successor of a is a natural number.)
3. $\forall a \in \mathbb{N}(S(a) \neq 0)$. (Zero is not the successor of any natural number.)
4. $\forall a, b \in \mathbb{N}(S(a) = S(b) \to a = b)$. (Two numbers whose successors are equal are themselves equal.)
5. $\forall K \subseteq \mathbb{N}[(0 \in K \land \forall n(n \in K \to S(n) \in K)) \to K = \mathbb{N}]$. (If a set S of numbers contains zero and also the successor of every number in S, then every number is in S. This is the **Axiom of Induction**.)

We denote the successor of a as $a + 1$.

The definition of the natural numbers has a very strong relationship with the concept of mathematical induction and the well-ordered set. In fact, they are logically equivalent (assuming the other axioms).

- Peano axioms ensure the validity of mathematical induction (Axiom 5 is MI).
- Peano axioms ensure the validity of the well-ordering principle.

Here, we will prove the logical equivalence of the well-ordering principle (WOP) and mathematical induction (MI).

MI implies Well-Ordering Principle. **To prove:** every non-empty subset of natural numbers has a least element. Assume, for contradiction, that there exists a non-empty subset $A \subseteq \mathbb{N}$ that has *no* least element. Define another set $B = \{n \in \mathbb{N} \mid n < a \text{ for all } a \in A\}$. ($B$ is the set of all numbers strictly smaller than everything in A).

Base Case: $0 \in B$? If $0 \notin B$, then there exists some $a \in A$ such that $0 \geq a$. Since $a \in \mathbb{N}$, this implies $a = 0$. So $0 \in A$. Since 0 is the smallest natural number, it would be the least element of A, contradicting the assumption that A has no least element. Thus, $0 \in B$.

Inductive Step: Assume $k \in B$ and prove $S(k) \in B$. Inductive Hypothesis: $k \in B$, meaning $k < a$ for all $a \in A$. If $S(k) \notin B$, then there exists some $a \in A$ such that $S(k) \geq a$. From the hypothesis, $k < a$. Combined with $S(k) \geq a$, and since numbers are discrete, the only possibility is $a = S(k)$. So $S(k) \in A$. Furthermore, since all numbers smaller than $S(k)$ (like k) are in B (and thus not in A), $S(k)$ would be the least element of A. This contradicts the assumption that A has no least element. Therefore, $S(k) \in B$.

Conclusion: By the principle of Mathematical Induction (Axiom 5), $B = \mathbb{N}$. Since A is non-empty, take any element $a \in A$. Because $B = \mathbb{N}$, $a \in B$. By the definition of B, this means $a < a$, which is a contradiction. The initial assumption that A has no least element must be false. Therefore, every non-empty subset of $\mathbb{N}$ has a least element. □

Well-Ordering Principle implies MI. (This part is left for the readers to practice.) □

Cardinality

The concept "amount" is clear for finite sets. But for an infinite set, how can we measure the "size" of the set?

Definition 2.1.2: Cardinality / Equinumerosity

Cardinality is an intrinsic property of sets. Two sets A and B are said to be **equinumerous** or have the same **cardinality**, denoted $A \approx B$ or $|A| = |B|$, if there exists a bijection $f : A \to B$.

Definition 2.1.3

For two sets A and B, if we can find an injection $f : A \to B$ but no bijection, we said the set A is strictly smaller than set B, denoted $A \prec B$ or $|A| < |B|$. If there is an injection $f : A \to B$, we write $A \preceq B$ or $|A| \leq |B|$.

Theorem 2.1.1: Schröder-Bernstein Theorem

If $A \preceq B$ and $B \preceq A$, then $A \approx B$.

Definition 2.1.4: Countable Set

A set is **countable** if either it is finite or it can be made in one-to-one correspondence with the set of natural numbers $\mathbb{N}$. If a set is countably infinite, we define its cardinality as $\aleph_0$ (aleph-nought).

Theorem 2.1.2: Cantor's Theorem

For every non-empty set A, $A \prec \mathcal{P}(A)$. (That is, $|A| < |\mathcal{P}(A)|$).

(This implies there is no "largest" infinity.)

In section 2.3, we will know that the cardinality of the real number set $\mathbb{R}$ is $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})| = 2^{\aleph_0}$, which is called the **Continuum**.

The Continuum Hypothesis (CH) Some students might be interested whether there exists any size of a set between $\aleph_0$ and $2^{\aleph_0}$. The Continuum Hypothesis (CH) is a famous conjecture that proposes there is no set S such that $\aleph_0 < |S| < 2^{\aleph_0}$.

So can we determine whether it's true or false? It was proven (by Kurt Gödel and Paul Cohen) that CH is "undecidable" within the standard ZFC foundation of mathematics. This means it can neither be proven true nor false using the accepted axioms, revealing a fundamental limitation of that system.

2.1.2 Integers and Rational Numbers

Why do we need integers and rational numbers? Why are natural numbers not enough?

Definition 2.1.5: Closure of operation

Closure of operation refers to the property that when an operation is performed on members of a set, the result is always a member of the same set.

In the set of natural numbers $\mathbb{N}$, operations like addition and multiplication are closed. But for subtraction $(3 - 5 =?)$ and division $(3/5 =?)$, $\mathbb{N}$ itself is not enough. Thus, we need to extend the number system.

Construction of Integers ($\mathbb{Z}$)

Definition 2.1.6

The relation R on $\mathbb{N} \times \mathbb{N}$ is defined as:

$$(a, b)R(c, d) \iff a + d = b + c$$

(This is the formal way of saying $a - b = c - d$).

Theorem 2.1.3

The relation R is an equivalence relation on $\mathbb{N} \times \mathbb{N}$.

Proof. (Left for readers to practice: check reflexivity, symmetry, transitivity.) □

Definition 2.1.7: Integers $\mathbb{Z}$

The **integers**, denoted as $\mathbb{Z}$, is the set of equivalence classes of $\mathbb{N} \times \mathbb{N}$ w.r.t the equivalence relation R. We denote the equivalence class $[(a, b)]$ as $a - b$.

We can now define operations on $\mathbb{Z}$:

Definition 2.1.8: Operations on $\mathbb{Z}$

- **Addition:** $[(a, b)] + [(c, d)] = [(a + c, b + d)]$
- **Negation:** $-[(a, b)] = [(b, a)]$
- **Subtraction:** $[(a, b)] - [(c, d)] = [(a, b)] + (-[(c, d)]) = [(a, b)] + [(d, c)] = [(a + d, b + c)]$

Because $a, b, c, d \in \mathbb{N}$, and $\mathbb{N}$ is closed under addition, $a + d$ and $b + c$ also belong to $\mathbb{N}$. This means that subtraction is closed for integers. The construction of $\mathbb{Z}$ from $\mathbb{N}$ via ordered pairs is a perfect example of the principles of number system expansion.

Construction of Rational Numbers ($\mathbb{Q}$) Though $\mathbb{Z}$ is closed under subtraction, it is not closed under division. That's why we still need to extend from integers to rational numbers.

Definition 2.1.9

Let $\mathbb{Z}^* = \mathbb{Z} - \{0\}$. The relation R on $\mathbb{Z} \times \mathbb{Z}^*$ is defined by:

$$(a, b)R(c, d) \iff ad = bc$$

(This is the formal way of saying $a/b = c/d$).

It's clear that this relation R is an equivalence relation.

Definition 2.1.10: Rational Numbers $\mathbb{Q}$

The set of **rational numbers**, denoted $\mathbb{Q}$, is the set of equivalence classes of $\mathbb{Z} \times \mathbb{Z}^*$ w.r.t the equivalence relation R. We denote the class $[(a, b)]$ as a/b.

Density of Rational Numbers There is a very different property of rational numbers compared to integers. The rational number set is a **dense order set**. We can't always find an intermediate number between two integers (e.g., 1 and 2), but we can always find a rational intermediate value between two rational numbers.

Proof. For two rational numbers a and b with $a < b$, their average $m = (a + b)/2$ is also a rational number, and $a < m < b$. Thus, we found an intermediate number between them. □

This property (the density of $\mathbb{Q}$ in $\mathbb{R}$) is essential for constructing and understanding the real number system. In topology, $\mathbb{Q}$ is a countable dense subset of $\mathbb{R}$, making $\mathbb{R}$ a separable space.

2.1.3 Real Numbers and Complex Numbers

The rational numbers $\mathbb{Q}$, while dense, are not *complete*. They contain "gaps". For example, the set $A = \{x \in \mathbb{Q} \mid x^2 < 2\}$ has upper bounds in $\mathbb{Q}$ (e.g., 1.5), but it has no *least* upper bound *within* $\mathbb{Q}$. The "number" $\sqrt{2}$ is missing. We construct the real numbers $\mathbb{R}$ to fill these gaps.

Construction of Real Numbers by Dedekind Cuts

Definition 2.1.11: Dedekind Cut

A **Dedekind cut** is a pair (A, B) of subsets of $\mathbb{Q}$ satisfying:

1. A and B are non-empty and form a partition of $\mathbb{Q}$ (i.e., $A \cup B = \mathbb{Q}$ and $A \cap B = \emptyset$).
2. Every element of A is less than every element of B.
3. A has no greatest element.

The set A is called the **lower class** and B the **upper class**. A real number is defined as a Dedekind cut.

For example, the real number $\sqrt{2}$ is represented by the cut where:

- $A = \{x \in \mathbb{Q} \mid x < 0 \text{ or } x^2 < 2\}$
- $B = \{x \in \mathbb{Q} \mid x > 0 \text{ and } x^2 > 2\}$

Definition 2.1.12: Order on Real Numbers

For two real numbers $\alpha = (A_1, B_1)$ and $\beta = (A_2, B_2)$, we define $\alpha < \beta$ if $A_1 \subset A_2$ (proper subset). We define $\alpha = \beta$ if $A_1 = A_2$.

Definition 2.1.13: Addition of Real Numbers

Let $\alpha = (A_1, B_1)$ and $\beta = (A_2, B_2)$ be real numbers. Define:

- $A = \{a_1 + a_2 \mid a_1 \in A_1, a_2 \in A_2\}$
- $B = \mathbb{Q} \setminus A$

Then $\alpha + \beta$ is defined as the cut (A, B).

Definition 2.1.14: Multiplication of Positive Real Numbers

For positive real numbers $\alpha = (A_1, B_1)$ and $\beta = (A_2, B_2)$ (where "positive" means they contain some positive rationals in their lower classes), define:

- $A = \{a_1 a_2 \mid a_1 \in A_1, a_2 \in A_2, a_1 > 0, a_2 > 0\} \cup \{q \in \mathbb{Q} \mid q \leq 0\}$
- $B = \mathbb{Q} \setminus A$

Then $\alpha \cdot \beta$ is defined as the cut (A, B). For other sign combinations, we adjust the definition accordingly.

Definition 2.1.15: Additive Inverse

For a real number $\alpha = (A, B)$, define its additive inverse $-\alpha$ by:

- $A' = \{-b \mid b \in B, b \text{ is not the smallest element of } B\}$
- $B' = \mathbb{Q} \setminus A'$

Then $-\alpha = (A', B')$.

These operations make $\mathbb{R}$ an ordered field. The multiplicative inverse can be defined similarly for non-zero elements.

Now we can define the fundamental concepts of analysis:

Definition 2.1.16: Upper Bound

Let $S \subseteq \mathbb{R}$. A number u is an **upper bound** of S if $s \leq u$ for all $s \in S$. A number l is a **lower bound** of S if $l \leq s$ for all $s \in S$.

Definition 2.1.17: Supremum

The **supremum** (or **least upper bound**) of S, denoted $\sup S$, is the smallest upper bound of S. That is:

1. $s \leq \sup S$ for all $s \in S$. ($\sup S$ is an upper bound.)
2. If v is any upper bound of S, then $\sup S \leq v$. (It is the *least* upper bound.)

The definition of **infimum** (or **greatest lower bound**), denoted $\inf S$, is similar.

Theorem 2.1.4: Completeness Axiom

Every non-empty subset of $\mathbb{R}$ that is bounded above has a supremum in $\mathbb{R}$.

(The case for the infimum is analogous: every non-empty subset of $\mathbb{R}$ that is bounded below has an infimum in $\mathbb{R}$.)

Proof. In the Dedekind cut construction, the supremum of a bounded set S of real numbers is given by the union of the lower classes of all elements in S. More precisely, if $S = \{\alpha_i = (A_i, B_i)\}$ is bounded above, then:

$$\sup S = \left(\bigcup_i A_i, \mathbb{Q} \setminus \bigcup_i A_i \right)$$

This pair forms a Dedekind cut and satisfies the definition of supremum. $\square$

The completeness axiom is fundamental to analysis and distinguishes $\mathbb{R}$ from $\mathbb{Q}$. It ensures that limits of Cauchy sequences exist, continuous functions attain their maximum and minimum on closed intervals, and many other essential properties.

Definition 2.1.18: Archimedean Property

The real numbers satisfy the **Archimedean property**: for any $x \in \mathbb{R}$, there exists a natural number n such that $n > x$. Equivalently, for any $\epsilon > 0$, there exists $n \in \mathbb{N}$ such that $1/n < \epsilon$.

Theorem 2.1.5: Density of Rationals

Between any two distinct real numbers, there exists a rational number. That is, $\mathbb{Q}$ is **dense** in $\mathbb{R}$.

While $\mathbb{R}$ solves the completeness problem of $\mathbb{Q}$, it is not **algebraically closed**. There are polynomial equations with real coefficients that have no real solutions, such as $x^2 + 1 = 0$. This motivates the extension to complex numbers.

Definition 2.1.19: Complex Numbers

The set of **complex numbers**, denoted $\mathbb{C}$, consists of all expressions of the form $a+bi$, where $a, b \in \mathbb{R}$ and i is the **imaginary unit** satisfying $i^2 = -1$. For $z = a + bi \in \mathbb{C}$:

- a is called the **real part**, denoted $\Re(z)$
- b is called the **imaginary part**, denoted $\Im(z)$

Two complex numbers $a + bi$ and $c + di$ are equal if and only if $a = c$ and $b = d$.

Definition 2.1.20: Operations on Complex Numbers

For complex numbers $z = a + bi$ and $w = c + di$, we define:
- **Addition**: $z + w = (a + c) + (b + d)i$
- **Multiplication**: $zw = (ac - bd) + (ad + bc)i$

With these operations, $\mathbb{C}$ forms a field. The real numbers $\mathbb{R}$ can be identified with the subset $\{a+0i : a \in \mathbb{R}\}$ of $\mathbb{C}$.

Definition 2.1.21: Complex Conjugate and Modulus

For $z = a + bi \in \mathbb{C}$:
- The **complex conjugate** is $\bar{z} = a - bi$
- The **modulus** (or absolute value) is $|z| = \sqrt{a^2 + b^2}$

Theorem 2.1.6: Properties of Conjugate and Modulus

For $z, w \in \mathbb{C}$:
1. $\overline{z + w} = \bar{z} + \bar{w}$
2. $\overline{zw} = \bar{z} \cdot \bar{w}$
3. $z\bar{z} = |z|^2$
4. $|zw| = |z||w|$
5. $|z + w| \leq |z| + |w|$ (Triangle Inequality)

Theorem 2.1.7: Fundamental Theorem of Algebra

Every non-constant polynomial with complex coefficients has at least one complex root. Equivalently, $\mathbb{C}$ is algebraically closed.

This theorem is profound: while we extended $\mathbb{R}$ to $\mathbb{C}$ to solve the equation $x^2 + 1 = 0$, we actually obtained a number system where every polynomial equation has a solution.

Definition 2.1.22: Polar Form of Complex Numbers

Any complex number $z = a + bi \neq 0$ can be written in **polar form** as:

$$z = r(\cos\theta + i\sin\theta)$$

where $r = |z| > 0$ is the modulus and θ is the **argument** of z, satisfying $\tan\theta = b/a$.

Using Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$, we can write $z = re^{i\theta}$.

Theorem 2.1.8: De Moivre's Theorem

For any integer n and complex number $z = r(\cos\theta + i\sin\theta)$:

$$z^n = r^n(\cos(n\theta) + i\sin(n\theta))$$

This theorem simplifies computations with powers and roots of complex numbers.

The complex numbers provide a powerful framework for many areas of mathematics, physics, and engineering. They allow us to:

- Solve all polynomial equations

- Represent periodic phenomena using complex exponentials
- Analyze signals and systems in electrical engineering
- Study fluid dynamics and electromagnetism
- Develop the mathematical foundation of quantum mechanics

2.2 Sequence Limit and The Properties of Real Numbers

As long as we finished the content about real numbers, we will now move on to the the core of the mathematical analysis: the **limit**. Why I claim that the limit is the core concept of mathematical analysis? In the following content you will realize that almost all the concept has some kind of connection with limit. I can say that the limit is the basis of many theory. And the language of limit represents a dynamic, approaching, and rigorous mathematical mindset. Let us begin.

2.2.1 Definitions and Basic Properties

The Definition of Limits

Definition 2.2.1: limit of a sequence

A sequence $\{a_n\}$ converges to a real number A if for all $\epsilon > 0$, there exists an integer N such that $|a_n - A| < \epsilon$ if $n \geq N$. The number A is the limit of the sequence and we write:

$$\lim_{n\to\infty} a_n = A$$

Naively speaking, if the sequence $\{a_n\}$ is a **convergent sequence** and A is the limit of the sequence, the value of a_n become arbitrarily close to a finite number A. You can get any value that is anyhow closer to the convergence A, once you pick a big enough n.

In a more commonly used language, if we pick an open interval in the real number line, whose center is a and radius is ϵ, written as $(a - \epsilon, a + \epsilon)$. We call this kind of intervals the neighborhood, denoted as $O(a, \epsilon)$:

$$O(a, \epsilon) = \{x | a - \epsilon < x < a + \epsilon\}$$

And what the definition said is that for all terms after a_n fall within the $O(a, \epsilon)$. Since the neighborhood is contractive, the sequence eventually converges to a.

However, in the contrary, if a sequence $\{a_n\}$ is not convergent, we say it is a **divergent sequence**. Rigorously speaking, if for all $\epsilon > 0$, $N \in \mathbb{N}^*$ and $A \in \mathbb{R}$, there exists at least one n_0, $|a_{n_0} - A| > \epsilon$.

For those sequences converges to 0, we call those sequences **infinitesimal**.

Remark 2.2.1

When we talk about infinitesimal, what we are discussing about is a sequence rather than a simple number. Be clear that infinitesimal is not a number.

Properties of Limits

After we define what is limit, let's take a look at it's properties.

Theorem 2.2.1: the uniqueness of limit

The limit of a convergent sequence must be unique.

Proof. Assume there exists a sequence $\{a_n\}$ that converges to two different values a and b, $a \neq b$. According to the definition of limit:

$$\forall \epsilon > 0, \exists N_1, n > N_1 : |x_n - a| < \epsilon/2$$

$$\forall \epsilon > 0, \exists N_2, n > N_2 : |x_n - b| < \epsilon/2$$

Pick $N = max\{N_1, N_2\}$, according to the triangle inequality, then $\forall n > N$ we have:

$$|a - b| = |a - x_n + x_n - b| \leq |x_n - a| + |x_n - b| < \epsilon$$

Since ϵ can get arbitrarily close to 0, we know that $a = b$ □

Theorem 2.2.2

A convergent sequence must be bounded.

Remark 2.2.2

However, the contrapositive is not always true. A bounded sequence may not be convergent.
Consider $a_n = (-1)^n$, the sequence $\{a_n\}$ bounded from -1 to 1, but the sequence won't converges to any value.
In the future may be we can add a stronger condition to make the $\{a_n\}$ convergent. If you are interested, please move on to the content in 2.2.2.

Proof. Assume $\{a_n\}$ is a convergent sequence, the limit is a. According to the definition of limit, we pick $\epsilon = 1$, thus $\exists N, \forall n > N : |x_n - a| < 1$, then $a - 1 < x_n < a + 1$

let $m = max\{a_1, a_2, ..., a_N, a + 1\}$, $M = min\{a_1, a_2, ..., a_N, a - 1\}$

Then we have $m \leq x_n \leq M$, which means that the sequence $\{x_n\}$ is bounded. □

Theorem 2.2.3: isotonicity

Assume we have two sequence $\{x_n\}$ and $\{y_n\}$, they converges to two different limits a and b, and $a < b$. There exists a N, $\forall n > N$, $x_n < y_n$

Proof. For two sequence $\{x_n\}$ and $\{y_n\}$ that converge to two different values a and b. Let's assume $a > b$

According to the definition of limit, we take $\epsilon = \frac{a-b}{2}$, and we have:

$$\exists N_1, \forall n > N_1, |x_n - a| < \epsilon$$

$$\exists N_2, \forall n > N_2, |y_n - b| < \epsilon$$

Thus we have: $x_n > (a + b)/2 > y_n$, pick $N = max\{N_1, N_2\}$, then we have $\forall n > N, x_n > y_n$. Finally we can claim that the limits have the property of isotonicity. □

Theorem 2.2.4: the calculations' law of limits

Assume that there exist two limits:

$$\lim_{n\to\infty} x_n = a, \lim_{n\to\infty} y_n = b$$

And we have:

- $\lim_{n\to\infty}(\alpha x_n + \beta y_n) = \alpha a + \beta b$, for two constants α and β.
- $\lim_{n\to\infty} x_n y_n = ab$
- $\lim_{n\to\infty}(\frac{x_n}{y_n}) = \frac{a}{b}, (b \neq 0)$

We recommend the readers finish the proof above themselves.

After we finished the definition and some basic properties of the convergent sequence, we will now move on to define a kind of not-convergent sequence: **the infinity**.

Infinity

Definition 2.2.2: Infinity

If we have a sequence $\{x_n\}$, for every given G, $G > 0$, we can find $N \in \mathbb{N}$, $\forall n > N$, we have $|x_n| > G$, we call the sequence $\{x_n\}$ is a infinity, denoted as:

$$\lim_{n\to\infty} x_n = \infty$$

Just like infinitesimal, infinity is also a sequence rather that a number. But in some conditions, we will deal with it as if it is a number.

If a infinity start to be positive from some point, we call this form of infinity the **positive infinity**. We can define what is negative infinity likewise. We denote them specially like this:

$$\lim_{n\to\infty} a_n = +\infty, (\lim_{n\to\infty} a_n = -\infty)$$

Theorem 2.2.5

The infinity has special relationship with the infinitesimal: The sequence $\{x_n\}(x_n \neq 0)$ is infinity iff $\{\frac{1}{x_n}\}$ is infinitesimal.

Using the definition of the limit will be enough to prove this theorem. We'll skip this part here.

Theorem 2.2.6

Assume $\{x_n\}$ is a infinity, if when $n > N_0$, $|y_n| \geq \delta > 0$, then $\{x_n y_n\}$ is infinity.

Corollary 2.2.1

Assume $\{x_n\}$ is infinity, $\lim_{n\to\infty} y_n = b \neq 0$, then $\{x_n y_n\}$ and $\{\frac{x_n}{y_n}\}$ are both infinity.

Stolz Theorem

With the help of the definitions of the limit, we can calculate various kinds of limits using algebraic techniques and the definition of limit. But when we face certain forms of limit like $\frac{0}{0}$ and $\frac{\infty}{\infty}$, they are especially tricky to deal with. But with the help of **Stolz theorem** we are going to introduced now, it will be much easier to deal with them (in some occasions).

Definition 2.2.3: Increasing Function

If a sequence $\{x_n\}$ satisfies: $x_n \leq x_{n+1}, n = 1, 2, 3\cdots$, we will call it **the monotone increasing function**.

Definition 2.2.4: Strict Monotone Increasing Function

If a sequence $\{x_n\}$ satisfies: $x_n < x_{n+1}, n = 1, 2, 3\cdots$, we will call it the **strict monotone increasing function**.

Theorem 2.2.7

Let $\{y_n\}$ be a **strict monotone increasing positive infinity**, and:

$$\lim_{n\to\infty} \frac{x_n - x_{n-1}}{y_n - y_{n-1}} = a$$

Then we have that:

$$\lim_{n\to\infty} \frac{x_n}{y_n} = a$$

Proof. Let's consider the condition when $a = 0$. Because $\lim_{n\to\infty} \frac{x_n - x_{n-1}}{y_n - y_{n-1}} = 0$, according to the definition of limit, we know that:

$$\forall\epsilon > 0, \exists N_1, \forall n > N_1 : |x_n - x_{n-1}| < \epsilon(y_n - y_{n-1})$$

Because $\{y_n\}$ is infinity, we can let $y_{N_1} > 0$ obviously. For the inequality above, we take everything from N_1 to n, and then we add them together, we have:

$$|x_n - x_{N_1}| \leq |x_n - x_{n-1}| + |x_{n-1} - x_{n-2}| + \cdots + |x_{N_1+1} - X_{N_1}|$$

$$< \epsilon(y_n - y_{n-1}) + \cdots + \epsilon(y_{N_1+1} - y_{N_1}) = \epsilon(y_n - y_{n-1})$$

Divide both side of the inequality by y_n, and we have:

$$|\frac{x_n}{y_n} - \frac{x_{N_1}}{y_n}| \leq \epsilon(1 - \frac{y_{N_1}}{y_n}) \leq \epsilon$$

And, for a fixed N_1, we can pick $N > N_1$, $\forall n > N : |\frac{x_{N_1}}{y_n}| < \epsilon$, then we have:

$$|\frac{x_n}{y_n}| < \epsilon + |\frac{x_{N_1}}{y_n}| < 2\epsilon$$

For other conditions: if a is a bounded value, and $a \neq 0$, let $x_n{}' = x_n - ay_n$, and with the help of the proof above, we can reach the conclusion.

When $a = +\infty$, We take the reciprocal of $\frac{x_n}{y_n}$. Similarly, it is not difficult to reach a conclusion.

$\square$

2.2.2 Convergence Criteria and the Properties of the Real Number System

Before we discuss more advanced concepts in analysis (such as derivatives and integrals), we must first firmly establish a fundamental property of the real number system $\mathbb{R}$: **Completeness**. It is this property that distinguishes the real numbers $\mathbb{R}$ from the rational numbers $\mathbb{Q}$ and serves as the bedrock for all important theorems in analysis.

The Completeness of the Real Number System

The completeness of the real number system can be expressed in several equivalent ways. Let's take a review. (The proof is in section 2.1.3)

Theorem 2.2.8: The Completeness Axiom

Every non-empty subset of $\mathbb{R}$ that is bounded above has a supremum in $\mathbb{R}$.

This axiom, while seemingly simple, directly leads to the first major convergence criterion in analysis.

The Monotone Convergence Theorem

We previously defined monotone increasing sequences. The Completeness Axiom guarantees that a bounded monotone sequence must converge.

Theorem 2.2.9: Monotone Convergence Theorem

A monotone sequence (either increasing or decreasing) that is bounded must converge.

- (i) If $\{x_n\}$ is a monotone increasing and bounded above, then $\lim_{n\to\infty} x_n = \sup\{x_n\}$.
- (ii) If $\{x_n\}$ is a monotone decreasing and bounded below, then $\lim_{n\to\infty} x_n = \inf\{x_n\}$.

Proof. We will prove (i); the proof for (ii) is analogous. Let $\{x_n\}$ be a monotone increasing sequence that is bounded above. Let $S = \{x_n \mid n \in \mathbb{N}\}$ be the set of its terms. By hypothesis, S is non-empty and bounded above. By the Completeness Axiom, S must have a supremum. Let $a = \sup S$.

We will now prove that $\lim_{n\to\infty} x_n = a$. According to the definition of a limit, we must show:

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n > N : |x_n - a| < \epsilon$$

This is equivalent to $a - \epsilon < x_n < a + \epsilon$.

1. First, by the definition of a supremum, a is an upper bound for S, so $x_n \le a$ for all n. It is clear that $x_n < a + \epsilon$.
2. Next, consider $a - \epsilon$. By the definition of a supremum, a is the *least* upper bound, which means $a - \epsilon$ (being smaller than a) *cannot* be an upper bound for S.
3. Since $a - \epsilon$ is not an upper bound, there must exist some element x_N in S such that $x_N > a - \epsilon$.
4. Because $\{x_n\}$ is monotone increasing, for any $n > N$, we have $x_n \ge x_N$.
5. Combining (1), (3), and (4), we have:

$$\forall n > N : a - \epsilon < x_N \le x_n \le a < a + \epsilon$$

This implies $\forall n > N : |x_n - a| < \epsilon$.

Therefore, $\lim_{n\to\infty} x_n = a$. $\square$

The Cauchy Convergence Criterion

The Monotone Convergence Theorem is powerful, but it requires the sequence to be monotone. For the general case, we need a criterion for convergence that does not depend on monotonicity, nor on knowing the value of the limit beforehand. This is the Cauchy Criterion.

Definition 2.2.5: Cauchy Sequence

A sequence $\{x_n\}$ is called a **Cauchy Sequence** if:

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall m, n > N : |x_m - x_n| < \epsilon$$

Intuitively, a Cauchy sequence is one whose terms become arbitrarily close to each other in the "tail" of the sequence.

Before proving the Cauchy Criterion, we need a key lemma, which is itself an important consequence of the Completeness Axiom.

Theorem 2.2.10: Bolzano-Weierstrass Theorem

Every bounded sequence in $\mathbb{R}$ must contain a convergent subsequence.

Proof. (Proof Sketch) Let $\{x_n\}$ be a bounded sequence, with all its terms contained in a closed interval $[a, b]$. We bisect $[a, b]$ into two subintervals $[a, \frac{a+b}{2}]$ and $[\frac{a+b}{2}, b]$. At least one of these must contain infinitely many terms of $\{x_n\}$. We choose such an interval and call it $I_1 = [a_1, b_1]$. Next, we bisect I_1 and again select a subinterval, $I_2 = [a_2, b_2]$, that contains infinitely many terms. We repeat this process, obtaining a **nest of closed intervals** $\{I_k = [a_k, b_k]\}$ such that:

1. $I_1 \supset I_2 \supset I_3 \supset \cdots$
2. The length of I_k, $len(I_k) = b_k - a_k = (b - a)/2^k \to 0$ as $k \to \infty$.

By the **Nested Intervals Property** of $\mathbb{R}$ (an equivalent form of completeness), there exists a unique real number c such that $c \in \bigcap_{k=1}^{\infty} I_k$.

Now, we construct a subsequence $\{x_{n_k}\}$ that converges to c:

- Choose $x_{n_1} \in I_1$.
- Since I_2 has infinitely many terms, we can choose $x_{n_2} \in I_2$ such that $n_2 > n_1$.
- ...
- Having chosen $x_{n_{k-1}} \in I_{k-1}$, we can choose $x_{n_k} \in I_k$ such that $n_k > n_{k-1}$ (as I_k has infinitely many terms).

This gives us a subsequence $\{x_{n_k}\}$. Since $c \in I_k$ and $x_{n_k} \in I_k$, we have:

$$|x_{n_k} - c| \leq len(I_k) = \frac{b-a}{2^k}$$

As $k \to \infty$, $\frac{b-a}{2^k} \to 0$. By the Squeeze Theorem, $\lim_{k\to\infty} |x_{n_k} - c| = 0$, which means $\lim_{k\to\infty} x_{n_k} = c$. □

Now we can prove the Cauchy Criterion.

Theorem 2.2.11: Cauchy Convergence Criterion

A sequence in $\mathbb{R}$ converges if and only if it is a Cauchy sequence.

Proof. ($\Rightarrow$) **Convergent $\implies$ Cauchy** Assume $\lim_{n\to\infty} x_n = L$. By the definition of a limit, $\forall\epsilon > 0, \exists N, \forall n > N : |x_n - L| < \frac{\epsilon}{2}$. Now, take any $m, n > N$. By the triangle inequality:

$$|x_m - x_n| = |(x_m - L) + (L - x_n)| \le |x_m - L| + |x_n - L|$$

$$|x_m - x_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Thus, $\{x_n\}$ is a Cauchy sequence.

($\Leftarrow$) **Cauchy $\implies$ Convergent** This direction relies critically on the completeness of $\mathbb{R}$.

1. **Step 1: Prove that a Cauchy sequence is bounded.** Let $\epsilon = 1$. By the Cauchy definition, $\exists N_1, \forall m, n > N_1 : |x_m - x_n| < 1$. Fix $m = N_1 + 1$. Then $\forall n > N_1 : |x_n - x_{N_1+1}| < 1$, which implies $x_{N_1+1} - 1 < x_n < x_{N_1+1} + 1$. This shows the "tail" of the sequence (terms with $n > N_1$) is bounded. The "head" of the sequence, $\{x_1, x_2, \ldots, x_{N_1}\}$, is a finite set and is thus bounded. Therefore, the entire sequence $\{x_n\}$ is bounded.

2. **Step 2: Apply the Bolzano-Weierstrass Theorem.** Since $\{x_n\}$ is bounded (by Step 1), the Bolzano-Weierstrass Theorem guarantees that it has a convergent subsequence, say $\{x_{n_k}\}$. Let $\lim_{k\to\infty} x_{n_k} = L$.

3. **Step 3: Prove the entire sequence $\{x_n\}$ converges to L.** We must show $\lim_{n\to\infty} x_n = L$. $\forall\epsilon > 0$:

   - Since $\{x_n\}$ is Cauchy, $\exists N_2, \forall m, n > N_2 : |x_m - x_n| < \frac{\epsilon}{2}$.
   - Since $\lim_{k\to\infty} x_{n_k} = L$, $\exists K, \forall k > K : |x_{n_k} - L| < \frac{\epsilon}{2}$.

   We need to find an N such that $\forall n > N : |x_n - L| < \epsilon$. Let's choose $N = N_2$. Then, we pick a single index n_k from the subsequence such that $k > K$ and $n_k > N_2$. (This is always possible since $n_k \to \infty$ as $k \to \infty$).

   Now, for any $n > N_2$, we have:

   $$|x_n - L| = |(x_n - x_{n_k}) + (x_{n_k} - L)| \le |x_n - x_{n_k}| + |x_{n_k} - L|$$

   Since $n > N_2$ and $n_k > N_2$, the first term is $< \frac{\epsilon}{2}$ by the Cauchy condition. Since $k > K$, the second term is $< \frac{\epsilon}{2}$ by the subsequence convergence.

   Thus, $\forall n > N_2 : |x_n - L| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$. This proves $\lim_{n\to\infty} x_n = L$.

□

2.3 Derivatives and Related Theorem

2.3.1 Derivatives and Differentials

Having studied limits of sequences and functions, we now turn to the central concept of differential calculus: the derivative. The derivative is the tool for studying the *rate of change* of a function.

The Concept of the Derivative

Definition 2.3.1: Derivative

Let the function f be defined in some neighborhood of a point x_0. If the limit

$$\lim_{\Delta x\to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

exists, we say the function f is **differentiable** at x_0, and this limit is called the **derivative** of f at x_0. It is denoted by $f'(x_0)$, $\frac{df}{dx}(x_0)$, or $y'|_{x=x_0}$.
Letting $\Delta y = f(x_0 + \Delta x) - f(x_0)$, the derivative can also be written as $\lim_{\Delta x\to 0} \frac{\Delta y}{\Delta x}$.

Geometric Meaning: $f'(x_0)$ is the slope of the tangent line to the curve $y = f(x)$ at the point $(x_0, f(x_0))$.
Physical Meaning: If $s(t)$ is the displacement as a function of time, then $s'(t)$ is the instantaneous velocity.

Differentiability is a stronger condition than continuity.

Theorem 2.3.1: Differentiability implies Continuity

If a function f is differentiable at x_0, then f must be continuous at x_0.

Proof. We want to prove $\lim_{x\to x_0} f(x) = f(x_0)$, which is equivalent to proving $\lim_{x\to x_0}[f(x) - f(x_0)] = 0$. Let $x = x_0 + \Delta x$, so $x \to x_0$ is equivalent to $\Delta x \to 0$.

$$\lim_{x\to x_0}[f(x) - f(x_0)] = \lim_{\Delta x\to 0}[f(x_0 + \Delta x) - f(x_0)]$$

We use the trick of multiplying and dividing by Δx (for $\Delta x \neq 0$):

$$= \lim_{\Delta x\to 0}\left[\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \cdot \Delta x\right]$$

By the product rule for limits:

$$= \left(\lim_{\Delta x\to 0}\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}\right) \cdot \left(\lim_{\Delta x\to 0}\Delta x\right)$$

Since f is differentiable at x_0, the first limit exists and is equal to $f'(x_0)$. The second limit is clearly 0.

$$= f'(x_0) \cdot 0 = 0$$

Thus $\lim_{x\to x_0} f(x) = f(x_0)$, so f is continuous at x_0. □

Uniform Continuity

The concept of continuity defined earlier is ”pointwise” continuity. A stronger and often more useful concept is uniform continuity.

Definition 2.3.2: Uniform Continuity

A function $f : D \to \mathbb{R}$ is **uniformly continuous** on D if for every $\epsilon > 0$, there exists a $\delta > 0$ such that for **all** $x, y \in D$:

$$|x - y| < \delta \implies |f(x) - f(y)| < \epsilon$$

The key difference: In standard continuity, δ can depend on both ϵ and the point x_0. In uniform continuity, δ depends *only* on ϵ and works for the entire domain simultaneously.

Theorem 2.3.2: Heine-Cantor Theorem

If a function f is continuous on a **closed and bounded** interval $[a, b]$, then f is uniformly continuous on $[a, b]$.

Example 2.3.1

$f(x) = x^2$ is uniformly continuous on $[0, 1]$ but *not* uniformly continuous on $[0, \infty)$. As x gets larger, we need a smaller and smaller δ to keep the change in $f(x)$ bounded, so no single δ works for the whole infinite domain.

Example 2.3.2: Continuous but not Differentiable

The converse is false. The function $f(x) = |x|$ is continuous at $x = 0$, but not differentiable. We check the derivative at $x = 0$:

$$\lim_{\Delta x \to 0} \frac{f(0 + \Delta x) - f(0)}{\Delta x} = \lim_{\Delta x \to 0} \frac{|\Delta x|}{\Delta x}$$

We check the left-hand and right-hand limits:

- Right-hand limit: $\lim_{\Delta x \to 0^+} \frac{|\Delta x|}{\Delta x} = \lim_{\Delta x \to 0^+} \frac{\Delta x}{\Delta x} = 1$
- Left-hand limit: $\lim_{\Delta x \to 0^-} \frac{|\Delta x|}{\Delta x} = \lim_{\Delta x \to 0^-} \frac{-\Delta x}{\Delta x} = -1$

Since the left and right limits are not equal, the limit does not exist. $f(x) = |x|$ is not differentiable at $x = 0$.

Using the definition to calculate the differentiation is complex. Here is a list for commonly used functions, showing their differentiation.

$$\frac{d}{dx} c = 0$$
$$\frac{d}{dx} x^n = nx^{n-1}$$
$$\frac{d}{dx} e^x = e^x$$
$$\frac{d}{dx} a^x = a^x \ln a$$
$$\frac{d}{dx} \ln x = \frac{1}{x}$$
$$\frac{d}{dx} \log_a x = \frac{1}{x \ln a}$$
$$\frac{d}{dx} \sin x = \cos x$$
$$\frac{d}{dx} \cos x = -\sin x$$
$$\frac{d}{dx} \tan x = \sec^2 x$$
$$\frac{d}{dx} \cot x = -\csc^2 x$$
$$\frac{d}{dx} \sec x = \sec x \tan x$$
$$\frac{d}{dx} \csc x = -\csc x \cot x$$
$$\frac{d}{dx} \arcsin x = \frac{1}{\sqrt{1 - x^2}}$$
$$\frac{d}{dx} \arccos x = -\frac{1}{\sqrt{1 - x^2}}$$
$$\frac{d}{dx} \arctan x = \frac{1}{1 + x^2}$$
$$\frac{d}{dx} \operatorname{arccot} x = -\frac{1}{1 + x^2}$$
$$\frac{d}{dx} \sinh x = \cosh x$$
$$\frac{d}{dx} \cosh x = \sinh x$$

Differentiation Rules

Here are some basic rules for differentiation:

- Sum: $(u \pm v)' = u' \pm v'$.
- Product: $(uv)' = u'v + uv'$.
- Quotient: $(\frac{u}{v})' = \frac{u'v - uv'}{v^2}$.
- Composition: $f'[g(x)] = f'(u) \cdot g'(x), u = g(x)$.

We (omit here) the proofs for basic differentiation rules (sum, product, quotient), but we will provide a rigorous proof for the Chain Rule.

Theorem 2.3.3: The Chain Rule

Let $u = g(x)$ be differentiable at x, and let $y = f(u)$ be differentiable at $u = g(x)$. Then the composite function $y = f(g(x))$ is differentiable at x, and

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} \quad \text{or} \quad (f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$

Proof. (A rigorous proof) Let $u_0 = g(x_0)$. Since $y = f(u)$ is differentiable at u_0, we define an auxiliary function $\phi(u)$:

$$\phi(u) = \begin{cases} \frac{f(u) - f(u_0)}{u - u_0} & \text{if } u \neq u_0 \\ f'(u_0) & \text{if } u = u_0 \end{cases}$$

Because $\lim_{u \to u_0} \phi(u) = \lim_{u \to u_0} \frac{f(u) - f(u_0)}{u - u_0} = f'(u_0) = \phi(u_0)$, the function $\phi(u)$ is continuous at $u = u_0$.

For all u (including $u = u_0$), we have $f(u) - f(u_0) = \phi(u)(u - u_0)$. Let $u = g(x_0 + \Delta x)$. Then $u - u_0 = g(x_0 + \Delta x) - g(x_0) = \Delta u$.

$$f(g(x_0 + \Delta x)) - f(g(x_0)) = \phi(g(x_0 + \Delta x)) \cdot (g(x_0 + \Delta x) - g(x_0))$$

Divide both sides by Δx (for $\Delta x \neq 0$):

$$\frac{f(g(x_0 + \Delta x)) - f(g(x_0))}{\Delta x} = \phi(g(x_0 + \Delta x)) \cdot \frac{g(x_0 + \Delta x) - g(x_0)}{\Delta x}$$

Now we take the limit as $\Delta x \to 0$:

$$\lim_{\Delta x \to 0} \frac{f(g(x_0 + \Delta x)) - f(g(x_0))}{\Delta x} = \lim_{\Delta x \to 0} \phi(g(x_0 + \Delta x)) \cdot \lim_{\Delta x \to 0} \frac{g(x_0 + \Delta x) - g(x_0)}{\Delta x}$$

The left side is the definition of $(f \circ g)'(x_0)$. On the right side, the second term is $g'(x_0)$. For the first term, since g is differentiable at x_0, it is continuous at x_0. Thus, as $\Delta x \to 0$, $g(x_0 + \Delta x) \to g(x_0) = u_0$. And since $\phi(u)$ is continuous at u_0, we have $\lim_{\Delta x \to 0} \phi(g(x_0 + \Delta x)) = \phi(u_0) = f'(u_0) = f'(g(x_0))$. Therefore,

$$(f \circ g)'(x_0) = f'(g(x_0)) \cdot g'(x_0)$$

□

Higher Derivative

Definition 2.3.3: Second Derivative

Likewise, if the differentiation of a function is differentiable, and we differentiate the differentiation, we get the second derivative of the function, and we call it differentiable for second order.

Similarly, we can define what is n-th derivative of the function $f(x)$, and we call the $f(x)$ n-th differentiable if the n-th derivative exists. The n-th derivative of $f(x)$ can be denoted as: $f^{(n)}(x)$ or $f^{(n)}$.

Theorem 2.3.4: Leibniz theorem

If u and v are two functions that are differentiable up to n times, the n-th derivative of their product can be expressed as:

$$(uv)^{(n)} = \sum_{r=0}^{n} C_n^r \cdot u^{(r)} \cdot v^{(n-r)}$$

The Leibniz formula solves the higher-order derivative of a product. For the derivatives of other operations, they can be easily derived from the preceding content.

The Differential

The derivative $f'(x_0)$ is a number, representing the rate of change. The differential provides a linear approximation.

Definition 2.3.4: Differential

Let $y = f(x)$ be differentiable at x_0. The increment Δy can be expressed as:

$$\Delta y = f(x_0 + \Delta x) - f(x_0) = f'(x_0)\Delta x + o(\Delta x)$$

where $\lim_{\Delta x \to 0} \frac{o(\Delta x)}{\Delta x} = 0$. We call the **linear principal part** of Δy, $f'(x_0)\Delta x$, the **differential** of f at x_0, denoted dy.

$$dy = f'(x_0)\Delta x$$

By convention, we define the differential of the independent variable dx to be equal to the increment $dx = \Delta x$. Therefore, the differential can be written as:

$$dy = f'(x_0)dx$$

This also provides the notation $f'(x) = \frac{dy}{dx}$, the derivative as a ratio of differentials.

Geometric Meaning:

- $\Delta y = f(x_0 + \Delta x) - f(x_0)$ is the **actual change** in y along the curve.
- $dy = f'(x_0)dx$ is the **change in y along the tangent line**.

When Δx is small, $dy \approx \Delta y$. This provides the basis for linear approximation: $f(x_0 + \Delta x) \approx f(x_0) + f'(x_0)\Delta x$.

2.3.2 Mean Value Theorems and L'Hôpital's Rule

The Mean Value Theorems are the bridge connecting the derivative of a function to its values, and they are the theoretical foundation for applications of differential calculus.

Mean Value Theorems

We begin with a necessary lemma.

Theorem 2.3.5: Fermat's Theorem

Let the function $f(x)$ satisfy at x_0:

1. f has a local extremum (max or min) at x_0.
2. f is differentiable at x_0.

Then $f'(x_0) = 0$.

Proof. Assume f has a local maximum at x_0. Then in some neighborhood $(x_0-\delta, x_0+\delta)$, $f(x) \le f(x_0)$ for all x.

- For $x \in (x_0, x_0+\delta)$, we have $x - x_0 > 0$ and $f(x) - f(x_0) \le 0$. Thus, the difference quotient $\frac{f(x)-f(x_0)}{x-x_0} \le 0$. The right-hand derivative $f'_+(x_0) = \lim_{x\to x_0^+} \frac{f(x)-f(x_0)}{x-x_0} \le 0$.
- For $x \in (x_0-\delta, x_0)$, we have $x - x_0 < 0$ and $f(x) - f(x_0) \le 0$. Thus, the difference quotient $\frac{f(x)-f(x_0)}{x-x_0} \ge 0$. The left-hand derivative $f'_-(x_0) = \lim_{x\to x_0^-} \frac{f(x)-f(x_0)}{x-x_0} \ge 0$.

Since f is differentiable at x_0, $f'(x_0) = f'_+(x_0) = f'_-(x_0)$. The only number that is both ≤ 0 and ≥ 0 is 0. Therefore, $f'(x_0) = 0$. $\square$

Theorem 2.3.6: Rolle's Theorem

Let the function $f(x)$ satisfy:
1. f is continuous on the closed interval $[a, b]$;
2. f is differentiable on the open interval (a, b);
3. $f(a) = f(b)$.

Then there exists at least one point $\xi \in (a, b)$ such that $f'(\xi) = 0$.

Proof.
- **Case 1:** $f(x)$ is a constant function on $[a, b]$. Then $f(x) = f(a)$ for all $x \in [a, b]$. In this case, $f'(x) = 0$ for all $x \in (a, b)$. We can choose any $\xi \in (a, b)$.
- **Case 2:** $f(x)$ is not a constant function. Since f is continuous on the closed interval $[a, b]$, by the **Extreme Value Theorem**, f must attain an absolute maximum M and an absolute minimum m on $[a, b]$. Since f is not constant, at least one of M or m must be different from $f(a)$ (and $f(b)$). Assume $M > f(a)$. Let $f(\xi) = M$ (where $\xi \in [a, b]$). Because $f(a) = f(b) < M$, ξ cannot be a or b. Thus, $\xi \in (a, b)$. At this point ξ, $f(x)$ attains a local (and global) maximum. By Fermat's Theorem, $f'(\xi) = 0$. (If $m < f(a)$, the same logic applies to the point ξ where the minimum occurs).

$\square$

Theorem 2.3.7: Lagrange's Mean Value Theorem

Let the function $f(x)$ satisfy:
1. f is continuous on the closed interval $[a, b]$;
2. f is differentiable on the open interval (a, b).

Then there exists at least one point $\xi \in (a, b)$ such that

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

or $f(b) - f(a) = f'(\xi)(b - a)$.

Geometric Meaning: There is at least one point $\xi \in (a, b)$ where the slope of the tangent line is equal to the slope of the secant line connecting $(a, f(a))$ and $(b, f(b))$.

Proof. The proof technique involves constructing an auxiliary function that satisfies Rolle's Theorem. Let $g(x)$ be the equation of the secant line connecting $(a, f(a))$ and $(b, f(b))$:

$$g(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

Now, construct the auxiliary function $h(x) = f(x) - g(x)$.

$$h(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a)$$

We check if $h(x)$ satisfies the conditions of Rolle's Theorem:

1. $h(x)$ is the difference of $f(x)$ and a linear function. Since both are continuous on $[a, b]$, $h(x)$ is continuous on $[a, b]$.
2. Similarly, $h(x)$ is differentiable on (a, b).
3. $h(a) = f(a) - f(a) - \frac{f(b)-f(a)}{b-a}(a - a) = 0$.
4. $h(b) = f(b) - f(a) - \frac{f(b)-f(a)}{b-a}(b - a) = f(b) - f(a) - (f(b) - f(a)) = 0$.

$h(a) = h(b) = 0$. $h(x)$ satisfies all conditions for Rolle's Theorem. Thus, $\exists \xi \in (a, b)$ such that $h'(\xi) = 0$. We compute $h'(x)$:

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

Setting $h'(\xi) = 0$:

$$f'(\xi) - \frac{f(b) - f(a)}{b - a} = 0$$

This gives $f'(\xi) = \frac{f(b)-f(a)}{b-a}$. □

Theorem 2.3.8: Cauchy's Mean Value Theorem

Let functions $f(x)$ and $g(x)$ satisfy:
1. f, g are continuous on the closed interval $[a, b]$;
2. f, g are differentiable on the open interval (a, b);
3. $g'(x) \neq 0$ for all $x \in (a, b)$.

Then there exists at least one point $\xi \in (a, b)$ such that

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

Proof. (Note: By (3) and Rolle's Theorem, $g(a) \neq g(b)$, otherwise $g'(\xi) = 0$ would hold for some ξ, which is forbidden.) We again construct an auxiliary function $h(x)$ for Rolle's Theorem:

$$h(x) = [f(b) - f(a)](g(x) - g(a)) - [g(b) - g(a)](f(x) - f(a))$$

(This form is chosen to ensure $h(a) = h(b) = 0$)

1. $h(x)$ is continuous on $[a, b]$.
2. $h(x)$ is differentiable on (a, b).
3. $h(a) = [f(b) - f(a)](g(a) - g(a)) - [g(b) - g(a)](f(a) - f(a)) = 0$.
4. $h(b) = [f(b) - f(a)](g(b) - g(a)) - [g(b) - g(a)](f(b) - f(a)) = 0$.

By Rolle's Theorem, $\exists \xi \in (a, b)$ such that $h'(\xi) = 0$. We compute $h'(x)$:

$$h'(x) = [f(b) - f(a)]g'(x) - [g(b) - g(a)]f'(x)$$

Setting $h'(\xi) = 0$:

$$[f(b) - f(a)]g'(\xi) - [g(b) - g(a)]f'(\xi) = 0$$
$$[f(b) - f(a)]g'(\xi) = [g(b) - g(a)]f'(\xi)$$

Since $g'(x) \neq 0$, we know $g'(\xi) \neq 0$. We also know $g(a) \neq g(b)$. We can safely divide:

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

□

L'Hôpital's Rule

Cauchy's Mean Value Theorem is the key to proving L'Hôpital's Rule, which is used to evaluate indeterminate forms of type $\frac{0}{0}$ and $\frac{\infty}{\infty}$.

Theorem 2.3.9: L'Hôpital's Rule ($\frac{0}{0}$ form)

Let c be a real number (or $\pm\infty$). Let f, g be differentiable on a (punctured) neighborhood of c, with $g'(x) \neq 0$. If

1. $\lim_{x\to c} f(x) = 0$ and $\lim_{x\to c} g(x) = 0$;
2. $\lim_{x\to c} \frac{f'(x)}{g'(x)} = L$ (where L can be a finite value or $\pm\infty$).

Then

$$\lim_{x\to c} \frac{f(x)}{g(x)} = L$$

Proof. We prove the case for $x \to c^+$ where c is a finite real number. We can define $f(c) = 0$ and $g(c) = 0$ (since the limits are 0), making f and g continuous at c. Now, for any x in a right-neighborhood of c, the functions f and g are continuous on $[c, x]$ and differentiable on (c, x). By Cauchy's Mean Value Theorem, there exists a $\xi_x \in (c, x)$ such that

$$\frac{f(x) - f(c)}{g(x) - g(c)} = \frac{f'(\xi_x)}{g'(\xi_x)}$$

Since $f(c) = 0$ and $g(c) = 0$, this simplifies to:

$$\frac{f(x)}{g(x)} = \frac{f'(\xi_x)}{g'(\xi_x)}$$

As $x \to c^+$, since $\xi_x \in (c, x)$, we must also have $\xi_x \to c^+$. By condition (2), $\lim_{\xi_x\to c^+} \frac{f'(\xi_x)}{g'(\xi_x)} = L$. Therefore,

$$\lim_{x\to c^+} \frac{f(x)}{g(x)} = \lim_{\xi_x\to c^+} \frac{f'(\xi_x)}{g'(\xi_x)} = L$$

The proofs for $x \to c^-$ and $x \to \infty$ are similar. □

Theorem 2.3.10: L'Hôpital's Rule ($\frac{\infty}{\infty}$ form)

Let c be a real number (or $\pm\infty$). Let f, g be differentiable on a (punctured) neighborhood of c, with $g'(x) \neq 0$. If

1. $\lim_{x\to c} |f(x)| = \infty$ and $\lim_{x\to c} |g(x)| = \infty$;
2. $\lim_{x\to c} \frac{f'(x)}{g'(x)} = L$ (where L can be a finite value or $\pm\infty$).

Then

$$\lim_{x\to c} \frac{f(x)}{g(x)} = L$$

Proof. (Sketch) This proof is more complex than the $\frac{0}{0}$ form. We consider $x \to c^+$ and L finite. By condition (2), $\forall\epsilon > 0, \exists\delta > 0$ such that $\forall x \in (c, c+\delta) : |\frac{f'(x)}{g'(x)} - L| < \frac{\epsilon}{2}$. We pick an $x_0 \in (c, c+\delta)$. Now, for any $x \in (c, x_0)$, we apply Cauchy's MVT on $[x, x_0]$. There exists $\xi \in (x, x_0)$ such that

$$\frac{f(x) - f(x_0)}{g(x) - g(x_0)} = \frac{f'(\xi)}{g'(\xi)}$$

Since $\xi \in (x, x_0) \subset (c, c+\delta)$, we know $|\frac{f'(\xi)}{g'(\xi)} - L| < \frac{\epsilon}{2}$.

$$\left|\frac{f(x) - f(x_0)}{g(x) - g(x_0)} - L\right| < \frac{\epsilon}{2}$$

We perform an algebraic manipulation of $\frac{f(x)}{g(x)}$:

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(x_0)}{g(x) - g(x_0)} \cdot \frac{g(x) - g(x_0)}{g(x)} + \frac{f(x_0)}{g(x)}$$

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(x_0)}{g(x) - g(x_0)} \cdot \left(1 - \frac{g(x_0)}{g(x)}\right) + \frac{f(x_0)}{g(x)}$$

We want to show $\frac{f(x)}{g(x)} \to L$ as $x \to c^+$. As $x \to c^+$, we know $f(x) \to \infty$ and $g(x) \to \infty$. The terms $f(x_0)$ and $g(x_0)$ are fixed constants. Thus, $\frac{g(x_0)}{g(x)} \to 0$ and $\frac{f(x_0)}{g(x)} \to 0$. This means $\left(1 - \frac{g(x_0)}{g(x)}\right) \to 1$. The limit behavior of $\frac{f(x)}{g(x)}$ is dominated by $\frac{f(x)-f(x_0)}{g(x)-g(x_0)}$, which we know is within $\epsilon/2$ of L. By choosing x sufficiently close to c (i.e., $x \to c^+$), the error terms $\frac{g(x_0)}{g(x)}$ and $\frac{f(x_0)}{g(x)}$ can be made arbitrarily small, and the full expression $\left|\frac{f(x)}{g(x)} - L\right|$ can be shown to be less than ϵ. □

Example 2.3.3

(1) Evaluate $\lim_{x\to 0} \frac{\sin x}{x}$ ($\frac{0}{0}$ form)

$$\lim_{x\to 0} \frac{\sin x}{x} \overset{L'H}{=} \lim_{x\to 0} \frac{(\sin x)'}{(x)'} = \lim_{x\to 0} \frac{\cos x}{1} = \cos 0 = 1$$

(2) Evaluate $\lim_{x\to +\infty} \frac{\ln x}{x}$ ($\frac{\infty}{\infty}$ form)

$$\lim_{x\to +\infty} \frac{\ln x}{x} \overset{L'H}{=} \lim_{x\to +\infty} \frac{(\ln x)'}{(x)'} = \lim_{x\to +\infty} \frac{1/x}{1} = \lim_{x\to +\infty} \frac{1}{x} = 0$$

Remark 2.3.1

Caution: The condition $\lim \frac{f'(x)}{g'(x)} = L$ is **sufficient** but not **necessary**. If $\lim \frac{f'(x)}{g'(x)}$ does not exist, we cannot conclude that $\lim \frac{f(x)}{g(x)}$ does not exist.

2.3.3 Taylor Expansion

The differential provides a *linear* approximation of a function $f(x)$ near a point x_0 (the tangent line). Taylor's Theorem generalizes this idea, providing a method to approximate a function with a polynomial of any arbitrary degree n.

Taylor Polynomials

We seek an n-th degree polynomial $P_n(x)$ that "best" approximates $f(x)$ near x_0. We do this by forcing the polynomial's value and its first n derivatives to match those of $f(x)$ at x_0.

$$P_n^{(k)}(x_0) = f^{(k)}(x_0) \quad \text{for } k = 0, 1, \ldots, n$$

Let the polynomial have the form:

$$P_n(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)^2 + \cdots + c_n(x - x_0)^n$$

We determine the coefficients c_k:

- $P_n(x_0) = c_0 \implies c_0 = f(x_0)$
- $P_n'(x) = c_1 + 2c_2(x - x_0) + 3c_3(x - x_0)^2 + \ldots$

- $P_n'(x_0) = c_1 \implies c_1 = f'(x_0)$
- $P_n''(x) = 2c_2 + 3 \cdot 2c_3(x - x_0) + \dots$
- $P_n''(x_0) = 2c_2 \implies c_2 = \frac{f''(x_0)}{2!}$
- $P_n'''(x_0) = 3 \cdot 2 \cdot 1c_3 \implies c_3 = \frac{f'''(x_0)}{3!}$

By induction, we find $P_n^{(k)}(x_0) = k!c_k$, which gives $c_k = \frac{f^{(k)}(x_0)}{k!}$.

Definition 2.3.5: Taylor Polynomial

Let f be a function with at least n derivatives at x_0. The **n-th degree Taylor polynomial** of f centered at x_0 is:

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

Definition 2.3.6: Maclaurin Polynomial

When the center is $x_0 = 0$, the Taylor polynomial is called the **Maclaurin polynomial**.

Taylor's Theorem and the Remainder

The Taylor polynomial $P_n(x)$ is an approximation of $f(x)$. The error of this approximation is called the remainder.

Definition 2.3.7: Remainder

The **remainder** $R_n(x)$ is defined as the difference:

$$R_n(x) = f(x) - P_n(x)$$

Thus, $f(x) = P_n(x) + R_n(x)$.

Taylor's Theorem gives us a precise formula for this remainder.

Theorem 2.3.11: Taylor's Theorem with Lagrange Remainder

Let f be a function such that $f^{(n+1)}$ (the $(n+1)$-th derivative) exists on an open interval I containing x_0. Then for any $x \in I$, there exists a number ξ (xi) strictly between x and x_0 such that

$$f(x) = P_n(x) + R_n(x)$$

where the **Lagrange form of the remainder** is

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}$$

Proof. This proof is a clever application of Rolle's Theorem. Fix x and x_0. For simplicity, let $x = b$. We are looking for $R_n(b) = f(b) - P_n(b)$. We want to find a constant K such that

$$f(b) = P_n(b) + K(b - x_0)^{n+1}$$

This means $K = \frac{f(b)-P_n(b)}{(b-x_0)^{n+1}}$. We must show that $K = \frac{f^{(n+1)}(\xi)}{(n+1)!}$ for some $\xi \in (x_0, b)$.

Define an auxiliary function $g(t)$ on the interval $[x_0, b]$:

$$g(t) = f(t) - P_n(t) - K(t - x_0)^{n+1}$$

where $P_n(t) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(t - x_0)^k$.

We check the values of $g(t)$ and its derivatives at $t = x_0$:

- $g(x_0) = f(x_0) - P_n(x_0) - K(x_0 - x_0)^{n+1} = f(x_0) - f(x_0) - 0 = 0$.
- $g'(t) = f'(t) - P_n'(t) - (n+1)K(t - x_0)^n$. $P_n'(t) = \sum_{k=1}^{n} \frac{f^{(k)}(x_0)}{(k-1)!}(t - x_0)^{k-1}$. $P_n'(x_0) = f'(x_0)$. So, $g'(x_0) = f'(x_0) - f'(x_0) - 0 = 0$.
- In general, for $k \leq n$, $P_n^{(k)}(x_0) = f^{(k)}(x_0)$. $g^{(k)}(t) = f^{(k)}(t) - P_n^{(k)}(t) - \frac{(n+1)!}{(n+1-k)!}K(t - x_0)^{n+1-k}$. So, $g^{(k)}(x_0) = f^{(k)}(x_0) - f^{(k)}(x_0) - 0 = 0$.

We have $g(x_0) = g'(x_0) = \cdots = g^{(n)}(x_0) = 0$.

Now we check $g(t)$ at $t = b$:

$$g(b) = f(b) - P_n(b) - K(b - x_0)^{n+1}$$

By our definition of K, $g(b) = 0$.

We are ready to apply Rolle's Theorem:

1. We have $g(x_0) = 0$ and $g(b) = 0$. By Rolle's Theorem, $\exists \xi_1 \in (x_0, b)$ s.t. $g'(\xi_1) = 0$.
2. We have $g'(x_0) = 0$ and $g'(\xi_1) = 0$. By Rolle's Theorem (applied to g'), $\exists \xi_2 \in (x_0, \xi_1)$ s.t. $g''(\xi_2) = 0$.
3. ...
4. We have $g^{(n)}(x_0) = 0$ and $g^{(n)}(\xi_n) = 0$. By Rolle's Theorem (applied to $g^{(n)}$), $\exists \xi \in (x_0, \xi_n)$ s.t. $g^{(n+1)}(\xi) = 0$. Note that $\xi \in (x_0, \xi_n) \subset \cdots \subset (x_0, b)$.

Finally, we compute $g^{(n+1)}(t)$:

$$g^{(n+1)}(t) = f^{(n+1)}(t) - P_n^{(n+1)}(t) - \frac{d^{n+1}}{dt^{n+1}}[K(t - x_0)^{n+1}]$$

Since $P_n(t)$ is a polynomial of degree n, $P_n^{(n+1)}(t) = 0$. The $(n+1)$-th derivative of $K(t-x_0)^{n+1}$ is $K \cdot (n+1)!$.

$$g^{(n+1)}(t) = f^{(n+1)}(t) - 0 - K(n+1)!$$

At $t = \xi$, we know $g^{(n+1)}(\xi) = 0$:

$$f^{(n+1)}(\xi) - K(n+1)! = 0$$

Solving for K: $K = \frac{f^{(n+1)}(\xi)}{(n+1)!}$. Substituting this back into $R_n(b) = K(b - x_0)^{n+1}$ (and replacing b with x):

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}$$

□

Taylor and Maclaurin Series

If the remainder $R_n(x) \to 0$ as $n \to \infty$, then the function $f(x)$ can be represented by its infinite series.

Definition 2.3.8: Taylor Series

If $\lim_{n\to\infty} R_n(x) = 0$ for x in an interval I, then $f(x)$ is equal to its **Taylor Series** on I:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

If $x_0 = 0$, this is called the **Maclaurin Series**.

Example 2.3.4: Common Maclaurin Series

1. $f(x) = e^x$ $f^{(k)}(x) = e^x$ for all k. So $f^{(k)}(0) = e^0 = 1$.

$$e^x = \sum_{k=0}^{\infty} \frac{1}{k!}x^k = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

The remainder is $R_n(x) = \frac{e^\xi}{(n+1)!}x^{n+1}$. For any fixed x, $\frac{x^{n+1}}{(n+1)!} \to 0$ as $n \to \infty$. Thus, this series converges to e^x for all $x \in \mathbb{R}$.

2. $f(x) = \sin x$ $f(0) = 0, f'(0) = 1, f''(0) = 0, f'''(0) = -1, f^{(4)}(0) = 0, \dots$ (Pattern: $0, 1, 0, -1, \dots$)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} \quad \text{(for all } x\text{)}$$

3. $f(x) = \cos x$ $f(0) = 1, f'(0) = 0, f''(0) = -1, f'''(0) = 0, f^{(4)}(0) = 1, \dots$ (Pattern: $1, 0, -1, 0, \dots$)

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} \quad \text{(for all } x\text{)}$$

4. $f(x) = \frac{1}{1-x}$ **(Geometric Series)** $f^{(k)}(x) = k!(1-x)^{-(k+1)}$. So $f^{(k)}(0) = k!$.

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} \frac{k!}{k!}x^k = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + \dots$$

This series is the geometric series, and it converges to $f(x)$ only when $|x| < 1$.

2.4 Integration

2.4.1 Indefinite Integration

Think about one question: if we have a differentiation of a function, how can we find its primitive function(s)?

We need to answer this question because in many cases, we need to figure out the primitive function from the differentiation. Considering a condition when population growth in the absence of predators or resource restrictions. In this case, the population growth rate will be proportional to the population size. In mathematical expressions, we can denote it like this:

$$\begin{cases} p'(t) = \lambda p(t) \\ p(t_0) = p_0 \end{cases}$$

If we want to know the expression of $p(t)$, we need to use the knowledge and methods of integration. But because it's too simple, many readers can simply guess the answer: that is $p(t) = p_0 \cdot e^{\lambda x}$. But how about a more complex condition? Considering the model of logistic growth:

$$\begin{cases} \frac{dN(t)}{dt} = \frac{rN(t)\cdot(K-N(t))}{K} \\ N(0) = N_0 \end{cases}$$

Guesses won't be enough to get the answers. Thus, we need to know how to do the work of integration.

Definition 2.4.1

If in a specific interval, the function $F(x)$ and $f(x)$ satisfy the following relationship:

$$F'(x) = f(x)$$

Or equivalently,

$$d[F(x)] = f(x) \cdot dx$$

Then we call F(x) is **one of** the **antiderivative** in this interval.

The reason why we said "one of" is because the antiderivative of a function is **not unique**. For example, if a function $F(x)$ is the antiderivative of $f(x)$, then $\forall [F(x) + C]$ C is a constant, $F(x) + C$ is also the antiderivative of $f(x)$. So we can say that there are infinity many antiderivatives of a function once it is integrable, and if we know one of the antiderivative of the function, we can use $G(x) = F(x)+C$ to represent all the primitive functions of the $f(x)$.

Definition 2.4.2

All the antiderivative of a function $f(x)$ is called the **indefinite integration** of this function. denoted as $\int f(x)d(x)$. The sign $\int$ is called the integral sign, $f(x)$ is called the integrant, and x is called the variable of integration.

In fact, the process of finding antiderivative is to find the primitive function of the derivative. And integration is the inverse operation of differentiation. And based on the table of derivative of commonly used function in section 2.3.1, we can deduce the integration of the commonly used functions.

One of the most important properties of the integration is its linear properties. We can express it in such way:

Theorem 2.4.1

If $f(x)$ and $g(x)$ are both integrable, then for every constant k_1 and k_2, the function $k_1 f(x) + k_2 g(x)$ is also integrable. and we have:

$$\int [k_1 f(x) + k_2 g(x)]dx = k_1 \int f(x)dx + k_2 \int g(x)dx$$

This is called the linear property of integration.

However, since our primitive purpose is to find ways to figure out the antiderivative of a function, using the definition won't be enough to cover all the needs of figure out the antiderivative of a function. So we will introduce several methods to help us figure out the antiderivative.

Integration By Substitutions

Substitution is one of the most popular methods in analysis. When figuring out the antiderivative of the function $f(x)$, we can use this protocol.

Definition 2.4.3: Integration by Substitution (First Part)

If the integrant can be transformed into the form: $f(x) = g(h(x)) \cdot h'(x)$, and the integration of $g(u)$ is easy to know. Then we have:

$$\int f(x)dx = \int g(h(x)) \cdot h'(x) \cdot dx = \int g(h(x)) \cdot d(h(x)) = G(h(x)) + C$$

(We denote the integration of the function $g(x)$ as $G(x)$)

Definition 2.4.4: Integration by Substitution (Second Part)

We can also perform substitution in a different manner. To evaluate $\int f(x)dx$, we can introduce a new variable t by setting $x = \phi(t)$, where $\phi(t)$ is a function with a continuous derivative and an inverse $t = \phi^{-1}(x)$.
If we substitute $x = \phi(t)$, then $dx = \phi'(t)dt$. The integration becomes:

$$\int f(x)dx = \int f(\phi(t)) \cdot \phi'(t)dt$$

If we can find the antiderivative of the right-hand side, say $H(t)$, we can then substitute back $t = \phi^{-1}(x)$ to express the final answer in terms of x.

$$\int f(x)dx = H(t) + C = H(\phi^{-1}(x)) + C$$

This method is particularly useful when the integrand $f(x)$ contains expressions that can be simplified by such a substitution.

The second method of substitution gives rise to several powerful, standardized techniques.

Trigonometric Substitution This method is used to eliminate square roots of quadratic expressions, specifically of the forms $\sqrt{a^2 - x^2}$, $\sqrt{a^2 + x^2}$, and $\sqrt{x^2 - a^2}$.

- **For integrands containing $\sqrt{a^2 - x^2}$:** We use the substitution $x = a\sin(\theta)$, with $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. Then $dx = a\cos(\theta)d\theta$. The expression becomes:

$$\sqrt{a^2 - x^2} = \sqrt{a^2 - a^2\sin^2(\theta)} = \sqrt{a^2(1 - \sin^2(\theta))} = \sqrt{a^2\cos^2(\theta)} = a\cos(\theta)$$

  (Note: $\cos(\theta) \geq 0$ in the specified range).

- **For integrands containing $\sqrt{a^2 + x^2}$:** We use the substitution $x = a\tan(\theta)$, with $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$. Then $dx = a\sec^2(\theta)d\theta$. The expression becomes:

$$\sqrt{a^2 + x^2} = \sqrt{a^2 + a^2\tan^2(\theta)} = \sqrt{a^2(1 + \tan^2(\theta))} = \sqrt{a^2\sec^2(\theta)} = a\sec(\theta)$$

  (Note: $\sec(\theta) > 0$ in the specified range).

- **For integrands containing $\sqrt{x^2 - a^2}$:** We use the substitution $x = a\sec(\theta)$, with $0 \le \theta < \frac{\pi}{2}$ or $\pi \le \theta < \frac{3\pi}{2}$. Then $dx = a\sec(\theta)\tan(\theta)d\theta$. The expression becomes:

$$\sqrt{x^2 - a^2} = \sqrt{a^2\sec^2(\theta) - a^2} = \sqrt{a^2(\sec^2(\theta) - 1)} = \sqrt{a^2\tan^2(\theta)} = a\tan(\theta)$$

  (Note: $\tan(\theta) \ge 0$ in the specified range).

After integration, one must convert the result from θ back to x using the original substitution, often by drawing a right-angled triangle.

Tangent Half-Angle Substitution (Weierstrass Substitution) This substitution is very powerful for integrals which are rational functions of $\sin(x)$ and $\cos(x)$. We introduce the substitution $t = \tan(x/2)$. Using trigonometric identities, we can express $\sin(x)$, $\cos(x)$, and dx in terms of t:

$$\sin(x) = \frac{2\tan(x/2)}{1 + \tan^2(x/2)} = \frac{2t}{1 + t^2}$$

$$\cos(x) = \frac{1 - \tan^2(x/2)}{1 + \tan^2(x/2)} = \frac{1 - t^2}{1 + t^2}$$

From $t = \tan(x/2)$, we have $dt = \frac{1}{2}\sec^2(x/2)dx = \frac{1}{2}(1 + \tan^2(x/2))dx = \frac{1}{2}(1 + t^2)dx$. This gives:

$$dx = \frac{2}{1 + t^2}dt$$

This substitution converts any rational function of $\sin(x)$ and $\cos(x)$ into a rational function of t, which can then be integrated using methods like partial fraction decomposition.

Remark 2.4.1

In trigonometric substitution, there are some commonly used equations:

$$\sin^2 x + \cos^2 x = 1$$

$$1 + \tan^2 x = \sec^2 x$$

$$1 + \cot^2 x = \csc^2 x$$

Integration By Parts

Integration by Parts is another fundamental technique, derived from the product rule for differentiation.

Definition 2.4.5: Integration by Parts

Let $u = u(x)$ and $v = v(x)$ be differentiable functions. The product rule for differentiation states:

$$\frac{d}{dx}(u(x)v(x)) = u'(x)v(x) + u(x)v'(x)$$

Integrating both sides with respect to x:

$$\int \frac{d}{dx}(u(x)v(x))dx = \int u'(x)v(x)dx + \int u(x)v'(x)dx$$

$$u(x)v(x) = \int v(x)u'(x)dx + \int u(x)v'(x)dx$$

Rearranging this gives the integration by parts formula. In the more compact differential notation, let $u = u(x)$ and $v = v(x)$, so $du = u'(x)dx$ and $dv = v'(x)dx$. The formula is:

$$\int u\,dv = uv - \int v\,du$$

The key to this method is to split the integrand into two parts, u and dv, such that the new integral, $\int v\,du$, is simpler to solve than the original.

Example: Let's compute $\int x\cos(x)dx$.

We must choose u and dv. A good choice is:

- Let $u = x$ (because its derivative, du, is simpler)
- Let $dv = \cos(x)dx$ (because it is easy to integrate)

Now we compute du (by differentiating u) and v (by integrating dv):

- $du = dx$
- $v = \int \cos(x)dx = \sin(x)$ (We omit the constant of integration until the final step)

Applying the formula $\int u\,dv = uv - \int v\,du$:

$$\int x\cos(x)dx = x \cdot \sin(x) - \int \sin(x)dx$$

The new integral is straightforward:

$$\int x\cos(x)dx = x\sin(x) - (-\cos(x)) + C = x\sin(x) + \cos(x) + C$$

Tips for Choosing u and dv:

- **The LIATE Rule:** A helpful mnemonic for choosing u is the acronym **LIATE**, which stands for:
  - **L**: Logarithmic functions (e.g., $\ln(x)$)
  - **I**: Inverse trigonometric functions (e.g., $\arctan(x)$, $\arcsin(x)$)
  - **A**: Algebraic functions (e.g., x^2, $x^3 + 1$)
  - **T**: Trigonometric functions (e.g., $\sin(x)$, $\cos(x)$)

  - **E**: Exponential functions (e.g., e^x, 2^x)

  You should choose u as the function that appears first in this list. The remaining part of the integrand becomes dv. This heuristic works because functions at the top of the list (like $\ln(x)$) generally become simpler upon differentiation, while functions at the bottom (like e^x) are easy to integrate.

- **Repeated Application:** Sometimes, integration by parts must be applied more than once. For example, to solve $\int x^2e^x dx$, you would first set $u = x^2$, which would lead to a new integral involving xe^x. You would then apply integration by parts a second time to solve that integral.
- **The "Boomerang" Technique:** For integrals like $\int e^x \cos(x)dx$, applying integration by parts twice (using consistent choices for u and dv) will result in the original integral appearing on the right-hand side of the equation. You can then algebraically solve for the value of the integral.

Also, For functions in forms like $\int P(x)e^x dx$, $\int P(x)\sin x dx$, $\int P(x)\cos x dx$, We can apply the integration by parts again and again. Then we will have that:

$$\int uv^{(n+1)}dx = uv^{(n)} - u'v^{(n-1)} + u''v^{(n-2)} - u^{(3)}v^{(n-3)} + \cdots + (-1)^n u^{(n)}v + (-1)^{n+1}\int u^{n+1}vdx$$

This formula is especially useful if u is a **polynomial function**.

General Method of Integrating the Rational Functions

There exists a form of function, that have really nice property and we can figure out the integration of every functions in that form. That is the rational function.

Definition 2.4.6: Rational Functions

Rational functions is a class of function, which is the fraction of two real coefficients polynomials, we can denote it as:

$$R(x) = \frac{P(x)}{Q(x)}$$

$P(x), Q(x)$ are both real coefficients polynomials. If the power of $P(x)$ is smaller than $Q(x)$, then we say $R(x)$ is a proper fraction. Else, we claim the $R(x)$ is a improper fraction.

And according to the **Fundamental Theorem of Algebra**, which we will introduce in later chapter, we know that the $Q(x)$ has the same amount of solutions of the number of the power.

Thus we can rewrite the $Q(x)$ in form of:

$$Q(x) = k(x-a)^\alpha(x-b)^\beta\cdots(x^2+px+q)^\mu(x^2+rx+s)^\delta\cdots$$

And now we will introduce a lemma, readers can try to prove it themselves.

Theorem 2.4.2

If we have a proper fraction $R(x) = \frac{P(x)}{Q(x)}$. and $Q(x)$ has factorization like above, then we can assert that:

$$R(x) = \sum_{j=1}^{\alpha}\frac{A_j}{(x-a)^j} + \sum_{j=1}^{\beta}\frac{B_j}{(x-b)^j} + \cdots + \sum_{j=1}^{\mu}\frac{2K_jx+L_j}{(x^2+px+q)^j} + \sum_{j=1}^{\delta}\frac{2M_jx+N_j}{(x^2+rx+s)^j} + \cdots$$

This factorization is unique for all proper rational fractions.

And it's easy to deduce that:

$$\int \frac{mx+n}{x^2+px+q}dx = \frac{m}{2}\ln|x^2+px+q| + \frac{2n-mp}{\sqrt{4q-p^2}}\arctan\frac{2x+p}{\sqrt{4q-p^2}} + C(q > \frac{p^2}{4})$$

$$\int \frac{mx+n}{x^2+px+q}dx = \frac{m}{2}\ln|x^2+px+q| + \frac{2n-mp}{2\sqrt{p^2-4q}}\ln\left|\frac{2x+p-\sqrt{p^2-4q}}{2x+p+\sqrt{p^2-4q}}\right| + C(q < \frac{p^2}{4})$$

Although the forms of factorization is complex, but at least it is integrable. No we are going to introduce the integration methods of each parts.

- $\int \frac{A}{(x-a)^n}dx = \frac{A}{1-n}(x-a)^{1-n} + C, (n \neq 1)$
- $\int \frac{A}{x-a}dx = A\ln|x-a| + C$
- $\int \frac{Mx+N}{x^2+px+q}dx = \int \frac{\frac{M}{2}(2x+p)+N-\frac{M}{2}p}{(x+\frac{p}{x})^2-\frac{p^2}{4}+q}dx$
- $I_r = \int \frac{1}{(x^2+px+q)^r}dx =, I_r = \frac{2x+p}{(4q-p^2)(r-1)(x^2+px+q)^{r-1}} + \frac{2(2r-3)}{(4q-p^2)(r-1)}I_{r-1}$

Integration Table

Integration by substitutions and parts are vital when facing different kinds of integration. But to accelerate the integration process, here we present some commonly used integration formula. We strongly suggest the readers to prove them one by one. They are practical for cultivating mathematical mindset.

$\int \frac{1}{a^2+x^2}\,dx = \frac{1}{a}\arctan\left(\frac{x}{a}\right) + C$

$\int \frac{1}{x^2-a^2}\,dx = \frac{1}{2a}\ln\left|\frac{x-a}{x+a}\right| + C$

$\int \frac{1}{a^2-x^2}\,dx = \frac{1}{2a}\ln\left|\frac{a+x}{a-x}\right| + C$

$\int \frac{x}{a^2+x^2}\,dx = \frac{1}{2}\ln(a^2+x^2) + C$

$\int \frac{1}{\sqrt{a^2-x^2}}\,dx = \arcsin\left(\frac{x}{a}\right) + C$

$\int \frac{1}{\sqrt{x^2+a^2}}\,dx = \ln|x+\sqrt{x^2+a^2}| + C$

$\int \frac{1}{\sqrt{x^2-a^2}}\,dx = \ln|x+\sqrt{x^2-a^2}| + C$

$\int \frac{1}{x\sqrt{x^2-a^2}}\,dx = \frac{1}{a}\mathrm{arcsec}\left|\frac{x}{a}\right| + C$

$\int \sqrt{a^2-x^2}\,dx = \frac{x}{2}\sqrt{a^2-x^2} + \frac{a^2}{2}\arcsin\left(\frac{x}{a}\right) + C$

$\int \frac{1}{x\sqrt{a^2-x^2}}\,dx = \frac{1}{a}\ln\left|\frac{a-\sqrt{a^2-x^2}}{x}\right| + C$

$\int \sqrt{x^2+a^2}\,dx = \frac{x}{2}\sqrt{x^2+a^2} + \frac{a^2}{2}\ln|x+\sqrt{x^2+a^2}| + C$

$\int \frac{1}{x\sqrt{a^2+x^2}}\,dx = -\frac{1}{a}\ln\left|\frac{a+\sqrt{a^2+x^2}}{x}\right| + C$

$\int \sqrt{x^2-a^2}\,dx = \frac{x}{2}\sqrt{x^2-a^2} - \frac{a^2}{2}\ln|x+\sqrt{x^2-a^2}| + C$

$\int \sin^2 x\,dx = \frac{x}{2} - \frac{\sin(2x)}{4} + C$

$\int \cos^2 x\,dx = \frac{x}{2} + \frac{\sin(2x)}{4} + C$

$\int \tan^2 x\,dx = \tan x - x + C$

$\int \cot^2 x\,dx = -\cot x - x + C$

$\int \sec^3 x\,dx = \frac{1}{2}(\sec x\tan x + \ln|\sec x + \tan x|) + C$

$\int \csc^3 x\,dx = \frac{1}{2}(-\csc x\cot x + \ln|\csc x - \cot x|) + C$

$\int \arcsin x\,dx = x\arcsin x + \sqrt{1-x^2} + C$

$\int \arccos x\,dx = x\arccos x - \sqrt{1-x^2} + C$

$\int \arctan x\,dx = x\arctan x - \frac{1}{2}\ln(1+x^2) + C$

$\int \mathrm{arcsec} x\,dx = x\mathrm{arcsec} x - \ln|x+\sqrt{x^2-1}| + C$

$\int \mathrm{arccot} x\,dx = x\mathrm{arccot} x + \frac{1}{2}\ln(1+x^2) + C$

$\int \sinh x\,dx = \cosh x + C$

$\int \cosh x\,dx = \sinh x + C$

$\int \tanh x\,dx = \ln(\cosh x) + C$

$\int \coth x\,dx = \ln|\sinh x| + C$

$\int \mathrm{sech}^2 x\,dx = \tanh x + C$

$\int \mathrm{csch}^2 x\,dx = -\coth x + C$

$\int \mathrm{sech} x\,dx = \arctan(\sinh x) + C$

$\int \mathrm{csch} x\,dx = \ln\left|\tanh\left(\frac{x}{2}\right)\right| + C$

$\int \sinh^2 x\,dx = \frac{\sinh(2x)}{4} - \frac{x}{2} + C$

$\int \cosh^2 x\,dx = \frac{\sinh(2x)}{4} + \frac{x}{2} + C$

$\int e^{ax}\sin(bx)\,dx = \frac{e^{ax}}{a^2+b^2}(a\sin(bx) - b\cos(bx)) + C$

$\int e^{ax}\cos(bx)\,dx = \frac{e^{ax}}{a^2+b^2}(a\cos(bx) + b\sin(bx)) + C$

2.4.2 Definite Integration

In previous content, we learned about the definition and calculation technique about indefinite integration. Here we are going to talk about definite integration. Unlike indefinite integration, definite integration have more intuitive geometric meaning. And when facing real world problems, definite integration have more direct connections with the physical background. Before we start to construct a rigorous theory and definition for definite integration, lets take a look at its geometric meaning first.

The geometric idea behind the definite integral is to compute the area under a curve $y = f(x)$ and above the x-axis, from $x = a$ to $x = b$. For a function that is non-negative and continuous, this "area" is an intuitive concept. The powerful idea of integration is to approximate this curved region by a collection of simple shapes—typically rectangles—whose areas are easy to calculate.

This approximation is achieved by first **partitioning** the interval $[a, b]$ into n subintervals. While a uniform partition is often used for simplicity, the general theory requires that our method must work for an **arbitrary**

partition, not just a regular one. On each subinterval, we construct a rectangle that approximates the area under the curve on that small segment. The key point is that the height of this rectangle is determined by the function's value at some **sample point** within the subinterval. The choice of this sample point (e.g., left endpoint, right endpoint, midpoint, or any point in between) is also arbitrary in the general formulation.

The total area of these rectangles, known as a **Riemann sum**, provides an approximation to the true area under the curve:

$$S_n = \sum_{i=1}^{n} f(c_i)\Delta x_i.$$

Intuitively, as we take thinner and thinner rectangles (i.e., as the maximum subinterval width, called the **norm** of the partition, approaches zero), the approximation becomes more accurate. The area under the curve is then defined as the **limit** of these Riemann sums, provided that this limit exists.

Crucially, for this definition to be meaningful and well-defined, the limit must converge to the same value **regardless** of how we choose the partition and the sample points. This requirement of independence from the arbitrary choices is what leads to the rigorous standard definition. For continuous functions, it can be proven that this is indeed the case, unifying the intuitive geometric concept with a precise analytical foundation.

That is the intuitive perspective of the definite integration. Now we can construct a rigorous definition.

Definition 2.4.7

Assume $f(x)$ is a bounded function defined on interval $[a, b]$, pick divisional points $\{x_i\}_{i=0}^{n}$ randomly on interval $[a, b]$, which forms a partition:

$$P : a = x_0 < x_1 < x_2 < \cdots < x_n = b$$

And we pick $\forall \xi_i \in [x_{i-1}, x_i]$, define the length of each intervals as $\Delta x_i = x_i - x_{i-1}$, and define $\lambda = \max_{1 \leq i \leq n}(\Delta x_i)$. If:

$$\lim_{\lambda \to 0} \sum_{i=1}^{n} f(\xi_i)\Delta x_i$$

exists, and the value is independent from the partition, then we claim that the function $f(x)$ is Riemann Integral on interval $[a, b]$. The expression of:

$$S_n = \sum_{i=1}^{n} f(\xi_i)\Delta x_i$$

is called the Riemann Sum. The limit is called the definite integration on $[a, b]$, denote as:

$$I = \int_a^b f(x)dx$$

a is called the lower limit of integral, and the b is called the upper limit of integral.

In the definition above, we require $a < b$. When $a > b$, we define:

$$\int_a^b f(x)dx = -\int_b^a f(x)dx$$

If $a = b$, we define the integral equals to 0.

When no confusion is likely to arise, a function that is Riemann integrable is generally simply referred to as integrable.

Example 2.4.1

Discuss the integrable of the Dirichlet function on $[0, 1]$:

$$D(x) = \begin{cases} 1, x \in \mathbb{Q} \\ 0, x \notin \mathbb{Q} \end{cases}$$

Proof. Because the set of rational number and the set of real number is dense, so whatever how you divide the interval $[0, 1]$, each small interval $[x_{i-1}, x_i]$ must include at least one rational number and real number.
So for the limit:

$$I = \lim_{\lambda \to 0} \sum_{i=1}^{n} f(\xi_i)\Delta x_i$$

If we take ξ_i as a rational number, then we have:

$$I = \lim_{\lambda \to 0} \sum_{i=1}^{n} f(\xi_i)\Delta x_i = \lim_{\lambda \to 0} \sum_{i=1}^{n} 1 \cdot \Delta x_i = 1$$

Likewise: if we take ξ_i as a irrational number, then we have:

$$I = \lim_{\lambda \to 0} \sum_{i=1}^{n} f(\xi_i)\Delta x_i = \lim_{\lambda \to 0} \sum_{i=1}^{n} 0 \cdot \Delta x_i = 0$$

Because the limits is reletaed with the value of ξ_i, the the limit $I = \lim_{\lambda \to 0} \sum_{i=1}^{n} f(\xi_i)\Delta x_i$ does not exists.
Which means the Dirichlet function cannot be integrated (from the perspective of Riemann integration).

□

Equivalent Integration Conditions

Using the definition to test whether a function is integrable is complex and sometimes inoperable. So the question is: can we find a method that is logical equivalent to the definition but is more useful? The answer is obviously yes!

Just take a look at the definition:

$$\int_a^b f(x)dx = \lim_{\lambda \to 0} \sum_{i=1}^{n} f(\xi_i)\Delta x_i$$

The key process and properties of the limit is the random selection of the value of ξ_i. So we consider the extreme condition: the supremum and the infimum of the function $f(x)$ in interval $[x_{i-1}, x_i]$, denoted as M_i and m_i.

Then we shall have two limits:

$$M = \lim_{\lambda \to 0} \sum_{i=1}^{n} M_i \cdot \Delta x_i$$

$$m = \lim_{\lambda \to 0} \sum_{i=1}^{n} m_i \cdot \Delta x_i$$

If both of the limits M and m are convergent and they converge to the same value, we can claim that the definite integration exists, because:

$$\lim_{\lambda\to 0}\sum_{i=1}^{n} m_i\cdot\Delta x_i = m \le \int_a^b f(x)dx = \lim_{\lambda\to 0}\sum_{i=1}^{n} f(\xi_i)\Delta x_i \le M = \lim_{\lambda\to 0}\sum_{i=1}^{n} M_i\cdot\Delta x_i$$

Then we can replace the random selection of ξ_i with the supremum and the infimum of the function $f(x)$, which is more operable compared to the original definition.

Now we will present a rigorous expression of the idea above, and give a proof of it.

Definition 2.4.8: Darboux Sum

For a partition P and each of its interval(s) $[x_{i-1}, x_i]$, we make the following denotions:

$$M_i = \sup\{f(x)|x\in[x_{i-1},x_i]\}, m_i = \inf\{f(x)|x\in[x_{i-1},x_i]\}$$

Obviously they are related to the choice of the partition. After we choose the partition P, we define:

$$\overline{S}(P) = \sum_{i=1}^{n} M_i\cdot\Delta x_i, \underline{S}(P) = \sum_{i=1}^{n} m_i\cdot\Delta x_i$$

The $\overline{S}(P)$ is called the **Darboux upper sum**, and the $\underline{S}(P)$ is called the **Darboux lower sum**.

It is very obvious that:

$$\underline{S}(P) \le \sum_{i=1}^{n} f(\xi_i)\Delta x_i \le \overline{S}(P)$$

The next step is to prove that if the limits $\lim_{\lambda\to 0}\overline{S}(P)$ and $\lim_{\lambda\to 0}\underline{S}(P)$ exist and converge to the same value, then the definite integral exist. ($\lambda = \max\{\Delta x_i\}$)

Theorem 2.4.3

Adding points to the original partition forms a new partition; the Darboux upper sum does not increase, and the Darboux lower sum does not decrease.

Proof. Assume $\overline{S}(P)$ and $\underline{S}(P)$ correspond to certain partition P, and $P:\{x_i\}_{i=1}^{n}$. And with a new divisional point added, we have a new partition P', whose Darboux supper sum and Darboux lower sum are $\overline{S}(P')$ and $\underline{S}(P')$ What we need to prove is that:

$$\overline{S}(P') \le \overline{S}(P), \underline{S}(P) \le \underline{S}(P')$$

Assume the added point x' falls into the interval (x_{i-1}, x_i), we denoted that:

$$M_i = \sup\{f(x)|x\in(x_{i-1},x_i)\}, M_i' = \sup\{f(x)|x\in(x_{i-1},x')\}, M_i'' = \sup\{f(x)|x\in(x',x_i)\}$$

Because $(x_{i-1}, x') \in (x_{i-1}, x_i)$, $(x', x_i) \in (x_{i-1}, x_i)$, then we have:

$$M_i' \le M_i, M_i'' \le M_i$$

$$M_i'(x'-x_{i-1}) + M_i''(x_i - x') \le M_i(x_i - x_{i-1})$$

Adding one divisional point won't interrupt other intervals, so now we have $\overline{S}(P') \le \overline{S}(P)$ Likewise, we can prove that $\underline{S}(P) \le \underline{S}(P')$

□

Now we can deduce that $m(b-a) \leq \underline{S}(P_2) \leq \overline{S}(P_1) \leq M(b-a)$.

According to The Monotone Convergence Theorem in 2.2.2.2, we can claim that the limits of $\lim_{\lambda\to 0} \overline{S}(P)$ and $\lim_{\lambda\to 0} \underline{S}(P)$ exists.

We denoted that:

$$\lim_{\lambda\to 0} \overline{S}(P) = L, \lim_{\lambda\to 0} \underline{S}(P) = l$$

And now we are going to prove that $L = \inf\{\overline{S}(P)|\overline{S}(P) \in \overline{\mathbf{S}}\}$, $l = \sup\{\underline{S}(P)|\underline{S}(P) \in \underline{\mathbf{S}}\}$ are established for all bounded function $f(x)$.

Theorem 2.4.4: Darboux Theorem

$$\lim_{\lambda\to 0} \overline{S}(P) = \inf\{\overline{S}(P)|\overline{S}(P) \in \overline{\mathbf{S}}\}$$
$$\lim_{\lambda\to 0} \underline{S}(P) = \sup\{\underline{S}(P)|\underline{S}(P) \in \underline{\mathbf{S}}\}$$

Proof. We shall only present the proof of the Darboux upper sum. The situation of the Darboux lower sum is likewise. The basic idea is to use the $\epsilon - \delta$ language, select a Darboux upper sum that satisfy the limit's condition, and prove that $\forall P, \lambda = \max_{1\leq i\leq n}(\Delta x_i) < \delta$, we have that $0 \leq \overline{S}(P) - L < \epsilon$. Now, assume we have partition P' that satisfy $0 \leq \overline{S}(P') - L < \frac{\epsilon}{2}$. And:

$$P' : a = x'_0 < x'_1 < x'_2 < \cdots < x'_p = b$$

We pick $\delta = \min\{\Delta x'_1, \Delta x'_2, \cdots, \Delta x'_p, \frac{\epsilon}{2(p-1)(M-m)}\}$. Now assume we have another partition P that satisfy $\lambda = \max_{1\leq i\leq n}(\Delta x_i) < \delta$:

$$P : a = x_0 < x_1 < x_2 < \cdots < x_n = b$$

And its Darboux upper sum is $\overline{S}(P)$, we insert $P' = \{x'_j\}_{j=0}^{p}$ into $P = \{x_i\}_{i=0}^{n}$ and form a new partition $P*$. Likewise, we denote its Darboux upper sum as $\overline{S}(P*)$. According to previous lemma, we have that:

$$\overline{S}(P*) - \overline{S}(P') \leq 0$$

For all the interval (x_{i-1}, x_i), we have at most $p-1$ intervals that have divisional points inserted. For other intervals, there won't be any changes. For the intervals that are inserted, take use of the notations previously used in the proof, we have:

$$M_i(x_i - x_{i-1}) - [M'_i(x'_j - x_{i-1}) + M''_i(x_i - x'_j)] \leq (M-m)(x_i - x_{i-1}) < (M-m)\delta$$

So now we have that:

$$0 \leq \overline{S}(P) - \overline{S}(P*) < (p-1)(M-m)\delta \leq \frac{\epsilon}{2}$$

So after all, we conclude:

$$0 \leq \overline{S}(P) - L = [\overline{S}(P) - \overline{S}(P*)] + [\overline{S}(P*) - \overline{S}(P')] + [\overline{S}(P') - L] < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

□

Now we have the necessary and sufficient condition of the integrable.

Theorem 2.4.5

The necessary and sufficient condition of bounded function $f(x)$ on interval $[a,b]$ is that: for every partition P, when $\lambda = \max_{1\le i\le n} \to 0$, and we have:

$$\lim_{\lambda\to 0} \overline{S}(P) = L = l = \lim_{\lambda\to 0} \underline{S}(P)$$

Proof. We now complete the proof of the theorem. Let f be a bounded function on $[a,b]$, and define:

$$L = \lim_{\lambda\to 0} \overline{S}(P), \quad l = \lim_{\lambda\to 0} \underline{S}(P).$$

By the Darboux Theorem, we have:

$$L = \inf\left\{\overline{S}(P)\right\}, \quad l = \sup\left\{\underline{S}(P)\right\}.$$

Necessity: If f is integrable on $[a,b]$, then there exists a number I such that for every $\epsilon > 0$, there exists $\delta > 0$ such that for any partition P with $\lambda < \delta$ and any choice of sample points $\xi_i \in [x_{i-1}, x_i]$, we have:

$$\left|\sum_{i=1}^{n} f(\xi_i)\Delta x_i - I\right| < \epsilon.$$

In particular, for any such partition P, we can choose sample points such that $f(\xi_i)$ is arbitrarily close to M_i, yielding:

$$\left|\overline{S}(P) - I\right| \le \epsilon.$$

Similarly, by choosing sample points where $f(\xi_i)$ is arbitrarily close to m_i, we obtain:

$$\left|\underline{S}(P) - I\right| \le \epsilon.$$

Hence, as $\lambda \to 0$, we have:

$$\overline{S}(P) \to I \quad \text{and} \quad \underline{S}(P) \to I,$$

which implies $L = l = I$.

Sufficiency: Conversely, suppose $L = l = I$. Then for every $\epsilon > 0$, there exists $\delta > 0$ such that for any partition P with $\lambda < \delta$, we have:

$$\left|\overline{S}(P) - I\right| < \epsilon \quad \text{and} \quad \left|\underline{S}(P) - I\right| < \epsilon.$$

For any Riemann sum $\sum_{i=1}^{n} f(\xi_i)\Delta x_i$ corresponding to P, we have:

$$\underline{S}(P) \le \sum_{i=1}^{n} f(\xi_i)\Delta x_i \le \overline{S}(P).$$

Therefore,

$$I - \epsilon < \underline{S}(P) \le \sum_{i=1}^{n} f(\xi_i)\Delta x_i \le \overline{S}(P) < I + \epsilon,$$

which implies:

$$\left|\sum_{i=1}^{n} f(\xi_i)\Delta x_i - I\right| < \epsilon.$$

Thus, f is integrable on $[a,b]$ with integral I.

This completes the proof of the theorem. $\square$

From the theorem above, we can derive a more practical criterion involving the oscillation of the function. Define the oscillation on the subinterval $[x_{i-1}, x_i]$ as $\omega_i = M_i - m_i$. The condition $L = l$ is equivalent to:

$$\lim_{\lambda \to 0} \sum_{i=1}^{n} \omega_i \Delta x_i = 0$$

This criterion allows us to identify broad classes of functions that are Riemann integrable.

Classes of Integrable Functions

While the definition involving Riemann sums or Darboux sums is necessary for theoretical rigor, we do not use it to check every specific function. Instead, we rely on established theorems regarding the integrability of common function classes.

Theorem 2.4.6: Integrability of Continuous Functions

If $f(x)$ is continuous on the closed interval $[a, b]$, then $f(x)$ is Riemann integrable on $[a, b]$.

Theorem 2.4.7: Integrability of Monotonic Functions

If $f(x)$ is bounded and monotonic on the closed interval $[a, b]$, then $f(x)$ is Riemann integrable on $[a, b]$.

These theorems cover the vast majority of functions encountered in physical and engineering problems.

The Fundamental Theorem of Calculus

So far, we have defined the definite integral as a limit of sums. However, calculating limits of sums for complex functions is incredibly tedious. The connection between the definite integral (area) and the indefinite integral (antiderivative) is provided by the Newton-Leibniz Formula.

Theorem 2.4.8: Newton-Leibniz Formula

If $f(x)$ is continuous on $[a, b]$, and $F(x)$ is an antiderivative of $f(x)$ on $[a, b]$ (i.e., $F'(x) = f(x)$), then:

$$\int_a^b f(x)dx = F(b) - F(a)$$

This theorem transforms the problem of summation into a problem of finding an antiderivative, unifying the geometric concept of integration with the algebraic operation of differentiation.

When applying the Newton-Leibniz Formula, we must ensure that the function is continuous on the interval. If there are discontinuities, we need to check whether the function is still integrable using the criteria discussed earlier.

There is some technique when applying the Newton-Leibniz Formula, such as substitution and integration by parts, which are similar to those used in indefinite integration.

For example, how can we calculate $\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{1+e^x} \cos^3 x dx$?

We spotted that the function is continuous on the interval and if we take $x = -t$, then we have:

$$I = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{1+e^x} \cos^3 x dx = \int_{\frac{\pi}{2}}^{-\frac{\pi}{2}} \frac{e^t}{1+e^t} \cos^3(-t) dt = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{e^x}{1+e^x} \cos^3 x dx$$

Then we have:

$$I + I = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (\frac{1}{1+e^x}\cos^3 x) + (\frac{e^x}{1+e^x}\cos^3 x)dx = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^3 x dx$$

$$I + I = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^3 x dx = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (1-\sin^2 x)d(\sin x) = (\sin x - \frac{\sin^3 x}{3})\Big|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = \frac{4}{3}$$

Thus, $\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{1+e^x}\cos^3 x dx = \frac{2}{3}$.

2.5 Improper Integrals

In the discussion of the Riemann integral $\int_a^b f(x)dx$, we relied on two fundamental restrictions:

1. The integration interval $[a, b]$ is finite (a closed, bounded interval).
2. The function $f(x)$ is bounded on $[a, b]$.

However, many problems in mathematics, physics, and probability theory require us to relax these conditions. For example, calculating the escape velocity of a rocket requires integrating gravitational force over an infinite distance, or calculating the mean lifetime of a particle involves an integral from 0 to $+\infty$. Furthermore, some physically relevant functions approach infinity (blow up) at certain points. Integrals that violate either of these two conditions are called **Improper Integrals** (or Generalized Integrals). We define them using limits. First, we consider the case where the interval of integration is infinite. These are often called improper integrals of the **first kind**.

2.5.1 Improper Integrals of the First Kind (Infinite Intervals)

Definition 2.5.1

Let $f(x)$ be defined on the infinite interval $[a, +\infty)$ and be integrable on every finite subinterval $[a, u]$ where $u > a$. We define the improper integral of f over $[a, +\infty)$ as:

$$\int_a^{+\infty} f(x)dx = \lim_{u\to+\infty} \int_a^u f(x)dx$$

- If the limit exists and is a finite number, we say the improper integral **converges**.
- If the limit does not exist (including becoming infinite), we say the improper integral **diverges**.

Geometrically, this represents the area of an unbounded region. Even though the region extends infinitely to the right, the total area can still be finite if the curve approaches the x-axis sufficiently fast.

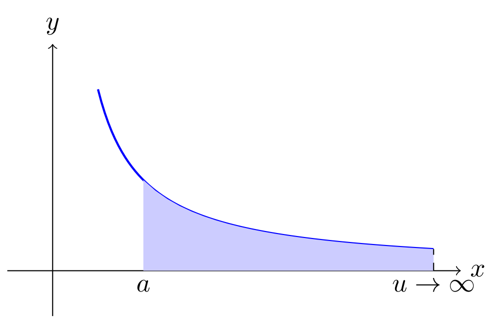

Figure 2.1: Visualizing an integral over an infinite interval

Similarly, we can define integrals for other infinite intervals:

$$\int_{-\infty}^{b} f(x)dx = \lim_{u \to -\infty} \int_{u}^{b} f(x)dx$$

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{c} f(x)dx + \int_{c}^{+\infty} f(x)dx$$

For the integral from $-\infty$ to $+\infty$ to converge, **both** constituent integrals must converge independently. The choice of the splitting point c does not affect the convergence.

Benchmark: The p-Integral (Type I)

To determine the convergence of complex functions, we compare them to the power function $1/x^p$.

Theorem 2.5.1

The integral $\int_1^{+\infty} \frac{1}{x^p}\, dx$:

- **Converges** if $p > 1$.
- **Diverges** if $p \leq 1$.

Proof. Evaluating $\int_1^u x^{-p}\, dx$:

- If $p = 1$, $\ln u \to \infty$ as $u \to \infty$.
- If $p \neq 1$, $\frac{u^{1-p}-1}{1-p}$. For convergence, we need $u^{1-p} \to 0$, which requires $1 - p < 0 \implies p > 1$.

$\square$

2.5.2 Improper Integrals of the Second Kind (Unbounded Functions)

These integrals occur on a finite interval $[a, b]$ where the integrand $f(x)$ becomes infinite (has a singularity) at one or more points.

Definition 2.5.2

If f is continuous on $[a, b)$ and discontinuous at b (e.g., $\lim_{x \to b^-} |f(x)| = \infty$), we define:

$$\int_a^b f(x)\, dx = \lim_{\epsilon \to 0^+} \int_a^{b-\epsilon} f(x)\, dx$$

Benchmark: The p-Integral (Type II)

Be careful: the convergence condition for singularities is the *reverse* of infinite intervals.

Theorem 2.5.2

For the interval $(0, 1]$, the integral $\int_0^1 \frac{1}{x^p}\, dx$:

- **Converges** if $p < 1$.
- **Diverges** if $p \geq 1$.

2.5.3 General Theory of Convergence

Before applying practical tests, we establish the rigorous necessary and sufficient conditions for convergence.

Cauchy Criterion

The Cauchy Criterion is fundamental because it allows us to prove convergence without knowing the limit's value.

Theorem 2.5.3: Cauchy Criterion for Improper Integrals

The integral $\int_a^{+\infty} f(x)\,dx$ converges **if and only if** for every $\epsilon > 0$, there exists an $M > a$ such that for all $u_2 > u_1 > M$:

$$\left| \int_{u_1}^{u_2} f(x)\,dx \right| < \epsilon$$

Absolute vs. Conditional Convergence

- **Absolute Convergence:** $\int |f(x)|\,dx$ converges.
- **Conditional Convergence:** $\int f(x)\,dx$ converges, but $\int |f(x)|\,dx$ diverges.

Theorem 2.5.4

If $\int_a^{+\infty} |f(x)|\,dx$ converges, then $\int_a^{+\infty} f(x)\,dx$ converges.

2.5.4 Convergence Tests

Direct Comparison Test

Theorem 2.5.5

Let $f(x)$ and $g(x)$ be continuous functions on $[a, +\infty)$ such that $0 \leq f(x) \leq g(x)$ for all $x \geq a$.

1. If $\int_a^{+\infty} g(x)dx$ converges, then $\int_a^{+\infty} f(x)dx$ also converges.
2. If $\int_a^{+\infty} f(x)dx$ diverges, then $\int_a^{+\infty} g(x)dx$ also diverges.

Intuition: If the area under the larger curve is finite, the area under the smaller curve must be finite. If the area under the smaller curve is infinite, the larger area must be infinite.

Example 2.5.1

Does $\int_1^{+\infty} \frac{\sin^2 x}{x^2} dx$ converge?

Solution: We know that $0 \leq \sin^2 x \leq 1$. Therefore:

$$0 \leq \frac{\sin^2 x}{x^2} \leq \frac{1}{x^2}$$

We know that $\int_1^{+\infty} \frac{1}{x^2} dx$ converges ($p = 2 > 1$). By the Direct Comparison Test, $\int_1^{+\infty} \frac{\sin^2 x}{x^2} dx$ converges.

Limit Comparison Test

Sometimes finding a direct inequality is difficult. The Limit Comparison Test is often more powerful.

Theorem 2.5.6

Let $f(x)$ and $g(x)$ be positive continuous functions on $[a, +\infty)$. If:

$$\lim_{x \to +\infty} \frac{f(x)}{g(x)} = L$$

where $0 < L < +\infty$, then $\int_a^{+\infty} f(x)dx$ and $\int_a^{+\infty} g(x)dx$ either both converge or both diverge.

If $L = 0$ and $g(x)$ converges, then $f(x)$ converges. If $L = +\infty$ and $g(x)$ diverges, then $f(x)$ diverges.

Example 2.5.2

Analyze $\int_1^{+\infty} \frac{x}{1+x^3} dx$.

Solution: For large x, the term 1 is negligible, so $\frac{x}{1+x^3} \approx \frac{x}{x^3} = \frac{1}{x^2}$.

Let $f(x) = \frac{x}{1+x^3}$ and $g(x) = \frac{1}{x^2}$.

$$\lim_{x \to +\infty} \frac{f(x)}{g(x)} = \lim_{x \to +\infty} \frac{x/(1+x^3)}{1/x^2} = \lim_{x \to +\infty} \frac{x^3}{1+x^3} = 1$$

Since $L = 1$ (finite and positive) and we know $\int_1^{+\infty} \frac{1}{x^2} dx$ converges, the original integral converges.

For positive functions, we use comparison tests. For oscillating functions, we need more advanced tools.

Cauchy's Limit Comparison Test (Order Analysis)

This is the most practical method for determining convergence. It formalizes the idea of comparing a function to $1/x^p$.

Theorem 2.5.7

Let $f(x)$ be a positive function defined on $[a, +\infty)$. Consider the limit of $f(x)$ multiplied by a test power x^p:

$$\lambda = \lim_{x \to +\infty} x^p f(x)$$

1. **Convergence Case:** If we can find a $p > 1$ such that $0 \leq \lambda < +\infty$, then $\int_a^{+\infty} f(x)\, dx$ **converges**. *(Meaning: $f(x)$ goes to zero faster than $1/x$, roughly like $1/x^p$)*
2. **Divergence Case:** If we can find a $p \leq 1$ such that $0 < \lambda \leq +\infty$, then $\int_a^{+\infty} f(x)\, dx$ **diverges**. *(Meaning: $f(x)$ goes to zero slower than or equal to $1/x$)*

Example 2.5.3

Test $\int_1^{+\infty} \frac{\ln x}{x^2}\, dx$. We suspect convergence because x^2 dominates. Let's compare with $p = 1.5$ (since $1 < 1.5 < 2$, giving us "room").

$$\lim_{x \to +\infty} x^{1.5} \frac{\ln x}{x^2} = \lim_{x \to +\infty} \frac{\ln x}{x^{0.5}} = 0 \quad \text{(by L'Hopital)}$$

Since $p = 1.5 > 1$ and the limit is finite, the integral converges.

Dirichlet and Abel Tests (For Conditional Convergence)

These tests are used for integrals of the product form $\int_a^{+\infty} f(x)g(x)\,dx$, typically where one part oscillates and the other decays.

Theorem 2.5.8: Dirichlet's Test

The integral $\int_a^{+\infty} f(x)g(x)\,dx$ converges if:

1. $f(x)$ has bounded partial integrals: $\exists M, \forall u > a, |\int_a^u f(t)\,dt| \le M$.
2. $g(x)$ is monotonic decreasing and $\lim_{x\to+\infty} g(x) = 0$.

Classic Application: $\int_0^{+\infty} \frac{\sin x}{x}\,dx$. Here $f(x) = \sin x$ (bounded integral) and $g(x) = 1/x$ (monotonic to 0).

Theorem 2.5.9: Abel's Test

The integral $\int_a^{+\infty} f(x)g(x)\,dx$ converges if:

1. $\int_a^{+\infty} f(x)\,dx$ converges (the integral itself).
2. $g(x)$ is monotonic and bounded.

Cauchy Principal Value (P.V.)

In some divergent integrals, the positive and negative areas might cancel each other out if limits are taken symmetrically. This value is called the Cauchy Principal Value.

Definition 2.5.3

1. For singularities at $c \in (a, b)$:

$$\text{P.V.} \int_a^b f(x)\,dx = \lim_{\epsilon\to0^+} \left(\int_a^{c-\epsilon} f(x)\,dx + \int_{c+\epsilon}^b f(x)\,dx \right)$$

2. For infinite intervals $(-\infty, +\infty)$:

$$\text{P.V.} \int_{-\infty}^{+\infty} f(x)\,dx = \lim_{R\to+\infty} \int_{-R}^{R} f(x)\,dx$$

Remark 2.5.1

Convergence $\implies$ **P.V. exists**, but the converse is false.

Example 2.5.4

Consider $f(x) = \frac{1}{x}$ on $[-1, 1]$.

- **Improper Integral:** $\int_{-1}^1 \frac{1}{x}\,dx = \int_{-1}^0 \frac{1}{x}\,dx + \int_0^1 \frac{1}{x}\,dx$. Since $\int_\epsilon^1 \frac{1}{x}\,dx = -\ln\epsilon \to \infty$, the standard integral **diverges**.
- **Principal Value:**

$$\text{P.V.} \int_{-1}^1 \frac{1}{x}\,dx = \lim_{\epsilon\to0^+} \left(\int_{-1}^{-\epsilon} \frac{1}{x}\,dx + \int_\epsilon^1 \frac{1}{x}\,dx \right) = 0$$

(Due to the odd symmetry of the function).

Comprehensive Example: The Gamma Function

$$\Gamma(s) = \int_0^{+\infty} x^{s-1}e^{-x}\,dx$$

This integral requires analysis of both singularity and infinite bounds.

- **At** 0**:** If $s < 1$, x^{s-1} has a singularity. Since $e^{-x} \approx 1$, it behaves like $\int_0^1 \frac{1}{x^{1-s}}dx$. By the Type II p-test, this converges if $1 - s < 1 \implies s > 0$.
- **At** $+\infty$**:** e^{-x} decays faster than any power x^p grows. Using the Limit Comparison Test with $1/x^2$:

  $$\lim_{x\to\infty} x^2(x^{s-1}e^{-x}) = 0$$

  Thus, it converges for all s.

Conclusion: The integral converges for $s > 0$.

At the end of this chapter, let's use the knowledge we learned to solve an interesting problem: the Gabriel's Horn.

Considering we have a horn, whose cross section is a circle but longitudinal section is function $y = \frac{1}{x}, x \in (1, \infty)$.

Let's first calculate the volume of the horn:

$$V = \pi\int_1^{+\infty}(\frac{1}{x})^2dx = \pi[-\frac{1}{x}]_1^{+\infty} = \pi$$

We can see that the horn have finite volume.

Then calculate the surface area of the horn.

$$S = 2\pi\int_1^{+\infty}\frac{1}{x}\sqrt{1+\frac{1}{x^4}}dx \geq 2\pi\int_1^{+\infty}\frac{1}{x}dx = +\infty$$

The integration product of the surface area is divergent.

Here comes a very interesting paradox. We can use finite many paint to fill the horn, but this horn of paint cannot coated the inner wall of the horn.

This consequence is contradicting to our naive understanding.

As we close this chapter, remember that this is not an end, but a gateway. The journey of mathematical analysis continues: in the **Chapter 4**, we will extend these foundations to higher dimensions, study curves and surfaces in greater depth, and encounter even more beautiful and unexpected results. The horn's call, echoing from the realm of the infinite, invites us to explore further.

In later chapters, we will move on to more advanced topics in analysis, such as multi-variables calculus, differential equations, complex analysis and functional analysis. Each of these areas builds upon the concepts we have developed here, and each offers its own unique insights and challenges. We encourage readers to continue their exploration, armed with the rigorous tools and deep understanding gained from this chapter.

References:

Mathematical Analysis (Third Edition), Chen, J., Higher Education Press
The Real Numbers and Real Analysis, Ethan D. Bloch, Springer

Chapter 3

Linear Algebra

So long after we finished the first part of the analysis section, we will now move on to the new chapter: Linear Algebra. It's a completely new part of mathematics.

Unlike analysis and calculus, linear algebra requires more geometric understanding. Here, proof is still important. But what is more crucial for learners is to construct geometric intuition for the subject. We would like to claim that intuition is different from imagination. Though we can only imagine the three dimensional Euclidean space, but our intuition can bring us forward to higher dimension, and more abstract linear spaces.

In this part, we will start from the topic of solving Systems of Linear Equations, then we will move on to study a special kind of linear space: the Vector Space. Finally, we will move on to a more abstract part: the Linear Space, this equipped us with a new tool to study more general forms of algebraic structure and system.

Specifically, in the final part focusing on abstract Linear Spaces, we will delve into key concepts such as linear independence, basis, dimension, and linear transformations. Understanding these foundational ideas allows us to unify seemingly disparate mathematical objects—like functions, polynomials, and matrices—under a single, powerful framework. This abstraction is not just an academic exercise; it provides the essential language and tools for tackling complex problems in fields ranging from differential equations and data science to quantum mechanics and engineering optimization. The geometric intuition developed in the study of $\mathbb{R}^n$ *will prove indispensable as we navigate these higher-dimensional, abstract realms, cementing linear algebra as one of the most fundamental and broadly applicable areas of modern mathematics.*

In latter chapters, we will introduce another branch of Algebra, the abstract algebra. Unlike Linear algebra that focused more on multivariate equations, abstract algebra study the solution structure of higher-degree equations, which is an even more abstract part of mathematics. But even in abstract algebra we still need the knowledge about linear algebra. So now, let's get started.

3.1 Linear Equations and Matrices

3.1.1 Systems of Linear Equations

I believe most of the readers have seen systems of equations like this:

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

They might have different numbers of variables and equations, but they all shared the same feature: **Each equation is linear.** Variables are raised only to the first power, with no products between variables (like xy), or nonlinear functions (e.g., $\sin(x)$ or $\sqrt{x}$).

For such kind of equation system, we call it the **system of linear equations**. Their general expression has m equations and n variables (also called an $m \times n$ system):

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases}$$

Here, $x_1, \dots, x_n$ are the **variables** (unknowns), the a_{ij} are the **coefficients**, and $b_1, \dots, b_m$ are the **constant terms**.

A **solution** to the system is a set of numbers $(s_1, s_2, \dots, s_n)$ that satisfies all m equations simultaneously when substituted for $(x_1, x_2, \dots, x_n)$. The set of all possible solutions is called the **solution set**.

Example 3.1.1: A 2x2 System

Consider the system:

$$\begin{cases} x_1 + x_2 = 3 \\ x_1 - x_2 = -1 \end{cases}$$

We can solve this using simple algebra.

1. **Substitution:** From the first equation, $x_2 = 3 - x_1$. Substitute this into the second: $x_1 - (3 - x_1) = -1$, which gives $2x_1 - 3 = -1$, so $2x_1 = 2$, and $x_1 = 1$. Then $x_2 = 3 - 1 = 2$. The unique solution is $(1, 2)$.
2. **Elimination:** Add the two equations: $(x_1 + x_2) + (x_1 - x_2) = 3 + (-1)$, which gives $2x_1 = 2$, so $x_1 = 1$. Subtract the second from the first: $(x_1 + x_2) - (x_1 - x_2) = 3 - (-1)$, which gives $2x_2 = 4$, so $x_2 = 2$.

The solution set is the single point $(1, 2)$.

In linear algebra, we are interested in three questions:

1. **Existence:** Does a solution exist? (Is the system **consistent**?)
2. **Uniqueness:** If a solution exists, is it the only one?
3. **Computation:** If solutions exist, how do we find them?

Geometrically, for a 2x2 system, each equation represents a line in the $\mathbb{R}^2$ plane. The solution set is the intersection of these lines.

- **Unique Solution:** The lines intersect at a single point.
- **No Solution:** The lines are parallel and distinct.
- **Infinitely Many Solutions:** The two equations represent the same line.

For a 3x3 system, each equation is a plane in $\mathbb{R}^3$. The solution set is the intersection of these three planes, which could be a point, a line, a plane, or empty.

The methods of substitution and elimination become extremely cumbersome for larger systems (e.g., 5 equations, 5 variables). We need a more systematic and efficient approach. This is where matrices come in.

3.1.2 Matrix Algebra

Matrix Notation and Special Matrices

The essence of our new method is to manipulate the equations without changing their solution set. We observe that all the information of the system is contained in the coefficients a_{ij} and the constant terms b_i. The variable names $x_1, x_2, \ldots$ are just placeholders. We can therefore encode the entire system into a compact rectangular array called a **matrix**.

Definition 3.1.1

A **matrix** is a rectangular array of numbers, called **entries** or **elements**. A matrix with m rows and n columns is called an $m \times n$ matrix (read "m by n").

For the general linear system, we define two key matrices.

The **coefficient matrix** is:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

We can denote this matrix as $A = [a_{ij}]$. a_{ij} represents the element in the i-th row and j-th column.

The **augmented matrix**, which includes the constant terms, is:

$$(A \mid \mathbf{b}) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right)$$

Solving the system is now equivalent to manipulating this augmented matrix.

Example 3.1.2

The system

$$\begin{cases} x_1 - 2x_2 + x_3 = 0 \\ 2x_2 - 8x_3 = 8 \\ 5x_1 - 5x_3 = 10 \end{cases}$$

has the coefficient matrix

$$A = \begin{pmatrix} 1 & -2 & 1 \\ 0 & 2 & -8 \\ 5 & 0 & -5 \end{pmatrix}$$

and the augmented matrix

$$(A \mid \mathbf{b}) = \left(\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ 0 & 2 & -8 & 8 \\ 5 & 0 & -5 & 10 \end{array}\right)$$

Matrix Terminology and Special Matrices

- **Size/Dimension:** A matrix A with m rows and n columns has size $m \times n$.
- **Square Matrix:** A matrix is **square** if its number of rows equals its number of columns ($m = n$).
- **Equality:** Two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ are **equal** if and only if they have the same size ($m \times n$) and all their corresponding entries are equal ($a_{ij} = b_{ij}$ for all i, j).

- **Principal Diagonal:** In a square matrix, the entries $a_{11}, a_{22}, \ldots, a_{nn}$ form the **principal diagonal** (or main diagonal).
- **Auxiliary Diagonal:** In a square matrix, the diagonal from the upper right to the lower left $(a_{1n}, a_{2,n-1}, \ldots, a_{n1})$ is the **auxiliary diagonal**.

There are several special types of matrices:

1. **Null matrix (Zero matrix):** A matrix (of any size) where all entries are 0. It is often denoted $\mathbf{0}$ or $\mathbf{0}_{m\times n}$.
$$\mathbf{0} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$
2. **Identity matrix:** A square matrix I_n (or just I) whose entries on the principal diagonal are all 1, and all other entries are 0.
$$I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
The identity matrix is the multiplicative identity: $AI = A$ and $IA = A$ (for compatible sizes).
3. **Diagonal matrix:** A square matrix where all entries *off* the principal diagonal are 0.
$$D = \begin{pmatrix} 3 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$
4. **Upper triangular matrix:** A square matrix whose entries *below* the principal diagonal are all 0 ($a_{ij} = 0$ for $i > j$).
$$U = \begin{pmatrix} 1 & 4 & -1 \\ 0 & 2 & 7 \\ 0 & 0 & 3 \end{pmatrix}$$
5. **Lower triangular matrix:** A square matrix whose entries *above* the principal diagonal are all 0 ($a_{ij} = 0$ for $i < j$).
$$L = \begin{pmatrix} 1 & 0 & 0 \\ 5 & 2 & 0 \\ -1 & 0 & 3 \end{pmatrix}$$
6. **Transpose:** The **transpose** of an $m \times n$ matrix A, denoted A^T (or A'), is the $n \times m$ matrix obtained by interchanging its rows and columns. That is, $(A^T)_{ij} = A_{ji}$.
$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \quad \Longrightarrow \quad A^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$$
7. **Symmetric matrix:** A square matrix A such that $A^T = A$. This means $a_{ij} = a_{ji}$ for all i, j.
$$S = \begin{pmatrix} 1 & 5 & -1 \\ 5 & 2 & 0 \\ -1 & 0 & 3 \end{pmatrix}$$
8. **Skew-symmetric matrix:** A square matrix A such that $A^T = -A$. This means $a_{ij} = -a_{ji}$ (and $a_{ii} = 0$).
$$K = \begin{pmatrix} 0 & 5 & -1 \\ -5 & 0 & 2 \\ 1 & -2 & 0 \end{pmatrix}$$

Matrix Operations

We can define algebraic operations on matrices.

Matrix Addition and Scalar Multiplication

Definition 3.1.2

Let $A = [a_{ij}]$ and $B = [b_{ij}]$ be two matrices of the **same size** $m \times n$.

1. **Addition:** Their sum $A + B$ is the $m \times n$ matrix $C = [c_{ij}]$ where $c_{ij} = a_{ij} + b_{ij}$.
2. **Scalar Multiplication:** Let c be a scalar (a real number). The scalar multiple cA is the $m \times n$ matrix $D = [d_{ij}]$ where $d_{ij} = c \cdot a_{ij}$.

Matrix subtraction is defined as $A - B = A + (-1)B$.

Example 3.1.3

Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $B = \begin{pmatrix} 5 & 0 \\ -1 & 7 \end{pmatrix}$. Then

$$A + B = \begin{pmatrix} 1+5 & 2+0 \\ 3-1 & 4+7 \end{pmatrix} = \begin{pmatrix} 6 & 2 \\ 2 & 11 \end{pmatrix}$$

$$3A = \begin{pmatrix} 3(1) & 3(2) \\ 3(3) & 3(4) \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 9 & 12 \end{pmatrix}$$

Note that $A + \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ is **undefined** as the sizes do not match.

These operations obey familiar properties:

Property 3.1.1: Properties of Addition and Scalar Multiplication

Let A, B, C be $m \times n$ matrices and c, d be scalars.

1. $A + B = B + A$ (Commutativity of Addition)
2. $(A + B) + C = A + (B + C)$ (Associativity of Addition)
3. $A + \mathbf{0} = A$ (Additive Identity)
4. $A + (-A) = \mathbf{0}$ (Additive Inverse)
5. $c(A + B) = cA + cB$ (Distributivity)
6. $(c + d)A = cA + dA$ (Distributivity)
7. $c(dA) = (cd)A$
8. $1A = A$

Remark 3.1.1

These 8 properties, plus closure, are precisely the axioms of a **Vector Space**. The set $M_{m\times n}$ of all $m \times n$ matrices is a prime example of a vector space.

Matrix Multiplication This operation is more complex and profoundly important.

Definition 3.1.3: Matrix Multiplication

Let A be an $m \times \mathbf{n}$ matrix and B be an $\mathbf{n} \times p$ matrix. Their **product** AB is an $m \times p$ matrix $C = [c_{ij}]$. The entry c_{ij} in the i-th row and j-th column of AB is computed by taking the **dot product** of the i-th row of A and the j-th column of B.

$$c_{ij} = (\text{Row } i \text{ of } A) \cdot (\text{Column } j \text{ of } B) = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$$

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$$

Crucial Note: The product AB is only defined if the **number of columns in A** equals the **number of rows in B**.

$$(m \times \mathbf{n}) \cdot (\mathbf{n} \times p) \to (m \times p)$$

Example 3.1.4

Let $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ (2×3) and $B = \begin{pmatrix} 7 & 8 \\ 9 & 0 \\ 1 & 2 \end{pmatrix}$ (3×2). The product AB will be a 2×2 matrix.

$$AB = \begin{pmatrix} (1 \cdot 7 + 2 \cdot 9 + 3 \cdot 1) & (1 \cdot 8 + 2 \cdot 0 + 3 \cdot 2) \\ (4 \cdot 7 + 5 \cdot 9 + 6 \cdot 1) & (4 \cdot 8 + 5 \cdot 0 + 6 \cdot 2) \end{pmatrix} = \begin{pmatrix} (7 + 18 + 3) & (8 + 0 + 6) \\ (28 + 45 + 6) & (32 + 0 + 12) \end{pmatrix} = \begin{pmatrix} 28 & 14 \\ 79 & 44 \end{pmatrix}$$

Now let's compute BA. This will be a 3×3 matrix.

$$BA = \begin{pmatrix} 7 & 8 \\ 9 & 0 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} (7 \cdot 1 + 8 \cdot 4) & (7 \cdot 2 + 8 \cdot 5) & (7 \cdot 3 + 8 \cdot 6) \\ (9 \cdot 1 + 0 \cdot 4) & (9 \cdot 2 + 0 \cdot 5) & (9 \cdot 3 + 0 \cdot 6) \\ (1 \cdot 1 + 2 \cdot 4) & (1 \cdot 2 + 2 \cdot 5) & (1 \cdot 3 + 2 \cdot 6) \end{pmatrix} = \begin{pmatrix} 39 & 54 & 69 \\ 9 & 18 & 27 \\ 9 & 12 & 15 \end{pmatrix}$$

Property 3.1.2: Properties of Matrix Multiplication

Let A, B, C be matrices of compatible sizes and c be a scalar.

1. **Warning:** $AB \neq BA$ in general. Matrix multiplication is **not commutative**. (See example above).
2. $A(BC) = (AB)C$ (Associativity)
3. $A(B + C) = AB + AC$ (Left Distributivity)
4. $(A + B)C = AC + BC$ (Right Distributivity)
5. $c(AB) = (cA)B = A(cB)$
6. $I_m A = A = AI_n$ (Multiplicative Identity)

But we still want the multiplication to have such kind of property like $AB = BA$, we will introduce a special kind of matrix in later chapters that can satisfy this property.

The prooves of the properties above are left for readers.

Remark 3.1.2

Remember, $\mathbf{Ab} = \mathbf{0}$ can not deduce $A = 0$ or $B = 0$. This is another counterexample against our common sense of algebraic calculations.

Definition 3.1.4

Assume we have a square matrix with n orders, and a number $m \in \mathbb{N}^*$, then we call the product of m copies of A "A to the m-th power", denoted as A^m Specifically, we define $A^0 = I_n$.

The calculations of power is **BASICALLY** the same with the regular rules. Except those have requirement with the orders of multiplication. Like:

$$(A+B)^2 = A^2 + AB + BA + B^2$$
$$(A+B)(A-B) = A^2 - AB + BA - B^2$$

Definition 3.1.5

Assume $A = (a_{ij})_{m\times n}$, we define $A^T = (b_{kl})_{n\times m}$ the transpose of matrix A, iff $b_{kl} = a_{lk}, (k = 1, 2, \cdots, n, l = 1, 2, \cdots, m)$

Property 3.1.3: Properties of the Transpose

1. $(A^T)^T = A$
2. $(A+B)^T = A^T + B^T$
3. $(cA)^T = cA^T$
4. **(Reversal Property)** $(AB)^T = B^T A^T$
5. $(A^m)^T = (A^T)^m$

Proof of $(AB)^T = B^T A^T$. Let A be $m \times n$ and B be $n \times p$. AB is $m \times p$, so $(AB)^T$ is $p \times m$. B^T is $p \times n$ and A^T is $n \times m$, so $B^T A^T$ is also $p \times m$. They have the same size. Let $C = AB$. The (i, j)-entry of C^T is C_{ji}. By definition, $C_{ji} = \sum_{k=1}^{n} A_{jk} B_{ki}$. Now let $D = B^T A^T$. The (i, j)-entry of D is:

$$D_{ij} = \sum_{k=1}^{n} (B^T)_{ik} (A^T)_{kj}$$

By definition of transpose, $(B^T)_{ik} = B_{ki}$ and $(A^T)_{kj} = A_{jk}$. So, $D_{ij} = \sum_{k=1}^{n} B_{ki} A_{jk} = \sum_{k=1}^{n} A_{jk} B_{ki}$. Thus, $D_{ij} = C_{ji} = (C^T)_{ij}$. Since all entries are equal, $B^T A^T = (AB)^T$. □

There is also another important value for square matrix, we will use it in later contents.

Definition 3.1.6: The trace of matrix

Assume a square matrix A with n orders $A = (a_{ij})$, $\sum_{i=1}^{n} a_{ii}$ is called the trace of matrix, denoted as $tr(A)$

Property 3.1.4

1. $tr(A+B) = tr(A) + tr(B)$
2. $tr(kA) = ktr(A)$
3. $tr(AB) = tr(BA)$
4. $tr(A^T) = tr(A)$

Definition of Block Matrices

A block matrix is a matrix that is partitioned into smaller submatrices called **blocks**. This is done by drawing horizontal and vertical lines that divide the matrix into rectangular blocks. Partitioning allows us to view a large matrix as composed of smaller, more manageable parts.

If A is an $m \times n$ matrix, we can partition it as follows:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1q} \\ A_{21} & A_{22} & \cdots & A_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pq} \end{bmatrix}$$

Here, each A_{ij} is a submatrix (block) of A, and the dimensions of the blocks must be consistent: the number of columns in A_{ik} must equal the number of columns in A_{jk} for all i, j, k, and similarly for rows.

Operations with Block Matrices

Block Matrix Addition If two matrices A and B are partitioned into blocks with the **same dimensions** for corresponding blocks, they can be added block-wise:

$$A + B = \begin{bmatrix} A_{11} + B_{11} & A_{12} + B_{12} & \cdots \\ A_{21} + B_{21} & A_{22} + B_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Each block A_{ij} and B_{ij} must have the same dimensions for the addition to be valid.

Block Matrix Scalar Multiplication Scalar multiplication is performed by multiplying each block by the scalar:

$$cA = \begin{bmatrix} cA_{11} & cA_{12} & \cdots \\ cA_{21} & cA_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Block Matrix Multiplication If A is an $m \times n$ block matrix and B is an $n \times p$ block matrix, and the partitions are such that the number of column blocks of A equals the number of row blocks of B, then the product $C = AB$ can be computed block-wise:

$$C_{ij} = \sum_{k=1}^{q} A_{ik} B_{kj}$$

This requires that the number of columns in A_{ik} equals the number of rows in B_{kj} for each k. The resulting block C_{ij} is the sum of products of corresponding blocks.

Example: If $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ and $B = \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix}$, then:

$$AB = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} \\ A_{21}B_{11} + A_{22}B_{21} \end{bmatrix}$$

Block Matrix Transpose The transpose of a block matrix is obtained by transposing each block and then transposing the block structure:

$$A^\top = \begin{bmatrix} A_{11}^\top & A_{21}^\top & \cdots \\ A_{12}^\top & A_{22}^\top & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Note that the positions of the blocks are also transposed (e.g., the block in the (1,2) position becomes the block in the (2,1) position after transposition).

Block Diagonal Matrices A block diagonal matrix is a square block matrix where all off-diagonal blocks are zero matrices:

$$A = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{pp} \end{bmatrix}$$

Operations on block diagonal matrices simplify because the blocks can be handled independently (e.g., the inverse of a block diagonal matrix is the block diagonal matrix of the inverses, if they exist).

Advantages of Using Block Matrices

- Simplifies operations on large matrices by breaking them into smaller parts.
- Facilitates parallel computation.
- Helps in proving theoretical results by induction on block structures.
- Commonly used in numerical linear algebra for efficient algorithms.

Definition 3.1.7

An $n \times n$ square matrix A is **invertible** (or **non-singular**) if there exists an $n \times n$ matrix B such that

$$AB = I_n \quad \text{and} \quad BA = I_n$$

This matrix B is unique and is called the **inverse** of A, denoted A^{-1}. If no such matrix B exists, A is **singular** (or **non-invertible**).

Example 3.1.5: Inverse of a 2x2 Matrix

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. If $ad - bc \neq 0$, then A is invertible and

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

If $ad - bc = 0$, A is singular. The quantity $ad - bc$ is the **determinant** of A.

Property 3.1.5: Properties of Inverses

Let A and B be invertible $n \times n$ matrices.

1. $(A^{-1})^{-1} = A$
2. $(AB)^{-1} = B^{-1}A^{-1}$ (Note the reversal of order)
3. $(A^T)^{-1} = (A^{-1})^T$
4. $(cA)^{-1} = \frac{1}{c}A^{-1}$ (for $c \neq 0$)

Proof of $(AB)^{-1} = B^{-1}A^{-1}$. We just need to check the definition.

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = A(I)A^{-1} = AA^{-1} = I$$

$$(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}(I)B = B^{-1}B = I$$

Since $B^{-1}A^{-1}$ satisfies the definition, it must be the inverse of AB. □

3.1.3 Solving Systems of Linear Equations

Now we return to our main problem: solving m equations in n variables. The general idea, **Gaussian Elimination**, is to transform the augmented matrix into a simpler form from which we can just read off the solution.

Elementary Row Operations (EROs)

We can manipulate the augmented matrix using operations that correspond to manipulating the original equations. These operations **do not change the solution set**.

Definition 3.1.8

The three **elementary row operations (EROs)** are:

1. **(Replacement)** Add to one row a multiple of another row. ($R_i \to R_i + cR_j$)
2. **(Interchange)** Interchange two rows. ($R_i \leftrightarrow R_j$)
3. **(Scaling)** Multiply all entries in a row by a non-zero constant. ($R_i \to cR_i, c \neq 0$)

Definition 3.1.9

Two matrices A and B are **row equivalent**, denoted $A \sim B$, if B can be obtained from A by a sequence of EROs.

Theorem 3.1.1

If the augmented matrices of two linear systems are row equivalent, then the two systems have the **same solution set**.

Justification.

- (Replacement) $R_i \to R_i + cR_j$ corresponds to adding c times equation j to equation i. This is a reversible step (by $R_i \to R_i - cR_j$), and any solution to the original system will also be a solution to the new one, and vice-versa.
- (Interchange) $R_i \leftrightarrow R_j$ corresponds to swapping the order of two equations, which clearly does not affect the solution set.
- (Scaling) $R_i \to cR_i$ (with $c \neq 0$) corresponds to multiplying an equation by c. This is reversible (by $R_i \to \frac{1}{c}R_i$), so it does not change the solution set.

$\square$

Row Echelon Form and Rank

The goal is to use EROs to simplify the matrix into a ”staircase” form.

Definition 3.1.10

A matrix is in **Row Echelon Form (REF)** if it satisfies:

1. All nonzero rows are above any rows of all zeros.
2. Each **leading entry** (or **pivot**), which is the leftmost nonzero entry of a row, is in a column to the right of the leading entry of the row above it.
3. All entries in a column *below* a leading entry are zeros.

Definition 3.1.11

A matrix is in **Reduced Row Echelon Form (RREF)** if it is in REF and also satisfies:
1. The leading entry in each nonzero row is 1.
2. Each leading 1 is the *only* nonzero entry in its column.

Example 3.1.6

REF:

$$\begin{pmatrix} \boxed{2} & 3 & 4 & 5 \\ 0 & \boxed{1} & 6 & 7 \\ 0 & 0 & 0 & \boxed{8} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

RREF:

$$\begin{pmatrix} \boxed{1} & 0 & -1 & 0 \\ 0 & \boxed{1} & 2 & 0 \\ 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(Pivots are boxed.)

Theorem 3.1.2

Every matrix is row equivalent to a **unique** Reduced Row Echelon Form (RREF).

This algorithm to get to REF/RREF is the core of our solution method.

Definition 3.1.12

- **Gaussian Elimination** is the process of using EROs to transform a matrix into REF.
- **Gauss-Jordan Elimination** is the process of using EROs to transform a matrix into RREF.

The Algorithm (Gauss-Jordan Elimination) Let's solve a system completely.

$$\begin{cases} x_2 + 3x_3 = 4 \\ x_1 + x_2 + x_3 = 1 \\ 2x_1 + 3x_2 + 4x_3 = 7 \end{cases}$$

The augmented matrix is:

$$\left(\begin{array}{ccc|c} 0 & 1 & 3 & 4 \\ 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 7 \end{array}\right)$$

Step 1: (Forward Phase - Get to REF) We need a pivot in the top-left (1,1) position. Swap with R2.

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 4 \\ 2 & 3 & 4 & 7 \end{array}\right) \quad (R_1 \leftrightarrow R_2)$$

Create zeros below the first pivot.

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 4 \\ 0 & 1 & 2 & 5 \end{array}\right) \quad (R_3 \to R_3 - 2R_1)$$

Now, move to the second pivot (2,2). It's already 1. Create a zero below it.

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 4 \\ 0 & 0 & -1 & 1 \end{array}\right) \quad (R_3 \to R_3 - R_2)$$

The matrix is now in **Row Echelon Form**. This completes Gaussian Elimination. We could stop here and use **back substitution**: From R_3: $-x_3 = 1 \implies x_3 = -1$. From R_2: $x_2 + 3x_3 = 4 \implies x_2 + 3(-1) = 4 \implies x_2 = 7$. From R_1: $x_1 + x_2 + x_3 = 1 \implies x_1 + 7 + (-1) = 1 \implies x_1 = -5$. The unique solution is $(-5, 7, -1)$.

Step 2: (Backward Phase - Get to RREF) Continue from the REF. Scale all pivots to 1.

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 4 \\ 0 & 0 & 1 & -1 \end{array}\right) \quad (R_3 \to -1 \cdot R_3)$$

Create zeros *above* the pivots, starting from the rightmost pivot.

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 1 & 0 & 7 \\ 0 & 0 & 1 & -1 \end{array}\right) \quad (R_1 \to R_1 - R_3, R_2 \to R_2 - 3R_3)$$

Create zero above the second pivot.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & -5 \\ 0 & 1 & 0 & 7 \\ 0 & 0 & 1 & -1 \end{array}\right) \quad (R_1 \to R_1 - R_2)$$

This is the **Reduced Row Echelon Form**. The corresponding system is:

$$\begin{cases} x_1 = -5 \\ x_2 = 7 \\ x_3 = -1 \end{cases}$$

This immediately gives the solution.

Rank of a Matrix A key concept emerges from the echelon form.

Definition 3.1.13

The **rank** of a matrix A, denoted $\text{rank}(A)$, is the number of leading entries (pivots) in its row echelon form. This number is unique for any given matrix.

Property 3.1.6: Properties of Rank

Let A be an $m \times n$ matrix.

1. $\text{rank}(A) \le \min(m, n)$.
2. $\text{rank}(A) = 0$ if and only if $A = \mathbf{0}$.
3. **(Major Theorem)** $\text{rank}(A) = \text{rank}(A^T)$. (The number of pivot rows equals the number of pivot columns).
4. $\text{rank}(AB) \le \min(\text{rank}(A), \text{rank}(B))$.
5. $\text{rank}(A + B) \le \text{rank}(A) + \text{rank}(B)$.
6. $\text{rank}(A) + \text{rank}(B) - n \le \text{rank}(AB)$
7. If P, Q are invertible, $\text{rank}(PAQ) = \text{rank}(A)$. EROs are equivalent to multiplying by an invertible matrix on the left, so row operations do not change the rank.

In short, we can denote them as:

$$\min\{r(A), r(B)\} \ge r(AB) \ge r(A) + r(B) - n$$

$$r(A + B) \le r(A, B) \le r\begin{pmatrix} A & O \\ O & B \end{pmatrix}$$

Solution Sets of Linear Systems

The RREF of the *augmented* matrix tells us everything about the solution set. A **pivot column** is a column in the coefficient matrix A that contains a pivot in its RREF. Variables corresponding to pivot columns are called **basic variables**. Variables corresponding to non-pivot columns are called **free variables**.

Let $r = \text{rank}(A)$ for an $m \times n$ coefficient matrix A. We analyze the RREF of the augmented matrix $[A \mid \mathbf{b}]$.

1. **No Solution (Inconsistent System)** This occurs if the RREF of $[A \mid \mathbf{b}]$ has a row of the form $\begin{pmatrix} 0 & 0 & \cdots & 0 & \mid & 1 \end{pmatrix}$. This corresponds to the impossible equation $0x_1 + \cdots + 0x_n = 1$, or $0 = 1$. In terms of rank, this means the last column (the augmented column) is a pivot column. **Condition:** $\text{rank}(A) < \text{rank}([A \mid \mathbf{b}])$**.**

2. **A Solution Exists (Consistent System)** This occurs if the augmented column is *not* a pivot column. **Condition:** $\text{rank}(A) = \text{rank}([A \mid \mathbf{b}])$**.** Let this rank be r.

   - **Unique Solution:** The system has a unique solution if there are *no free variables*. This means every variable is a basic variable, so every column of A is a pivot column. **Condition:** $r = n$ **(the number of variables).**

   - **Infinitely Many Solutions:** The system has infinitely many solutions if there is *at least one free variable*. This means some columns of A are not pivot columns. **Condition:** $r < n$ **(the number of variables).** The $n - r$ free variables can be set to any arbitrary value (parameters), and the basic variables can be expressed in terms of them.

Parametric Vector Form When we have infinitely many solutions, we write the solution set in **parametric vector form**.

Example 3.1.7: Infinitely Many Solutions

Find the general solution to the system with augmented matrix:

$$\left(\begin{array}{ccc|c} 1 & 0 & -5 & 1 \\ 0 & 1 & 1 & 4 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

This matrix is already in RREF. The corresponding system is:

$$\begin{cases} x_1 - 5x_3 = 1 \\ x_2 + x_3 = 4 \\ 0 = 0 \end{cases}$$

The pivot columns are 1 and 2. So, x_1 and x_2 are **basic variables**. Column 3 is not a pivot column. So, x_3 is a **free variable**. We introduce a parameter, t, for the free variable. Let $x_3 = t$, where t can be any real number. Now, we express the basic variables in terms of the free variables:

$$x_1 = 1 + 5x_3 = 1 + 5t$$

$$x_2 = 4 - x_3 = 4 - t$$

$$x_3 = t$$

The general solution $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ is:

$$\mathbf{x} = \begin{pmatrix} 1+5t \\ 4-t \\ t \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \\ 0 \end{pmatrix} + t \begin{pmatrix} 5 \\ -1 \\ 1 \end{pmatrix}$$

This is the **parametric vector form**. Geometrically, this is the equation of a **line** in $\mathbb{R}^3$ passing through the point $(1, 4, 0)$ and parallel to the vector $(5, -1, 1)$.

Homogeneous and Non-homogeneous Systems

Definition 3.1.14

A system of linear equations is called **homogeneous** if it is of the form $A\mathbf{x} = \mathbf{0}$, where $\mathbf{0}$ is the zero vector (all $b_i = 0$).

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = 0 \\ \quad \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = 0 \end{cases}$$

A system $A\mathbf{x} = \mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$ is called **non-homogeneous**.

A homogeneous system $A\mathbf{x} = \mathbf{0}$ is **always** consistent, because $\mathbf{x} = \mathbf{0}$ (the zero vector) is always a solution, known as the **trivial solution**. The important question is whether a **non-trivial solution** exists. This happens if and only if there is at least one free variable, which is equivalent to $\text{rank}(A) < n$ (the number of variables).

Theorem 3.1.3

The homogeneous system $A\mathbf{x} = \mathbf{0}$ has a non-trivial solution if and only if $\text{rank}(A) < n$.

Corollary 3.1.1

If A is $m \times n$ with $m < n$ (fewer equations than variables, a "wide" matrix), then $A\mathbf{x} = \mathbf{0}$ **always** has infinitely many solutions, because $\text{rank}(A) \leq m < n$.

There is a fundamental connection between the solution sets of the two systems.

Theorem 3.1.4: Structure of Solutions

Suppose the non-homogeneous system $A\mathbf{x} = \mathbf{b}$ is consistent and has a particular solution $\mathbf{x}_p$. Then the general solution $\mathbf{x}_g$ of $A\mathbf{x} = \mathbf{b}$ is the set of all vectors of the form

$$\mathbf{x}_g = \mathbf{x}_p + \mathbf{x}_h$$

where $\mathbf{x}_h$ is any solution to the corresponding homogeneous system $A\mathbf{x} = \mathbf{0}$.

Proof. Let $\mathbf{x}_g$ be any solution to $A\mathbf{x} = \mathbf{b}$. Let $\mathbf{x}_h = \mathbf{x}_g - \mathbf{x}_p$. Then $A\mathbf{x}_h = A(\mathbf{x}_g - \mathbf{x}_p) = A\mathbf{x}_g - A\mathbf{x}_p = \mathbf{b} - \mathbf{b} = \mathbf{0}$. So, $\mathbf{x}_h$ is a solution to the homogeneous system. This shows any solution $\mathbf{x}_g$ can be written in the form $\mathbf{x}_p + \mathbf{x}_h$. Conversely, let $\mathbf{x}_h$ be any homogeneous solution. Then $A(\mathbf{x}_p + \mathbf{x}_h) = A\mathbf{x}_p + A\mathbf{x}_h = \mathbf{b} + \mathbf{0} = \mathbf{b}$. So, $\mathbf{x}_p + \mathbf{x}_h$ is a solution to the non-homogeneous system. $\square$

Remark 3.1.3

Look back at our last example:

$$\mathbf{x} = \underbrace{\begin{pmatrix} 1 \\ 4 \\ 0 \end{pmatrix}}_{\mathbf{x}_p} + \underbrace{t \begin{pmatrix} 5 \\ -1 \\ 1 \end{pmatrix}}_{\mathbf{x}_h}$$

Here $\mathbf{x}_p = (1, 4, 0)$ is one *particular solution* to $A\mathbf{x} = \mathbf{b}$. $\mathbf{x}_h = t(5, -1, 1)$ is the *general solution* to the corresponding homogeneous system $A\mathbf{x} = \mathbf{0}$. Geometrically, the solution set to $A\mathbf{x} = \mathbf{b}$ is a *translation* (by $\mathbf{x}_p$) of the solution set to $A\mathbf{x} = \mathbf{0}$.

3.2 Determinants

We now study a powerful tool associated with **square** matrices: the determinant. The determinant is a single number that reveals a wealth of information about a matrix, most notably whether it is invertible.

The calculation of determinants require familiarity and patience, and once we can find other ways to avoid using determinants, we shall do so.

3.2.1 The Determinant of a Matrix

For a 1×1 matrix $A = (a)$, $\det(A) = a$. For a 2×2 matrix, the determinant is simple:

$$\det(A) = \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

This value $ad - bc$ is non-zero if and only if the matrix is invertible.

For larger $n \times n$ matrices, we define the determinant recursively using **cofactor expansion**.

Definition 3.2.1

Let A be an $n \times n$ matrix.

- The **minor** M_{ij} of the entry a_{ij} is the determinant of the $(n-1) \times (n-1)$ matrix obtained by deleting the i-th row and j-th column of A.
- The **cofactor** C_{ij} is given by $C_{ij} = (-1)^{i+j} M_{ij}$.

The "checkerboard" pattern of signs for $(-1)^{i+j}$ is $\begin{pmatrix} + & - & + & \cdots \\ - & + & - & \cdots \\ + & - & + & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$.

Theorem 3.2.1: Cofactor Expansion

The determinant of an $n \times n$ matrix A can be found by expanding along **any** row i:

$$\det(A) = a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in} = \sum_{j=1}^{n} a_{ij}C_{ij}$$

Alternatively, we can expand down **any** column j:

$$\det(A) = a_{1j}C_{1j} + a_{2j}C_{2j} + \cdots + a_{nj}C_{nj} = \sum_{i=1}^{n} a_{ij}C_{ij}$$

Example 3.2.1: Cofactor Expansion of 3x3

Let $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$. Let's expand along Row 1.

$$\det(A) = 1 \cdot C_{11} + 2 \cdot C_{12} + 3 \cdot C_{13}$$

$$C_{11} = (-1)^{1+1} \begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} = +1(5 \cdot 9 - 6 \cdot 8) = 45 - 48 = -3$$

$$C_{12} = (-1)^{1+2} \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} = -1(4 \cdot 9 - 6 \cdot 7) = -(36 - 42) = 6$$

$$C_{13} = (-1)^{1+3} \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix} = +1(4 \cdot 8 - 5 \cdot 7) = 32 - 35 = -3$$

$$\det(A) = 1(-3) + 2(6) + 3(-3) = -3 + 12 - 9 = 0$$

Since $\det(A) = 0$, this matrix is **singular** (not invertible).

A more formal definition of determinants relies on the concept of negative sequence. And it is logically equivalent to the definition above so we won't present it again. But here is a more axiomatized definitions about determinant I would like to share with you:

Definition 3.2.2: The axiomatic definition of determinants

The determinant is the unique function det: $M_n(\mathbb{R}) \leftarrow \mathbb{R}$ satisfying the following three axioms:

1. Multilinearity: It is a linear function of each column when the other columns are held fixed.
2. Alternating: If two columns of the matrix are identical, then its determinant is zero. This also implies that swapping two columns changes the sign of the determinant.
3. Normalization: The determinant of the identity matrix is 1.

Another definition is more **modern**, it's about exterior product

Definition 3.2.3: Exterior Product (Wedge Product)

Let V be a vector space over field $\mathbb{K}$. The **exterior product** (or **wedge product**) is a bilinear map:

$$\wedge : V \times V \to \Lambda^2(V)$$

satisfying:

1. **Anticommutativity**: $u \wedge v = -v \wedge u$ for all $u, v \in V$
2. **Nilpotence**: $v \wedge v = 0$ for all $v \in V$

The k-th exterior power $\Lambda^k(V)$ is spanned by elements of the form $v_1 \wedge v_2 \wedge \cdots \wedge v_k$ where $v_i \in V$.

Definition 3.2.4: Determinant via Exterior Algebra

Let V be an n-dimensional vector space with basis $\{e_1, \ldots, e_n\}$. A linear operator $T : V \to V$ induces $\Lambda^n T : \Lambda^n(V) \to \Lambda^n(V)$ on the top exterior power:

$$\Lambda^n T(e_1 \wedge \cdots \wedge e_n) = T(e_1) \wedge \cdots \wedge T(e_n)$$

Since $\Lambda^n(V)$ is 1-dimensional, there exists a unique scalar $\det(T) \in \mathbb{K}$ such that:

$$T(e_1) \wedge \cdots \wedge T(e_n) = \det(T) \cdot (e_1 \wedge \cdots \wedge e_n)$$

This scalar $\det(T)$ is called the **determinant** of T.
For matrix $A = (a_{ij})$ with column vectors $a_1, \ldots, a_n \in \mathbb{R}^n$:

$$a_1 \wedge a_2 \wedge \cdots \wedge a_n = \det(A) \cdot (e_1 \wedge e_2 \wedge \cdots \wedge e_n)$$

Theorem 3.2.2

The exterior algebra definition implies the axiomatic definition of determinant.

Proof. We verify the three axioms:

1. **Multilinearity**: The wedge product is linear in each argument:

$$(\lambda u + \mu v) \wedge w = \lambda(u \wedge w) + \mu(v \wedge w)$$

Thus $(a_1, \ldots, a_n) \mapsto a_1 \wedge \cdots \wedge a_n$ is multilinear, and so is $\det(A)$.

2. **Alternating property**: If $a_i = a_j$ $(i \neq j)$, then:

$$a_1 \wedge \cdots \wedge a_i \wedge \cdots \wedge a_j \wedge \cdots \wedge a_n = 0$$

since $v \wedge v = 0$. Hence $\det(A) = 0$. Swapping columns introduces a sign change due to anticommutativity.

3. **Normalization**: For identity matrix I:

$$e_1 \wedge \cdots \wedge e_n = \det(I) \cdot (e_1 \wedge \cdots \wedge e_n) \Rightarrow \det(I) = 1$$

□

Corollary 3.2.1

The multiplicative property $\det(AB) = \det(A)\det(B)$ follows naturally.

Proof. Consider the composition on $\Lambda^n(V)$:

$$\Lambda^n(AB)(e_1 \wedge \cdots \wedge e_n) = \Lambda^n A(\Lambda^n B(e_1 \wedge \cdots \wedge e_n)) = \det(A)\det(B)(e_1 \wedge \cdots \wedge e_n)$$

But also equals $\det(AB)(e_1 \wedge \cdots \wedge e_n)$, so $\det(AB) = \det(A)\det(B)$. □

Remark 3.2.1: Sarrus's Rule for 3x3

For 3×3 matrices *only*, there is a shortcut. Write the first two columns again to the right:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{matrix} a & b \\ d & e \\ g & h \end{matrix}$$

Sum the products of the down-right diagonals and subtract the products of the up-right diagonals:

$$\det = (aei + bfg + cdh) - (gec + hfa + idb)$$

Using our example: $(1\cdot5\cdot9+2\cdot6\cdot7+3\cdot4\cdot8)-(7\cdot5\cdot3+8\cdot6\cdot1+9\cdot4\cdot2) = (45+84+96)-(105+48+72) = 225 - 225 = 0$. **Warning:** This *does not* work for 4x4 or larger.

3.2.2 Properties of Determinants

Calculating determinants via cofactors is computationally slow ($O(n!)$). A more efficient method ($O(n^3)$) uses row operations.

Theorem 3.2.3: Determinants and EROs

Let A be an $n \times n$ matrix.

1. **(Replacement)** If B is obtained from A by $R_i \to R_i + cR_j$, then $\det(B) = \det(A)$.
2. **(Interchange)** If B is obtained from A by $R_i \leftrightarrow R_j$, then $\det(B) = -\det(A)$.
3. **(Scaling)** If B is obtained from A by $R_i \to cR_i$, then $\det(B) = c \cdot \det(A)$.

This allows us to row-reduce A to an echelon form U (which is triangular) while keeping track of the changes.

Theorem 3.2.4

If A is a triangular matrix (upper or lower), its determinant is the product of its diagonal entries.

$$\det(A) = a_{11}a_{22}\cdots a_{nn}$$

Proof. Expand cofactors along the first row (if lower triangular) or first column (if upper triangular) repeatedly. □

Example 3.2.2: Calculating $\det$ with EROs

$$A = \begin{pmatrix} 0 & 1 & 5 \\ 3 & -6 & 9 \\ 2 & 6 & 1 \end{pmatrix}$$

$$\det(A) = -\begin{vmatrix} 3 & -6 & 9 \\ 0 & 1 & 5 \\ 2 & 6 & 1 \end{vmatrix} \quad (R_1 \leftrightarrow R_2)$$

$$\det(A) = -3\begin{vmatrix} 1 & -2 & 3 \\ 0 & 1 & 5 \\ 2 & 6 & 1 \end{vmatrix} \quad (R_1 \to \frac{1}{3}R_1, \text{pull out } 3)$$

$$\det(A) = -3\begin{vmatrix} 1 & -2 & 3 \\ 0 & 1 & 5 \\ 0 & 10 & -5 \end{vmatrix} \quad (R_3 \to R_3 - 2R_1)$$

$$\det(A) = -3\begin{vmatrix} 1 & -2 & 3 \\ 0 & 1 & 5 \\ 0 & 0 & -55 \end{vmatrix} \quad (R_3 \to R_3 - 10R_2)$$

The matrix is now triangular.

$$\det(A) = -3 \cdot (1 \cdot 1 \cdot -55) = 165$$

Property 3.2.1: More Properties of Determinants

Let A, B be $n \times n$ matrices.

1. **(Major Theorem)** A is invertible if and only if $\det(A) \neq 0$.
2. **(Multiplicative Property)** $\det(AB) = \det(A)\det(B)$.
3. $\det(A^T) = \det(A)$. (This implies all ERO properties also work for *columns*).
4. If A is invertible, $\det(A^{-1}) = \frac{1}{\det(A)}$.
5. $\det(cA) = c^n \det(A)$ (where A is $n \times n$).
6. If A has a zero row (or column), $\det(A) = 0$.
7. If A has two identical rows (or columns), $\det(A) = 0$.

Proof of $\det(A^{-1}) = 1/\det(A)$. $AA^{-1} = I$. $\det(AA^{-1}) = \det(I)$. $\det(A)\det(A^{-1}) = 1$. $\det(A^{-1}) = \frac{1}{\det(A)}$. (This requires $\det(A) \neq 0$, which is true since A is invertible). □

3.2.3 Cramer's Rule and Adjoint Formula

Determinants provide explicit formulas for solving $A\mathbf{x} = \mathbf{b}$ and finding A^{-1}. While elegant, they are computationally *inefficient* for large matrices compared to elimination.

Theorem 3.2.5: Cramer's Rule

Let A be an invertible $n \times n$ matrix. For any $\mathbf{b}$ in $\mathbb{R}^n$, the unique solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$ has entries given by

$$x_i = \frac{\det(A_i(\mathbf{b}))}{\det(A)}, \quad \text{for } i = 1, 2, \ldots, n$$

where $A_i(\mathbf{b})$ is the matrix obtained from A by replacing its i-th column with the vector $\mathbf{b}$.

Example 3.2.3

Solve $\begin{cases} 2x_1 + 5x_2 = -1 \\ 3x_1 + 7x_2 = 4 \end{cases}$ $A = \begin{pmatrix} 2 & 5 \\ 3 & 7 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} -1 \\ 4 \end{pmatrix}$ $\det(A) = 2(7) - 5(3) = 14 - 15 = -1$. $A_1(\mathbf{b}) = \begin{pmatrix} -1 & 5 \\ 4 & 7 \end{pmatrix}, \det(A_1(\mathbf{b})) = -7 - 20 = -27$. $A_2(\mathbf{b}) = \begin{pmatrix} 2 & -1 \\ 3 & 4 \end{pmatrix}, \det(A_2(\mathbf{b})) = 8 - (-3) = 11$. $x_1 = \frac{-27}{-1} = 27$. $x_2 = \frac{11}{-1} = -11$.

Definition 3.2.5: Adjoint Matrix

Let $C = [C_{ij}]$ be the matrix of cofactors of A. The **adjoint** (or **adjugate**) of A, denoted $\text{adj}(A)$, is the **transpose** of the cofactor matrix.

$$\text{adj}(A) = C^T$$

Theorem 3.2.6: Inverse Formula

Let A be an invertible matrix. Then

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$$

Remark 3.2.2

This theorem explains the 2×2 inverse formula. For $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$: $C_{11} = d, C_{12} = -c, C_{21} = -b, C_{22} = a$. Cofactor Matrix $C = \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$. Adjoint Matrix $\text{adj}(A) = C^T = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. $A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

3.2.4 The Invertible Matrix Theorem (IMT)

This is one of the most important theorems in linear algebra. It links all the major concepts we have seen so far for a **square** $n \times n$ matrix A.

Theorem 3.2.7: The Invertible Matrix Theorem

Let A be a square $n \times n$ matrix. The following statements are equivalent (that is, if any one is true, they are all true, and if any one is false, they are all false).

1. A is an invertible matrix.
2. A is row equivalent to the identity matrix I_n.
3. A has n pivot positions.
4. The equation $A\mathbf{x} = \mathbf{0}$ has only the trivial solution ($\mathbf{x} = \mathbf{0}$).
5. The columns of A form a linearly independent set.
6. The linear transformation $T(\mathbf{x}) = A\mathbf{x}$ is one-to-one.
7. The equation $A\mathbf{x} = \mathbf{b}$ has at least one solution for *each* $\mathbf{b}$ in $\mathbb{R}^n$.
8. The columns of A span $\mathbb{R}^n$.
9. The linear transformation $T(\mathbf{x}) = A\mathbf{x}$ maps $\mathbb{R}^n$ *onto* $\mathbb{R}^n$.
10. There is an $n \times n$ matrix C such that $CA = I_n$.
11. There is an $n \times n$ matrix D such that $AD = I_n$.
12. A^T is an invertible matrix.
13. $\det(A) \neq 0$.
14. $\text{rank}(A) = n$.
15. $\text{Nul}(A) = \{\mathbf{0}\}$ (The null space is the zero vector).
16. $\text{Col}(A) = \mathbb{R}^n$ (The column space is all of $\mathbb{R}^n$).
17. 0 is not an eigenvalue of A. (We will see this later).

This theorem is a powerful diagnostic tool. To check if a square matrix is invertible, we only need to verify *one* of these conditions. For example, checking if $\det(A) \neq 0$ is often the fastest way.

3.3 Vectors in $\mathbb{R}^n$

We now introduce a new and fundamental object: the vector. This allows us to re-interpret systems of equations in a powerful, geometric way.

3.3.1 Vectors and Operations

Geometrically, in two ($\mathbb{R}^2$) or three ($\mathbb{R}^3$) dimensions, we can think of a vector as an arrow with a specific length and direction.

Algebraically, we define a **vector** in $\mathbb{R}^n$ (read: "R-n") as an ordered n-tuple of real numbers. We typically write it as a **column vector**:

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

The set $\mathbb{R}^n$ is the collection of all such n-dimensional vectors. $\mathbb{R}^2$ is the set of all vectors $\begin{pmatrix} x \\ y \end{pmatrix}$, which we identify with the 2D Cartesian plane. A vector $\mathbf{v} = (v_1, \ldots, v_n)$ can also be written as a **row vector**, but column vectors are standard when working with matrix equations.

We define two fundamental operations on vectors. Let $\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ be vectors in $\mathbb{R}^n$ and let

c be a real number (a **scalar**).

1. **Vector Addition:** $\mathbf{u}+\mathbf{v}$ is found by adding corresponding components:

$$\mathbf{u}+\mathbf{v} = \begin{pmatrix} u_1+v_1 \\ \vdots \\ u_n+v_n \end{pmatrix}$$

   Geometrically, this corresponds to the **Parallelogram Law**.

2. **Scalar Multiplication:** $c\mathbf{v}$ is found by multiplying each component by c:

$$c\mathbf{v} = \begin{pmatrix} cv_1 \\ \vdots \\ cv_n \end{pmatrix}$$

   Geometrically, this scales the length of the vector by $|c|$ and reverses its direction if $c < 0$.

These operations satisfy the 8 properties (associativity, commutativity, etc.) listed in Section 2.2.1, making $\mathbb{R}^n$ a prime example of a vector space.

3.3.2 Dot Product, Norm, and Orthogonality

Beyond addition and scaling, we can define a product that gives a scalar in $\mathbb{R}^n$.

Definition 3.3.1: Dot Product

The **dot product** (or **inner product**) of $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$ is:

$$\mathbf{u}\cdot\mathbf{v} = u_1v_1 + u_2v_2 + \cdots + u_nv_n = \sum_{i=1}^{n} u_iv_i$$

Note: $\mathbf{u}\cdot\mathbf{v}$ is a **scalar**, not a vector. We can also write this using matrix multiplication: $\mathbf{u}\cdot\mathbf{v} = \mathbf{u}^T\mathbf{v}$.

Property 3.3.1: Properties of the Dot Product

1. $\mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u}$ (Commutative)
2. $(\mathbf{u}+\mathbf{v})\cdot\mathbf{w} = \mathbf{u}\cdot\mathbf{w} + \mathbf{v}\cdot\mathbf{w}$ (Distributive)
3. $(c\mathbf{u})\cdot\mathbf{v} = c(\mathbf{u}\cdot\mathbf{v})$
4. $\mathbf{u}\cdot\mathbf{u} \geq 0$, and $\mathbf{u}\cdot\mathbf{u} = 0 \iff \mathbf{u} = \mathbf{0}$.

Definition 3.3.2: Norm and Distance

1. The **norm** (or **length**) of a vector $\mathbf{v}$ is:

$$\|\mathbf{v}\| = \sqrt{\mathbf{v}\cdot\mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

2. A vector $\mathbf{u}$ with $\|\mathbf{u}\| = 1$ is called a **unit vector**.
3. **Normalizing** a vector $\mathbf{v} \neq \mathbf{0}$ means finding the unit vector in its direction: $\mathbf{u} = \frac{1}{\|\mathbf{v}\|}\mathbf{v}$.
4. The **distance** between $\mathbf{u}$ and $\mathbf{v}$ is $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|$.

However, when dealing with abstract vector spaces, we may not have a natural dot product. In such cases, we can define an **inner product** that satisfies the same properties as the dot product. We shall see how to do this in later sections.

Definition 3.3.3: Orthogonality

Two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ are **orthogonal** (perpendicular) if their dot product is zero:

$$\mathbf{u} \perp \mathbf{v} \iff \mathbf{u} \cdot \mathbf{v} = 0$$

The zero vector $\mathbf{0}$ is orthogonal to every vector in $\mathbb{R}^n$.

Likewise, we can define orthogonality in inner product spaces and weighted dot product spaces by replacing the dot product with the inner product or weighted dot product.

Theorem 3.3.1: Pythagorean Theorem

$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$ if and only if $\mathbf{u} \cdot \mathbf{v} = 0$.

Proof. $\|\mathbf{u} + \mathbf{v}\|^2 = (\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} + \mathbf{v}) = \mathbf{u} \cdot \mathbf{u} + \mathbf{u} \cdot \mathbf{v} + \mathbf{v} \cdot \mathbf{u} + \mathbf{v} \cdot \mathbf{v}$ $\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + 2(\mathbf{u} \cdot \mathbf{v}) + \|\mathbf{v}\|^2$. The equality holds iff $2(\mathbf{u} \cdot \mathbf{v}) = 0$, which means $\mathbf{u} \cdot \mathbf{v} = 0$. □

The dot product also defines the angle between two vectors.

Theorem 3.3.2

For $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$,

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta$$

where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$.

This leads to a famous inequality:

Theorem 3.3.3: Cauchy-Schwarz Inequality

For all $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$,

$$|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\|\|\mathbf{v}\|$$

Theorem 3.3.4: Triangle Inequality

For all $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$,

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$$

3.3.3 Linear Combinations and Span

This is one of the most important ideas in the entire subject.

Definition 3.3.4

A **linear combination** of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ in $\mathbb{R}^n$ is any vector $\mathbf{y}$ of the form:

$$\mathbf{y} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_p\mathbf{v}_p$$

where $c_1, \ldots, c_p$ are any scalars (also called weights).

Example 3.3.1

In $\mathbb{R}^3$, let $\mathbf{v}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}$. $\mathbf{y} = 3\mathbf{v}_1 + (-2)\mathbf{v}_2 = 3\begin{pmatrix}1\\0\\0\end{pmatrix} - 2\begin{pmatrix}0\\1\\0\end{pmatrix} = \begin{pmatrix}3\\-2\\0\end{pmatrix}$ is a linear combination of $\mathbf{v}_1, \mathbf{v}_2$. But $\begin{pmatrix}1\\1\\1\end{pmatrix}$ is *not* a linear combination of $\mathbf{v}_1, \mathbf{v}_2$.

Definition 3.3.5

The set of **all possible** linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_p$ is called the **Span** of these vectors, denoted $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$.

Geometrically, the span has a simple interpretation:

- $\text{Span}\{\mathbf{v}\}$ (for $\mathbf{v} \neq \mathbf{0}$) is the line through the origin and $\mathbf{v}$.
- $\text{Span}\{\mathbf{u}, \mathbf{v}\}$ (for non-collinear $\mathbf{u}, \mathbf{v}$) is the plane containing the origin, $\mathbf{u}$, and $\mathbf{v}$.
- $\text{Span}\{\mathbf{0}\}$ is just the set $\{\mathbf{0}\}$, the origin.

3.3.4 The Matrix Equation $A\mathbf{x} = \mathbf{b}$

We can now connect our topics. Let A be an $m \times n$ matrix. We can view its columns as n vectors in $\mathbb{R}^m$: $A = \begin{pmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n\end{pmatrix}$. Let $\mathbf{x} = \begin{pmatrix}x_1\\\vdots\\x_n\end{pmatrix}$ be a vector in $\mathbb{R}^n$.

Definition 3.3.6: Matrix-Vector Product

The product of the $m \times n$ matrix A and the $n \times 1$ vector $\mathbf{x}$, denoted $A\mathbf{x}$, is defined as the **linear combination of the columns of A using the entries of x as weights**:

$$A\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n$$

This product results in an $m \times 1$ vector (a vector in $\mathbb{R}^m$).

Example 3.3.2

$\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}\begin{pmatrix}5\\-1\end{pmatrix} = 5\begin{pmatrix}1\\3\end{pmatrix} - 1\begin{pmatrix}2\\4\end{pmatrix} = \begin{pmatrix}5\\15\end{pmatrix} - \begin{pmatrix}2\\4\end{pmatrix} = \begin{pmatrix}3\\11\end{pmatrix}$ Note: This matches the row-column rule for matrix multiplication: $\begin{pmatrix}1(5)+2(-1)\\3(5)+4(-1)\end{pmatrix} = \begin{pmatrix}3\\11\end{pmatrix}$.

Now look at our original system of equations:

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \quad\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \end{cases}$$

The left side can be written as a vector equation:

$$x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ \vdots \\ a_{m2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

Using our new definitions, this is precisely:

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$$

Which is identical to the **matrix equation**:

$$A\mathbf{x} = \mathbf{b}$$

This gives us three equivalent ways to view the same problem:

1. A system of m linear equations in n variables.
2. A vector equation $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$.
3. A matrix equation $A\mathbf{x} = \mathbf{b}$.

This is a profound re-interpretation! The question "Does the system $A\mathbf{x} = \mathbf{b}$ have a solution?" is identical to the question:

"Is the vector b a linear combination of the column vectors of A?"

Or, more simply: **"Is b in** $\text{Span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$**?"**

Theorem 3.3.5

The equation $A\mathbf{x} = \mathbf{b}$ has a solution if and only if $\mathbf{b}$ is in the span of the columns of A. This span is called the **Column Space** of A, denoted $\text{Col}(A)$.

3.3.5 Linear Independence

We now ask a related question. What if $\mathbf{b} = \mathbf{0}$? The equation $A\mathbf{x} = \mathbf{0}$ (or $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{0}$) is the **homogeneous equation**. We know this *always* has the **trivial solution** $\mathbf{x} = \mathbf{0}$ (i.e., $x_1 = 0, \ldots, x_n = 0$). But does it have *only* the trivial solution?

Definition 3.3.7

A set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ in $\mathbb{R}^n$ is said to be **linearly independent** if the vector equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_p\mathbf{v}_p = \mathbf{0}$$

has **only** the trivial solution ($c_1 = c_2 = \cdots = c_p = 0$).
The set is **linearly dependent** if there exist weights c_i, *not all zero*, such that the equation holds. This is called a **linear dependence relation**.

Example 3.3.3

Check if $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 3 \end{pmatrix} \right\}$ is linearly independent. We must solve $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$. This is the matrix equation $A\mathbf{c} = \mathbf{0}$ where $A = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{pmatrix}$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 2 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 1 & 3 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 0 & 2 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) \quad (R_3 \to R_3 - R_2)$$

c_3 is a free variable! So there are non-trivial solutions. Let $c_3 = t$. Then $c_2 = -3t$ and $c_1 = -2t$. For $t = 1$, we get $c_1 = -2, c_2 = -3, c_3 = 1$. This gives the linear dependence relation:

$$-2\mathbf{v}_1 - 3\mathbf{v}_2 + 1\mathbf{v}_3 = \mathbf{0} \quad \text{or} \quad \mathbf{v}_3 = 2\mathbf{v}_1 + 3\mathbf{v}_2$$

The set is **linearly dependent**.

Property 3.3.2: Secondary Conclusions on Independence

- A set of two vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$ is linearly dependent if and only if one is a scalar multiple of the other.
- A set is linearly dependent if and only if at least one vector in the set is a linear combination of the others.
- Any set containing the zero vector ($\{\mathbf{v}_1, \dots, \mathbf{0}, \dots, \mathbf{v}_p\}$) is linearly dependent.
- **(Key Theorem)** If a set contains *more vectors than entries* in each vector (e.g., p vectors in $\mathbb{R}^n$ where $p > n$), the set is **linearly dependent**.

Proof of $p > n$ implies dependent. Let the set be $\{\mathbf{v}_1, \dots, \mathbf{v}_p\}$ in $\mathbb{R}^n$. Form the $n \times p$ matrix $A = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_p \end{pmatrix}$. We want to solve $A\mathbf{x} = \mathbf{0}$. This is a homogeneous system with n equations and p variables. Since $p > n$ (more variables than equations), there must be at least $p - n > 0$ free variables. The existence of free variables guarantees a non-trivial solution. Therefore, the columns are linearly dependent. $\square$

Connecting this to matrices, we see that:

- The columns of a matrix A are linearly independent if and only if the homogeneous system $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.
- This happens if and only if there are no free variables, i.e., $\operatorname{rank}(A) = n$ (every column is a pivot column).

This forms several more lines of the Invertible Matrix Theorem.

3.4 Linear Transformations

The matrix-vector product $A\mathbf{x}$ can be viewed as an *action* or *function*. The matrix A *transforms* the vector $\mathbf{x}$ into a new vector $A\mathbf{x}$.

3.4.1 Matrix Transformations

A **transformation** (or function, or mapping) T from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a rule that assigns to each vector $\mathbf{x}$ in $\mathbb{R}^n$ a vector $T(\mathbf{x})$ in $\mathbb{R}^m$.

$$T : \mathbb{R}^n \to \mathbb{R}^m$$

- $\mathbb{R}^n$ is the **domain** of T.

- $\mathbb{R}^m$ is the **codomain** of T.
- $T(\mathbf{x})$ is the **image** of $\mathbf{x}$ under T.
- The set of all images $T(\mathbf{x})$ is the **range** of T.

An important class of transformations are matrix transformations. For an $m \times n$ matrix A, the transformation $T(\mathbf{x}) = A\mathbf{x}$ maps $\mathbb{R}^n \to \mathbb{R}^m$.

Example 3.4.1

Let $A = \begin{pmatrix} 1 & -3 \\ 3 & 5 \\ -1 & 7 \end{pmatrix}$. This A defines $T : \mathbb{R}^2 \to \mathbb{R}^3$. Let $\mathbf{x} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$. $T(\mathbf{x}) = A\mathbf{x} = \begin{pmatrix} 1 & -3 \\ 3 & 5 \\ -1 & 7 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \end{pmatrix} = \begin{pmatrix} 1(2) - 3(-1) \\ 3(2) + 5(-1) \\ -1(2) + 7(-1) \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ -9 \end{pmatrix}$. The image of $\begin{pmatrix} 2 \\ -1 \end{pmatrix}$ is $\begin{pmatrix} 5 \\ 1 \\ -9 \end{pmatrix}$.

3.4.2 Linearity

Matrix transformations $T(\mathbf{x}) = A\mathbf{x}$ have special properties that come from the properties of matrix multiplication:

1. $A(\mathbf{u} + \mathbf{v}) = A\mathbf{u} + A\mathbf{v}$
2. $A(c\mathbf{u}) = c(A\mathbf{u})$

Definition 3.4.1

A transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is **linear** if for all $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$ and all scalars c:
1. $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ (Preserves addition)
2. $T(c\mathbf{u}) = cT(\mathbf{u})$ (Preserves scalar multiplication)

These two rules imply $T(\mathbf{0}) = \mathbf{0}$ and the "superposition principle": $T(c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p) = c_1T(\mathbf{v}_1) + \cdots + c_pT(\mathbf{v}_p)$.

Theorem 3.4.1

Every matrix transformation $T(\mathbf{x}) = A\mathbf{x}$ is a linear transformation.

The more powerful fact is that the reverse is also true.

3.4.3 The Standard Matrix

Theorem 3.4.2

Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then there exists a **unique** $m \times n$ matrix A such that

$$T(\mathbf{x}) = A\mathbf{x} \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

This matrix A is called the **standard matrix** for T and is given by:

$$A = \begin{pmatrix} T(\mathbf{e}_1) & T(\mathbf{e}_2) & \cdots & T(\mathbf{e}_n) \end{pmatrix}$$

where $\mathbf{e}_j = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$ (1 in j-th position) is the j-th standard basis vector for $\mathbb{R}^n$.

Proof. Any vector $\mathbf{x} \in \mathbb{R}^n$ can be written as $\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$. Since T is linear:

$$T(\mathbf{x}) = T(x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n) = x_1T(\mathbf{e}_1) + \cdots + x_nT(\mathbf{e}_n)$$

This is a linear combination of the vectors $T(\mathbf{e}_j)$. By the definition of $A\mathbf{x}$, this is exactly:

$$T(\mathbf{x}) = \begin{pmatrix} T(\mathbf{e}_1) & T(\mathbf{e}_2) & \cdots & T(\mathbf{e}_n) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = A\mathbf{x}$$

□

3.4.4 Geometric Transformations in $\mathbb{R}^2$

This section allows us to find the matrix for geometric operations.

Example 3.4.2: Rotation

Find the standard matrix for $T : \mathbb{R}^2 \to \mathbb{R}^2$ that rotates a vector counter-clockwise by an angle θ. We just need to find $T(\mathbf{e}_1)$ and $T(\mathbf{e}_2)$. $\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Rotating this by θ gives $T(\mathbf{e}_1) = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$. $\mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Rotating this by θ gives $T(\mathbf{e}_2) = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$. The standard matrix is $A = \begin{pmatrix} T(\mathbf{e}_1) & T(\mathbf{e}_2) \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.

Example 3.4.3: Reflection

Find the standard matrix for $T : \mathbb{R}^2 \to \mathbb{R}^2$ that reflects a vector across the x-axis. $T(\mathbf{e}_1) = T\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. $T(\mathbf{e}_2) = T\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$. The standard matrix is $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

Definition 3.4.2: Kernel and Range

Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation.

- The **Kernel** of T, $\text{Ker}(T)$, is the set of all $\mathbf{x}$ in $\mathbb{R}^n$ such that $T(\mathbf{x}) = \mathbf{0}$.
- The **Range** of T, $\text{Range}(T)$, is the set of all $\mathbf{y}$ in $\mathbb{R}^m$ such that $\mathbf{y} = T(\mathbf{x})$ for some $\mathbf{x}$ in $\mathbb{R}^n$.

T is **one-to-one** if $\text{Ker}(T) = \{\mathbf{0}\}$. T is **onto** if $\text{Range}(T) = \mathbb{R}^m$.

If $T(\mathbf{x}) = A\mathbf{x}$, these are just our old subspaces:

- $\text{Ker}(T)$ is the solution set of $A\mathbf{x} = \mathbf{0}$. This is the **Null Space** of A, $\text{Nul}(A)$.
- $\text{Range}(T)$ is the set of all linear combinations of the columns of A. This is the **Column Space** of A, $\text{Col}(A)$.

3.5 Abstract Linear Spaces and Subspaces

In the previous sections, we studied $\mathbb{R}^n$ and its algebraic properties. We observed that matrices ($M_{m\times n}$) and polynomials ($\mathcal{P}_n$) also have similar properties (we can add them, scale them). We will now **abstract** these properties to define a more general concept.

3.5.1 The Formal Definition

Definition 3.5.1

A **Linear Space** (or **Vector Space**) V is a non-empty set of objects, called **vectors**, on which two operations are defined: vector addition ($\mathbf{u}+\mathbf{v}$) and scalar multiplication ($c\mathbf{u}$) (over a field F, usually $\mathbb{R}$). These operations must satisfy the following ten axioms for all vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ in V and all scalars c, d in $\mathbb{R}$:

1. $\mathbf{u}+\mathbf{v}$ is in V. (Closure under addition)
2. $\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}$. (Commutativity)
3. $(\mathbf{u}+\mathbf{v})+\mathbf{w} = \mathbf{u}+(\mathbf{v}+\mathbf{w})$. (Associativity of addition)
4. There is a **zero vector** $\mathbf{0}$ in V such that $\mathbf{u}+\mathbf{0} = \mathbf{u}$.
5. For each $\mathbf{u}$ in V, there is an **additive inverse** $-\mathbf{u}$ in V such that $\mathbf{u}+(-\mathbf{u}) = \mathbf{0}$.
6. $c\mathbf{u}$ is in V. (Closure under scalar multiplication)
7. $c(\mathbf{u}+\mathbf{v}) = c\mathbf{u}+c\mathbf{v}$. (Distributivity)
8. $(c+d)\mathbf{u} = c\mathbf{u}+d\mathbf{u}$. (Distributivity)
9. $c(d\mathbf{u}) = (cd)\mathbf{u}$. (Associativity of multiplication)
10. $1\mathbf{u} = \mathbf{u}$. (Scalar identity element)

3.5.2 Examples of Linear Spaces

The power of this definition comes from the variety of sets that satisfy these axioms.

- **Example 1:** $\mathbb{R}^n$ As we've just seen, $\mathbb{R}^n$ with standard component-wise operations is our prototype vector space.
- **Example 2: The Space of Polynomials** $\mathcal{P}_n$ Let $V = \mathcal{P}_n$ be the set of all polynomials of degree **at most** n. A "vector" in this space is a polynomial $\mathbf{p}(t) = a_0 + a_1 t + \cdots + a_n t^n$. Standard polynomial addition and scalar multiplication satisfy all ten axioms. The "zero vector" is the zero polynomial, $\mathbf{0}(t) = 0$.
- **Example 3: The Space of Matrices** $M_{m\times n}$ The set $V = M_{m\times n}$ of all $m\times n$ matrices, with standard matrix addition and scalar multiplication (as defined in Section 2.2), forms a vector space. The "zero vector" is the $m \times n$ zero matrix.

- **Example 4: The Space of Functions** $C[a,b]$ Let $V = C[a,b]$ be the set of all *continuous* real-valued functions on an interval $[a,b]$. We define operations "pointwise": $(f+g)(x) = f(x) + g(x)$ $(cf)(x) = c \cdot f(x)$ Since the sum of continuous functions is continuous, and a scalar multiple is continuous, the set is closed. The "zero vector" is the constant function $f(x) = 0$. This forms a vector space.

3.5.3 Subspaces

Often, a vector space is contained inside a larger one.

Definition 3.5.2

A **subspace** of a vector space V is a subset H of V that satisfies three properties:

1. The zero vector of V is in H. ($\mathbf{0} \in H$)
2. H is closed under vector addition: For all $\mathbf{u}, \mathbf{v}$ in H, $\mathbf{u} + \mathbf{v}$ is in H.
3. H is closed under scalar multiplication: For all $\mathbf{u}$ in H and scalar c, $c\mathbf{u}$ is in H.

These three properties guarantee that H is itself a vector space (it inherits the other 7 axioms from V).

Example 3.5.1: A subspace

Let $V = \mathbb{R}^3$. Let H be the xy-plane, i.e., $H = \left\{ \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \mid x, y \in \mathbb{R} \right\}$. Is H a subspace?

1. Is $\mathbf{0} \in H$? Yes, $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ has $z = 0$.
2. Let $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix}, \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ 0 \end{pmatrix}$ be in H. Is $\mathbf{u} + \mathbf{v} \in H$? $\mathbf{u} + \mathbf{v} = \begin{pmatrix} u_1 + v_1 \\ u_2 + v_2 \\ 0 \end{pmatrix}$. Yes, its third component is 0.
3. Let $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix}$ be in H. Is $c\mathbf{u} \in H$? $c\mathbf{u} = \begin{pmatrix} cu_1 \\ cu_2 \\ c \cdot 0 \end{pmatrix} = \begin{pmatrix} cu_1 \\ cu_2 \\ 0 \end{pmatrix}$. Yes, it is in H.

Thus, H is a subspace of $\mathbb{R}^3$.

Example 3.5.2: A non-subspace

Let $V = \mathbb{R}^2$. Let H be the first quadrant, $H = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \mid x \geq 0, y \geq 0 \right\}$.

1. $\mathbf{0} \in H$. (Pass)
2. H is closed under addition. (Pass: $x_1 + x_2 \geq 0, y_1 + y_2 \geq 0$)
3. Is H closed under scalar multiplication? Let $c = -1$ and $\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in H$. $c\mathbf{u} = -1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$.

   This is *not* in H.

H is **not** a subspace.

Theorem 3.5.1

If $\mathbf{v}_1, \ldots, \mathbf{v}_p$ are in a vector space V, then $H = \text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is **always** a subspace of V.

Proof.

1. $\mathbf{0} = 0\mathbf{v}_1 + \cdots + 0\mathbf{v}_p$, so $\mathbf{0}$ is in the span.

2. Let $\mathbf{u} = c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p$ and $\mathbf{v} = d_1\mathbf{v}_1 + \cdots + d_p\mathbf{v}_p$. Then $\mathbf{u} + \mathbf{v} = (c_1 + d_1)\mathbf{v}_1 + \cdots + (c_p + d_p)\mathbf{v}_p$, which is a linear combination, so it is in the span.
3. Let k be a scalar. $k\mathbf{u} = k(c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p) = (kc_1)\mathbf{v}_1 + \cdots + (kc_p)\mathbf{v}_p$, which is also in the span.

Thus, any span is a subspace. □

3.5.4 Null Spaces and Column Spaces

There are two fundamental subspaces associated with any $m \times n$ matrix A.

Definition 3.5.3

- The **Null Space** of A, Nul(A), is the set of all solutions to the homogeneous equation $A\mathbf{x} = \mathbf{0}$.

$$\text{Nul}(A) = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\}$$

  This is a subspace of $\mathbb{R}^n$.
- The **Column Space** of A, Col(A), is the span of the columns of A.

$$\text{Col}(A) = \text{Span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = \{\mathbf{b} \in \mathbb{R}^m \mid \mathbf{b} = A\mathbf{x} \text{ for some } \mathbf{x} \in \mathbb{R}^n\}$$

  This is a subspace of $\mathbb{R}^m$.

Nul(A) describes the structure of the homogeneous solution set. Col(A) describes the set of all $\mathbf{b}$ for which $A\mathbf{x} = \mathbf{b}$ is consistent.

3.5.5 Basis and Dimension

We now unify the ideas of spanning and linear independence.

Definition 3.5.4

A **basis** for a vector space V is a set of vectors $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ in V such that:
1. $\mathcal{B}$ is a linearly independent set.
2. $\mathcal{B}$ spans V (i.e., Span$\{\mathcal{B}\} = V$).

A basis is the "smallest" possible spanning set and the "largest" possible linearly independent set.

Example 3.5.3

The set of standard vectors $\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is the **standard basis** for $\mathbb{R}^n$. The set $\{1, t, t^2, \ldots, t^n\}$ is the **standard basis** for $\mathcal{P}_n$.

Theorem 3.5.2

All bases for a vector space V have the same number of vectors.

Definition 3.5.5

The **dimension** of a non-zero vector space V, denoted dim(V), is the number of vectors in any basis for V. The dimension of the zero subspace $\{\mathbf{0}\}$ is defined to be 0.

Examples of Dimension:

- $\dim(\mathbb{R}^n) = n$.
- $\dim(\mathcal{P}_n) = n + 1$ (because of the $t^0 = 1$ term).
- $\dim(M_{m\times n}) = m \times n$.
- $C[a, b]$ is **infinite-dimensional**.

There is an interesting conclusion. The cardinality of $C[0, 1]$ equals to the cardinality of $\mathbb{R}$.

Proof. Let $D = \mathbb{Q} \cap [0, 1]$ be the countable dense set of rationals in $[0, 1]$.

Upper bound ($\#C[0, 1] \leq \mathfrak{c}$): Define $\Phi : C[0, 1] \to \mathbb{R}^{\mathbb{N}}$ by

$$\Phi(f) = (f(q_1), f(q_2), f(q_3), \dots)$$

where $\{q_i\}$ enumerates D. If $\Phi(f) = \Phi(g)$, then $f(q) = g(q)$ for all $q \in D$. By continuity and density, $f = g$ on $[0, 1]$, so Φ is injective. Thus

$$\#C[0, 1] \leq \#(\mathbb{R}^{\mathbb{N}}) = \mathfrak{c}^{\aleph_0} = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0} = \mathfrak{c}.$$

Lower bound ($\#C[0, 1] \geq \mathfrak{c}$): The constant functions $\{f_r(x) = r : r \in \mathbb{R}\}$ form a subset of $C[0, 1]$ with cardinality $\mathfrak{c}$.

By Cantor-Bernstein theorem, $\#C[0, 1] = \mathfrak{c}$. $\square$

We can now find bases for our two favorite subspaces.

- **Basis for** $\mathrm{Col}(A)$**:** The pivot columns of the *original* matrix A form a basis for $\mathrm{Col}(A)$. (Do not use the RREF columns, as EROs change the column space).
- **Basis for** $\mathrm{Nul}(A)$**:** The vectors found when writing the solution of $A\mathbf{x} = \mathbf{0}$ in parametric vector form form a basis for $\mathrm{Nul}(A)$.

Example 3.5.4

$A = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \\ 0 & 1 & 3 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix}$ **Column Space:** Pivots are in columns 1 and 2. Basis for $\mathrm{Col}(A) = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right\}$. $\dim(\mathrm{Col}(A)) = 2$.

Null Space: Solve $A\mathbf{x} = \mathbf{0}$. $x_1 + 2x_3 = 0 \implies x_1 = -2x_3$ $x_2 + 3x_3 = 0 \implies x_2 = -3x_3$ x_3 is free.

Let $x_3 = t$. $\mathbf{x} = \begin{pmatrix} -2t \\ -3t \\ t \end{pmatrix} = t \begin{pmatrix} -2 \\ -3 \\ 1 \end{pmatrix}$. Basis for $\mathrm{Nul}(A) = \left\{ \begin{pmatrix} -2 \\ -3 \\ 1 \end{pmatrix} \right\}$. $\dim(\mathrm{Nul}(A)) = 1$.

Notice the connection to rank:

- $\dim(\mathrm{Col}(A)) =$ (Number of pivot columns) $= \mathrm{rank}(A)$.
- $\dim(\mathrm{Nul}(A)) =$ (Number of free variables) $= n - \mathrm{rank}(A)$.

This leads to one of the most important theorems in linear algebra.

Theorem 3.5.3: The Rank-Nullity Theorem

For an $m \times n$ matrix A,

$$\dim(\mathrm{Col}(A)) + \dim(\mathrm{Nul}(A)) = n$$

or, equivalently,

$$\mathrm{rank}(A) + \mathrm{nullity}(A) = n$$

where n is the number of **columns** and $\mathrm{nullity}(A) = \dim(\mathrm{Nul}(A))$.

In our last example, $n = 3$. $\mathrm{rank}(A) = 2$, $\mathrm{nullity}(A) = 1$. $2 + 1 = 3$. The theorem holds. This theorem beautifully ties together the dimensions of the two fundamental subspaces associated with a matrix.

3.5.6 Coordinate Systems

A basis $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ for $\mathbb{R}^n$ acts like a new coordinate system. Because $\mathcal{B}$ spans $\mathbb{R}^n$ and is linearly independent, every $\mathbf{x} \in \mathbb{R}^n$ can be written *uniquely* as

$$\mathbf{x} = c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n$$

Definition 3.5.6

The scalars $c_1, \ldots, c_n$ are the **coordinates of x relative to the basis** $\mathcal{B}$. The **coordinate vector** of $\mathbf{x}$ (relative to $\mathcal{B}$) is

$$[\mathbf{x}]_{\mathcal{B}} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$$

Let $P_{\mathcal{B}}$ be the **change-of-coordinates matrix** $P_{\mathcal{B}} = \begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{pmatrix}$. The equation $\mathbf{x} = c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n$ is just the matrix equation

$$\mathbf{x} = P_{\mathcal{B}}[\mathbf{x}]_{\mathcal{B}}$$

Since the columns of $P_{\mathcal{B}}$ are a basis, $P_{\mathcal{B}}$ is invertible (by the IMT).

$$[\mathbf{x}]_{\mathcal{B}} = P_{\mathcal{B}}^{-1}\mathbf{x}$$

This provides a way to ”translate” between the standard coordinate system $\mathcal{E}$ and the new system $\mathcal{B}$.

3.5.7 Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors are fundamental concepts in linear algebra that provide crucial insights into the structure of linear transformations. They play a central role in many applications, including vibration analysis, quantum mechanics, and data analysis.

The reason why we wants to study eigenvalues and eigenvectors is that they help us understand how a linear transformation (represented by a matrix) acts on certain special directions in space. Specifically, an eigenvector is a direction that remains unchanged (up to scaling) when the transformation is applied, and the corresponding eigenvalue indicates how much the vector is stretched or compressed.

Definition 3.5.7: Eigenvalues and Eigenvectors

Let A be an $n \times n$ square matrix. A scalar λ is called an **eigenvalue** of A if there exists a nonzero vector $\mathbf{v} \in \mathbb{R}^n$ such that

$$A\mathbf{v} = \lambda\mathbf{v}$$

The vector $\mathbf{v}$ is called an **eigenvector** corresponding to the eigenvalue λ.

Geometrically, an eigenvector $\mathbf{v}$ is a vector whose direction remains unchanged when transformed by A; it is only scaled by the factor λ.

To find eigenvalues, we rewrite the equation $A\mathbf{v} = \lambda\mathbf{v}$ as

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

This is a homogeneous system of linear equations. Since $\mathbf{v} \neq \mathbf{0}$, this system must have nontrivial solutions, which requires that the matrix $A - \lambda I$ be singular, i.e., its determinant must be zero.

Definition 3.5.8: Characteristic Polynomial

The **characteristic polynomial** of a matrix A is defined as

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

This is an nth-degree polynomial in λ. The eigenvalues of A are the roots of the characteristic equation $p(\lambda) = 0$.

The roots of the characteristic polynomial may be real or complex. Repeated roots are called eigenvalues with algebraic multiplicity greater than 1. Each eigenvalue corresponds to an eigenspace.

Definition 3.5.9: Eigenspace

For an eigenvalue λ, the corresponding **eigenspace** is the solution space of the homogeneous system $(A - \lambda I)\mathbf{v} = \mathbf{0}$, i.e., $\mathrm{Nul}(A - \lambda I)$. The dimension of the eigenspace is called the **geometric multiplicity** of λ.

Example 3.5.5

Find the eigenvalues and eigenvectors of $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.
The characteristic polynomial is:

$$p(\lambda) = \det \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} = (2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3)$$

The eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 3$.
For $\lambda_1 = 1$, solve $(A - I)\mathbf{v} = \mathbf{0}$:

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies v_1 + v_2 = 0$$

Thus the eigenvectors are $\mathbf{v}_1 = t \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $t \neq 0$.
For $\lambda_2 = 3$, solve $(A - 3I)\mathbf{v} = \mathbf{0}$:

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies -v_1 + v_2 = 0$$

Thus the eigenvectors are $\mathbf{v}_2 = s \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $s \neq 0$.

3.5.8 Diagonalization

Diagonalization is the process of transforming a matrix into diagonal form, which greatly simplifies computations involving matrix powers and exponentials. The geometric interpretation is that diagonalization aligns the coordinate system with the eigenvectors of the matrix, making the transformation represented by the matrix easier to understand.

Definition 3.5.10: Diagonalizable Matrix

An $n \times n$ matrix A is said to be **diagonalizable** if there exists an invertible matrix P and a diagonal matrix D such that

$$A = PDP^{-1}$$

Equivalently, $P^{-1}AP = D$.

The diagonal entries of D are the eigenvalues of A, and the columns of P are the corresponding linearly independent eigenvectors.

Theorem 3.5.4

A matrix A is diagonalizable if and only if it has n linearly independent eigenvectors. This is equivalent to the condition that the geometric multiplicity of each eigenvalue equals its algebraic multiplicity (the multiplicity as a root of the characteristic polynomial).

Example 3.5.6

Continuing the previous example, $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 3$ with corresponding eigenvectors $\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. These eigenvectors are linearly independent, so we can take

$$P = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

It's easy to verify that $A = PDP^{-1}$.

Once diagonalized, computing powers of A becomes straightforward:

$$A^k = (PDP^{-1})^k = PD^kP^{-1}$$

since D^k is simply obtained by raising each diagonal element to the kth power.

3.5.9 Inner Product Spaces

An inner product generalizes the dot product and provides a framework for defining lengths, angles, and orthogonality in vector spaces.

Definition 3.5.11: Inner Product

Let V be a real vector space. An **inner product** is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ satisfying the following properties for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and all scalars c:

1. $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$ (Symmetry)
2. $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$ (Linearity)
3. $\langle c\mathbf{u}, \mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle$
4. $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$, and $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = \mathbf{0}$ (Positive definiteness)

The most common example is the dot product (standard inner product) on $\mathbb{R}^n$:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$$

Definition 3.5.12: Norm and Distance

The **norm** (length) induced by an inner product is defined as

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$$

The **distance** between vectors $\mathbf{u}$ and $\mathbf{v}$ is $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$.

Definition 3.5.13: Orthogonality

Two vectors $\mathbf{u}$ and $\mathbf{v}$ are **orthogonal** if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. A set of vectors is orthogonal if all pairs of distinct vectors in the set are orthogonal. If, in addition, each vector has unit norm, the set is **orthonormal**.

3.5.10 Orthogonal Bases and the Gram-Schmidt Process

In inner product spaces, orthogonal bases simplify many computations.

Theorem 3.5.5

If $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is an orthogonal basis for a subspace H, then for any $\mathbf{y} \in H$,

$$\mathbf{y} = c_1 \mathbf{v}_1 + \cdots + c_k \mathbf{v}_k, \quad \text{where } c_i = \frac{\langle \mathbf{y}, \mathbf{v}_i \rangle}{\langle \mathbf{v}_i, \mathbf{v}_i \rangle}$$

If the basis is orthonormal, then $c_i = \langle \mathbf{y}, \mathbf{v}_i \rangle$.

The Gram-Schmidt process converts any linearly independent set into an orthogonal basis.

Theorem 3.5.6: Gram-Schmidt Orthogonalization Process

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ be a basis for a subspace H. Define:

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{x}_1 \\
\mathbf{v}_2 &= \mathbf{x}_2 - \frac{\langle \mathbf{x}_2, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 \\
\mathbf{v}_3 &= \mathbf{x}_3 - \frac{\langle \mathbf{x}_3, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 - \frac{\langle \mathbf{x}_3, \mathbf{v}_2 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} \mathbf{v}_2 \\
&\vdots \\
\mathbf{v}_p &= \mathbf{x}_p - \sum_{i=1}^{p-1} \frac{\langle \mathbf{x}_p, \mathbf{v}_i \rangle}{\langle \mathbf{v}_i, \mathbf{v}_i \rangle} \mathbf{v}_i
\end{aligned}$$

Then $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is an orthogonal basis for H. Normalizing each vector yields an orthonormal basis.

3.5.11 Symmetric Matrices and Quadratic Forms

Real symmetric matrices have particularly nice properties that make them important in many applications.

Theorem 3.5.7: Spectral Theorem

Let A be an $n \times n$ real symmetric matrix. Then:

1. All eigenvalues of A are real.
2. A has n linearly independent eigenvectors, and eigenvectors corresponding to distinct eigenvalues are orthogonal.
3. A is orthogonally diagonalizable: there exists an orthogonal matrix Q (satisfying $Q^T = Q^{-1}$) and a diagonal matrix D such that
$$A = QDQ^T$$

Quadratic forms are homogeneous polynomials of degree 2 that can be represented in matrix form as $\mathbf{x}^T A\mathbf{x}$, where A is a symmetric matrix.

Definition 3.5.14: Quadratic Form

A **quadratic form** is a function $Q : \mathbb{R}^n \to \mathbb{R}$ defined by

$$Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

where A is a symmetric matrix.

Through the orthogonal transformation $\mathbf{x} = Q\mathbf{y}$, a quadratic form can be reduced to its canonical form:

$$Q(\mathbf{x}) = \mathbf{y}^T D\mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2$$

where the λ_i are the eigenvalues of A.

Quadratic forms are classified based on the signs of their eigenvalues:

- **Positive definite**: All eigenvalues positive; $Q(\mathbf{x}) > 0$ for $\mathbf{x} \neq \mathbf{0}$.
- **Negative definite**: All eigenvalues negative.
- **Indefinite**: Eigenvalues have mixed signs.

This classification has important applications in optimization and the study of critical points in multivariable calculus.

3.5.12 Singular Value Decomposition (SVD)

The Diagonalization Theorem ($A = PDP^{-1}$) applies only to square, diagonalizable matrices. The Spectral Theorem applies only to symmetric matrices. The SVD is the ultimate generalization: it applies to **any** $m \times n$ matrix.

Theorem 3.5.8: Singular Value Decomposition

Let A be an $m \times n$ matrix with rank r. Then there exists an $m \times n$ factorization of the form:

$$A = U\Sigma V^T$$

where:

- U is an $m \times m$ orthogonal matrix ($U^T U = I$). The columns of U are called the **left singular vectors**.
- V is an $n \times n$ orthogonal matrix ($V^T V = I$). The columns of V are called the **right singular vectors**.
- Σ is an $m \times n$ rectangular diagonal matrix with non-negative entries on the diagonal:

$$\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

  where $D = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

The scalars σ_i are called the **singular values** of A. They are the square roots of the non-zero eigenvalues of $A^T A$.

Geometric Interpretation: Any linear transformation $T(\mathbf{x}) = A\mathbf{x}$ maps the unit sphere in the domain to a hyperellipse in the codomain. The singular values are the lengths of the semi-axes of this hyperellipse.

Construction of SVD:

1. Compute the eigenvalues of the symmetric matrix $A^T A$. Let them be $\lambda_1 \geq \cdots \geq \lambda_n$.
2. The singular values are $\sigma_i = \sqrt{\lambda_i}$.
3. Find orthonormal eigenvectors of $A^T A$; these form the matrix V.
4. The first r columns of U are given by $\mathbf{u}_i = \frac{1}{\sigma_i} A\mathbf{v}_i$. Extend this set to an orthonormal basis for $\mathbb{R}^m$ to fill the rest of U.

3.6 Conclusions

In this chapter, we have learned a lot of concepts. From matrix to determinant, from rank to eigenvalue $\cdots$. We can spot very beautiful symmetry between each part, we are actually using one single language to describe different things from the same perspective but have varying results. This is what makes mathematics attractive. Now we will focus on two concepts about Linear Algebra, to conclude what we have learned through out the journey:

3.6.1 Interpretations of Rank

Let A be an $m \times n$ matrix over a field $\mathbb{F}$ (e.g., $\mathbb{R}$ or $\mathbb{C}$). The rank of A, denoted as $\text{rank}(A)$ or $\rho(A)$, can be defined and interpreted in the following equivalent ways:

Vector Space Interpretations

- **Column Rank:** The dimension of the column space of A (the vector space spanned by its columns).

$$\text{rank}(A) = \dim(\text{Col}(A))$$

- **Row Rank:** The dimension of the row space of A (the vector space spanned by its rows). A fundamental property is that row rank equals column rank:

$$\dim(\text{Row}(A)) = \dim(\text{Col}(A))$$

- **Linear Independence:** The maximum number of linearly independent column vectors (or row vectors) in the matrix.

Computational/Algebraic Interpretations

- **Pivot Definition:** The number of pivots (leading 1s) in the Reduced Row Echelon Form (RREF) of A.
- **Determinantal Rank:** The order of the largest non-zero square minor of A. That is, r is the rank if there exists an $r \times r$ submatrix with a non-zero determinant, and every $(r+1) \times (r+1)$ minor is zero.
- **Decomposition Rank:** The smallest integer k such that A can be factored as $A = CR$, where C is $m \times k$ and R is $k \times n$.

Geometric and Mapping Interpretations

- **Image Dimension:** If we view A as a linear transformation $T : \mathbb{F}^n \to \mathbb{F}^m$ defined by $T(\mathbf{x}) = A\mathbf{x}$, the rank is the dimension of the image (range) of T:

$$\text{rank}(A) = \dim(\text{Im}(T))$$

- **Singular Value Decomposition (SVD):** The number of non-zero singular values of A.

3.6.2 The Rank-Nullity Theorem

The Rank-Nullity Theorem (often called the Fundamental Theorem of Linear Algebra) relates the dimensions of the domain, the image, and the kernel. Below are its expressions in different contexts.

1. Matrix Context

For an $m \times n$ matrix A:

Theorem 3.6.1: Matrix Rank-Nullity

The number of columns equals the sum of the rank and the nullity.

$$\text{rank}(A) + \text{nullity}(A) = n$$

- **rank(A):** The number of pivot columns (basic variables).
- **nullity(A):** The dimension of the null space ($\dim(\text{Null}(A))$), which corresponds to the number of free columns (free variables).
- **Interpretation:** Total Variables = Pivot Variables + Free Variables.

2. Linear Transformation Context

Let V and W be vector spaces, where V is finite-dimensional. Let $T : V \to W$ be a linear transformation.

Theorem 3.6.2: Linear Map Rank-Nullity

$$\dim(\text{Im}(T)) + \dim(\ker(T)) = \dim(V)$$

- $\dim(\mathrm{Im}(T))$ is the rank of the transformation.
- $\dim(\ker(T))$ is the nullity (dimension of the kernel).
- Note that the sum equals the dimension of the *domain*, not the codomain.

3. Abstract Algebra Context (Isomorphism Theorems)

The theorem is a direct consequence of the **First Isomorphism Theorem** for vector spaces (or modules).

Theorem 3.6.3

$$V/\ker(T) \cong \mathrm{Im}(T)$$

Taking dimensions of both sides:

$$\dim(V) - \dim(\ker(T)) = \dim(\mathrm{Im}(T))$$

Rearranging this yields the standard Rank-Nullity equation.

4. Systems of Linear Equations

Consider the homogeneous system $A\mathbf{x} = \mathbf{0}$, where A is $m \times n$.

- The dimension of the solution space is $k = n - r$, where $r = \mathrm{rank}(A)$.
- If $r = n$ (full column rank), the only solution is the trivial solution ($\mathbf{0}$), so nullity is 0.
- If $r < m$ (for the augmented system $A\mathbf{x} = \mathbf{b}$), existence of solutions depends on column space consistency.

3.6.3 The Axiom of Linear Algebra

Afterall, we need to answer the question: what is the core axiom of the linear algebra? We believe it is **Axiomatic Definition of a Vector Space**.

The axiomatic definition of vector space is central to linear algebra because it captures the essence of linearity through just two fundamental operations—addition and scalar multiplication—and the eight axioms that govern them. This simple yet powerful abstract framework unifies countless mathematical objects, from geometric vectors to functions and matrices, and provides the common foundation for all core theories, such as linear transformations and solving linear systems. In this way, it serves as the universal language that bridges mathematical theory and scientific application.

We shall present the definition again here.

Let V be a nonempty set whose elements are called **vectors**, and let $\mathbb{F}$ be a **field** (such as the real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$). Two operations are defined on V:

- **Vector addition**: $+ : V \times V \to V$, denoted by $(\mathbf{u}, \mathbf{v}) \mapsto \mathbf{u} + \mathbf{v}$
- **Scalar multiplication**: $\cdot : \mathbb{F} \times V \to V$, denoted by $(c, \mathbf{v}) \mapsto c\mathbf{v}$

These operations must satisfy the following 8 axioms (sometimes listed as 10 by including closure explicitly):

1. Axioms for Vector Addition

1. **Closure under addition**: For all $\mathbf{u}, \mathbf{v} \in V$, $\mathbf{u} + \mathbf{v} \in V$.
2. **Associativity of addition**: For all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$,

$$\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}.$$

3. **Commutativity of addition**: For all $\mathbf{u}, \mathbf{v} \in V$,
$$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}.$$

4. **Existence of a zero vector**: There exists a vector $\mathbf{0} \in V$ such that for all $\mathbf{v} \in V$,
$$\mathbf{v} + \mathbf{0} = \mathbf{v}.$$

5. **Existence of additive inverses**: For each $\mathbf{v} \in V$, there exists a vector $-\mathbf{v} \in V$ such that
$$\mathbf{v} + (-\mathbf{v}) = \mathbf{0}.$$

2. Axioms for Scalar Multiplication

6. **Closure under scalar multiplication**: For all $c \in \mathbb{F}$ and $\mathbf{v} \in V$, $c\mathbf{v} \in V$.

7. **Associativity of scalar multiplication**: For all $a, b \in \mathbb{F}$ and $\mathbf{v} \in V$,
$$a(b\mathbf{v}) = (ab)\mathbf{v}.$$

8. **Multiplicative identity**: For all $\mathbf{v} \in V$,
$$1\mathbf{v} = \mathbf{v},$$
where 1 is the multiplicative identity in $\mathbb{F}$.

3. Distributive Laws

9. **Distributivity of scalar multiplication over vector addition**: For all $a \in \mathbb{F}$ and $\mathbf{u}, \mathbf{v} \in V$,
$$a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}.$$

10. **Distributivity of scalar multiplication over scalar addition**: For all $a, b \in \mathbb{F}$ and $\mathbf{v} \in V$,
$$(a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}.$$

Then V is called a **vector space** (or **linear space**) over the field $\mathbb{F}$.

That's what make the whole system works perfectly.

3.7 Summary and Outlook

As we close this chapter on linear algebra, we recognize that we have acquired more than just a collection of techniques for solving equations or manipulating matrices. We have learned a new language—the language of linearity—that reveals hidden structures throughout mathematics and science. From the elegant abstraction of vector spaces to the powerful diagonalization of transformations, linear algebra provides a universal framework for understanding relationships that are, at their heart, proportional and additive. The concepts of basis, dimension, and linear transformation form a conceptual toolkit that will serve as indispensable preparation for the deeper mathematical landscapes ahead—from the infinite-dimensional spaces of functional analysis to the curved geometries of differential manifolds. Linear algebra reminds us that simplicity and structure often underlie apparent complexity, and that the most powerful mathematics is that which provides not just answers, but clarity.

Linear algebra is fundamental not only for its elegant theoretical structure but also as a universal language-with ubiquitous applications across science and engineering. In computer science, it underpins 3D graphics and search algorithms; in data science, techniques like PCA and SVD are core to data reduction. In physics, quantum mechanics is formulated on Hilbert spaces, and in economics, models rely on linear systems. This cross-disciplinary relevance makes linear algebra an indispensable foundation.

Theoretical development in mathematics deeply relies on linear algebraic concepts. Vector spaces generalize to modules, manifolds, and Banach spaces; linear transformations lead to operator and representation theory. Eigenvalues and eigenvectors form the basis for stability analysis in dynamical systems, network science, and quantum mechanics. Mastering linear algebra provides a key to understanding modern mathematics and theoretical science.

In advanced studies, these ideas extend into numerical linear algebra (solving large-scale systems), abstract algebra (modules over rings), and calculus (Jacobian matrices as linear approximations). From signal processing to control theory, linear algebra offers essential models and tools. Ultimately, it represents a mindsetfor uncovering linear structure within complexity, providing a powerful language for modeling, analysis, and solving problems across disciplines.

Keywords: Eigenvalues, Eigenvectors, Diagonalization, Inner Product, Orthogonal Bases, Gram-Schmidt Process, Symmetric Matrices, Quadratic Forms
References:
Linear Algebra and Its Applications, 4th ed., Gilbert Strang, Cengage Learning, 2005.
Shanghai Jiao Tong University, School of Mathematical Sciences. Linear Algebra. China Machine Press.

Chapter 4

Abstract Algebra

Abstract algebra is the study of algebraic structures defined by axiomatic systems. Unlike elementary algebra, which focuses on solving equations involving real or complex numbers, abstract algebra generalizes these concepts to analyze structures that obey specific algebraic laws. It abstracts the common properties of diverse mathematical systems—such as integers, symmetry transformations, matrices, and polynomials—allowing us to reason about them in a unified framework.

This chapter provides a rigorous exploration of four pillars of algebra: **Groups**, **Rings**, **Fields**, and **Modules**. We emphasize axiomatic definitions, structural theorems, and the interplay between these systems.

4.1 Groups

Groups are the fundamental structures for studying symmetry. A group abstracts the notion of invertible operations, whether they are geometric rotations, permutations of a set, or arithmetic addition.

4.1.1 Definition and Examples

Definition 4.1.1: Group

A **group** is a set G equipped with a binary operation $\cdot : G \times G \to G$ satisfying the following axioms:

1. **Associativity**: For all $a, b, c \in G$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.
2. **Identity Element**: There exists a unique element $e \in G$ (often denoted 1 or 0 depending on context) such that for all $a \in G$, $a \cdot e = e \cdot a = a$.
3. **Inverses**: For every $a \in G$, there exists a unique element $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$.

(Note: The closure property is implicit in the definition of a binary operation $G \times G \to G$).

Definition 4.1.2: Abelian Group

If the operation is commutative (i.e., $a \cdot b = b \cdot a$ for all $a, b \in G$), the group is called **abelian**.

Example 4.1.1: Fundamental Examples

1. **Integers**: $(Z, +)$ is an infinite abelian group with identity 0 and inverse $-a$.
2. **General Linear Group**: The set $GL_n(R)$ of invertible $n \times n$ matrices with real entries is a non-abelian group under matrix multiplication.
3. **Symmetric Group**: S_n, the set of all bijections from $\{1, \ldots, n\}$ to itself, is a group under composition. $|S_n| = n!$. It is non-abelian for $n \geq 3$.
4. **Cyclic Groups**: Z_n (integers modulo n) under addition is a cyclic group of order n.
5. **Dihedral Groups**: D_{2n} represents the symmetries of a regular n-gon, containing n rotations and n reflections. $|D_{2n}| = 2n$.

4.1.2 Elementary Properties

The axioms imply strong structural regularities.

Proposition 4.1.1: Cancellation Laws

Let G be a group and $a, b, c \in G$.
1. If $ab = ac$, then $b = c$ (Left Cancellation).
2. If $ba = ca$, then $b = c$ (Right Cancellation).
3. $(ab)^{-1} = b^{-1}a^{-1}$ (The "Shoe-Sock" Property).

Definition 4.1.3: Order

The **order of a group** G, denoted $|G|$, is the cardinality of the set G. The **order of an element** $g \in G$, denoted $|g|$, is the smallest positive integer n such that $g^n = e$. If no such n exists, g has infinite order.

4.1.3 Subgroups and Cosets

Definition 4.1.4: Subgroup

A subset $H \subseteq G$ is a **subgroup** (denoted $H \leq G$) if H is a group under the restricted operation of G.

Theorem 4.1.1: Subgroup Test

A non-empty subset $H \subseteq G$ is a subgroup if and only if for all $x, y \in H$, $xy^{-1} \in H$.

Definition 4.1.5: Cosets

Let $H \leq G$. For any $g \in G$:
- The **left coset** of H containing g is $gH = \{gh \mid h \in H\}$.
- The **right coset** of H containing g is $Hg = \{hg \mid h \in H\}$.

Cosets partition the group G. Importantly, all cosets of a subgroup H have the same cardinality as H. This leads to one of the most famous theorems in finite group theory.

Theorem 4.1.2: Lagrange's Theorem

If G is a finite group and $H \leq G$, then $|H|$ divides $|G|$. Furthermore,

$$|G| = [G : H] \cdot |H|,$$

where $[G : H]$ is the number of distinct left cosets of H in G, called the **index**.

Corollary 4.1.1

If $|G| = p$ where p is a prime number, then G is cyclic and essentially unique (isomorphic to Z_p).

4.1.4 Normal Subgroups and Quotient Groups

Not all subgroups are created equal. To construct a new group from cosets, we require the subgroup to be "normal."

Definition 4.1.6: Normal Subgroup

A subgroup $N \leq G$ is **normal**, denoted $N \trianglelefteq G$, if it is invariant under conjugation. That is, for all $g \in G$ and $n \in N$, $gng^{-1} \in N$.

Proposition 4.1.2

The following are equivalent:

1. $N \trianglelefteq G$.
2. $gN = Ng$ for all $g \in G$ (Left cosets equal right cosets).
3. The operation $(aN)(bN) := (ab)N$ is well-defined.

Definition 4.1.7: Quotient Group

If $N \trianglelefteq G$, the set of cosets G/N forms a group under the operation defined above. This is called the **quotient group**. The order is $|G/N| = [G : N]$.

4.1.5 Homomorphisms and Isomorphisms

Definition 4.1.8: Homomorphism

A function $\phi : G \to H$ is a **homomorphism** if $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in G$.

Associated with any homomorphism are two structural components:

- **Kernel**: $Ker(\phi) = \{g \in G \mid \phi(g) = e_H\}$. This is always a *normal* subgroup of G.
- **Image**: $Img(\phi) = \{\phi(g) \mid g \in G\}$. This is a subgroup of H.

Theorem 4.1.3: First Isomorphism Theorem

Let $\phi : G \to H$ be a homomorphism. Then there is an isomorphism:

$$G/Ker(\phi) \cong Img(\phi).$$

This theorem essentially states that the image of a group looks exactly like the group "modulo" the elements that are sent to identity.

Theorem 4.1.4: Second and Third Isomorphism Theorems

1. Let $H \leq G$ and $N \trianglelefteq G$. Then $H \cap N \trianglelefteq H$ and $H/(H \cap N) \cong HN/N$.
2. Let $N \trianglelefteq G$ and $K \trianglelefteq G$ with $N \leq K$. Then $(G/N)/(K/N) \cong G/K$.

4.1.6 Group Actions and Sylow Theorems

Group actions provide a dynamic view of groups as "doers" rather than just static structures.

Definition 4.1.9: Group Action

A group G **acts** on a set X if there is a map $G \times X \to X$, denoted $g \cdot x$, such that $e \cdot x = x$ and $g \cdot (h \cdot x) = (gh) \cdot x$.

Key concepts include the **Orbit** $Orb(x) = \{g \cdot x \mid g \in G\}$ and the **Stabilizer** $Stab(x) = \{g \in G \mid g \cdot x = x\}$.

Theorem 4.1.5: Orbit-Stabilizer

For a finite group G acting on X, $|Orb(x)| = |G|/|Stab(x)|$.

Theorem 4.1.6: The Class Equation

Let G act on itself by conjugation ($g \cdot x = gxg^{-1}$). The orbits are called conjugacy classes. We have:

$$|G| = |Z(G)| + \sum_i [G : C_G(g_i)],$$

where $Z(G)$ is the center of G, and the sum runs over representatives of distinct non-central conjugacy classes.

Theorem 4.1.7: Sylow Theorems

Let $|G| = p^k m$ with $p \nmid m$.

1. **Existence**: G has a subgroup of order p^k (Sylow p-subgroup).
2. **Conjugacy**: All Sylow p-subgroups are conjugate.
3. **Number**: Let n_p be the number of Sylow p-subgroups. Then $n_p \equiv 1 \pmod p$ and $n_p \mid m$.

4.2 Rings

Rings are sets equipped with two binary operations, usually modeling "arithmetic" where we can add, subtract, and multiply, but not necessarily divide.

4.2.1 Fundamentals

Definition 4.2.1: Ring

A **ring** R is a set with operations $(+, \cdot)$ such that:

1. $(R, +)$ is an abelian group (identity 0).
2. $(R, \cdot)$ is associative.
3. The Distributive Laws hold: $a(b + c) = ab + ac$ and $(a + b)c = ac + bc$.

If there is a multiplicative identity $1 \neq 0$, R is a **ring with unity**. If $ab = ba$, R is **commutative**.

Definition 4.2.2: Types of Elements

- **Unit**: An element u is a unit if it has a multiplicative inverse.
- **Zero Divisor**: A non-zero element a is a zero divisor if $\exists b \neq 0$ such that $ab = 0$.
- **Integral Domain**: A commutative ring with unity and no zero divisors.

4.2.2 Ideals and Homomorphisms

Ideals are to rings what normal subgroups are to groups: they allow the construction of quotients.

Definition 4.2.3: Ideal

A subset $I \subseteq R$ is a (two-sided) **ideal** if:

1. $(I, +)$ is a subgroup of $(R, +)$.
2. Absorbency: For all $r \in R$ and $x \in I$, both $rx \in I$ and $xr \in I$.

Definition 4.2.4: Prime and Maximal Ideals

Let R be a commutative ring with unity.

- An ideal $P \subsetneq R$ is **prime** if $ab \in P \implies a \in P$ or $b \in P$.
- An ideal $M \subsetneq R$ is **maximal** if there is no ideal I such that $M \subsetneq I \subsetneq R$.

Theorem 4.2.1: Quotients by Special Ideals

1. R/P is an Integral Domain $\iff$ P is a prime ideal.
2. R/M is a Field $\iff$ M is a maximal ideal.

4.2.3 Polynomial Rings and Divisibility

Let R be an integral domain.

- **Euclidean Domain (ED)**: A domain with a division algorithm (e.g., $Z, F[x]$).
- **Principal Ideal Domain (PID)**: A domain where every ideal is generated by one element ($I = \langle a \rangle$).
- **Unique Factorization Domain (UFD)**: A domain where every non-zero non-unit factors uniquely into irreducibles.

Theorem 4.2.2: Hierarchy of Domains

$$\text{Fields} \subset \text{Euclidean Domains} \subset \text{PIDs} \subset \text{UFDs} \subset \text{Integral Domains}$$

Theorem 4.2.3: Gauss's Lemma

If R is a UFD, then the polynomial ring $R[x]$ is also a UFD. Consequently, $Z[x]$ is a UFD, even though it is not a PID.

4.3 Fields

Fields are commutative rings where division (by non-zero elements) is always defined. They are the setting for linear algebra and Galois theory.

4.3.1 Extensions

Definition 4.3.1

If $F \subseteq K$ are fields, K is an **extension** of F, denoted K/F. The **degree** $[K : F]$ is the dimension of K as an F-vector space.

Theorem 4.3.1: Tower Law

If $F \subseteq L \subseteq K$, then $[K : F] = [K : L][L : F]$.

Definition 4.3.2

Let $\alpha \in K$.

- α is **algebraic** over F if it is the root of some polynomial $f(x) \in F[x]$.
- The **minimal polynomial** of α is the unique monic irreducible polynomial in $F[x]$ having α as a root.

If all elements of K are algebraic over F, K/F is an **algebraic extension**.

4.3.2 Splitting Fields and Algebraic Closure

Definition 4.3.3: Splitting Field

The splitting field of a polynomial $f(x) \in F[x]$ is the smallest extension K/F in which $f(x)$ decomposes into linear factors $(x - \alpha_1) \dots (x - \alpha_n)$.

Definition 4.3.4: Algebraic Closure

A field $\bar{F}$ is algebraically closed if every non-constant polynomial in $\bar{F}[x]$ has a root in $\bar{F}$. Every field F has a unique (up to isomorphism) algebraic closure.

4.3.3 Finite Fields

Finite fields are fully classified.

Theorem 4.3.2

Let F be a finite field.

1. The characteristic of F is a prime p.
2. The number of elements is $|F| = p^n$ for some $n \geq 1$.
3. For every prime p and integer n, there is a unique finite field of order p^n, denoted F_{p^n} or $GF(p^n)$.
4. F_{p^n} is the splitting field of $x^{p^n} - x$ over F_p.

4.4 Galois Theory

Galois Theory relates field extensions to groups of automorphisms, solving ancient problems like the impossibility of trisecting an angle or solving quintic equations by radicals.

4.4.1 The Galois Correspondence

Definition 4.4.1

Let K/F be an extension. The **Galois Group**, $Gal(K/F)$, is the set of all automorphisms $\sigma : K \to K$ such that $\sigma(a) = a$ for all $a \in F$.

Definition 4.4.2: Galois Extension

An extension K/F is **Galois** if it is:

1. **Normal**: Irreducible polynomials in $F[x]$ with a root in K split completely in K.
2. **Separable**: Irreducible polynomials over F have distinct roots in algebraic closure (no multiple roots).

Theorem 4.4.1: Fundamental Theorem of Galois Theory

Let K/F be a finite Galois extension with Galois group $G = Gal(K/F)$. There is a one-to-one inclusion-reversing correspondence between subgroups $H \leq G$ and intermediate fields $F \subseteq E \subseteq K$. Specifically:

1. The fixed field of H is E.
2. E is a normal extension of F if and only if H is a normal subgroup of G. In this case, $Gal(E/F) \cong G/H$.

4.4.2 Solvability by Radicals

Definition 4.4.3

A group G is **solvable** if there is a chain $1 = G_0 \trianglelefteq G_1 \trianglelefteq \cdots \trianglelefteq G_n = G$ where G_{i+1}/G_i is abelian.

Theorem 4.4.2

A polynomial $f(x)$ is solvable by radicals (using n-th roots) if and only if its Galois group is a solvable group. Since S_n is not solvable for $n \geq 5$, there is no general quintic formula.

4.5 Modules

Modules are generalizations of vector spaces where the scalars come from a ring R rather than a field. This seemingly small change adds significant complexity (e.g., lack of bases).

4.5.1 Definitions

Definition 4.5.1: R-Module

Let R be a ring. A left **R-module** M is an abelian group $(M, +)$ equipped with an action $R \times M \to M$ such that for all $r, s \in R$ and $m, n \in M$:

1. $r(m + n) = rm + rn$.
2. $(r + s)m = rm + sm$.
3. $(rs)m = r(sm)$.
4. $1m = m$ (if R has unity).

Example 4.5.1

- Any vector space over F is an F-module.
- Any abelian group G is a Z-module ($n \cdot g$ is repeated addition).
- R itself is an R-module.
- An ideal $I \subseteq R$ is an R-submodule of R.

4.5.2 Module Homomorphisms and Exact Sequences

Definition 4.5.2

An R-module homomorphism is a map $f : M \to N$ respecting addition and scalar multiplication.

Definition 4.5.3: Exact Sequence

A sequence of modules and homomorphisms

$$\dots \xrightarrow{f_{i-1}} M_i \xrightarrow{f_i} M_{i+1} \xrightarrow{f_{i+1}} \dots$$

is **exact** at M_i if $Img(f_{i-1}) = Ker(f_i)$.

A **Short Exact Sequence** $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ implies A embeds into B and $C \cong B/A$.

4.5.3 Finitely Generated Modules over PIDs

This is the crowning theorem of basic module theory, generalizing the Fundamental Theorem of Finite Abelian Groups and the Jordan Canonical Form.

Theorem 4.5.1: Structure Theorem

Let R be a PID and M a finitely generated R-module. Then M decomposes uniquely as:

$$M \cong R^k \oplus R/\langle d_1 \rangle \oplus R/\langle d_2 \rangle \oplus \cdots \oplus R/\langle d_m \rangle$$

where $k \geq 0$ is the **rank**, and $d_1 \mid d_2 \mid \cdots \mid d_m$ are non-zero non-units called the **invariant factors**.