ROYAL
MILITARY ACADEMY

CALCULUS

Ben Lauwens, PhD MScEng
Colonel IMM
Professor

2025

ES111

Preface

1 Who are we?

- Lecturer: Col IMM Ben LAUWENS, PhD MScEng / D30.19 / ben.lauwens@mil.be
- Assistant: Mr Indy VAN DEN BROECK, MSc / D30.17 / indy.vandenbroeck@mil.be

2 Course

- Theory: 24 Hr → 72 Hr
- Exercises: 22 Hr → 44 Hr

3 Documentation

- Handbook: Calculus, a complete course - 10th edition
  Chapter P-9: theory and exercises
- Handouts: supplements, summaries and other proofs

4 Evaluations

- Test: 13/11 - 2 Hr: Exercises
- Exam:
  - Theory: oral exam - 5 min access to notes, 55 min preparation without notes and 30 min presentation
  - Exercises: written exam - 4 Hr without notes

5 *Schedule*

Lectures	Theory	Exercises	Topic
06/10	1-2		Introduction and Preliminaries
09/10	3-4		Real Numbers and Functions
13/10	5-6		Limits
14/10		1-2	Real Numbers and Functions
20/10	7-8		Continuity
21/10	9-10		Continuity and Derivatives
27/10		3-4	Limits
28/10		5-6	Continuity
03/11	11-12		Derivatives and Differentials
04/11		7-8	Derivatives and Differentials
13/11	13-14		Applications of Differentiation
17/11	15-16		Trancendental Functions
18/11		9-10	Applications of Differentiation
24/11		11-12	Applications of Differentiation
25/11		13-14	Trancendental Functions
01/12	17-18		Integrals
02/12		15-16	Integrals - Integration Techniques
08/12	19-20		Integration Techniques - Applications of Integration
08/12-12/12		17	Integration Techniques
09/12		18-19	Integration Techniques - Applications of Integration
15/12	21-22		Polar Coordinates and Parametric Curves
16/12		20-21	Polar Coordinates and Parametric Curves
22/12	23-24		*Sequences, Infinite Series and Power Series*
23/12		22-23	*Sequences, Infinite Series and Power Series*

TABLE OF CONTENTS

LEXICON

Chapter 1

axiom of induction: axiome de l'induction (m), het inductie axioma

axioms of Peano: axiomes de Peano (m), de axioma's van Peano

bijection: bijection (f), de bijectie

cardinal number: nombre cardinal (m), het cardinaalgetal

cartesian product: produit cartésien (m), het cartesisch product

complement: complément (m), het complement

composite proposition: proposition composée (f), de samengestelde bewering

conditional proposition: proposition conditionelle (f), de voorwaardelijke bewering

conjunction: conjonction (f), de conjunctie

connective: connecteur (m), de connector

contradiction: contradiction (f), de tegenstrijdigheid

converse: inverse (m), het omgekeerde

countable: denombrable, aftelbaar

counterexample: contre-exemple (m), het tegenvoorbeeld

difference: différence (f), het verschil

direct proof: démonstration directe (f), het rechtstreeks bewijs

disjunction: disjonction (f), de disjunctie

domain: domaine (m), het domein

element: élément (m), het element

empty set: ensemble vide (m), de lege verzameling

equal: égal, gelijk

equivalent: équivalent, gelijkwaardig

existence proof: démonstration de l'existence (f), het bestaansbewijs

existential quantifier: quantificateur existentiel, de existentiële kwantificator

expression: expression (f), de uitdrukking

if and only if: si et seulement si, als en slechts als

image: image (f), het beeld

implication: implication (f), de gevolgtrekking

injection: injection (f): de injectie

intersection: intersection (f), de doorsnede

inverse image: image réciproque (f), het inverse beeld

inverse function: fonction réciproque (f), de inverse functie

finite: fini, eindig

function: fonction (f), de functie

laws of De Morgan: lois de De Morgan (f), de wetten van De Morgan

natural number: nombre naturel (m), het natuurlijk getal

negation: négation (f), de ontkenning

ordered pair: paire ordonnée (f), het geordend paar

prime number: nombre premier (m), het priemgetal

primitive proposition: proposition primitive (f), de enkelvoudige bewering

proof: démonstration (f), het bewijs

proof by contradiction: démonstration par l’absurde (f), het bewijs uit het ongeruime

proof by contraposition: démonstration par contraposition (f), het bewijs door contrapositie

proof by induction: démonstration par induction, het bewijs door inductie

proposition: proposition (f), de bewering

propositional function: fonction propositionnelle (f), de propositiefunctie

quantifier: quantificateur (m), de kwantificator

range: étendue (f), het bereik

set: ensemble (m), de verzameling

subset: sousensemble (m), de deelverzameling

successor: successeur (m), de opvolger

supposition: supposition (f), de veronderstelling

surjection: surjection (f), de surjectie

tautology: tautologie (f), de tautologie

template: template (m), het sjabloon

theorem: théorème (m), de stelling

to assert the hypothesis: affirmer l’hypothèse, de hypothese bevestigen

to list the implications: énumérer les implications, de gevolgen oplijsten

to prove: démontrer, bewijzen

to set the context: introduire le contexte, de context inleiden

to state the conclusion: formuler la conclusion, het besluit verwoorden

truth set: ensemble de vérité (m), de waarheidsverzameling

union: union (f), de vereniging

universal quantifier: quantificateur universel (m), de universele kwantificator

universe: univers (m), het universum

variabel: variable (f), de veranderlijke

Chapter 2

absolute value: valeur absolue (f), de absolute waarde

acute angle: angle aigu (m), de scherpe hoek

addition: addition (f), de optelling

arc length: longueur d'arc (f); de booglengte

Archimedean principle: principe d'Archimède (m), het principe van Archimedes

associativity: associativité (f), de associativiteit

binary search: recherche par dichotomie (f), de binaire zoektocht

bisection: dichotomie (f), de bisectie

boundary point: borne (f), het grenspunt

bounded: borné, begrensd

capture theorem: théorème de capture (m), de vangstelling

Cartesian plane: plan cartésien (m), het cartesisch vlak

closed interval: intervalle fermé (m), het gesloten interval

commutativity: commutativité (f), de commutativiteit

completeness: complétude (f), de volledigheid

composite function: fonction composée (f), de samengestelde functie

composition: composition (f), de samenstelling

connected set: ensemble connecté (m), de samenhangende verzameling

coordinate: coordonnée (f), de coördinaat

coordinate axis: axe de coordonnées (m), de coördinaatsas

coordinate system: système de coordonnées (m), het assenstelsel

counterclockwise: contre le sens des aiguilles, tegenuurwijzerszin

cosecans: cosécante (f), de cosecans

cosine: cosinus (m), de cosinus

cotangent: cotangente (f), de cotangens

curve: courbe (f), de kromme

decimal expansion: expansion décimale (f), de decimale expansie

denominator: dénominateur (m), de noemer

dense: dense, dicht

distributivity: distributivité (f), de distributiviteit

divison: division (f), de deling

even function: fonction paire (f), de even functie

field: champ (m), het veld

floating point number: nombre à virgule flottante (m), het drijvende-kommagetal

fraction: rapport (m), de breuk

graph: graphe (m), de grafiek

Hein-Borel theorem: théorème de Heine-Borel (m), de stelling van Heine-Borel identity: identité (f), de identiteit

infimum: borne inférieure (f), het infimum

infinity: infinité (f), de oneidigheid

integer: nombre entier (m), het geheel getal

interval: intervalle (m), het interval

inverse number: nombre inverse (m), het omgekeerde getal

irreducible: irréductible, onvereenvoudigbaar

lower bound: minorant (m), de ondergrens

metric space: espace métrique (m), de metrische ruimte

multiplication: multiplication (f), het product

numerator: numérateur (m), de teller

odd function: fonction impaire (f), de oneven functie

open interval: intervalle ouvert (m), het open interval

operation: opération (f), de operatie

opposite number: nombre opposé (m), het tegengestelde getal

ordered: ordonné, geordend

origin: origine (f), de oorsprong

polynomial: polynôme (m), de veelterm

product: produit (m), het product

proof by bisection: démonstration par dichotomie (f), het bewijs door bisectie

Pythagorean theorem: théorème de Pythagore (m), de stelling van Pythagoras

radian: radian (m), de radiaal

rational function: fonction rationnelle (f), de rationale functie

rational number: nombre rationnel (m), het rationaal getal

real line: droite réelle (f), de reële as

real number: nombre réel (m), het reëel getal

relation: relation (f), de relatie

upper bound: majorant (m), de bovengrens

secans: sécante (f), de secant

sector area: aire de secteur (f), de sectoroppervlakte

sine: sinus (m), de sinus

smooth curve: courbe lisse (f), de gladde kromme

subtraction: soustraction (f), het verschil

sum: sommation (f), de som

supremum: borne supérieure (f), het supremum

tangent: tangente (f), de tangens

theorem of nested intervals: théorème des intervalles imbriqués (m), de stelling van de geneste intervallen

transitivity: transitivité (f), de transitiviteit

triangle inequality: inégalité triangulaire (f), de driehoeksongelijkheid trigonometric function: fonction trigonométrique (f), de goniometrische functie

CHAPTER 1

PRELIMINARIES

This chapter introduces mathematical logic as a tool for rigorously proving statements. An elementary introduction to set theory and functions is also included.

1–1 Propositions

Definition 1. A *proposition* is a declarative sentence which is true T or false F, but not both.

Many propositions are *composite*, that is, composed of *subpropositions* and various *connectives* discussed subsequently. Such composite propositions are called *compound propositions*. A proposition is said to be *primitive* if it cannot be broken down into simpler propositions.

Example. Some primitive propositions:

- Brussels is in Belgium: T
- $2 + 2 = 3$: F
- $x = 2$ is a solution of $x^2 = 4$: T

The fundamental property of a compound proposition is that its truth value is completely determined by the truth values of its subpropositions together with the way in which they are connected to form the compound proposition.

1–2 Connectives

Any two propositions $\mathcal{P}$ and $\mathcal{Q}$ can be combined by the word "and" to form a compound proposition called the *conjunction* of $\mathcal{P}$ and $\mathcal{Q}$, denoted $\mathcal{P} \wedge \mathcal{Q}$ and read "$\mathcal{P}$ and $\mathcal{Q}$". If $\mathcal{P}$ and $\mathcal{Q}$ are true, then $\mathcal{P} \wedge \mathcal{Q}$ is true; otherwise $\mathcal{P} \wedge \mathcal{Q}$ is false. The truth value of $\mathcal{P} \wedge \mathcal{Q}$ may be defined equivalently by the following truth table:

$\mathcal{P}$	$\mathcal{Q}$	$\mathcal{P} \wedge \mathcal{Q}$
T	T	T
T	F	F
F	T	F
F	F	F

Example. Some conjunctions:

- Brussels is the capital of Belgium and the home town of the Royal Military Academy: T
- $2 + 2 = 3$ and $x = 2$ is a solution of $x^2 = 4$: F

Any two propositions $\mathcal{P}$ and $\mathcal{Q}$ can be combined by the word "or" to form a compound proposition called the *disjunction* of $\mathcal{P}$ and $\mathcal{Q}$, denoted $\mathcal{P} \vee \mathcal{Q}$ and read "$\mathcal{P}$ or $\mathcal{Q}$". If $\mathcal{P}$ and $\mathcal{Q}$ are false, then $\mathcal{P} \vee \mathcal{Q}$ is false; otherwise $\mathcal{P} \vee \mathcal{Q}$ is true. The truth value of $\mathcal{P} \vee \mathcal{Q}$ may be defined equivalently by the following truth table:

$\mathcal{P}$	$\mathcal{Q}$	$\mathcal{P} \vee \mathcal{Q}$
T	T	T
T	F	T
F	T	T
F	F	F

Example. Some disjunctions:

- Belgium is the capital of Brussels or Brussels is the biggest city in Europe: F
- $2 + 2 = 3$ or $x = 2$ is a solution of $x^2 = 4$: T

Any proposition $\mathcal{P}$ can be preceded by the word "not" to form a new proposition called the *negation* of $\mathcal{P}$, denoted $\neg\mathcal{P}$ and read "not $\mathcal{P}$". If $\mathcal{P}$ is true, then $\neg\mathcal{P}$ is false; and if If $\mathcal{P}$ is false, then $\neg\mathcal{P}$ is true. The truth value of $\neg\mathcal{P}$ may be defined equivalently by the following truth table:

$\mathcal{P}$	$\neg\mathcal{P}$
T	F
F	T

Example. Some negations:

- Belgium is not the capital of Brussels: T
- $x = 2$ is not a solution of $x^2 = 4$: F

A proposition containing only T in the last column of its truth table is called a *tautology* denoted $\top$, eg.

$\mathcal{P}$	$\neg\mathcal{P}$	$\mathcal{P} \vee \neg\mathcal{P}$
T	F	T
F	T	T

Example. A person is a combattant or a non-combattant. (*Law of Armed Conflicts*)

A proposition containing only F in the last column of its truth table is called a *contradiction* denoted $\perp$, eg.

$\mathcal{P}$	$\neg\mathcal{P}$	$\mathcal{P} \wedge \neg\mathcal{P}$
T	F	F
F	T	F

Example. The cat is dead and alive. (*Quantum Mechanics*)

1–3 Logical Equivalence

The propositions $\mathcal{P}$ and $\mathcal{Q}$ are said to be *logically equivalent*, denoted by $\mathcal{P} \equiv \mathcal{Q}$ if they have identical truth tables.

Example. Show that $\neg(\mathcal{P} \wedge \mathcal{Q}) \equiv \neg\mathcal{P} \vee \neg\mathcal{Q}$.

Column 4 and 7 of the following truth table are identical, so the propositions are logically equivalent.

$\mathcal{P}$	$\mathcal{Q}$	$\mathcal{P} \wedge \mathcal{Q}$	$\neg(\mathcal{P} \wedge \mathcal{Q})$	$\neg\mathcal{P}$	$\neg\mathcal{Q}$	$\neg\mathcal{P} \vee \neg\mathcal{Q}$
T	T	T	F	F	F	F
T	F	F	T	F	T	T
F	T	F	T	T	F	T
F	F	F	T	T	T	T

The basic rules, called laws, governing propositions are stated in the following theorem.

Theorem 1. *Let $\mathcal{P}$, $\mathcal{Q}$ and $\mathcal{R}$ be propositions.*

1. $\mathcal{P} \vee \mathcal{P} \equiv \mathcal{P}$ *and* $\mathcal{P} \wedge \mathcal{P} \equiv \mathcal{P}$ *(idempotent laws)*
2. $\mathcal{P} \vee \mathcal{Q} \equiv \mathcal{Q} \vee \mathcal{P}$ *and* $\mathcal{P} \wedge \mathcal{Q} \equiv \mathcal{Q} \wedge \mathcal{P}$ *(commutative laws)*
3. $\mathcal{P} \vee (\mathcal{Q} \vee \mathcal{R}) \equiv (\mathcal{P} \vee \mathcal{Q}) \vee \mathcal{R} \equiv \mathcal{P} \vee \mathcal{Q} \vee \mathcal{R}$ *and*
   $\mathcal{P} \wedge (\mathcal{Q} \wedge \mathcal{R}) \equiv (\mathcal{P} \wedge \mathcal{Q}) \wedge \mathcal{R} \equiv \mathcal{P} \wedge \mathcal{Q} \wedge \mathcal{R}$ *(associative laws)*
4. $\mathcal{P} \wedge (\mathcal{Q} \vee \mathcal{R}) \equiv (\mathcal{P} \wedge \mathcal{Q}) \vee (\mathcal{P} \wedge \mathcal{R})$ *and*
   $\mathcal{P} \vee (\mathcal{Q} \wedge \mathcal{R}) \equiv (\mathcal{P} \vee \mathcal{Q}) \wedge (\mathcal{P} \vee \mathcal{R})$ *(distributive laws)*
5. $\mathcal{P} \vee \bot \equiv \mathcal{P}$ *and* $\mathcal{P} \wedge \bot \equiv \bot$ *(identity laws)*
6. $\mathcal{P} \vee \top \equiv \top$ *and* $\mathcal{P} \wedge \top \equiv \mathcal{P}$ *(identity laws)*
7. $\mathcal{P} \vee \neg\mathcal{P} \equiv \top$ *and* $\mathcal{P} \wedge \neg\mathcal{P} \equiv \bot$ *(complement laws)*
8. $\neg\top \equiv \bot$ *and* $\neg\bot \equiv \top$ *(complement laws)*
9. $\neg(\neg\mathcal{P}) \equiv \mathcal{P}$ *(involution law)*
10. $\neg(\mathcal{P} \vee \mathcal{Q}) \equiv \neg\mathcal{P} \wedge \neg\mathcal{Q}$ *and* $\neg(\mathcal{P} \wedge \mathcal{Q}) \equiv \neg\mathcal{P} \vee \neg\mathcal{Q}$ *(De Morgan's laws)*

Exercise. Use a truth table to show the logical equivalence of the other propositions.

1–4 Conditional Propositions

Many statements are of the form "If $\mathcal{P}$ then $\mathcal{Q}$". Such statements are called *conditional propositions*, and are denoted by $\mathcal{P} \implies \mathcal{Q}$. The conditional $\mathcal{P} \implies \mathcal{Q}$ is frequently read "$\mathcal{P}$ implies $\mathcal{Q}$", or "$\mathcal{P}$ only if $\mathcal{Q}$".

Example. If I have exercised then I am hungry.

Another common statement is of the form "$\mathcal{P}$ if and only if $\mathcal{Q}$". Such statements are called *biconditional propositions*, and are denoted by $\mathcal{P} \iff \mathcal{Q}$.

Example. I am hungry if and only if I have skipped a meal.

Their truth values are defined by following truth tables:

$\mathcal{P}$	$\mathcal{Q}$	$\mathcal{P} \implies \mathcal{Q}$	$\neg\mathcal{P}$	$\neg\mathcal{P} \vee \mathcal{Q}$	$\mathcal{P} \iff \mathcal{Q}$	$\mathcal{Q} \implies \mathcal{P}$	$(\mathcal{P} \implies \mathcal{Q}) \wedge (\mathcal{Q} \implies \mathcal{P})$
T	T	T	F	T	T	T	T
T	F	F	F	F	F	T	F
F	T	T	T	T	F	F	F
F	F	T	T	T	T	T	T

Observe that:

- The conditional $\mathcal{P} \implies \mathcal{Q}$ is false only when the first part $\mathcal{P}$ is true and the second part $\mathcal{Q}$ is false. Accordingly, when $\mathcal{P}$ is false, the conditonal $\mathcal{P} \implies \mathcal{Q}$ is true regardless of the truth value of $\mathcal{Q}$.
- $\mathcal{P} \implies \mathcal{Q}$ is logically equivalent to $\neg\mathcal{P} \vee \mathcal{Q}$.
- The biconditional $\mathcal{P} \iff \mathcal{Q}$ is true whenever $\mathcal{P}$ and $\mathcal{Q}$ have the same truth values and false otherwise.
- $\mathcal{P} \iff \mathcal{Q}$ is logically equivalent to $(\mathcal{P} \implies \mathcal{Q}) \wedge (\mathcal{Q} \implies \mathcal{P})$. $\mathcal{Q} \implies \mathcal{P}$ is called the *converse* of $\mathcal{P} \implies \mathcal{Q}$.

1–5 Sets

Definition 2. A *set* is a collection of objects called *elements* of the set.

In general, we denote a set by a capital letter and an element by a lower case letter. If an element a belongs to a set S we write $a \in S$. If a does not belong to S we write $a \notin S$.

A set can be described by listing its elements in braces separated by commas, eg. $\{a, e, i, o, u\}$, or by describing some property held by all elements, eg. $\{x \mid x \text{ is a vowel}\}$.

If each element of a set A belongs to a set B we call A a *subset* of B, written $A \subset B$. If $A \subset B$ and $B \subset A$ we call A and B *equal* and write $A = B$.

Often we restrict our discussion to subsets of a particular set called the *universe* or the *universal set* denoted by Ω, eg. the set of the letters of the roman alphabet.

It is useful to consider a set having no elements at all. This is called the *empty set* and is denoted by $\varnothing$. It is a subset on any set.

A universe Ω can be represented geometrically by the set of points inside a rectangle. In such case subsets of Ω such as A and B are represented by sets of points inside ellipses. Such diagrams, called *Venn diagrams*, often serve to provide geometric intuition regarding possible relationships between sets.

Theorem 2. *If $A \subset B$ and $B \subset C$, then $A \subset C$.*

Proof. *SET THE CONTEXT:* Let A, B and C be sets for which $A \subset B$ and $B \subset C$.

ASSERT THE HYPOTHESIS: Suppose $x \in A$.

LIST IMPLICATIONS:

1. Since $x \in A$, it is true that $x \in B$ by the definition of subset.
2. Since $x \in B$, it is true that $x \in C$ by the definition of subset.

STATE THE CONCLUSION: Therefore, by the definition of subset, $A \subset C$.

1–6 Set Operations

The set of all elements which belong to A or B is called the *union* of A and B and is denoted by $A \cup B$.

The set of all elements which belong to A and B is called the *intersection* of A and B and is denoted by $A \cap B$. Two sets A and B such that $A \cap B = \varnothing$ are called *disjoint sets*.

The set consisting of all elements of A which do not belong to B is called the *difference* of A and B denoted by $A \setminus B$.

The set consisting of all elements of Ω which do not belong to A is called the *complement* of A denoted by $A^c = \Omega \setminus A$.

Theorem 3. *Let A, B and C be sets.*

1. $A \cup A = A$ *and* $A \cap A = A$ *(idempotent laws)*
2. $A \cup B = B \cup A$ *and* $A \cap B = B \cap A$ *(commutative laws)*
3. $A \cup (B \cup C) = (A \cup B) \cup C = A \cup B \cup C$ *and*
   $A \cap (B \cap C) = (A \cap B) \cap C = A \cap B \cap C$ *(associative laws)*
4. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ *and*
   $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ *(distributive laws)*
5. $A \cup \varnothing = A$ *and* $A \cap \varnothing = \varnothing$ *(identity laws)*
6. $A \cup \Omega = \Omega$ *and* $A \cap \Omega = A$ *(identity laws)*
7. $A \cup A^c = \Omega$ *and* $A \cap A^c = \varnothing$ *(complement laws)*
8. $\Omega^c = \varnothing$ *and* $\varnothing^c = \Omega$ *(complement laws)*
9. $\left(A^c\right)^c = A$ *(involution law)*
10. $(A \cup B)^c = A^c \cap B^c$ *and* $(A \cap B)^c = A^c \cup B^c$ *(De Morgan's laws)*
11. $A \setminus B = A \cap B^c$
12. *If* $A \subset B$*, then* $B^c \subset A^c$
13. $A = (A \cap B) \cup \left(A \cap B^c\right)$

Example. Proof the first law of De Morgan.

Proof. *SET THE CONTEXT:* Let A and B be any two sets.

PART 1: $(A \cup B)^c \subset A^c \cap B^c$

ASSERT THE HYPOTHESIS: Suppose $x \in (A \cup B)^c$.

LIST IMPLICATIONS:

1. By the definition of set complement, $x \notin A \cup B$.
2. If $x \in A$ or $x \in B$, then $x \in A \cup B$ which is false.
3. Thus, $x \notin A$ and $x \notin B$, so by the definition of set complement $x \in A^c$ and $x \in B^c$.
4. By the definition of set intersection $x \in A^c \cap B^c$.

CONCLUSION PART 1: Hence, from the definition of subset, it follows that $(A \cup B)^c \subset A^c \cap B^c$.

PART 2: $A^c \cap B^c \subset (A \cup B)^c$

ASSERT THE HYPOTHESIS: Suppose $x \in A^c \cap B^c$.

LIST IMPLICATIONS:

1. By the definition of set intersection, $x \in A^c$ and $x \in B^c$.
2. Thus by the definition of set complement, $x \notin A$ and $x \notin B$.
3. If $x \in A \cup B$, then by the definition of the union, it would follow that $x \in A$ or $x \in B$ which is false.
4. Thus, $x \notin A \cup B$, and by definition of set complement $x \in (A \cup B)^c$.

CONCLUSION PART 2: Hence, from the definition of subset, it follows that $A^c \cap B^c \subset (A \cup B)^c$.

STATE THE CONCLUSION: Therefore, because $(A \cup B)^c$ and $A^c \cap B^c$ are subsets of each other, by the definition of set equality $(A \cup B)^c = A^c \cap B^c$.

Exercise. Proof the other theorems.

1–7 Cartesian Product

The set of all *ordered pairs* of elements (x, y) where $x \in A$ and $y \in B$ is called the *Cartesian product* or *product set* of A and B and is denoted by $A \times B$. In general, $A \times B \neq B \times A$.

The notion of Cartesian product can be generalized to ordered tuples of element $(x, y, z, \ldots)$.

1–8 Functions

Definition 3.
A function f from a set X to a set Y, often written $f : X \to Y$, is a rule which assigns to each $x \in X$ a unique element $y \in Y$.

The element y is called the *image* of x under f and is denoted by $f(x)$. If $A \subset X$, then $f(A)$ is the set of all elements $f(x)$ where $x \in A$ and is called the *image* of A under f. Symbols x and y are called *variables*.

A function $f : X \to Y$ can also be defined as a subset of the Cartesian product $X \times Y$ such that if (x_1, y_1) and (x_2, y_2) are in this subset and $x_1 = x_2$, then $y_1 = y_2$.

The set X is called the *domain* of f and $f(X)$ is called the *range* of f. If $Y = f(X)$ we say that f is from X onto Y and refer to f as a *surjective* function.

If an element $a \in A \subset X$ maps into an element $b \in B \subset Y$, then a is called the *inverse image* of b under f and is denoted by $f^{-1}(b)$. The set of all $x \in X$ for which $f(x) \in B$ is called the *inverse image* of B under f and is denoted by $f^{-1}(B)$.

If $f(a_1) = f(a_2)$ only when $a_1 = a_2$, we say that f is an *injective* function.

If a function $f : X \to Y$ is both surjective and injective, we say there is a one to one correspondence between X and Y and call f a *bijective* function. Given any element $y \in Y$, there will be only one element $f^{-1}(y)$ in X. In such case f^{-1} will define a function from Y to X called the *inverse function*.

1–9 Cardinal Numbers

Two sets A and B are called *equivalent* and we write $A \sim B$ if there exists a one to one correspondence between A and B.

Theorem 4. *If $A \sim B$ and $B \sim C$, then $A \sim C$.*

To determine the number of elements belonging to a set, we need a definition for counting.

Axiom 1 (Peano). Following axioms define the natural numbers.

1. 0 is a natural number.
2. Every natural number has a successor which is also a natural number.
3. 0 is not the successor of any natural number.
4. If the successor of x equals the successor of y, then x equals y.
5. If a statement is true for 0, and if the truth of that statement for a natural number implies its truth for the successor of that natural number, then the statement is true for every natural number. (*Axiom of induction*)

The set of natural number is denoted by $\mathbb{N}$.

A set which is equivalent to the set $\{0,1,2,3,\ldots,n\}$ for some $n \in \mathbb{N}$ is called *finite*; otherwise it is called *infinite*.

An infinite set which is equivalent to the set of natural numbers is called *denumerable*; otherwise it is called *non-denumerable*.

> **Definition 4.**
> The *cardinal number* of the set $\{1,2,3,\ldots,n\}$ as well as any set equivalent to it is defined to be n. The cardinal number of any denumerable set is defined as $\aleph_0$, *aleph null*. The cardinal number of the empty set $\varnothing$ is defined as 0.

The cardinal number of a set S is denoted by $\#S$.

1–10 Quantifiers

Let A be a given set. A *propositional function* defined on A is a function $\mathcal{P} : A \to \{T, F\}$ which has the property that $\mathcal{P}(a)$ is true or false for each $a \in A$. That is, $\mathcal{P}(x)$ becomes a proposition (with a truth value) whenever any element $a \in A$ is substituted for the variable x. The set $T_{\mathcal{P}}$ of all elements of $a \in A$ for which $\mathcal{P}(a)$ is true is called the *truth set* of $\mathcal{P}(x)$.

> **Example.** Find the truth set $T_{\mathcal{P}}$ of each propositional function $\mathcal{P}(x)$ defined on $\mathbb{N}$.
>
> - Let $\mathcal{P}(x)$ be "$x+2>7$". Then $T_{\mathcal{P}} = \{x \mid x \in \mathbb{N}, x+2>7\} = \{6,7,8,\ldots\}$.
> - Let $\mathcal{P}(x)$ be "$x+5<3$". Then $T_{\mathcal{P}} = \{x \mid x \in \mathbb{N}, x+5<3\} = \varnothing$. In other words, $\mathcal{P}(x)$ is false for any natural number.
> - Let $\mathcal{P}(x)$ be "$x+5>1$". Then $T_{\mathcal{P}} = \{x \mid x \in \mathbb{N}, x+5>1\} = \mathbb{N}$. Thus $\mathcal{P}(x)$ is true for every natural number.

Let $\mathcal{P}(x)$ be a propositional function defined on a set A. Consider the expression "$\forall x \in A : \mathcal{P}(x)$" which reads "For every x in A, $\mathcal{P}(x)$ is a true statement". The symbol $\forall$ which reads "for all" or "for every" is called the *universal quantifier*. The proposition $\forall x \in A : \mathcal{P}(x)$ expresses that the truth set of $\mathcal{P}(x)$ is the entire set A, or symbolically, $T_{\mathcal{P}} = \{x \mid x \in A, \mathcal{P}(x)\} = A$.

> **Example.** Some propositions using the universal quantifier:
>
> - The proposition $\forall n \in \mathbb{N} : n+4>3$ is true since $\{n \mid n \in \mathbb{N}, n+4>3\} = \mathbb{N}$.
> - The proposition $\forall n \in \mathbb{N} : n+2>8$ is false since $\{n \mid n \in \mathbb{N}, n+2>8\} = \{7,8,9,\ldots\}$.
> - The symbol $\forall$ can be used to define the intersection of an indexed collection $\{A_i \mid i \in I\}$ of sets A_i as follows: $\bigcap_{i \in I} A_i = \{x \mid \forall i \in I : x \in A_i\}$.

Let $\mathcal{P}(x)$ be a propositional function defined on a set A. Consider the expression "$\exists x \in A : \mathcal{P}(x)$" which reads "There exists an x in A such that $\mathcal{P}(x)$ is a true statement". The symbol $\exists$ which reads "there exists" or "for some" or "for at least one" is called the *existential quantifier*. The proposition $\exists x \in A : \mathcal{P}(x)$ expresses that the truth set of $\mathcal{P}(x)$ is not the emptyset, or symbolically, $T_{\mathcal{P}} = \{x \mid x \in A, \mathcal{P}(x)\} \neq \varnothing$.

> **Example.** Some propositions using the existential quantifier:
>
> - The proposition $\exists n \in \mathbb{N} : n+4<7$ is true since $\{n \mid n \in \mathbb{N}, n+4<7\} = \{0,1,2\} \neq \varnothing$.
> - The proposition $\exists n \in \mathbb{N} : n+6<4$ is false since $\{n \mid n \in \mathbb{N}, n+6<4\} = \varnothing$.
> - The symbol $\exists$ can be used to define the intersection of an indexed collection $\{A_i \mid i \in I\}$ of sets A_i as follows: $\bigcup_{i \in I} A_i = \{x \mid \exists i \in I : x \in A_i\}$.

Consider the proposition: “All officers are engineers”. Its negation is either of the following equivalent statements:

- “It is not the case that all officers are engineers”.
- “There exists at least one officer who is not an engineer”.

Symbolically, using M to denote the set of officers, the above can be written as

$$\neg(\forall x \in M : x \text{ is an engineer }) \equiv \exists x \in M : x \text{ is not an engineer}\,,$$

or, when $\mathcal{P}(x)$ denotes “x is an engineer”,

$$\neg(\forall x \in M : \mathcal{P}(x)) \equiv \exists x \in M : \neg\mathcal{P}(x)\,.$$

The above is true for any proposition $\mathcal{P}(x)$.

Theorem 5 (De Morgan 1). $\neg(\forall x \in A : \mathcal{P}(x)) \equiv \exists x \in A : \neg\mathcal{P}(x)$.

In other words, the following two statements are equivalent:

- It is not true that, for all $a \in A$, $\mathcal{P}(a)$ is true.
- There exists an $a \in A$ such that $\mathcal{P}(a)$ is false.

There is an analogous theorem for the negation of a proposition which contains the existential quantifier.

Theorem 6 (De Morgan 2). $\neg(\exists x \in A : \mathcal{P}(x)) \equiv \forall x \in A : \neg\mathcal{P}(x)$.

That is, the following two statements are equivalent:

- It is not true that, for some $a \in A$, $\mathcal{P}(a)$ is true.
- For all $a \in A$, $\mathcal{P}(a)$ is false.

Previously, $\neg$ was used as an operation on propositions, here $\neg$ is used as an operation on propositional functions. The operations $\vee$ and $\wedge$ can also be applied to propositional functions.

1. The truth set of $\neg\mathcal{P}(x)$ is the complement of $T_{\mathcal{P}}$, that is $T_{\mathcal{P}}^{c}$.
2. The truth set of $\mathcal{P}(x) \vee \mathcal{Q}(x)$ is the union of $T_{\mathcal{P}}$ and $T_{\mathcal{Q}}$, that is $T_{\mathcal{P}} \cup T_{\mathcal{Q}}$.
3. The truth set of $\mathcal{P}(x) \wedge \mathcal{Q}(x)$ is the intersection of $T_{\mathcal{P}}$ and $T_{\mathcal{Q}}$, that is $T_{\mathcal{P}} \cap T_{\mathcal{Q}}$.

A propositional function of 2 variables defined over a product set $A = A_1 \times A_2$ is a function $A_1 \times A_2 \to \{T, F\} : \mathcal{P}(x_1, x_2)$ which has the property that $\mathcal{P}(a_1, a_2)$ is true or false for any pair (a_1, a_2) in A.

A propositional function can be generalized over a product set of more than 2 sets. A propositional function preceded by a quantifier for each variable denotes a proposition and has a truth value.

Example.
Let $B = \{1,2,3,\ldots,9\}$ and let $\mathcal{P}(x, y)$ denotes “$x + y = 10$”. Then $\mathcal{P}(x, y)$ is a propositional function on $A = B \times B$.

1. The following is a proposition since there is a quantifier for each variable: $\forall x \in B, \exists y \in B : \mathcal{P}(x, y)$ that is, “For every x in B, there exists a y in B such that $x + y = 10$”. This statement is true.
2. The following is also a proposition: $\exists y \in B, \forall x \in B : \mathcal{P}(x, y)$ that is, “There exists a y in B such that, for every x in B, we have $x + y = 10$”. No such y exists; hence the statement is false.

Observe that the only difference between both examples is the order of the quantifiers. Thus a different ordering of the quantifiers may yield a different statement!

Theorem 7. *For any propositional function* $\mathcal{P}(x,y)$ *:*

1. $\forall x \in A, \forall y \in B : \mathcal{P}(x,y) \iff \forall y \in B, \forall x \in A : \mathcal{P}(x,y)$.
2. $\exists x \in A, \exists y \in B : \mathcal{P}(x,y) \iff \exists y \in B, \exists x \in A : \mathcal{P}(x,y)$.
3. $\exists x \in A, \forall y \in B : \mathcal{P}(x,y) \implies \forall y \in B, \exists x \in A : \mathcal{P}(x,y)$.
4. $\forall x \in A, \exists y \in B : \mathcal{P}(x,y) \not\Longrightarrow \exists y \in B, \forall x \in A : \mathcal{P}(x,y)$.

Quantified statements with more than one variable may be negated by successively applying the theorems of De Morgan. Thus each $\forall$ is changed to $\exists$, and each $\exists$ is changed to $\forall$ as the negation symbol $\neg$ passes through the statement from left to right.

Example. Some examples of the negation of quantified statement with more than one variable:

- $$\begin{aligned}&\neg(\forall x \in A, \exists y \in B, \exists z \in C : \mathcal{P}(x,y,z)) \equiv \exists x \in A, \neg(\exists y \in B, \exists z \in C : \mathcal{P}(x,y,z))\\ &\equiv \exists x \in A, \forall y \in B, \neg(\exists z \in C : \mathcal{P}(x,y,z)) \equiv \exists x \in A, \forall y \in B, \forall z \in C : \neg\mathcal{P}(x).\end{aligned}$$
- Consider the proposition: "Every student has at least one course where the lecturer is an officer". Its negation is the statement: "There is a student such that all his courses have a lecturer which is not an officer".

1–11 Proofs

Many proofs can be written by following a simple *template* that suggests guidelines to follow when writing the proof.

Direct Proof

To prove $\mathcal{P} \implies \mathcal{Q}$, we can proceed by looking at the truth table. The table shows that if $\mathcal{P}$ is false, the statement $\mathcal{P} \implies \mathcal{Q}$ is automatically true. This means that if we are concerned with showing $\mathcal{P} \implies \mathcal{Q}$ is true, we don't have to worry about the situations where $\mathcal{P}$ is false because the statement $\mathcal{P} \implies \mathcal{Q}$ will be automatically true in those cases. But we must be very careful about the situations where $\mathcal{P}$ is true. We must show that the condition of $\mathcal{P}$ being true forces $\mathcal{Q}$ to be true also.

Template 1 (direct proof). *SET THE CONTEXT*

ASSERT THE HYPOTHESIS

LIST IMPLICATIONS

STATE THE CONCLUSION

Example. Prove the proposition "The sum of any two odd natural numbers is even".

Proof. *SET THE CONTEXT:* Let m and n be two natural numbers.

ASSERT THE HYPOTHESIS: Suppose m and n are odd.

LIST IMPLICATIONS:

1. From the definition of odd natural numbers, there is a natural number k_1 such that $m = 2k_1 + 1$ and a natural number k_2 such that $n = 2k_2 + 1$.
2. Then $m + n = (2k_1 + 1) + (2k_2 + 1) = 2(k_1 + k_2 + 1)$.
3. Since k_1 and k_2 are natural numbers, so is $k_1 + k_2 + 1$.
4. Thus, the sum $m + n$ is equal to twice a natural number, so by the definition of even natural numbers, $m + n$ is even.

STATE THE CONCLUSION: Therefore, the sum of any two odd natural number is always even.

In proving a statement is true, we sometimes have to examine multiple cases before showing the statement is true in all possible scenarios.

Template 2 (direct proof with multiple cases). *SET THE CONTEXT*

CASE 1: ASSERT THE HYPOTHESIS

CASE 1: LIST IMPLICATIONS

CASE 1: STATE THE CONCLUSION

CASE 2: ASSERT THE HYPOTHESIS

CASE 2: LIST IMPLICATIONS

CASE 2: STATE THE CONCLUSION

...

STATE THE GENERAL CONCLUSION

Example. Prove the proposition "If $n \in \mathbb{N}$, then $1 + (-1)^n(2n - 1)$ is a multiple of 4".

Proof. Let n be a natural number.

CASE 1: Suppose n is even. Then $n = 2k$ for some $k \in \mathbb{N}$, and $(-1)^n = 1$. Thus $1 + (-1)^n(2n - 1) = 1 + (1)(2 \cdot 2k - 1) = 4k$, which is a multiple of 4.

CASE 2: Suppose n is odd. Then $n = 2k + 1$ for some $k \in \mathbb{N}$, and $(-1)^n = -1$. Thus $1 + (-1)^n(2n - 1) = 1 - (2(2k + 1) - 1) = -4k$, which is a multiple of 4.

These cases show that $1 + (-1)^n(2n - 1)$ is always a multiple of 4.

Proof by Contraposition

Sometimes a direct proof of $\mathcal{P} \implies \mathcal{Q}$ is very hard. The proposition $\neg\mathcal{Q} \implies \neg\mathcal{P}$ is logically equivalent to $\mathcal{P} \implies \mathcal{Q}$. This is called the *contraposition* of the initial proposition.

Exercise. Show $(\neg\mathcal{Q} \implies \neg\mathcal{P}) \equiv (\mathcal{P} \implies \mathcal{Q})$ using a truth table.

Template 3 (proof by contraposition). *SET THE CONTEXT*

ASSERT THE HYPOTHESIS: $\neg\mathcal{Q}$ is true

LIST IMPLICATIONS

STATE THE CONCLUSION: $\neg\mathcal{P}$ is true

Example. Prove by contraposition the proposition "Let $x \in \mathbb{N}$. If x^2 is even, then x is even".

A direct proof would be problematic. We will proof the logically equivalent proposition "If x is not even, then x^2 is not even".

Proof (by contraposition). Let $x \in \mathbb{N}$.

Suppose x is not even.

Then x is odd and $x = 2k + 1$ for some $k \in \mathbb{N}$. Thus $x^2 = (2k+1)^2 = 2\left(2k^2 + 2k\right) + 1$. Since k is a natural number, $2k^2 + 2k$ is also a natural number. Consequently, x^2 is odd.

Therefore, x^2 is not even.

Proof by Contradiction

A proof by *contradiction* is not limited to proving just conditional statements—it can be used to prove any kind of statement whatsoever. The basic idea is to assume that the statement we want to prove is false, and then show that this assumption leads to nonsense. We are then led to conclude that we were wrong to assume the statement was false, so the statement must be true.

Exercise. Show $\mathcal{P} \equiv (\neg\mathcal{P} \implies (\mathcal{C} \wedge \neg\mathcal{C}))$ using a truth table.

Template 4 (proof by contradiction). *SET THE CONTEXT*

ASSERT THE HYPOTHESIS: $\mathcal{P}$ is false.

LIST IMPLICATIONS

STATE THE CONCLUSION: $\mathcal{C} \wedge \neg\mathcal{C}$.

A slightly unsettling feature of this method is that we may not know at the beginning of the proof what the statement $\mathcal{C}$ is going to be.

Example. Prove by contradiction the proposition "If $a, b \in \mathbb{N}$, then $a^2 - 4b \neq 2$".

Proof (by contradiction). Let $a, b \in \mathbb{N}$.

Suppose there exist a and b for which $a^2 - 4b = 2$. From this equation we get $a^2 = 4b + 2 = 2(2b + 1)$, so a^2 is even.

Because a^2 is even, it follows that a is even, so $a = 2c$ for some natural number c. Now plug $a = 2c$ back into the boxed equation to get $(2c)^2 - 4b = 2$, so $4c^2 - 4b = 2$. Dividing by 2, we get $2c^2 - 2b = 1$.

Therefore, $1 = 2\left(c^2 - b\right)$, and because $c^2 - b \in \mathbb{N}$, it follows that 1 is even.

We know 1 is **not** even, so something went wrong. But all the logic after the first line of the proof is correct, so it must be that the first line was incorrect. In other words, we were wrong to assume the proposition was false. Thus the proposition is true.

The previous two proof methods dealt exclusively with proving conditional statements, we now formalize the procedure in which contradiction is used to prove a conditional statement. Thus we need to prove that $\mathcal{P} \implies \mathcal{Q}$ is true. Proof by contradiction begins with the assumption that $\neg(\mathcal{P} \implies Q)$ is true, that is, that $\mathcal{P} \implies \mathcal{Q}$ is false. But we know that $\mathcal{P} \implies \mathcal{Q}$ being false means that it is possible that $\mathcal{P}$ can be true while $\mathcal{Q}$ is false. Thus the first step in the proof is to assume $\mathcal{P}$ and $\neg\mathcal{Q}$.

Template 5 (proof by contradiction of a conditional proposition). *SET THE CONTEXT*

ASSERT THE HYPOTHESIS: $\mathcal{P}$ and $\neg\mathcal{Q}$ are true.

LIST IMPLICATIONS

STATE THE CONCLUSION: $\mathcal{C} \wedge \neg\mathcal{C}$.

Example. Prove by contradiction the proposition "Let $a \in \mathbb{N}$. If a^2 is even, then a is even".

Proof (by contradiction). Let $a \in \mathbb{N}$.

Suppose a^2 is even and a is not even.

Since a is odd, there exists a natural number c for which $a = 2c + 1$. Then $a^2 = (2c+1)^2 = 4c^2 + 4c + 1 = 2\left(2c^2 + 2c\right) + 1$, so a^2 is odd.

Thus a^2 is even and a^2 is not even, a contradiction.

If-and-only-if Proof

Some propositions have the form $\mathcal{P} \iff \mathcal{Q}$. We know that this is logically equivalent to $(\mathcal{P} \implies \mathcal{Q}) \wedge (\mathcal{Q} \implies \mathcal{P})$. So to prove "$\mathcal{P}$ if and only if $\mathcal{Q}$" we must prove **two** conditional statements. Recall that $\mathcal{Q} \implies \mathcal{P}$ is called the *converse* of $\mathcal{P} \implies \mathcal{Q}$. Thus we need to prove both $\mathcal{P} \implies \mathcal{Q}$ and its converse. These are both conditional statements, so we may prove them with either direct, contrapositive or contradiction proof.

Example. Prove the proposition "The natural number n is odd if and only if n^2 is odd".

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

First we show that n being odd implies that n^2 is odd.

Suppose n is odd.

Then, by definition of an odd number, $n = 2a + 1$ for some natural number. Thus $n^2 = (2a+1)^2 = 4a^2 + 4a + 1 = 2\left(2a^2 + 2a\right) + 1$. This expresses n^2 as twice a natural number, plus 1.

Therefore, n^2 is odd.

Conversely, we need to prove that n^2 being odd implies that n is odd. We use contraposition and proof the proposition "n not odd implies that n^2 is not odd".

Suppose n is not odd.

Then n is even, so $n = 2a$ for some natural number a by definition of an even number. Thus $n^2 = (2a)^2 = 2\left(2a^2\right)$, so n^2 is even because it's twice a natural number.

Therefore, n^2 is not odd.

Existence Proof

Up until this point, we have dealt with proving conditional statements or with statements that can be expressed with two or more conditional statements. Generally, these conditional statements have form $\mathcal{P}(x) \implies \mathcal{Q}(x)$(Possibly with more than one variable). We saw that this can be interpreted as a universally quantified statement $\forall x : \mathcal{P}(x) \implies \mathcal{Q}(x)$.

But how would we prove an *existentially* quantified statement? What technique would we employ to prove a theorem of the form $\exists x : \mathcal{P}(x)$. This statement asserts that there exists some specific object x for which $\mathcal{P}(x)$ is true. To prove $\exists x : \mathcal{P}(x)$ is true, all we would have to do is find and display an *example* of a specific x that makes $\mathcal{P}(x)$ true.

Example.
There exists a natural number that can be expressed as the sum of two perfect cubes in two different ways.

Proof. Consider the number 1729.

Note that $1^3 + 12^3 = 1729$ and $9^3 + 10^3 = 1729$.

Therefore, the number 1729 can be expressed as the sum of two perfect cubes in two different ways.

Counterexamples

How to disprove a universally quantified statement such as $\forall x : \mathcal{P}(x)$? To disprove this statement, we must prove its negation. Its negation is $\neg(\forall x : \mathcal{P}(x) \implies \mathcal{Q}(x)) \equiv \exists x : \neg(\mathcal{P}(x) \implies \mathcal{Q}(x))$. The negation is an existence statement. To prove the negation is true, we just need to produce an example of an x that makes $\mathcal{P}(x)$ false.

Example.
Disprove the proposition "For every $n \in \mathbb{N}$, the natural number $f(n) = n^2 - n + 11$ is prime".

Proof.
The statement "For every $n \in \mathbb{N}$, the natural number $f(n) = n^2 - n + 11$ is prime," is false. For a counterexample, note that for $n = 11$, the natural number $f(11) = 121 = 11 \cdot 11$ is not prime.

Proof by Induction

Suppose the variable n represents any natural number, and there is a propositional function $\mathcal{P}(n)$ that includes this variable as an argument. *Mathematical induction* is a proof technique that uses the axiom of induction to show that $\mathcal{P}(n)$ is true for all n greater than or equal to some base value $b \in \mathbb{N}$.

Template 6 (proof by induction).
SET THE CONTEXT: The statement will be proved by mathematical induction on n for all $n \geq b$.

PROVE $\mathcal{P}(b)$ *:* Prove that the statement is true when the variable n is equal to the base value, b.

STATE THE INDUCTION HYPOTHESIS: Assume that $\mathcal{P}(n)$ is true for some natural number $n = k \geq b$.

PERFORM THE INDUCTION STEP: Using the fact that $\mathcal{P}(k)$ is true, prove that $\mathcal{P}(k+1)$ is true.

STATE THE CONCLUSION: Therefore, by mathematical induction, $\mathcal{P}(n)$ is true for all natural numbers $n \geq b$.

Example.
Prove by induction the proposition "If $n \in \mathbb{N} \setminus \{0\}$, then $1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$".

Proof (by induction). Let $n \in \mathbb{N}_0$.

Observe that if $n = 1$, this statement is $1 = \frac{1}{6}(1)(2)(3)$, which is obviously true.

Suppose that $1^2 + 2^2 + 3^2 + \cdots + k^2 = \frac{1}{6}k(k+1)(2k+1)$ for some natural number $k \geq 1$.

Then,

$$\begin{aligned}
1^2 + 2^2 + 3^2 + \cdots + k^2 + (k+1)^2 &= \frac{1}{6}k(k+1)(2k+1) + (k+1)^2 \\
&= \frac{1}{6}\left(2k^3 + (3+6)k^2 + (1+12)k + 6\right) \\
&= \frac{1}{6}\left(2k^3 + 9k^2 + 13k + 6\right) \\
&= \frac{1}{6}(k+1)(k+2)(2k+3) \\
&= \frac{1}{6}(k+1)((k+1)+1)(2(k+1)+1)
\end{aligned}$$

Therefore, by mathematical induction, $1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$ for all natural numbers $n \geq 1$.

CHAPTER 2

Real Numbers and Functions

In this chapter, the real numbers are formally defined and some basic real functions are explored.

2–1 Natural Numbers and Extensions

The axioms of Peano define the *natural numbers*, denoted by $\mathbb{N}$.

To close the set of numbers under the subtraction operation, the natural numbers are extended with the negative integers to form the set of *integers*, denoted by $\mathbb{Z}$.

To close the set of numbers under the division operation, numbers that can be expressed as an irreducible fraction $\frac{m}{n}$, where $m \in \mathbb{Z}$ and $n \in \mathbb{Z} \setminus \{0, 1\}$ are added to the integers to form the set of *rational numbers*, denoted by $\mathbb{Q}$.

This definition comes with the understanding that the two rational numbers $\frac{m}{n}$ and $\frac{a}{b}$ are equal whenever $mb = na$.

2–2 Algebraic Properties of Rational Numbers

The set of rational number is more than a set of fractions with integers for numerators and denominators. It also comes with the two binary operations of *addition* and *multiplication*.

Axiom 2 (Algebraic properties).
A set $\mathbb{F}$ together with the binary operations of addition and multiplication form a *field* if $\mathbb{F}$ contains two elements 0 and 1 with $0 \neq 1$ such that $\forall a, b, c \in \mathbb{F}$:

1. $a + b = b + a$ and $ab = ba$ (*commutativity*)
2. $a + (b + c) = (a + b) + c = a + b + c$ and $a(bc) = (ab)c = abc$ (*associativity*)
3. $a + 0 = a$ and $a \cdot 1 = a$ (*identity*)
4. There exists $-a \in \mathbb{F}$ such that $a + (-a) = 0$ (*opposite*) and for $a \neq 0$ there exists $a^{-1} = \frac{1}{a} \in \mathbb{F}$ such that $a \cdot a^{-1} = 1$ (*inverse*)
5. $a(b + c) = ab + ac$ (*distributivity*)

Notice that the rational numbers do satisfy the field axioms.

Following properties can be derived from the algebraic properties.

Theorem 8. *Let $a, b, c \in \mathbb{Q}$.*

1. $a + c = b + c \implies a = b$
2. $a \cdot 0 = 0$
3. $(-a) \cdot b = -ab$
4. $c \neq 0 \wedge ac = bc \implies a = b$
5. $ab = 0 \implies a = 0 \vee b = 0$

Example. Prove the second property.

Proof. Let $a \in \mathbb{Q}$.

Since 0 is the additive identity, $0 = 0 + 0$. Then $a \cdot 0 = a \cdot (0 + 0)$. By the distributive property, $a \cdot 0 = a \cdot 0 + a \cdot 0$. By adding $-a \cdot 0$ to each side of this equality, one gets

$$0 = a \cdot 0 - a \cdot 0 = (a \cdot 0 + a \cdot 0) - a \cdot 0 = a \cdot 0 + (a \cdot 0 - a \cdot 0) = a \cdot 0 + 0 = a \cdot 0 .$$

Therefore, $a \cdot 0 = 0$

Exercise. Prove the other properties.

The associative property allows to write in an unambiguous way $a + b + c$ and abc. More generally, the sum of $a_1, a_2, \ldots, a_n \in \mathbb{Q}$ is unambiguously defined. To present this sum concisely, we use the *sigma notation*:

$$\sum_{i=1}^{n} a_i \equiv \begin{cases} 0 & \text{if } n = 0 \\ a_1 + a_2 + \cdots + a_n & \text{if } n \in \mathbb{N}_0 \end{cases} .$$

We call the first case $n = 0$ the *empty sum*.

In a similar way we can define and represent the product of $a_1, a_2, \ldots, a_n \in \mathbb{Q}$:

$$\prod_{i=1}^{n} a_i \equiv \begin{cases} 1 & \text{if } n = 0 \\ a_1 \cdot a_2 \cdot \cdots \cdot a_n & \text{if } n \in \mathbb{N}_0 \end{cases} .$$

We call the first case $n = 0$ the *empty product*.

If all factors of the product have the same value $a \in \mathbb{Q}$, we use the *exponential notation*:

$$\Pi_{i=1}^{n} a \equiv a^n .$$

For $a \in \mathbb{Q}_0$, we can write

$$\left(\frac{1}{a}\right)^n = \left(a^{-1}\right)^n = a^{-n} = (a^n)^{-1} = \frac{1}{a^n}$$

and we have the following property

$$\forall m, n \in \mathbb{Z} : a^{m+n} = a^m a^n \text{ and } (a^m)^n = a^{mn} .$$

2–3 Order Properties of Rational Numbers

Rational numbers are not only determined by the operations of addition and multiplication, they are also *ordered* in a way that is compatible with addition and multiplication.

Axiom 3 (Order properties). A field $\mathbb{F}$ is an *ordered field* with total order relation $\leq$ if $\forall a, b, c \in \mathbb{F}$:

1. $a \leq a$ (*reflectivity*)
2. $a \leq b \vee b \leq a$ (*totality*)
3. $a \leq b \wedge b \leq a \implies a = b$ (*antisymmetry*)
4. $a \leq b \wedge b \leq c \implies a \leq c$ (*transitivity*)

Moreover, addition and multiplication are compatible with the total order relation:

1. $a \leq b \implies a + c \leq b + c$
2. $a \leq b \wedge 0 \leq c \implies ac \leq bc$

Notice that the rational numbers do satisfy the ordered field axioms.

We say that a is *positive* if $0 \leq a$ and *negative* if $a \leq 0$. The *strict order relation* $a < b$ means that $a \leq b$ and $a \neq b$.

Following properties can be derived from the order properties.

Theorem 9. *Let* $a, b, c \in \mathbb{Q}$.

1. $a \leq b \implies -b \leq -a$
2. $a \leq b \wedge c \leq 0 \implies bc \leq ac$
3. $0 \leq a \wedge 0 \leq b \implies 0 \leq ab$
4. $0 \leq a^2$
5. $0 < 1$
6. $0 < a \implies 0 < a^{-1}$
7. $0 < a < b \implies 0 < b^{-1} < a^{-1}$

Example. Prove the fifth property.

Proof (by contradiction). By the totality property is $0 \leq 1 \vee 1 \leq 0$. Suppose $1 \leq 0$.

By the compatibility of the addition is $0 = 1 + (-1) \leq 0 + (-1) = -1$.

By the compatibility of the multiplication is $0 = 0 \cdot (-1) \leq -1 \cdot (-1) = 1$.

Because $0 \neq 1$, $0 \leq 1$ is in contradiction with $1 \leq 0$. We supposed wrongly that $1 \leq 0$.

Therefore, $0 \leq 1$ and $0 \neq 1$, that is, $0 < 1$.

Exercise. Prove the other properties.

Decimal Expansions

The rational numbers have a *decimal expansion* that is either:

1. terminating, that is, ending with an infinite string of zeros, eg. $\frac{3}{4} = 0.75000\ldots$, or
2. repeating, that is, ending with a string of digits that repeats over and over, eg. $\frac{23}{11} = 1.090909\ldots = 2.\overline{09}$.

Example.
Show that $1.\overline{32}$ and $.3\overline{405}$ are rational numbers by expressing them as a quotient of two irreducible integers.

1. Let $x = 1.323232\ldots$. Then $x - 1 = 0.323232\ldots$ and
$$100x = 132.323232\ldots = 132 + 0.323232\ldots = 132 + x - 1\,.$$
Therefore, $99x = 131$ and $x = \dfrac{131}{99}$.

2. Let $y = 0.3405405405\ldots$. Then $10y = 3.405405405\ldots$ and $10y - 3 = 0.405405405\ldots$. Also,
$$10000y = 3405{,}405405405\ldots = 3405 + 10y - 3\,.$$
Therefore, $9990y = 3402$ and $y = \dfrac{3402}{9990} = \dfrac{63}{185}$.

2–4 Not all Numbers are Rational

Geometrically, we represent the integers as points on a horizontal line by defining a starting point (the number 0) and a unit distance (between two consecutive integers). The unit distance can be further subdivided to represent rational numbers such as $\dfrac{1}{n}$ with $n \in \mathbb{N}_0$. In this way we can uniquely represent each rational number on the *number line*.

Although the rational numbers have a rich structure, there are limitations. Not every point on the number line corresponds to a rational number, eg., we can indicate the number $\sqrt{2}$ using a simple geometric construction, based on the *Pythagorean theorem*, as a single point between the numbers 1 and 2 but it is not a rational number.

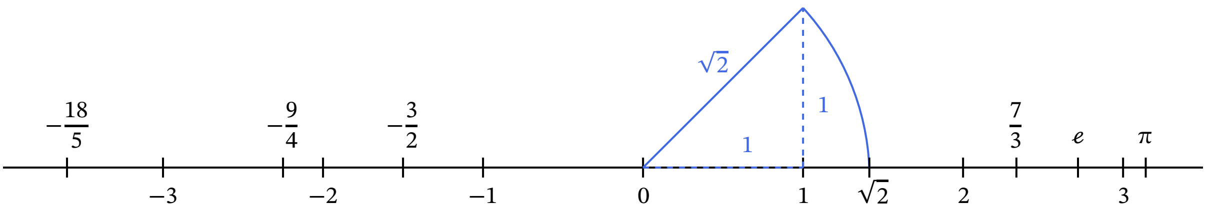

Figure 2–1: The real line.

Theorem 10. $\sqrt{2} \notin \mathbb{Q}$.

Proof (by contradiction). Suppose $\sqrt{2} \in \mathbb{Q}$.

Since $\sqrt{2}$ as a rational number can be represented by an irreducible fraction $\frac{m}{n}$, we have $\frac{m^2}{n^2} = \left(\sqrt{2}\right)^2 = 2$, that is, $m^2 = 2n^2$. Consequently, m^2 is even and so is m (see previous chapter). Then $m^2 = (2k)^2 = 4k^2 = 2n^2$ or $n^2 = 2k^2$, that is, n^2 is even and so is n.

Both m and n are even contradicts that $\frac{m}{n}$ is an irreducible fraction. We supposed wrongly that $\sqrt{2} \in \mathbb{Q}$.

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

Exercise. Prove $\sqrt{3} \notin \mathbb{Q}$. What goes wrong if you try to prove $\sqrt{4} \notin \mathbb{Q}$?

Hint: If a prime number divides the square of an integer, it also divides the integer itself.

2–5 Completeness of the Real Numbers

To express that there are no holes in the *real line*, that is, the line is *complete*, another axiom is needed.

Definition 5.
A number u of an ordered field $\mathbb{F}$ is said to be an *upper bound* for a nonempty set $S \subset \mathbb{F}$ if $\forall x \in S : x \le u$.

The number $u^\star$ is called the *least upper bound* or *supremum* of S if $u^\star$ is an upper bound for S and $u^\star \le u$ for every upper bound u of S. The supremum of S is denoted $\sup\, S$.

Similarly, l is a *lower bound* for S if $\forall x \in S : l \le x$. The number $l^\star$ is called the *greatest lower bound* or *infimum* of S if $l^\star$ is a lower bound for S and $l \le l^\star$ for every lower bound l of S. The infimum of S is denoted $\inf\, S$.

Example. Let $a > 0$. Show that $\sup\, aS = a \sup\, S$.

1. Since $\forall x \in S : x \le \sup\, S$, so $ax \le a \sup\, S$ and $\sup\, aS \le a \sup\, S$.
2. Since $\forall x \in S : ax \le \sup\, aS$, so $x \le \dfrac{\sup\, aS}{a}$ and $a \sup\, S \le \sup\, aS$.
   Therefore, $\sup\, aS = a \sup\, S$.

Exercise. Let $a > 0$. Show that $\inf\, aS = a \inf\, S$.

If a set has a maximum (minimum) then this maximum (minimum) is also the supremum (infimum). If this is not the case, the supremum (infimum), not belonging to the set, are the next-best thing.

Axiom 4 (Completeness).
A nonempty set of real number that has an upper bound must have a least upper bound.

Equivalently, a nonempty set of real numbers having a lower bound must have a greatest lower bound.

We stress that this is an axiom to be assumed without proof. It cannot be deduced from the algebraic and order properties. These properties are shared by the rational numbers, a set that is not complete.

The completeness axiom has massive consequences.

Lemma 11. *The set of natural numbers is not bounded above or $\forall r \in \mathbb{R}, \exists n \in \mathbb{N} : r \le n$.*

Proof (by contradiction). Suppose that there is an $r \in \mathbb{R}$ such that $\forall n \in \mathbb{N} : r > n$.

Then the set $\mathbb{N}$ is a nonempty subset of $\mathbb{R}$ with an upper bound, so by the completeness axiom, $\mathbb{N}$ has a least upper bound M.

Then $M - 1 < M$, so $M - 1$ is not an upper bound for $\mathbb{N}$.

Thus, there is a $k \in \mathbb{N}$ with the property that $k > M - 1$.

But then $k + 1$ is also in $\mathbb{N}$, yet $k + 1 > (M - 1) + 1 = M$ where M is an upper bound for $\mathbb{N}$.

This is a contradiction since no element of a set can be greater than an upper bound for that set.

Therefore, the assumption that $r > n$ for every $n \in \mathbb{N}$ must be false, and for every $r \in \mathbb{R}$ there must be at least one $n \in \mathbb{N}$ with $n > r$.

The Archimedean principle expresses geometrically that any line segment, no matter how long, may be covered by a finite number of line segments of a given positive length, no matter how small. This is a fundamental property of the real line.

Theorem 12 (Archimedean Principle). $\forall a \in \mathbb{R}_0^+, \forall b \in \mathbb{R}, \exists n \in \mathbb{N} : na > b$.

This principle is a direct consequence of the previous lemma, just replace r by $\frac{b}{a}$.

Example. Let $S = \left\{\frac{1}{2^n} \mid n \in \mathbb{N}\right\} = \left\{1, \frac{1}{2}, \frac{1}{4}, \dots\right\}$. Show that $\inf\ S = 0$.

1. 0 is a lower bound of S, $\forall n \in \mathbb{N} : 0 < \frac{1}{2^n}$ and
2. $\forall \varepsilon \in \mathbb{R}_0^+ : 0 + \varepsilon$ is no lower bound, $\forall \varepsilon \in \mathbb{R}_0^+, \exists n \in \mathbb{N} : \frac{1}{2^n} \leq \frac{1}{n+1} < 0 + \varepsilon$ (Archimedean principle with $x = 1$ and $y = \varepsilon$), so 0 is the greatest lower bound.

2–6 Absolute Values

The concepts that separates the area of mathematics known as analysis (the fundaments of calculus) from other branches such as algebra, set theory, ... is the idea of *distance*. In the real numbers, one can measure distance by using the *absolute value*.

Definition 6. The absolute value of a real number x, denoted by $|x|$, is defined by the formula

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

Geometrically, $|x|$ represents the (nonnegative) distance from x to 0 on the real line. $|x - y|$ represents the (nonnegative) distance between the points x and y on the real line, since this distance is the same as that from the point $x - y$ to 0.

The absolute value has the following properties.

Theorem 13. *Let $a, b \in \mathbb{R}$*

- $|-a| = |a|$
- $|ab| = |a||b|$
- $|a \pm b| \leq |a| + |b|$ *(triangle inequality)*

Proof.
The first two of these properties can be checked by considering the cases where either a or b is either positive or negative. The third property follows from the first two.

Let $a, b \in \mathbb{R}$.

$$\begin{aligned} |a \pm b|^2 &= (a \pm b)^2 = a^2 \pm 2ab + b^2 \\ &\leq |a|^2 + 2|a||b| + |b|^2 = (|a| + |b|)^2 \end{aligned}$$

Taking the (positive) square root of both sides, we obtain $|a \pm b| \leq |a| + |b|$.

The notion of an absolute value will be generalized by a *metric* in the setting of a *metric space*[1].

2–7 Intervals

Definition 7.
An *interval* is a subset of the real numbers that contains all real numbers lying between any two numbers of the subset.

The notion of an interval will be generalized by a *connected set* in the setting of a metric space.

[1]A metric space is an ordered pair (M, d) where M is a set and d is a metric on M, i.e., a function $d : M \times M \to \mathbb{R}$ satisfying the following axioms for all points $x, y, z \in M$:

- $d(x, x) = 0$ and $x \neq y \Rightarrow d(x, y) > 0$ (*positivity*)
- $d(x, y) = d(y, x)$ (*symmetry*)
- $d(x, z) \leq d(x, y) + d(y, z)$ (*triangle inequality*)

Example.
The set of real numbers x such that $x > 6$ is an interval, but the set of real numbers y such that $y \neq 0$ is not an interval.

If a and b are real numbers and $a < b$ we often refer to

1. the *open interval* from a to b, denoted $]a, b[$, consisting of all real numbers x satisfying $a < x < b$
2. the *closed interval* from a to b, denoted $[a, b]$, consisting of all real numbers x satisfying $a \leq x \leq b$
3. the *half-open interval* from a to b, denoted $[a, b[$, consisting of all real numbers x satisfying $a \leq x < b$
4. the *half-open interval* from a to b, denoted $]a, b]$, consisting of all real numbers x satisfying $a < x \leq b$

In a figure, hollow dots indicate endpoints of intervals that are not included in the intervals, and solid dots indicate endpoints that are included. The endpoints of an interval are also called *boundary points*.

The already defined intervals are *finite intervals*, that is, each of them has finite length $b - a$. Intervals can also have infinite length, in which case they are called *infinite intervals*. Note that the whole real line $\mathbb{R}$ is an interval, denoted by $]-\infty, \infty[$. The symbol ∞ ("infinity") does **not** denote a real number.

Theorem 14 (Theorem of Nested intervals). *If $\forall n \in \mathbb{N}$, we have a closed interval*

$$I_n = [a_n, b_n] = \{x \mid x \in \mathbb{R} \wedge a_n \leq x \leq b_n\}$$

such that

$$a_n \leq a_{n+1} \text{ and } b_{n+1} \leq b_n \text{, so that } \cdots \subset I_n \subset \cdots \subset I_2 \subset I_1 \subset I_0 \,,$$

then

$$\bigcap_{n \in \mathbb{N}} I_n \neq \varnothing$$

and if, in addition,

$$\inf\{b_n - a_n\} = 0 \,,$$

then

$$\bigcap_{n \in \mathbb{N}} I_n = \{x\} \text{ where } x = \sup\{a_n\} = \inf\{b_n\} \,.$$

Proof. We split the proof in four parts.

- We prove first that $\bigcap_{n\in\mathbb{N}} I_n \neq \emptyset$.

  Since the intervals are nested, $a_m \leq b_n$ for all m and n. This shows that every b_n is an upper bound for the set $\{a_m\}$ and every a_m is a lower bound for the set $\{b_n\}$. Let $a = \sup\{a_n\}$ and $b = \inf\{b_n\}$.

  By definition, $a_n \leq a$ for all n, and since b_n is an upper bound for $\{a_n\}$, we have $a_n \leq a \leq b_n$, which says that $a \in I_n$ for every n and so $a \in \bigcap_{n\in\mathbb{N}} I_n$.

  Similarly, $b \in \bigcap_{n\in\mathbb{N}} I_n$.

  Therefore, $\bigcap_{n\in\mathbb{N}} I_n \neq \emptyset$

- Secondly, we prove by contradiction that $a \leq b$.

  Suppose $a > b$. We have $b < \frac{a+b}{2} < a$.

  Since $\frac{a+b}{2} < a$, $\frac{a+b}{2}$ is not a upper bound for $\{a_n\}$, so there exists some $a_k > \frac{a+b}{2}$.

  Similarly, since $b < \frac{a+b}{2}$, $\frac{a+b}{2}$ is not a lower bound for $\{b_n\}$, so there exists some $b_l < \frac{a+b}{2}$.

  But then $b_l < a_k$ which is a contradiction since $a_m \leq b_n$ for all m and n.

  Therefore, $a \leq b$.

- Given the condition $\inf\{b_n - a_n\} = 0$, we prove that $a = b$.

  Note that $a_n \leq a \leq b \leq b_n$ implies $0 \leq b - a \leq b_n - a_n$ for all n, so $b - a$ is a lower bound for $\{b_n - a_n\}$ and $0 \leq b - a \leq \inf\{b_n - a_n\} = 0$

  Therefore, $a = b$.

- Finally, we prove that $\bigcap_{n\in\mathbb{N}} I_n = \{a\}$.

  Let $y \in \bigcap_{n\in\mathbb{N}} I_n$. Then, $a_n \leq y \leq b_n$ for all n.

  The first inequality implies that y is an upper bound for a_n, hence $a \leq y$.

  The second inequality implies that y is an lower bound for b_n, hence $y \leq b = a$.

  Combining both inequalities $a \leq y \leq a$, we have $y = a$, so $\bigcap_{n\in\mathbb{N}} I_n \subset \{a\}$.

  We also have $\{a\} \subset \bigcap_{n\in\mathbb{N}} I_n$.

  Therefore, $\bigcap_{n\in\mathbb{N}} I_n = \{a\}$.

A direct corollary of the nested intervals theorem is the uncountability of the real numbers.

Corollary 15 (The real numbers are not countable). *There is no surjection $f : \mathbb{N} \to \mathbb{R}$.*

Proof. Let $f : \mathbb{N} \to \mathbb{R}$ be a function.

For each $n \in \mathbb{N}$, let $f(n) = x_n$

Choose a closed interval $I_0 = [a_0, b_0]$ not containing x_0.

Choose a closed interval $I_1 = [a_1, b_1] \subset I_0$ not containing x_1.

Choose a closed interval $I_2 = [a_2, b_2] \subset I_1 \subset I_0$ not containing x_2.

And so on.

By the nested intervals theorem, $\bigcap_{n\in\mathbb{N}} I_n \neq \varnothing$. But no x_k can be in the intersection since $x_k \notin I_k$, so there exists some real number $y \in \bigcap_{n\in\mathbb{N}} I_n$ such that $y \neq f(n)$ for any $n \in \mathbb{N}$.

Therefore, f is not a surjection.

The cardinal number of the set of real numbers is called the *continuum*, denoted by $\#\mathbb{R} = \mathfrak{c} = \aleph_1 = 2^{\aleph_0} > \aleph_0$.

2–8 Bisection Method

When we have to find an element in a sorted array, we apply intuitively a *binary search algorithm*, eg. looking up a name in an alphabetically ordered list.

1. We start by inspecting the first and the last name in the list.
2. If the name we are looking for, is not the first or the last name. We check the name that is located half way between the first and the last name.
3. If this is the name we are looking for, we stop.

   If this name is alphabetically after the name we are looking for, we reduce the list from the first name to the name in the center position and go back to the first step.

   If this name is alphabetically before the name we are looking for, we reduce the list from the name in the center position to the last name and go back to the first step.

The same methodology can be used to prove some important theorems in calculus. We call such a procedure, the *bisection method*.

Template 7 (Prove by bisection). Begin with a closed interval $I_0 = [a_0, b_0]$.

STEP 1: Bisect I_0 to obtain two closed intervals $\left[a_0, \frac{a_0+b_0}{2}\right]$ and $\left[\frac{a_0+b_0}{2}, b_0\right]$.

STEP 2: Select one of the two subintervals above, and call it $I_1 = [a_1, b_1]$.

Keep repeating this process to obtain a sequence of intervals $I_0, I_1, I_2, I_3, \ldots$.

The sequence of intervals $I_0, I_1, I_2, I_3, \ldots$ satisfies both hypotheses of the nested intervals theorem:

1. Since for each n we have $a_n < \frac{a_n+b_n}{2} < b_n$ and either $a_{n+1} = a_n$ or $b_{n+1} = b_n$. Therefore $a_n \leq a_{n+1}$ and $b_{n+1} \leq b_n$ for all $n \in \mathbb{N}$, so that $\cdots \subset I_n \subset \cdots \subset I_2 \subset I_1 \subset I_0$.
2. We have that $b_n - a_n = \frac{b_0-a_0}{2^n}$, so

$$\inf\{b_n - a_n\} = \inf\left\{\frac{b_0 - a_0}{2^n}\right\} = (b_0 - a_0)\inf\left\{\frac{1}{2^n}\right\} = 0$$

   by previous examples.

Let's prove by the bisection method that there is a rational between any two reals.

We need following lemma.

Lemma 16 (Capture theorem).
Let A be a nonempty subset of $\mathbb{R}$. If A is bounded above, then any open interval containing $\sup A$ *contains an element of A.*

Similarly, if A is bounded below, than any open interval containing $\inf A$ *contains an element of A.*

Proof (by contradiction). Let $]x, y[$ be an open interval such that $x < \sup A < y$.

Suppose $]x, y[$ doesn't contain an element of A. So x is an upper bound for A which is a contradiction since $x < \sup A$.

Therefore, $]x, y[$ contains an element of A.

Exercise. Prove the second statement.

Theorem 17 (The rational numbers are dense in the real numbers).
Let $x \in \mathbb{R}$ and $\varepsilon \in \mathbb{R}^+$. The interval $]x - \delta, x + \delta[$ contains a rational number.

Proof (by bisection). If x is rational we are done, so let x be irrational.

Let b_0 be the smallest integer greater than x, and let $a_0 = b_0 - 1$.

Then $I_0 = [a_0, b_0]$ contains x and has rational endpoints. It follows that x is contained in either $\left[a_0, \frac{a_0 + b_0}{2}\right]$ or $\left[\frac{a_0 + b_0}{2}, b_0\right]$. Let I_1 be the closed subinterval containing x.

Continuing this way, we obtain a sequence of closed intervals $\cdots \subset I_n \subset \cdots \subset I_2 \subset I_1 \subset I_0$ satisfying the hypotheses of the nested intervals theorem, where each I_n contains x and has rational endpoints.

By the nested intervals theorem, $\bigcap_{n \in \mathbb{N}} I_n = \{y\}$ where $y = \sup\{a_n\} = \inf\{b_n\}$. Since $x \in I_n$ for all $n \in \mathbb{N}$, $y = x$.

Since $x = \sup\{a_n\}$, by the previous lemma, the open interval $]x - \delta, x + \delta[$ contains $a_m \in \mathbb{Q}$ for some $m \in \mathbb{N}$.

The practical use of this theorem is the possibility to approximate with a given tolerance a real number by a rational number, eg. in computer science *floating point* values are rational approximations of real numbers.

Something stronger than the Capture theorem is actually true: for $x = \sup\{a_n\} = \inf\{b_n\}$ the open interval $]x - \delta, x + \delta[$ actually contains an entire interval I_N, for some N. To see this, note that there are tree possibilities when $a_k \in\ ]x - \delta, x + \delta[$ and $b_l \in\ ]x - \delta, x + \delta[$:

- if $k = l$, then the open interval contains $I_k = I_l$;
- if $k < l$, then the open interval contains $a_k \leq a_{k+1} \leq \ldots a_l \leq b_l$, so the open interval contains I_l;
- if $k > l$, then the open interval contains $a_k \leq b_k \leq \cdots \leq \ldots b_{l+1} \leq b_l$, so the open interval contains I_k.

Theorem 18 (Heine-Borel Theorem).
Let $[a, b]$ be a closed interval and $\mathcal{O} = \{]c_i, d_i[\mid i \in I\}$ be an infinite set of open intervals. If $[a, b] \subset \bigcup_{i \in I}]c_i, d_i[$, then there exists $n \in \mathbb{N}$ such that $[a, b] \subset \bigcup_{k=0}^{n} \left]c_{i_k}, d_{i_k}\right[$.

Proof (by bisection and contradiction). Suppose that no finite subset of $\mathcal{O}$ covers $[a,b]$.

Let $I_0 = [a,b] = [a_0, b_0]$.

At least one of the intervals $\left[a_0, \frac{a_0+b_0}{2}\right]$ or $\left[\frac{a_0+b_0}{2}, b_0\right]$ cannot be covered by a finite subset of $\mathcal{O}$. If both could be covered by finite subsets, their union would cover I_0.

Let the interval that can't be covered by a finite subset of $\mathcal{O}$ be $I_1 = [a_1, b_1]$.

Continuing this way, we obtain a sequence of closed intervals $\cdots \subset I_n \subset \cdots \subset I_2 \subset I_1 \subset I_0$ satisfying the hypotheses of the nested intervals theorem, where each I_n can't be covered by a finite subset of $\mathcal{O}$.

By the nested intervals theorem, $\bigcap_{n\in\mathbb{N}} I_n = \{x\}$ where $x = \sup\{a_n\} = \inf\{b_n\}$.

Since $x \in [a,b]$ and $[a,b]$ is covered by the union of $\mathcal{O}$, there exists an open interval $]c_i, d_i[$ such that $x \in]c_i, d_i[$.

Since $]c_i, d_i[$ is open, there exists an $\delta > 0$ such that $]x-\delta, x+\delta[\subset]c_i, d_i[$.

Since $x \in \bigcap_{n\in\mathbb{N}} I_n$, there exists $N \in \mathbb{N}$ such that $I_N \subset]x-\delta, x+\delta[$ by the extension of the Capture theorem.

This means that for $n \geq N$, $I_n \subset]c_i, d_i[$, contradicting our assumption that no I_n can be covered by a finite subset of $\mathcal{O}$.

Therefore, our initial assumption must be false, and there must exist a finite subset of $\mathcal{O}$ that covers $[a,b]$.

The Heine-Borel theorem let us replace an infinite set of open intervals with a finite set. Something that will be very useful in the next chapters.

2–9 *Functions, Cartesian Plane and Graphs*

Remember the definition of a function.

Definition 8.
A function f from a set X to a set Y, often written $f : X \to Y$, is a rule which assigns to each $x \in X$ a unique element $f(x) = y \in Y$.

In this course, we will only consider functions having $X \subset \mathbb{R}$ and $Y \subset \mathbb{R}$.

There are several ways to represent a function symbolically:

1. by a formula such as $y = x^2$;
2. by a formula such as $f(x) = x^2$;
3. by a mapping rule such as $x \mapsto x^2$.

Strictly speaking we should call a function f and not $f(x)$ since the latter denotes the value of the function at the point x.

A function is not properly defined until its domain is specified. For instance the function $f(x) = x^2$ defined for all real numbers $x \geq 0$ is different from the function $g(x) = x^2$ defined for all real numbers x.

When a function f is defined without specifying its domain, we assume that the domain consists of all real numbers x for which the value $f(x)$ of the function is a real number.

Example.
The domain of $f : x \mapsto \sqrt{x}$ is the interval $[0, \infty[$ since negative numbers do not have a real square root. Note that the *square root function* $\sqrt{x}$ always denotes the positive square root of x.

Functions can be represented in the *Cartesian plane*.

Definition 9.
A *Cartesian coordinate system* in a plane is a coordinate system that specifies each point uniquely by a pair of real numbers called coordinates, which are the signed distances to the point from two fixed perpendicular oriented lines, called *coordinate axes* of the system. The point where they meet is called the *origin* denoted by O and has $(0,0)$ as coordinates.

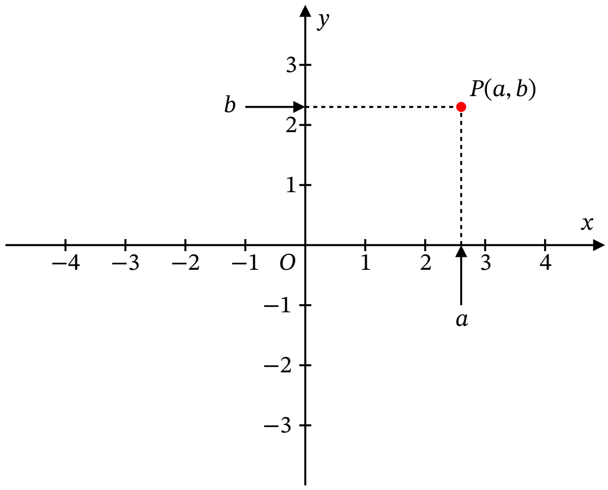

Figure 2–2: The Cartesian Plane.

Definition 10.
The *graph of a function* f consists of those points in the Cartesian plane whose coordinates (x, y) are pairs of input-output values for f.

Thus, (x, y) lies on the graph of f provided x is in the domain of f and $y = f(x)$.

Example. Graph the function $f(x) = x^2$.

Make a table of (x, y) pairs that satisfy $y = x^2$. Now plot the points and join them with a smooth curve.

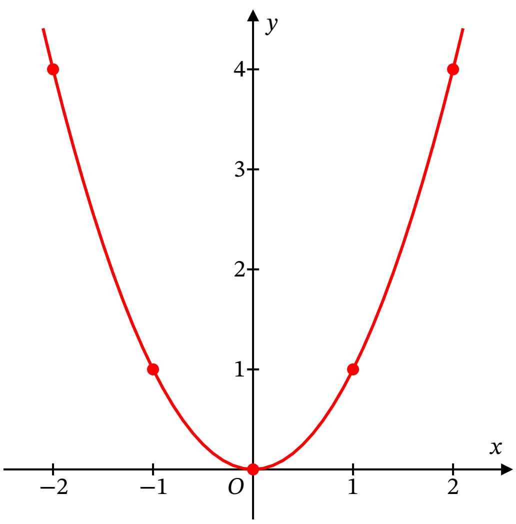

Figure 2–3: The graph of $y = x^2$

Not every curve you can draw is the graph of a function. A function f can have only one value $f(x)$ for each x is its domain, so no vertical line can intersect the graph of a function at more than one point.

Example.
The circle $x^2 + y^2 = 1$ cannot be the graph of a function since some vertical lines intersect it twice. It is, however, the union of the graphs of two functions, namely,

$$y = \sqrt{1 - x^2} \quad \text{and} \quad y = -\sqrt{1 - x^2},$$

which are, respectively, the upper and the lower halves of the given circle.

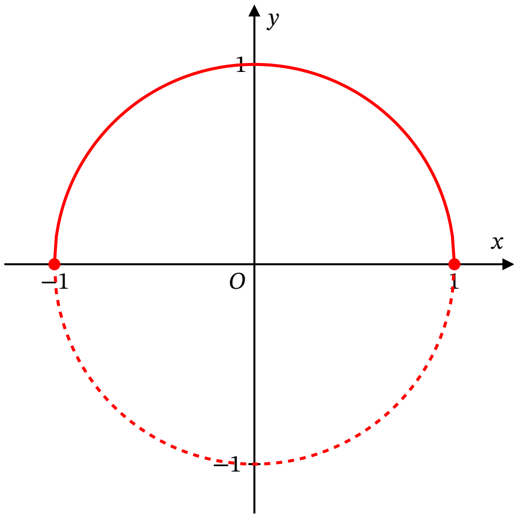

Figure 2–4: The circle $x^2 + y^2 = 1$ is not the graph of a function.

2–10 Combining Functions

Functions can be combined in a variety of ways to produce new functions.

Like numbers, functions can be added, subtracted, multiplied, and divided (except when the denominator is zero) to produce new functions.

Definition 11.
If f and g are functions, then for every x that belongs to the domains of both f and g we define functions $f + g$, $f - g$, fg, and $\frac{f}{g}$ by the formulas:

$$\begin{aligned}
(f + g)(x) &= f(x) + g(x) \\
(f - g)(x) &= f(x) - g(x) \\
(fg)(x) &= f(x)g(x) \\
\left(\frac{f}{g}\right)(x) &= \frac{f(x)}{g(x)}\,, \text{ where } g(x) \neq 0\,.
\end{aligned}$$

Example.
Draw the graphs of $f(x) = x^2$, $g(x) = x - 1$ and their sum $(f + g)(x) = x^2 + x + 1$. Observe that the height of the graph of $f + g$ at any point x is the sum of the heights of the graphs of f and g at that point.

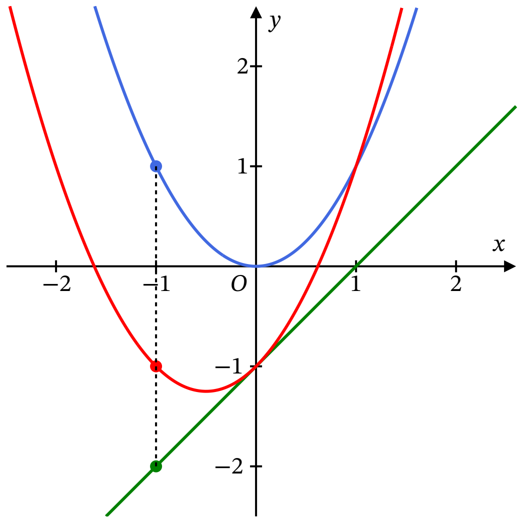

Figure 2–5: $(f+g)(x) = f(x) + g(x)$.

Example. The functions f and g are defined by the formulas

$$y = \sqrt{x} \quad \text{and} \quad y = \sqrt{1-x}\,.$$

Find formulas for the values of $3f$, $f+g$, $f-g$, fg, $\frac{f}{g}$, and $\frac{g}{f}$ at x, and specify the domains of each of these functions.

Function	**Formula**	**Domain**
f	$f(x) = \sqrt{x}$	$[0, \infty[$
g	$g(x) = \sqrt{1-x}$	$]\infty, 1]$
$3f$	$(3f)(x) = 3\sqrt{x}$	$[0, \infty[$
$f+g$	$(f+g)(x) = \sqrt{x} + \sqrt{1-x}$	[0,1]
$f-g$	$(f-g)(x) = \sqrt{x} - \sqrt{1-x}$	[0,1]
fg	$(fg)(x) = \sqrt{x(1-x)}$	[0,1]
$\frac{f}{g}$	$\left(\frac{f}{g}\right)(x) = \sqrt{\frac{x}{1-x}}$	[0,1[
$\frac{g}{f}$	$\left(\frac{g}{f}\right)(x) = \sqrt{\frac{1-x}{x}}$	]0,1]

There is another method, called *composition*, by which two functions can be combined to form a new function

Definition 12. If f and g are two functions, the *composite function* $f \circ g$ is defined by

$$(f \circ g)(x) = f(g(x))\,.$$

The domain of $f \circ g$ consists of those numbers x in the domain of g for which $g(x)$ is in the domain of f. In particular, if the range of g is contained in the domain of f, then the domain of $f \circ g$ is just the domain of g.

In calculating $(f \circ g)(x) = f(g(x))$, we first calculate $g(x)$ and then calculate f of the result.

Example.
Given $f(x) = \sqrt{x}$ and $g(x) = x + 1$ calculate the following composite functions $(f \circ g)$, $(g \circ f)$, $(f \circ f)$, and $(g \circ g)$, and specify the domain of each.

Function	Formula	Domain
f	$f(x) = \sqrt{x}$	$[0, \infty[$
g	$g(x) = x + 1$	$\mathbb{R}$
$f \circ g$	$(f \circ g)(x) = \sqrt{x+1}$	$[-1, \infty[$
$g \circ f$	$(f \circ g)(x) = \sqrt{x} + 1$	$[0, \infty[$
$f \circ f$	$(f \circ g)(x) = x^{1/4}$	$[0, \infty[$
$g \circ g$	$(f \circ g)(x) = x + 2$	$\mathbb{R}$

To see why, for example, the domain of $f \circ g$ is $[-1, \infty[$, observe that $g(x) = x + 1$ is defined for all real x but belongs to the domain of f only if $x + 1 \geq 0$, that is, if $x \geq -1$.

Sometimes it is necessary to define a function by using different formulas on different parts of its domain. One example is the absolute value function

$$\operatorname{abs}(x) = |x| = \begin{cases} x & \text{if } x \geq 0\,, \\ -x & \text{if } x < 0\,. \end{cases}$$

Example. The *signum function* is defined as follows:

$$\operatorname{sgn}(x) = \frac{x}{|x|} = \begin{cases} 1 & \text{if } x > 0\,, \\ -1 & \text{if } x < 0\,, \\ \text{undefined} & \text{if } x = 0\,. \end{cases}$$

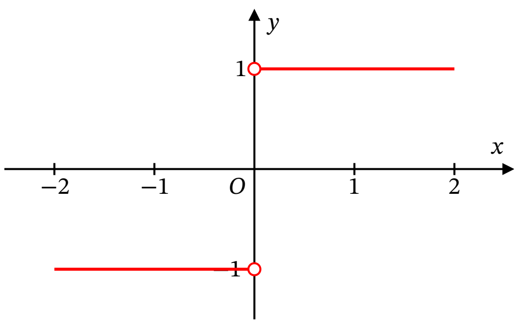

Figure 2–6: The signum function.

Note how we use a hollow dot in the graphs to indicate which endpoints do not lie on various parts of the graph. Similarly, we use a solid dot to indicate which endpoint do lie on various parts of the graph.

2–11 Inverse Functions

Remember a function f is bijective if $f(x_1) \neq f(x_2)$ whenever x_1 and x_2 belong to the domain of f and $x_1 \neq x_2$. So, not only a vertical line intersects the graph of the function in one point, also a horizontal line intersects the graph at only one point.

Definition 13.
If f is a bijective function, then it has an inverse f^{-1}. The value of $f^{-1}(x)$ is the unique member y in the domain of f for which $f(y) = x$. Thus,

$$y = f^{-1}(x) \iff x = f(y)\,.$$

There are several things you should remember about the relation between a function f and its inverse f^{-1}:

1. The domain of f^{-1} is the range of f.
2. The range of f^{-1} is the domain of f.
3. $\left(f^{-1} \circ f\right)(x) = f^{-1}(f(x)) = x$ for all x in the domain of f.
4. $\left(f \circ f^{-1}\right)(x) = f\left(f^{-1}(x)\right) = x$ for all x in the domain of f^{-1}.
5. $\left(f^{-1}\right)^{-1}(x) = f(x)$ for all x in the domain of f.
6. The graph of f^{-1} is the reflection of the graph of f in the line $x = y$.

Example. Show that $g(x) = \sqrt{2x+1}$ is invertible and find its inverse.

If $g(x_1) = g(x_2)$ then $\sqrt{2x_1+1} = \sqrt{2x_2+1}$. Squaring both sides we get $2x_1 + 1 = 2x_2 + 1$, which implies $x_1 = x_2$. Thus g is bijective and invertible.

Let $y = g^{-1}(x)$ then $x = g(y) = \sqrt{2y+1}$. It follows that $x \geq 0$ and $x^2 = 2y + 1$. Therefore, $y = \frac{x^2-1}{2}$ and

$$g^{-1}(x) = \frac{x^2-1}{2} \quad \text{for } x \geq 0 .$$

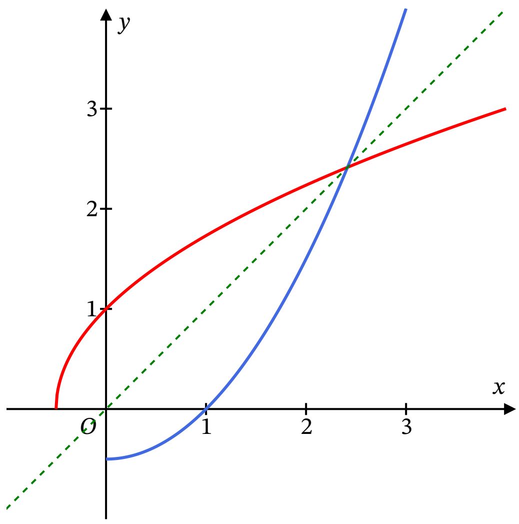

Figure 2–7: The graphs of g in red and its inverse in blue.

Many important functions are not bijective on their whole domain. It is possible to define an inverse for such a function, but we have to restrict the domain of the function artificially so that the restricted function is bijective.

Example.
Consider the function $f(x) = x^2$. Unrestricted, its domain is the whole real line and it is not bijective since $f(-a) = f(a)$ for any a. Let us define a new function $F(x)$ equal to $f(x)$ but having a smaller domain, so that it is bijective:

$$F : [0, \infty[\; \to [0, \infty[: x \mapsto x^2 .$$

F is bijective, so it has an inverse F^{-1}. Let $y = F^{-1}(x)$, then $x = F(y) = y^2$ and $y \geq 0$. Thus, $y = \sqrt{x}$. Hence,

$$F : [0, \infty[\; \to [0, \infty[: x \mapsto \sqrt{x} .$$

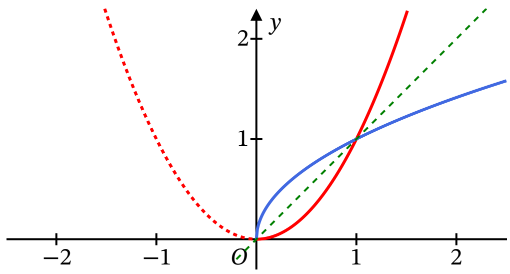

Figure 2–8: The graphs of F in red and its inverse in blue.

2–12 Polynomial and Rational Functions

Among the easiest functions to deal with in calculus are *polynomials*.

Definition 14. A polynomial is a function P whose value at x is

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 ,$$

where a_n, a_{n-1}, ..., a_2, a_1, and a_0 called the *coefficients* of the polynomial, are constants and, if $n > 0$, then $a_n \neq 0$. The number n, the degree of the highest power of x in the polynomial, is called the *degree* of the polynomial.

Polynomials play a role in the study of functions somewhat analogous to the role played by integers in the study of numbers.

The following definition is analogous to the definition of a rational number as the quotient of two integers.

Definition 15.
If $P(x)$ and $Q(x)$ are two polynomials and $Q(x)$ is not the zero polynomial the the function

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

is called a *rational function*. By the domain convention, the domain of $R(x)$ consists of all real numbers x except those for which $Q(x) = 0$.

Example.

$$R(x) = \frac{2x^3 - 3x^2 + 3x + 4}{x^2 + 1} \text{ with domain } \mathbb{R} .$$

$$S(x) = \frac{1}{x^2 - 4} \text{ with domain all real numbers except } \pm 2 .$$

If the numerator and the denominator of a rational function have a common factor, that factor can be cancelled out just as with integers. However, the resulting simpler rational function may not have the same domain as the orginal one, so it should be regarded as a different rational function even though it is equal to the orginal one at all points of the original domain.

Example.

$$\frac{x^2 - x}{x^2 - 1} = \frac{x(x-1)}{(x+1)(x-1)} = \frac{x}{x+1} \quad \text{only if } x \neq \pm 1$$

even though $x = 1$ is in the domain of $\frac{x}{x+1}$.

2–13 Trigonometric Functions

Most students first encounter the quantities $\cos t$ and $\sin t$ as ratios of sides in a right-angled triangle having t as one of the acute angles. If the sides of the triangle are labelled "hyp" for hypotenuse, "adj" for the side adjacent to angle t, and "opp" for the side opposite angle t, then

$$\cos t = \frac{\text{adj}}{\text{hyp}} \quad \text{and} \quad \sin t = \frac{\text{opp}}{\text{hyp}}$$

These ratios depend only on the angle t, not on the particular triangle, since all right-angled triangles having an acute angle t are similar.

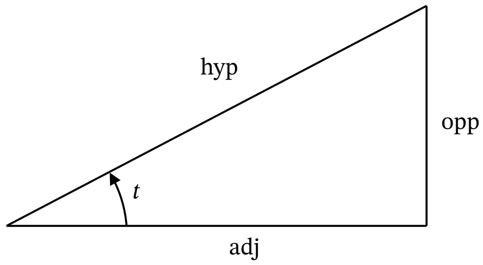

Figure 2–9: Basic definition of sinus and cosinus.

In calculus we need more general definitions of $\cos t$ and $\sin t$ as functions defined for all real numbers t, not just acute angles.

Let C be the circle with centre at the origin O and radius 1; its equation is $x^2 + y^2 = 1$. Let A be the point (1,0) on C. For any real number t, let P_t be the point on C at distance $|t|$ from A, measured along C in the counterclockwise direction if $t > 0$ and the clockwise direction if $t < 0$. For example, since C has circumference 2π, the point $P_{\pi/2}$ is one-quarter of the way counterclockwise around C from A; it is the point (0,1).

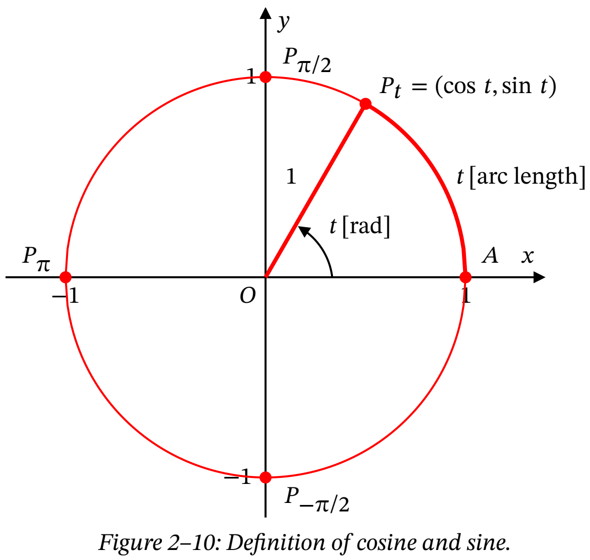

Figure 2–10: Definition of cosine and sine.

Definition 16. The *radian measure* of $\angle AOP_t$ is t radians.

In calculus it is assumed that all angles are measured in radians unless degrees or other units are stated explicitly. When we talk about the angle $\frac{\pi}{3}$ we mean $\frac{\pi}{3}$ radians (which is 60°), not $\frac{\pi}{3}$ degrees.

Example. *Arc length* and *sector area*.

An arc of a circle of radius r subtends an angle t at the centre of the circle. Find the length s of the arc and the area A of the sector lying between the arc and the centre of the circle.

The length s of the arc is the same fraction of the circumference $2\pi r$ of the circle that the angle t is to a complete revolution 2π radians. Thus,

$$s = \frac{t}{2\pi}(2\pi r) = rt\,.$$

Similarly, the area A of the circular sector is the same fraction of the area πr^2 of the whole circle:

$$A = \frac{t}{2\pi}(\pi r^2) = \frac{r^2 t}{2}\,.$$

Using the procedure described above, we can find the point P_t corresponding to any real number t.

Definition 17.
For any real t, the *cosine* of t denoted by $\cos t$ and the *sine* of t denoted by $\sin t$ are the x- and the y-coordinates of the point P_t.

Because they are defined this way, cosine and sine are often called the *circular functions*. Note that these definitions agree with the ones given earlier for an acute angle.

Many important properties of $\cos t$ and $\sin t$ follow from the fact that they are coordinates of the point P_t on the circle C with equation $x^2 + y^2 = 1$.

Theorem 19. *Let $t \in \mathbb{R}$.*

1. $-1 \le \cos t \le 1$ *and* $-1 \le \sin t \le 1$
2. $\cos^2 t + \sin^2 t = 1$ *(Pythagorean identity)*
3. $\cos(t + 2\pi) = \cos t$ *and* $\sin(t + 2\pi) = \sin t$ *(periodicity)*
4. $\cos(-t) = \cos(t)$ *and* $\sin(-t) = -\sin t$ *(cosine is an even function, sine is an odd function)*
5. $\cos\left(\frac{\pi}{2} - t\right) = \sin t$ *and* $\sin\left(\frac{\pi}{2} - t\right) = \cos t$ *(complementary angles, symmetry about $y = x$)*
6. $\cos(\pi - t) = -\cos(t)$ *and* $\sin(\pi - t) = \sin t$ *(supplementary angles, symmetry about $x = 0$)*

The next table summarizes the most important values of cosine and sine.

Radians	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{3\pi}{4}$	$\frac{5\pi}{6}$	π
Cosine	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{3}}{2}$	-1
Sine	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0

Observe that the graph of $\sin x$ is the graph of $\cos x$ shifted to the right a distance $\frac{\pi}{2}$.

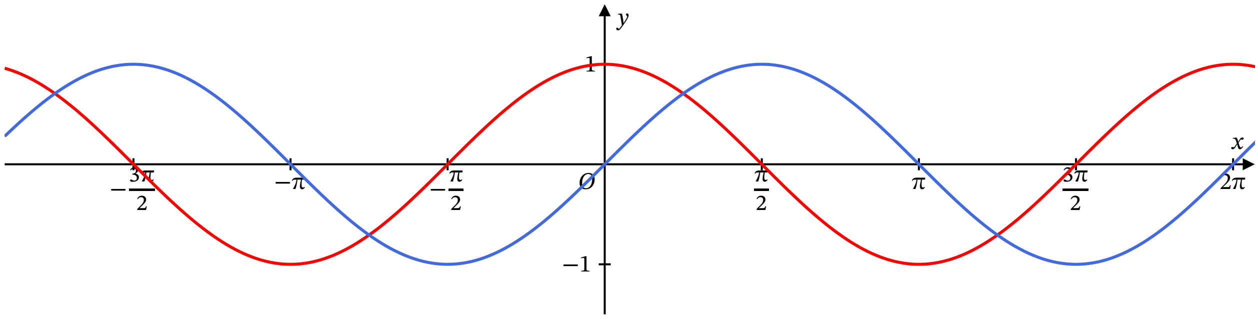

Figure 2–11: The graph of the cosine function in red and the sine function in blue.

The following formulas enable us to determine the cosine and sine of a sum or difference of two angles in terms of the cosines and sines of those angles.

Theorem 20. *Let $s, t \in \mathbb{R}$.*

$$\cos(s+t) = \cos s \cos t - \sin s \sin t$$
$$\sin(s+t) = \sin s \cos t + \cos s \sin t$$
$$\cos(s-t) = \cos s \cos t + \sin s \sin t$$
$$\sin(s-t) = \sin s \cos t - \cos s \sin t$$

Proof. Let $s, t \in \mathbb{R}$ and consider the points

$$P_t = (\cos t, \sin t) \qquad P_{s-t} = (\cos(s-t), \sin(s-t))$$
$$P_s = (\cos s, \sin s) \qquad A = (1,0)\,.$$

The angle $\angle P_t O P_s = s - t = \angle AOP_{s-t}$, so the distance $P_s P_t$ is equal to the distance $P_{s-t}A$.

Therefore, $(P_s P_t)^2 = (P_{s-t}A)^2$:

$$(\cos s - \cos t)^2 + (\sin s - \sin t)^2 = (\cos(s-t) - 1)^2 + \sin^2(s-t)\,,$$
$$\cos^2 s - 2\cos s \cos t + \cos^2 t + \sin^2 s - 2\sin s \sin t + \sin^2 t$$
$$= \cos^2(s-t) - 2\cos(s-t) + 1 + \sin^2(s-t)\,.$$

Since $\cos^2 x + \sin^2 x = 1$, this reduces to $\cos(s-t) = \cos s \cos t + \sin s \sin t$.

Exercise.
Prove the other formulas using the even/odd behavior of the cosine and sine functions and the complementary angles relations.

From the addition formulas, we obtain as special cases certain useful formulas called *double-angle formulas*.

Corollary 21. *Let $t \in \mathbb{R}$.*

1. $\sin 2t = 2\sin t \cos t$ *and*
2. $\cos 2t = \cos^2 t - \sin^2 t = 2\cos^2 t - 1 = 1 - 2\sin^2 t$.

Solving the last two formulas for $\cos^2 t$ and $\sin^2 t$, we obtain

$$\cos^2 t = \frac{1+\cos 2t}{2} \quad \text{and} \quad \sin^2 t = \frac{1-\cos 2t}{2}\,,$$

which are sometimes calles the *half-angle formulas*.

There are four other trigonometric functions, each defined in terms of cosine and sine.

Definition 18.

$$\tan\ : \mathbb{R} \setminus \left\{\frac{\pi}{2} + k\pi \mid k \in \mathbb{Z}\right\} \to \mathbb{R} : x \mapsto \frac{\sin\ x}{\cos\ x} \quad \text{(tangent)}$$

$$\cot\ : \mathbb{R} \setminus \{k\pi \mid k \in \mathbb{Z}\} \to \mathbb{R} : x \mapsto \frac{\cos\ x}{\sin\ x} \quad \text{(cotangent)}$$

$$\sec\ : \mathbb{R} \setminus \left\{\frac{\pi}{2} + k\pi \mid k \in \mathbb{Z}\right\} \to \mathbb{R} : x \mapsto \frac{1}{\cos\ x} \quad \text{(secans)}$$

$$\csc\ : \mathbb{R} \setminus \{k\pi \mid k \in \mathbb{Z}\} \to \mathbb{R} : x \mapsto \frac{1}{\sin\ x} \quad \text{(cosecans)}$$

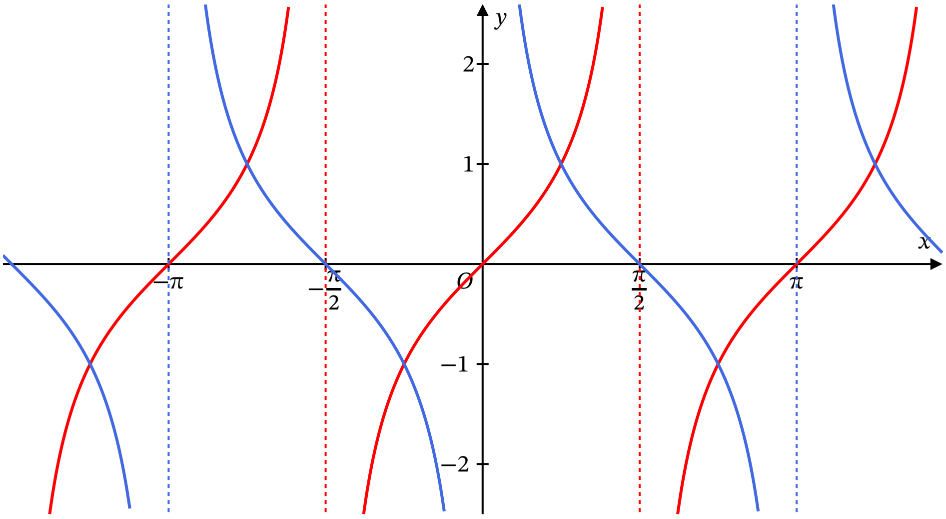

Figure 2–12: The graph of the tangent function in red and the cotangent function in blue.

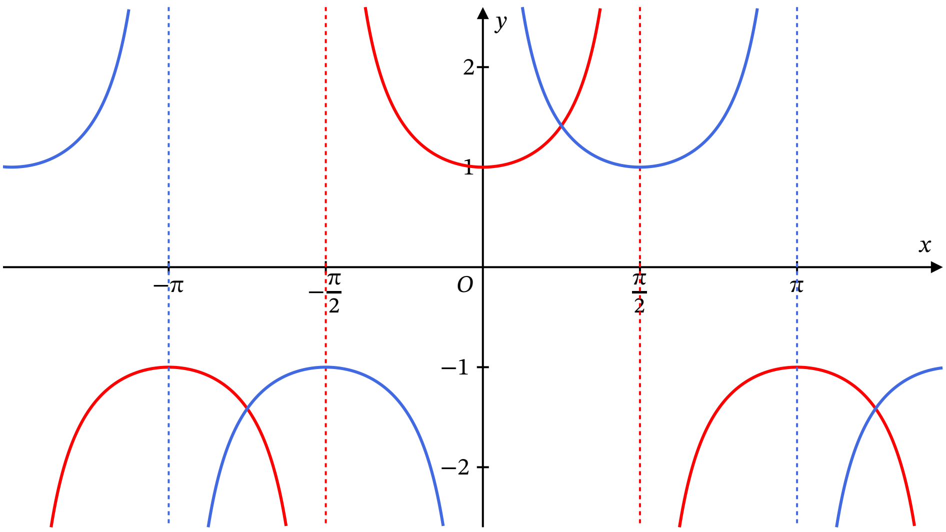

Figure 2–13: The graph of the secans function in red and the cosecans function in blue.

Observe that each of these functions is undefined (and its graph approaches vertical asymptotes) at points where the function in the denominator of its defining fraction has values 0.

CHAPTER 3

Limits and Continuity

Calculus was created to describe how quantities change:

- *differentiation*, for finding the rate of change of a given function, and
- *integration*, for finding a function having a given rate of change.

Both of these procedure are based on the fundamental concept of *limit* of a function.

In this chapter we will introduce the limit concept and develop some of its properties, including the nice behaviour of some functions that is called *continuity*.

3–1 Average and Instantaneous Velocity

The position of a moving object is a function of time. The average velocity of the object over a time interval is found by dividing the change in the object's position by the length of the time interval.

Example. The average velocity of a falling rock.

Physical experiments show that if a rock is dropped from rest near the surface of the earth, in the first $t\,[\mathrm{s}]$ if will fall a distance

$$y = \frac{gt^2}{2}\,[\mathrm{s}],$$

with $g \approx 9.8$ a constant representing the combined action of the gravition (from mass distribution within earth) and centrifugal forces (from the earth's rotation).

1. What is the average velocity of the falling rock during the first $2\,[\mathrm{s}]$ (time interval $[0,2]$)?

$$\frac{\Delta y}{\Delta t} = \frac{g}{2}\frac{t_2^2 - t_1^2}{t_2 - t_1} = \frac{g}{2}\frac{2^2 - 0^2}{2 - 0} = 9.8\left[\frac{\mathrm{m}}{\mathrm{s}}\right].$$

2. What is the average velocity of the falling rock in the time interval $[1,2]$?

$$\frac{\Delta y}{\Delta t} = \frac{g}{2}\frac{t_2^2 - t_1^2}{t_2 - t_1} = \frac{g}{2}\frac{2^2 - 1^2}{2 - 1} = 14.7\left[\frac{\mathrm{m}}{\mathrm{s}}\right].$$

The instantaneous velocity of the object at the instant t can be estimated by evaluating the average velocity in a small time interval containing t.

Example. How fast is the rock of the previous example falling

1. at time $t = 1\,[\mathrm{s}]$?

Time Interval	Average Velocity
[1,1.1]	10.29
[1,1.01]	9.849
[1,1.001]	9.8049
[1,1.0001]	9.8005

2. at time $t = 2\,[\mathrm{s}]$?

Time Interval	Average Velocity
[2,2.1]	20.09
[2,2.01]	19.649
[2,2.001]	19.6049
[2,2.0001]	19.6005

In the second example the average velocity of the falling rock over the time interval $[t, t+h]$ is

$$\frac{\Delta y}{\Delta t} = \frac{g}{2}\,\frac{(t+h)^2 - t^2}{(t+h)-t} = \frac{g}{2}\,\frac{2th + h^2}{h}$$

We examined the values of this average velocity for time intervals whose lengths h became smaller and smaller. We were finding the *limit of the average velocity as h tends to zero*. This is expressed symbolically in the form

$$\lim_{h\to 0} \frac{\Delta y}{\Delta t} = \frac{g}{2}\,\frac{2th + h^2}{h}\,.$$

We can't find the limit of the fraction by just substituting $h = 0$ because that would involve dividing by zero. However, we can calculate the limit by first performing som algebraic simplifications on the expression:

$$\lim_{h\to 0} \frac{\Delta y}{\Delta t} = \frac{g}{2}(2t + h)\,.$$

The final form no longer involves division by h. It approaches $gt + \frac{g}{2}0 = gt$. In particular, at $t = 1\,[\mathrm{s}]$ and $t = 2\,[\mathrm{s}]$, the instantaneous velocities are $v_1 = 9.8\left[\frac{\mathrm{m}}{\mathrm{s}}\right]$ and $v_2 = 19.6\left[\frac{\mathrm{m}}{\mathrm{s}}\right]$, respectively.

3–2 The Area of a Circle

All circles are similar geometric figures; they all have the same shape and differ only in size. The ratio of the circumference C to the diameter $2r$ has the same value for all circles. The number π is defined to be this common ratio:

$$\frac{C}{2r} = \pi \quad \text{or} \quad C = 2\pi r\,.$$

We were taught that the area A of a circle is this same number π times the square of the radius:

$$A = \pi r^2\,.$$

Can we deduce this area formula from the formula for the circumference?

The answer to this question lies in regarding the circle as a "limit" of regular polygons, which are in turn made up of triangles.

Suppose a regular polygon having n sides is inscribed in a circle of radius r.

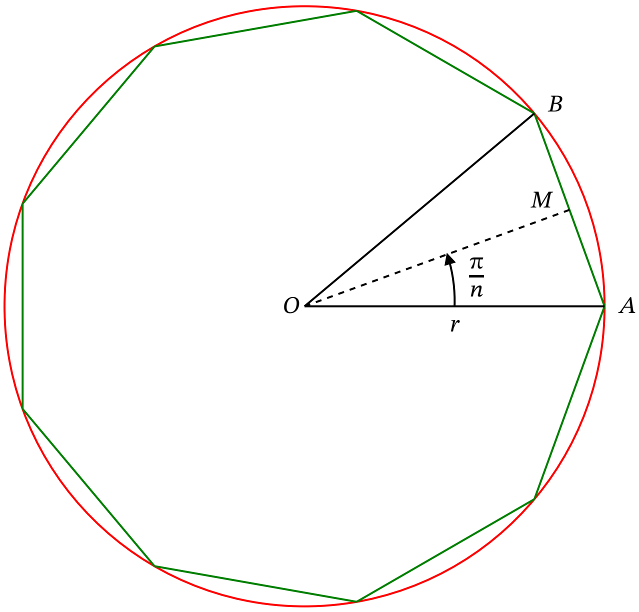

Figure 3–1: A regular polygon (green) of n sides inscribed in a red circle. Here $n = 9$.

The perimeter P_n and the area A_n of the polygon are, respectively, less than the circumference C and the area A of the circle, but if n is large P_n is close to C and A_n is close to A. We could expect P_n to approach the limit C and A_n to approach the limit A as n tends to infinity.

Since the total angle around the point O is 2π radians, $\angle AOB$ is $\frac{2\pi}{n}$ radians. If M is the midpoint of AB, then O bisects $\angle AOB$. We can write the length of AB and the area of $\triangle OAB$ in terms of the radius:

$$|AB| = 2|AM| = 2r\sin\frac{\pi}{n}$$

$$\triangle OAB = \frac{1}{2}|AB||OM| = r^2\sin\frac{\pi}{n}\cos\frac{\pi}{n}$$

The perimeter P_n and area A_n of the polygon are n times these expressions:

$$P_n = 2rn\sin\frac{\pi}{n}$$

$$A_n = r^2 n\sin\frac{\pi}{n}\cos\frac{\pi}{n}$$

Solving the first equation for $rn\sin\frac{\pi}{n} = \frac{P_n}{2}$ and substituting into the second equation, we get

$$A_n = \frac{1}{2}P_n r\cos\frac{\pi}{n}$$

Now $\angle AOM = \frac{\pi}{n}$ approaches 0 as n tends to infinity, so its cosine, $\cos\frac{\pi}{n} = \frac{|OM|}{|OA|}$, approaches 1. Since P_n approaches $C = 2\pi r$ as n tends to infinity, the expression for A_n approaches $\frac{1}{2}(2\pi r)r(1) = \pi r^2$, which must therefore be the area of the circle.

3–3 Limits Defined

In order to speak meaningfully about rate of change, tangent lines, and areas bounded by curves, we have to investigate the process of finding limits. Let us look at some examples.

Example. Describe the behaviour of the function $f(x) = \frac{x^2-1}{x-1}$ near $x = 1$.

Note that $f(x)$ is defined for all real numbers x except $x = 1$. (We can't divide by zero.) For any $x \neq 1$ we can simplify the expression for $f(x)$ by factoring the numerator and cancelling common factors:

$$f(x) = \frac{(x-1)(x+1)}{x-1} = x + 1 \quad \text{for } x \neq 1\,.$$

The graph of f is the line $y = x + 1$ with one point removed, namely, the point (1,2).

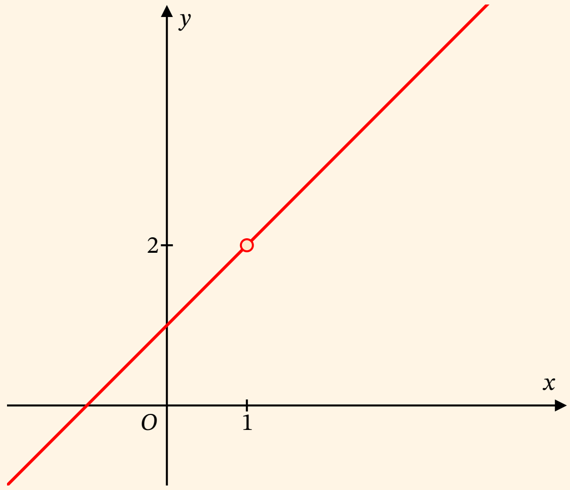

Figure 3–2: The graph of $f(x) = \frac{x^2-1}{x-1}$.

This removed point is shown as a *hole* in the graph. Even though (1) is not defined, it is clear that we can make the value of $f(x)$ *as close as we want* to 2 by choosing x *close enough* to 1. Therefore, we say f approaches the limit 2 as x tends to 1. We write this as

$$\lim_{x\to 1} f(x) = \lim_{x\to 1} \frac{x^2-1}{x-1} = 2\,.$$

Example. What happens to the function $g(x) = \left(1+x^2\right)^{\frac{1}{x^2}}$ as x approaches zero?

Note that g is not defined at $x = 0$. In fact, for the moment it does not appear to be defined for any x whose square x^2 is not a rational number. Let us ignore for now the problem of deciding what $g(x)$ means if x^2 is irrational and consider only rational values of x. There is no obvious way to simplify the expression for $g(x)$ as we did in previous example. However, we can use a scientific calculator to obtain approximate values of $g(x)$ for some rational values of x approaching 0.

x	$g(x)$
± 0.1	2.704813829
± 0.01	2.718145927
± 0.001	2.718280469
± 0.0001	2.718281798

Except the last value in the table, the values of $g(x)$ seem to be approaching a certain number, 2.71828..., as x tends to 0. We will see in Chapter 5 that

$$\lim_{x\to 0} g(x) = \lim_{x\to 0} \left(1+x^2\right)^{\frac{1}{x^2}} = e = 2.718281828459045\ldots\,.$$

The examples and the previous sections suggest the following informal definition of limit.

Definition 19 (informal).
If f is defined for all x near a, except possibly at a itself, and if we can ensure that $f(x)$ is as close as we want to L by taking x close enough to a, but not equal to a, we say that the function f approaches the *limit* L as x tends to a, and we write

$$\lim_{x \to a} f(x) = L$$

This definition is informal because phrases such as *close as we want* and *close enough* are imprecise; their meaning depends on the context. If we want to prove results about limits a more precise definition is needed. This precise definition is based on the idea of controlling the input x of a function f so that the output $f(x)$ will lie in a specific interval.

Example.
The area of a circular disk of radius $r\,[\mathrm{cm}]$ is $A = \pi r^2\,[\mathrm{cm}^2]$. A machinist is required to manufacture a circular metal disk having area $400\pi\,[\mathrm{cm}^2]$ within an error tolerance of $\pm 5\,[\mathrm{cm}^2]$. How close to $20\,[\mathrm{cm}]$ must the machinist control the radius of the disk to achieve this?

The machinist wants $\left|\pi r^2 - 400\pi\right| < 5$, that is,

$$400\pi - 5 < \pi r^2 < 400\pi + 5$$

or, equivalently,

$$\sqrt{400 - \frac{5}{\pi}} < r < \sqrt{400 + \frac{5}{\pi}}$$

$$19.96017 < r < 20.03975\,.$$

Thus, the machinist needs $|r - 20| < 0.03975$.

When we say that f has limit L as x tends to a, we are really saying that we can ensure that the error $|f(x) - L|$ will be less than *any* allowed tolerance, no matter how small, by taking x *close enough* to a (but not equal to a). It is traditional to use ε, the Greek letter "epsilon", for the size of the allowable *error* and δ, the Greek letter "delta" for the difference $|x - a|$ that measures how close x must be to a to ensure that the error is within that tolerance.

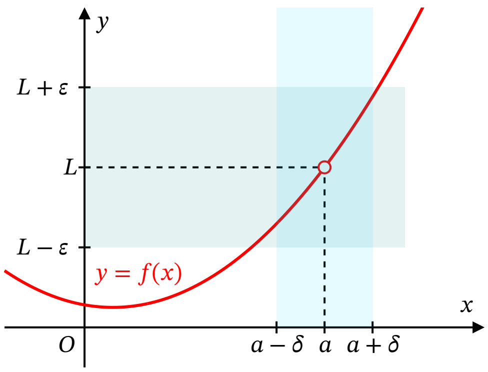

Figure 3–3: If $x \neq a$ and $|x - a| < \delta$, then $|f(x) - L| < \varepsilon$.

If ε is any strict positive number, *no matter how small*, we must be able to ensure that $|f(x) - L| < \varepsilon$ by restricting x to be *close enough* to (but not equal to) a. How close is close enough? It is sufficient that the distance $|x - a|$ from x to a be less than a positive number δ that depends on ε.

Definition 20 (formal). We say that $f : X \mapsto Y$ approaches the limit L as x tends to a, and we write

$$\lim_{x\to a} f(x) = L$$

if the following condition is satisfied:

$$\forall\varepsilon > 0, \exists\delta(\varepsilon) > 0 : 0 < |x - a| < \delta \implies x \in X \wedge |f(x) - L| < \varepsilon\,.$$

Note the possible dependency of δ on ε and the fact that x belongs to the domain of f.

The formal definition of limit does not tell you how to find the limit of a function, but it does enable you to verify that a suspected limit is correct.

Example. Verify that:

1. $\lim_{x\to a} x = a$.

   Let $\varepsilon > 0$ be given. We must find $\delta > 0$ so that

   $$0 < |x - a| < \delta \implies |x - a| < \varepsilon\,.$$

   Clearly, we can take $\delta = \varepsilon$ and the implication above will be true. This proves that $\lim_{x\to a} x = a$

2. $\lim_{x\to a} c = c$ (c is a constant).

   Let $\varepsilon > 0$ be given. We must find $\delta > 0$ so that

   $$0 < |x - a| < \delta \implies |c - c| < \varepsilon\,.$$

   Since $c - c = 0$, we can use any positive number for δ and the implication above will be true. This proves that $\lim_{x\to a} c = c$.

3. $\lim_{x\to 2} x^2 = 4$.

   Here $a = 2$ and $L = 4$. Let $\varepsilon > 0$ be given. We must find $\delta > 0$ so that

   $$0 < |x - 2| < \delta \implies \left|x^2 - 4\right| < \varepsilon\,.$$

   Now,

   $$\left|x^2 - 4\right| = |(x + 2)(x - 2)| = |x + 2||x - 2|\,.$$

   We want the expression above to be less than ε. We can make the factor $|x - 2|$ as small as we wish by choosing δ properly, but we need to control the factor $|x + 2|$ so that it does not become too large.

   If we first assume $\delta \le 1$ and require that $|x - 2| < \delta$, then we have

   $$|x - 2| < 1 \implies 1 < x < 3 \implies 3 < x + 2 < 5 \implies |x + 2| < 5$$

   Hence,

   $$|f(x) - 4| < 5|x - 2| \quad \text{if} \quad |x - 2| < \delta \le 1\,.$$

   But $5|x - 2| < \varepsilon$ if $|x - 2| < \frac{\varepsilon}{5}$. Therefore, if we take $\delta = \min\left\{1, \frac{\varepsilon}{5}\right\}$, the *minimum* of the two numbers 1 and $\frac{\varepsilon}{5}$, then

   $$|f(x - 4)| < 5|x - 2| < 5 \times \frac{\varepsilon}{5} = \varepsilon \quad \text{if} \quad |x - 2| < \delta\,.$$

   This proves that $\lim_{x\to 2} x^2 = 4$.

We do not usually rely on the formal definition of limit to verify specific limits such as those in the last example. Rather, we appeal to general theorems about limits in particular the theorems of the next section.

If a function has a limit at a point, this limit is *unique*.

Theorem 22. *If* $\lim_{x\to a} f(x) = L$ *and* $\lim_{x\to a} f(x) = M$*, then* $L = M$.

Proof (by contradiction). Suppose $L \neq M$.

Let $\varepsilon = \frac{|L-M|}{2} > 0$.

Since $\lim_{x\to a} f(x) = L$, there exists $\delta_1 > 0$ such that

$$0 < |x-a| < \delta_1 \implies |f(x) - L| < \varepsilon$$

and since $\lim_{x\to a} f(x) = M$, there exists $\delta_2 > 0$ such that

$$0 < |x-a| < \delta_2 \implies |f(x) - M| < \varepsilon\,.$$

Let $\delta = \min\{\delta_1, \delta_2\}$. If $0 < |x-a| < \delta$, then $|x-a| < \delta_1$, so $|f(x)-L| < \varepsilon$, and $|x-a| < \delta_2$, so $|f(x) - M| < \varepsilon$. Therefore,

$$\begin{aligned} |L-M| &= |L - f(x) + f(x) - M| \\ &\leq |L - f(x)| + |f(x) - M| \\ &< 2\varepsilon = 2\frac{|L-M|}{2} = |L-M| \end{aligned}$$

by the triangle inequality.

This is a contradiction, so $L = M$.

Although a function f can only have one limit at any particular point, it is, nevertheless, useful to be able to describe the behaviour of functions that approach different numbers a x tends to a from one side or the other.

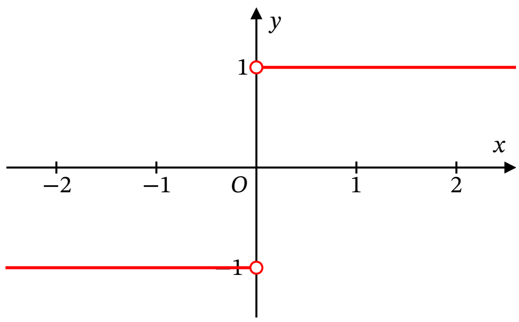

Figure 3–4: $\operatorname{sgn} \lim_{x\to 0^-} \operatorname{sgn} x = -1$ *and* $\lim_{x\to 0^+} \operatorname{sgn} x = 1$.

Definition 21. We say that $f : X \mapsto Y$ has *right limit* L at a, and we write

$$\lim_{x\to a^+} f(x) = L\,,$$

if the following condition is satisfied:

$$\forall \varepsilon > 0, \exists \delta(\varepsilon) > 0 : a < x < a + \delta \implies x \in X \wedge |f(x) - L| < \varepsilon\,.$$

Similarly, we say that $f : X \mapsto Y$ has *left limit* L at a, and we write

$$\lim_{x\to a^-} f(x) = L\,,$$

if the following condition is satisfied:

$$\forall \varepsilon > 0, \exists \delta(\varepsilon) > 0 : a - \delta < x < a \implies x \in X \wedge |f(x) - L| < \varepsilon\,.$$

Note again the dependency of δ on ε and the fact that x belongs to the domain of f.

Example. Show that $\lim_{x\to 0^+} \sqrt{x} = 0$.

Let $\varepsilon > 0$ be given. If $x > 0$, then $\left|\sqrt{x} - 0\right| = \sqrt{x}$. We can ensure that $\sqrt{x} < \varepsilon$ by requiring $x < \varepsilon^2$. Thus, we can take $\delta = \varepsilon^2$ and the condition of the definition will be satisfied:

$$0 < x < \delta = \varepsilon^2 \implies \left|\sqrt{x} - 0\right| < \varepsilon\,.$$

Therefore, $\lim_{x\to 0^+} \sqrt{x} = 0$.

The existence of different right and left limits of a function at a point excludes the existence of a limit at that point.

Theorem 23.
A function f has limit L at $x = a$ if and only if it has both left and right limits there and these one-sided limits are both equal to L:

$$\lim_{x\to a} f(x) = L \iff \lim_{x\to a^-} f(x) = \lim_{x\to a^+} f(x) = L\,.$$

Example. What one-sided limits does $g(x) = \sqrt{1 - x^2}$ have at $x = -1$ and $x = 1$?

The domain of g is $[-1,1]$, so g is defined only to the right of $x = -1$ and only to the left of $x = 1$.

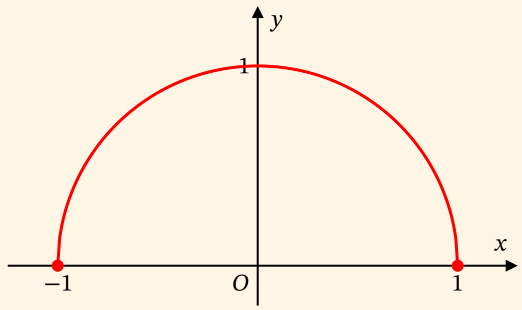

Figure 3–5: $\sqrt{1-x^2}$ *has right limit 0 at -1 and left limit 0 at 1.*

As can be seen in the figure,

$$\lim_{x\to -1^+} g(x) = 0 \quad \text{and} \quad \lim_{x\to 1^-} g(x) = 0$$

g has no left limit or limit at $x = -1$, no right limit or limit at $x = 1$.

3–4 *Rules and Theorems about Limits*

The following rules make it easy to calculate limits and one-sided limits of many kinds of functions when we know some elementary limits.

Theorem 24 (Limit Rules). *If* $\lim_{x\to a} f(x) = L$, $\lim_{x\to a} g(x) = M$ *and* k *is a constant, then*

1. $\lim_{x\to a}(f(x)+g(x)) = L+M$ *(limit of a sum)*
2. $\lim_{x\to a}(f(x)-g(x)) = L-M$ *(limit of a difference)*
3. $\lim_{x\to a} f(x)g(x) = LM$ *(limit of a product)*
4. $\lim_{x\to a} kf(x) = kL$ *(limit of a multiple)*
5. $\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{L}{M}$ if $M \neq 0$ *(limit of quotient)*
6. *If* $m \in \mathbb{Z}$ *and* $n \in \mathbb{N}_0$, *then* $\lim_{x\to a}(f(x))^{\frac{m}{n}} = L^{\frac{m}{n}}$, *provided* $L > 0$ *if* n *is even and* $L \neq 0$ *if* $m < 0$. *(limit of a power)*
7. *If* $f(x) \leq g(x)$ *on an interval containing* a *in its interior, then* $L \leq M$. *(order is preserved)*

Rules 1—6 are also valid for one-sided limits. So is rule 7, under the assumption that $f(x) \leq g(x)$ on an open interval extending from a in the appropriate direction.

Proof (limit of a sum). Let $\varepsilon > 0$ be given.

We want to find a strict positive number δ such that

$$0 < |x-a| < \delta \implies |(f(x)+g(x)) - (L+M)| < \varepsilon$$

Observe that

$$\begin{aligned}|(f(x)+g(x)) - (L+M)| &= |(f(x)-L) + (g(x)-M)| \\ &\leq |f(x)-L| + |g(x)-M|\end{aligned}$$

by the triangle inequality.

Since $\lim_{x\to a} f(x) = L$ and $\frac{\varepsilon}{2}$ is a strict positive number, there exists a number δ_1 such that

$$0 < |x-a| < \delta_1 \implies |f(x)-L| < \frac{\varepsilon}{2}\,.$$

Similarly, since $\lim_{x\to a} g(x) = M$ and $\frac{\varepsilon}{2}$ is a strict positive number, there exists a number δ_2 such that

$$0 < |x-a| < \delta_2 \implies |g(x)-M| < \frac{\varepsilon}{2}\,.$$

Let $\delta = \min\{\delta_1, \delta_2\}$. If $0 < |x-a| < \delta$, then $|x-a| < \delta_1$, so $|f(x)-L| < \frac{\varepsilon}{2}$, and $|x-a| < \delta_2$, so $|g(x)-M| < \frac{\varepsilon}{2}$. Therefore,

$$|(f(x)+g(x)) - (L+M)| \leq |f(x)-L| + |g(x)-M| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\,.$$

This shows that $\lim_{x\to a}(f(x)+g(x)) = L+M$.

Example. Find $\lim_{x\to a} \frac{x^2+x+4}{x^3-2x^2+7}$.

The expression $\frac{x^2+x+4}{x^3-2x^2+7}$ is formed by combining the basic functions x and c (constant) using addition, subtraction, multiplication, and division. The previous theorem assures us that the limit of such a combination is the same combination of the limits a and c of the basic functions, provided the denominator does not have limit zero. Thus,

$$\lim_{x\to a} \frac{x^2+x+4}{x^3-2x^2+7} = \frac{a^2+a+4}{a^3-2a^2+7} \quad \text{provided } a^3-2a^2+7 \neq 0$$

The result of the example can be generalized as a direct corollary.

Corollary 25. *If $a \in \mathbb{R}$ and*

1. *$P(x)$ is a polynomial, then*

$$\lim_{x \to a} P(x) = P(a)$$

2. *$P(x)$ and $Q(x)$ are polynomials and $Q(a) \neq 0$, then*

$$\lim_{x \to a} \frac{P(x)}{Q(x)} = \frac{P(a)}{Q(a)} \,.$$

The following theorem will enable us to calculate some very important limits in subsequent chapters. It is called the *squeeze theorem* because it refers to a function g whose values are squeezed between the values of two other functions that have the same limit L at a point a. Being trapped between the values of two functions that approach L, the values of g must also approach L.

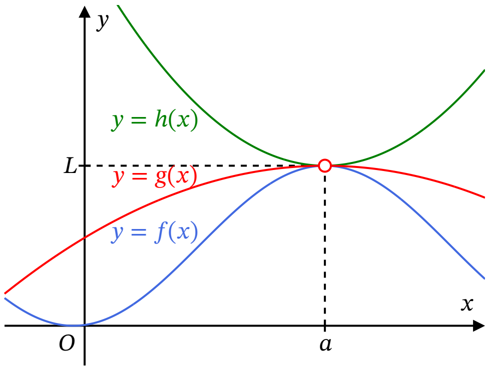

Figure 3–6: The graph of g is squeezed between those of f and h.

Theorem 26 (Squeeze theorem).
If $f(x) \leq g(x) \leq h(x)$ holds for all x in some open interval containing a, except possibly at $x = a$, and

$$\lim_{x \to a} f(x) = \lim_{x \to a} h(x) = L\,,$$

then,

$$\lim_{x \to a} g(x) = L\,.$$

Similar statements hold for one-sided limits.

Proof. Let $\varepsilon > 0$ be given.

We want to find a strict positive number δ such that

$$0 < |x-a| < \delta \implies |g(x) - L| < \varepsilon$$

Observe that

$$\begin{aligned} |g(x) - L| &= |g(x) - f(x) + f(x) - L| \\ &\leq |g(x) - f(x)| + |f(x) - L| \\ &\leq |h(x) - f(x)| + |f(x) - L| = |h(x) - L + L - f(x)| + |f(x) - L| \\ &\leq |h(x) - L| + |L - f(x)| + |f(x) - L| \end{aligned}$$

by the triangle inequality and the squeezing of g between f and h.

Since $\lim_{x\to a} f(x) = L$ and $\frac{\varepsilon}{3}$ is a strict positive number, there exists a number δ_1 such that

$$0 < |x-a| < \delta_1 \implies |f(x) - L| < \frac{\varepsilon}{3}\,.$$

Similarly, since $\lim_{x\to a} h(x) = L$ and $\frac{\varepsilon}{3}$ is a strict positive number, there exists a number δ_2 such that

$$0 < |x-a| < \delta_2 \implies |h(x) - L| < \frac{\varepsilon}{3}\,.$$

Let $\delta = \min\{\delta_1, \delta_2\}$. If $0 < |x-a| < \delta$, then $|x-a| < \delta_1$, so $|f(x) - L| < \frac{\varepsilon}{3}$, and $|x-a| < \delta_2$, so $|h(x) - L| < \frac{\varepsilon}{3}$. Therefore,

$$|g(x) - L| \leq |h(x) - L| + |L - f(x)| + |f(x) - L| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon\,.$$

This shows that $\lim_{x\to a} g(x) = L$.

Example. Show that if $\lim_{x\to a}|f(x)| = 0$, then $\lim_{x\to a} f(x) = 0$.

Since $-|f(x)| \leq f(x) \leq |f(x)|$, and $-|f(x)|$ and $|f(x)|$ both have limit 0 as x tends to a, so does $f(x)$ by the squeeze theorem.

3–5 Limits at Infinity and Infinite Limits

We will extend the concept of limit to allow for two situations not covered by the definitions of limit and one-sided limit in the previous section:

1. *limits at infinity*, where x becomes arbitrarily large, positive or negative;
2. *infinite limits*, which are not real limits at all but provide usefull symbolism for describing the behaviour of functions whose values become arbitrarily large, positive or negative.

Example. How behaves the function

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

whose graph is shown in the next figure and for which some values are given in the following table for values of x that becomes arbitrarly large, positive and negative?

x	$f(x) = \frac{x}{\sqrt{x^2+1}}$	x	$f(x) = \frac{x}{\sqrt{x^2+1}}$
−1000	−0.9999995	1	0.707106781
−100	−0.999950004	10	0.99503719
−10	−0.99503719	100	0.999950004
1	0.707106781	1000	0.9999995

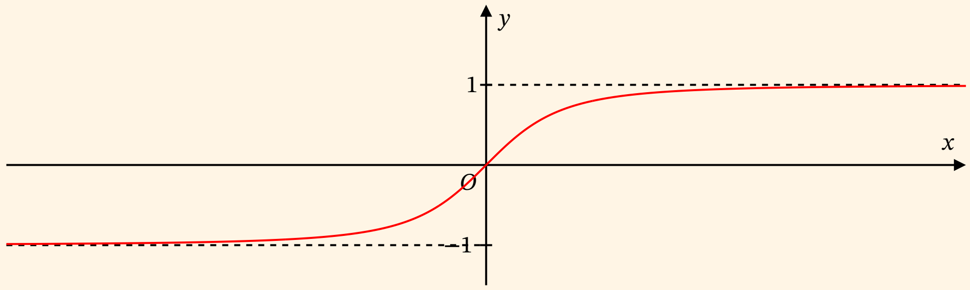

Figure 3–7: The graph of $\frac{x}{\sqrt{x^2+1}}$.

The values of $f(x)$ seem to approach 1 as x takes on larger and larger positive values, and -1 as x takes on negative values that get larger and larger in absolute value. We express this behaviour by writing

$$\lim_{x\to\infty} f(x) = 1 \quad \text{"}f(x)\text{ approaches 1 as }x\text{ approaches infinity."}$$

$$\lim_{x\to-\infty} f(x) = -1 \quad \text{"}f(x)\text{ approaches }-1\text{ as }x\text{ approaches negative infinity."}$$

The graph of f conveys this limiting behaviour by approaching the horizontal lines $y = 1$ as x moves far to the right and $y = -1$ as x moves far to the left. These lines are called *horizontal asymptotes` of the graph. In general, if a curve approaches a straight line as it recedes very far away from the origin, that line is called an *asymptote* of the curve.

This example suggest the following definition of a limit at infinity.

Definition 22. We say that $f : X \mapsto Y$ approaches the limit L as x tends to infinity, and we write

$$\lim_{x\to\infty} f(x) = L\,,$$

if the following condition is satisfied:

$$\forall\varepsilon > 0, \exists R(\varepsilon) : x > R \implies x \in X \wedge |f(x) - L| < \varepsilon\,.$$

Similarly, we say that $f : X \mapsto Y$ approaches the limit L as x tends to negative infinity, and we write

$$\lim_{x\to-\infty} f(x) = L\,,$$

if the following condition is satisfied:

$$\forall\varepsilon > 0, \exists R(\varepsilon) : x < R \implies x \in X \wedge |f(x) - L| < \varepsilon\,.$$

Example. Show that

$$\lim_{x \to \infty} \frac{1}{x} = 0\,.$$

Let ε be a given positive number. For $x > 0$, we have

$$\left|\frac{1}{x} - 0\right| = \left|\frac{1}{x}\right| = \frac{1}{x} < \varepsilon \quad \text{provided} \quad x > \frac{1}{\varepsilon}\,.$$

Therefore, the condition of the definition is satisfied with $R = \frac{1}{\varepsilon}$. We have shown that $\lim_{x \to \infty} \frac{1}{x} = 0$.

The rules and theorems of previous section have suitable counterparts for limits at infinity.

Example. Evaluate $\lim\limits_{x \to \infty} f(x)$ and $\lim\limits_{x \to -\infty} f(x)$ for $f(x) = \dfrac{x}{\sqrt{x^2+1}}$.

Rewrite the expression for $f(x)$ as follows:

$$\begin{aligned} f(x) &= \frac{x}{\sqrt{x^2+1}} = \frac{x}{\sqrt{x^2\left(1+\frac{1}{x^2}\right)}} \\ &= \frac{x}{\sqrt{x^2}\sqrt{1+\frac{1}{x^2}}} = \frac{x}{|x|\sqrt{1+\frac{1}{x^2}}} \\ &= \frac{\operatorname{sgn} x}{\sqrt{1+\frac{1}{x^2}}}\,. \end{aligned}$$

The factor $\sqrt{1+\frac{1}{x^2}}$ approaches 1 as x approaches ∞ or $-\infty$, so $f(x)$ must have the same limits as $x \to \pm\infty$ as does $\operatorname{sgn} x$. Therefore,

$$\lim_{x \to \infty} f(x) = 1 \quad \text{and} \quad \lim_{x \to -\infty} f(x) = -1$$

The only polynomials that have limits at infinity are constant ones. The situation is more interesting for rational functions. The following examples show how to render such a function in a form where its limits at infinity (if they exist) are apparent. The way to do this is to *divide the numerator and the denominator by the highest power of x in the denominator.

Example. Evaluate

$$\lim_{x \to \infty} \frac{2x^2 - x + 3}{3x^2 + 5}\,.$$

Divide the numerator and the denominator by x^2, the highest power of x appearing in the denominator:

$$\lim_{x \to \infty} \frac{2x^2 - x + 3}{3x^2 + 5} = \lim_{x \to \infty} \frac{2 - \frac{1}{x} + \frac{3}{x^2}}{3 + \frac{5}{x^2}} = \frac{2 - 0 + 0}{3 + 0} = \frac{2}{3}\,.$$

The technique used in the previous example can also be applied to more general kinds of functions.

Example. Find $\lim\limits_{x\to\infty}\left(\sqrt{x^2+x}-x\right)$.

We can rationalize the expression by multiplying the numerator and the denominator (which is 1) by the conjugate expression $\sqrt{x^2+x}+x$.

$$\lim_{x\to\infty}\left(\sqrt{x^2+x}-x\right)=\frac{\left(\sqrt{x^2+x}-x\right)\left(\sqrt{x^2+x}+x\right)}{\left(\sqrt{x^2+x}+x\right)}=\frac{x^2+x-x^2}{\left(\sqrt{x^2\left(1+\frac{1}{x}\right)}+x\right)}$$

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

A function whose values grow arbitrarily large can sometimes said to have an infinite limite. Since infinity is not a number, infinite limits are not really limits at all, but they provide a way of describing the behaviour of functions that grow arbitrarily large positive of negative.

Example. Describe the behaviour of the function $f(x)=\dfrac{1}{x^2}$ near $x=0$.

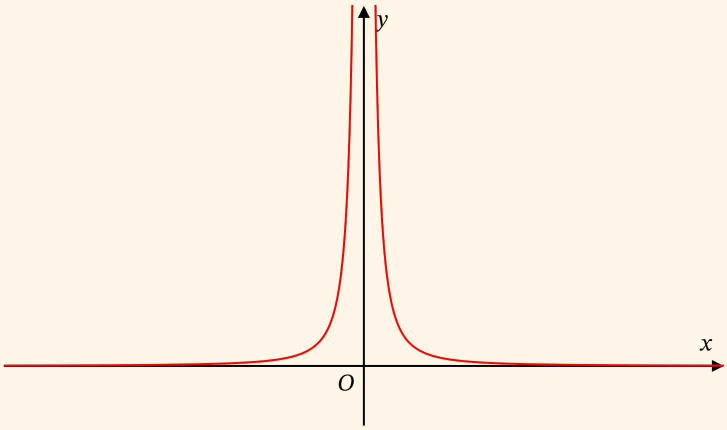

Figure 3–8: The graph of $\frac{1}{x^2}$.

As x approaches 0 from either side, the values of $f(x)$ are positive and grow larger and larger, so the limit of $f(x)$ as x approaches 0 *does not exist*. it is nevertheless convenient to describe the behaviour of f near 0 by saying that $f(x)$ approaches ∞ as x tends to zero. We write

$$\lim_{x\to 0}f(x)=\lim_{x\to 0}\frac{1}{x^2}=\infty\,.$$

Note that in writing this we are *not* saying that $\lim_{x\to 0}\frac{1}{x^2}$ *exists*. Rather, we are saying that the limit *does not exist because $\frac{1}{x^2}$ *becomes arbitrarily large near* $x=0$. Observe how the graph of f approaches the y-axis as x tends to 0. the y-axis is a **vertical asymptote** of the graph.

Definition 23. We say that $f:X\mapsto Y$ approaches infinity as x tends to a and write

$$\lim_{x\to a}f(x)=\infty\,.$$

if the following condition is satisfied:

$$\forall B>0,\exists\delta(B)>0:0<|x-a|<\delta\implies x\in X\wedge f(x)>B\,.$$

Example. Verify that

$$\lim_{x\to 0} \frac{1}{x^2} = \infty\,.$$

Let B be any positive number. We have

$$\frac{1}{x^2} > B \quad \text{provided that} \quad x^2 < \frac{1}{B}\,.$$

If $\delta = \frac{1}{\sqrt{B}}$, then

$$0 < |x| < \delta \quad\implies\quad x^2 < \delta^2 = \frac{1}{B} \quad\implies\quad \frac{1}{x^2} > B\,.$$

Therefore, $\lim_{x\to 0} \frac{1}{x^2} = \infty$.

Example. Describe the behaviour of the function $f(x) = \frac{1}{x}$ near $x = 0$.

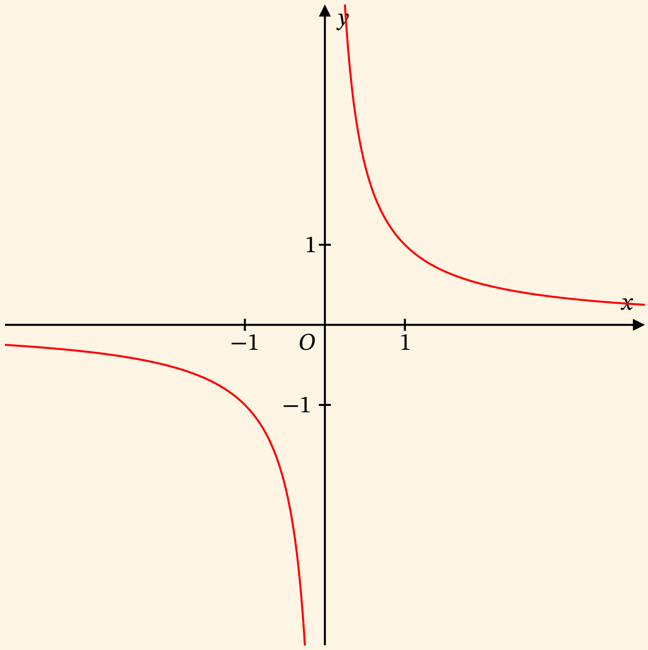

Figure 3–9: The graph of $\frac{1}{x}$.

As x approaches 0 from the right, the values of $f(x)$ become larger and larger positive numbers, and we say that f has right-hand limit infinity at $x = 0$:

$$\lim_{x\to 0^+} f(x) = \infty$$

Similarly, the values of $f(x)$ become larger and larger negative numbers as x approaches 0 from the left, so f has left-hand limit $-\infty$ at $x = 0$:

$$\lim_{x\to 0^-} f(x) = -\infty$$

These statements do not say that the one-sided limits *exist*; they do not exist because ∞ and $-\infty$ are not numbers. Since the one-sided limits are not equal even as infinite symbols, all we can say about the two-sided $\lim_{x\to 0} f(x)$ is that it does not exist.

We can now say a bit more about the limits at infinity and negative infinity of a rational function whose nuberator has higher degree than the denominator. Earlier we said that such a limit *does not exist*. This is true, but we can assign ∞ or $-\infty$ to such limits, as the following example shows.

Example. Evaluate

$$\lim_{x\to\infty} \frac{x^3+1}{x^2+1}\,.$$

Divide the numerator and the denominator by x^2, the largest power of the denominator:

$$\lim_{x\to\infty} \frac{x^3+1}{x^2+1} = \lim_{x\to\infty} \frac{x+\frac{1}{x^2}}{1+\frac{1}{x^2}} = \frac{\lim_{x\to\infty} x+\frac{1}{x^2}}{1} = \infty\,.$$

A polynomial $Q(x)$ of degree $n > 0$ can have at most n *zeros*; that is, there are at most n different real numbers r for which $Q(r) = 0$. If $Q(r)$ is the denominator of a rational function $R(x) = \dfrac{P(x)}{Q(x)}$, that function will be defined for all x except those finitely many zeros of Q. At each of those zeros, $R(x)$ may have limits, infinite limits, or one-sided infinite limits. Here are some examples.

Example.

1. $\displaystyle\lim_{x\to 2} \frac{(x-2)^2}{x^2-4} = \lim_{x\to 2} \frac{(x-2)^2}{(x-2)(x+2)} = \lim_{x\to 2} \frac{x-2}{x+2} = 0.$
2. $\displaystyle\lim_{x\to 2} \frac{x-2}{x^2-4} = \lim_{x\to 2} \frac{x-2}{(x-2)(x+2)} = \lim_{x\to 2} \frac{1}{x+2} = \frac{1}{4}.$
3. $\displaystyle\lim_{x\to 2^+} \frac{x-3}{x^2-4} = \lim_{x\to 2^+} \frac{x-3}{(x-2)(x+2)} = -\infty.$
4. $\displaystyle\lim_{x\to 2^-} \frac{x-3}{x^2-4} = \lim_{x\to 2^-} \frac{x-3}{(x-2)(x+2)} = \infty.$
5. $\displaystyle\lim_{x\to 2} \frac{x-3}{x^2-4} = \lim_{x\to 2} \frac{x-3}{(x-2)(x+2)}$ does not exist.
6. $\displaystyle\lim_{x\to 2} \frac{2-x}{(x-2)^3} = \lim_{x\to 2} \frac{-(x-2)}{(x-2)^3} = \lim_{x\to 2} \frac{-1}{(x-2)^2} = -\infty.$

3–6 Continuity Defined

When a car is driven along a highway, its distance from its starting point depends on time in a *continuous* way, changing by small amounts over short intervals of time. But not all quantities change in this way. When the car is parked in a parking lot where the rate is quoted as "€2,00 per hour or portion," the parking charges remain at €2,00 for the first hour and then suddenly jump to €4,00 as soon as the first hour has passed. The function relating parking charges to parking time will be called *discontinuous* at each hour.

Most functions that we encounter have domains that are intervals, or unions of separate intervals. A point P in the domain of such a function is called an *interior point* of the domain if it belongs to some open interval contained in the domain. If it is not an interior point, then P is called an *endpoint* of the domain. For example, the domain of the function $f(x) = \sqrt{4-x^2}$ is the closed interval $[-2{,}2]$, which consists of interior points in the interval $]-2{,}2[$, a left endpoint -2, and a right endpoint 2. The domain of the function $g(x) = \dfrac{1}{x}$ is the union of open intervals $]-\infty,0[\,\cup\,]0,\infty[$ and consists entirely of interior points. Note that although 0 is an endpoint of each of those intervals, it does not belong to the domain of g and so is not an endpoint of that domain.

Definition 24 (Continuity of a function at a point).
A function f, defined on an open interval containing the point c, an interior point, is said to be continuous at the point c if

$$\lim_{x \to c} f(x) = f(c);$$

that is,

$$\forall \varepsilon > 0, \exists \delta(\varepsilon) > 0 : |x - c| < \delta \implies |f(x) - f(c)| < \varepsilon .$$

If either $\lim_{x \to c} f(x)$ fails to exist or it exists but is not equal to $f(c)$, then we will say that f is discontinuous at c.

In graphical terms, f is continuous at an interior point c of its domain if its graph has no break in it at the point $(c, f(x))$; in other words, if you can draw the graph through that point without lifting your pen from the paper.

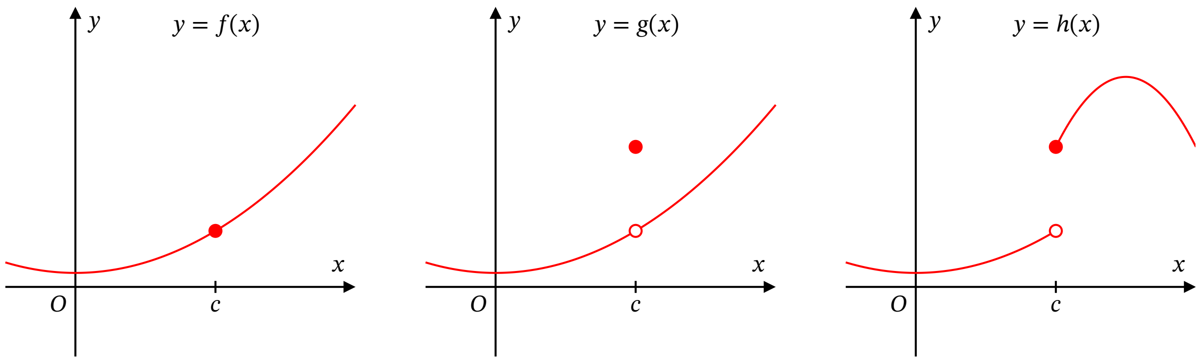

Figure 3–10: f is continuous at c; $\lim_{x \to c} g(x) \neq g(c)$; $\lim_{x \to c} h(x)$ does not exist.

Although a function cannot have a limit at an endpoint of its domain, it can still have a one-sided limit there. We extend the definition of continuity to provide for such situations.

Definition 25. We say that f is *right continuous* at c if $\lim_{x \to c^+} f(x) = f(c)$.

We say that f is *left continuous* at c if $\lim_{x \to c^-} f(x) = f(c)$.

Example.
The Heaviside function $H(x)$ is continuous at every number x except 0. It is right continuous at 0 but is not left continuous or continuous there.

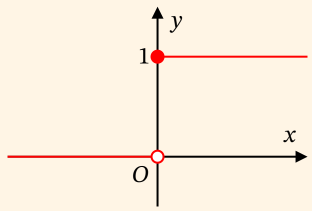

Figure 3–11: The Heaviside function

The relationship between continuity and one-sided continuity is summarized in the following theorem.

Theorem 27.
Function f is continuous at c if and only if it is both right continuous and left continuous at c.

We have defined the concept of continuity at a point. Of greater importance is the concept of continuity on an interval.

Definition 26 (Continuity of a function on an interval).
A function f is continuous on an interval if it is continuous at every point of that interval. In the case of an endpoint of a closed interval, f need only be continuous on one side. Thus, f is continuous on the interval $[a, b]$ if

$$\lim_{x \to t} f(x) = f(t) \quad \forall t : a < t < b\,,$$

and

$$\lim_{x \to a^+} f(x) = f(a) \quad \text{and} \quad \lim_{x \to b^-} f(x) = f(b)\,.$$

These concepts are illustrated in following figure.

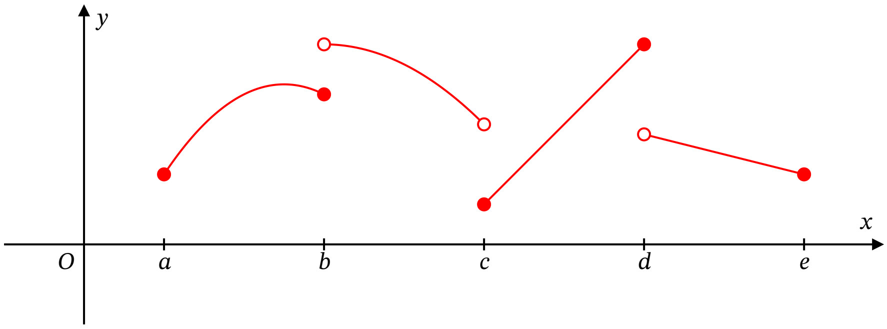

Figure 3–12: f is continuous on the intervals $[a, b]$, $]b, c[$, $[c, d]$, $]d, e]$.

3–7 Continuous Functions

Definition 27 (Continuous function).
We say that f is a *continuous function* if f is continuous at every point of its domain.

Example.
The function $f(x) = \sqrt{x}$ is a continuous function. Its domain is $[0, \infty[$. It is continuous at the left endpoint 0 because it is right continuous there. Also, f is continuous at every number $c > 0$ since $\lim_{x \to c} \sqrt{x} = \sqrt{c}$.

Example.
The function $g(x) = \dfrac{1}{x}$ is also a continuous function. This may seem wrong to you at first glance because its graph is broken at $x = 0$. However, the number 0 is not in the domain of g, so we will prefer to say that g is undefined rather than discontinuous there.

The following functions are continuous wherever they are defined:

1. all polynomials;
2. all rational functions;
3. all rational powers $x^{\frac{m}{n}} = \sqrt[n]{x^m}$;
4. the sine, cosine, tangent, secant, cosecant, and cotangent functions; and
5. the absolute value function $|x|$.

Corollary 4 of this chapter assures us that every polynomial is continuous everywhere on the real line, and every rational function is continuous everywhere on its domain (which consists of all real numbers except the finitely many where its denominator is zero). If m and n are integers and $n \neq 0$, the rational power function $x^{\frac{m}{n}}$ is defined for all positive numbers x, and also for all negative numbers x if n is odd. The domain includes 0 if and only if $\frac{m}{n} \geq 0$.

The following theorems show that if we combine continuous functions in various ways, the results will be continuous.

Theorem 28. *If f and g are continuous at the point c, then so are $f+g$, $f-g$, fg, and, if $g(c) \neq 0$, $\frac{f}{g}$.*

Proof. This is just a restatement of various rules for combining limits; for example,

$$\lim_{x \to c} f(x)g(x) = \Big(\lim_{x \to c} f(x)\Big)\Big(\lim_{x \to c} g(x)\Big) = f(c)g(c)\,.$$

Theorem 29. *If f is a continuous function at the point L and if $\lim_{x \to c} g(x) = L$, then we have*

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

Proof. Let $\varepsilon > 0$ be given.

Since f is continuous at L, there exists $\kappa > 0$ such that $|f(g(x)) - f(L)| < \varepsilon$ whenever $|g(x) - L| < \kappa$.

Since $\lim_{x \to c} g(x) = L$, there exists $\delta > 0$ such that if $0 < |x - c| < \delta$, then $|g(x) - L| < \kappa$.

Hence, if $0 < |x - c| < \delta$, then $|f(g(x)) - f(L)| < \varepsilon$, and $\lim_{x \to c} f(g(x)) = f(L)$.

3–8 Continuous Extensions

As we have seen a rational function may have a limit even at a point where its denominator is zero. If $f(c)$ is not defined, but $\lim_{x \to c} f(x) = L$ exists, we can define a new function $F(x)$ by

$$F(x) = \begin{cases} f(x) & \text{if } x \text{ is in the domain of } f \\ L & \text{if } x = c\,. \end{cases}$$

$F(x)$ is continuous at $x = c$. It is called the *continuous extension* of $f(x)$ to $x = c$. For rational functions, continuous extensions are usually found by cancelling common factors.

Example. Show that $f(x) = \frac{x^2 - x}{x^2 - 1}$ has a continuous extension to $x = 1$, and find that extension.

Although $f(1)$ is not defined, if $x \neq 1$ we have

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

The function $F(x) = \dfrac{x}{x+1}$ is equal to $f(x)$ for $x \neq 1$ but is also continuous at $x = 1$, having there the value $\frac{1}{2}$. The continuous extension of $f(x)$ to $x = 1$ is $F(x)$. It has the same graph as $f(x)$ except with no hole at $\left(1, \frac{1}{2}\right)$.

If a function f is undefined or discontinuous at a point c but can be (re)defined at that single point so that it becomes continuous there, then we say that f has a *removable discontinuity* at c. The function f in the above example has a removable discontinuity at $x = 1$. To remove it, define $f(1) = \frac{1}{2}$.

Example.

The function $g(x) = \begin{cases} x & \text{if } x \neq 2 \\ 1 & \text{if } x = 2 \end{cases}$ has a removable discontinuity at $x = 2$. To remove it, redefine $g(2) = 2$
.

3–9 *The Intermediate-Value Theorem*

Continuous functions that are defined on closed, finite intervals have special properties that make them particularly useful in mathematics and its applications. We will discuss two of these properties here. Although they may appear obvious, these properties are much more subtle than the results about limits stated earlier in this chapter; their proofs require a careful study of the implications of the completeness property of the real numbers and are based on the Nested Intervals theorem.

The first property of a continuous function defined on a closed, finite interval is that the function takes on all real values between any two of its values. This property is called the intermediate-value property.

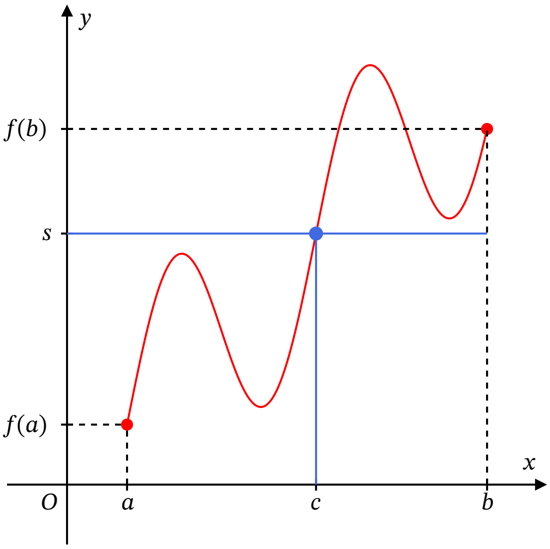

Figure 3–13: The continuous function f takes on the value s at some point c between a and b.

The figure shows a typical situation. The points $(a, f(a))$ and $(b, f(b))$ are on opposite sides of the horizontal line $y = s$. Being unbroken, the graph $y = f(x)$ must cross this line in order to go from one point to the other. In the figure, it crosses the line only once, at $x = c$. If the line $y = s$ were somewhat lower or higher, there might have been three crossings and three possible values for c.

We need the following lemma to prove the intermediate-value property.

Lemma 30 (Aura theorem). *Let f be continuous at c.*

1. *If $f(c) > 0$, then $f(x) > 0$ for all x in some open interval containing c.*
2. *If $f(c) < 0$, then $f(x) < 0$ for all x in some open interval containing c.*

Proof. Let $f(c) > 0$.

Then, corresponding to $\varepsilon = \dfrac{f(c)}{2} > 0$ there exists a corresponding $\delta > 0$ such that $|x - c| < \delta$ implies

$$|f(x) - f(c)| < \frac{f(c)}{2} \iff -\frac{f(c)}{2} < f(x) - f(c) < \frac{f(c)}{2} \iff 0 < \frac{f(c)}{2} < f(x) < \frac{3f(c)}{2}.$$

Hence, $f(x) > 0$ for all $x \in \left]c - \delta, c + \delta\right[$.

Exercise. Prove the second part of the Aura theorem.

We will first prove a special case from which the general case follows easily.

Theorem 31 (Bolzano's theorem).
Let f be a continuous function defined on $[a,b]$. If $f(a) < 0$ and $f(b) > 0$, then there exists $c \in\,]a,b[$ such that $f(c) = 0$.

Proof (by contradiction). Let $I_0 = [a_0, b_0] = [a,b]$.

If $f\left(\dfrac{a_0+b_0}{2}\right) = 0$, we are done. Otherwise, f changes sign on either $\left[a_0, \dfrac{a_0+b_0}{2}\right]$ or $\left[\dfrac{a_0+b_0}{2}, b_0\right]$.

Let $I_1 = [a_1, b_1]$ be the subinterval on which f changes sign and repeat.

By the Nested Intervals theorem, $\bigcap_{n\in\mathbb{N}} I_n = \{c\}$, where $c = \sup\{a_n\} = \inf\{b_n\}$.

Suppose $f(c) > 0$. By the Aura theorem, f must be positive on an open interval containing c. Since $c = \sup\{a_n\}$, by the Capture theorem this open interval must contain some a_m. But $f(a_m) < 0$ which is a contradiction.

Suppose $f(c) < 0$. By the Aura theorem, f must be negative on an open interval containing c. Since $c = \inf\{b_n\}$, by the Capture theorem this open interval must contain some b_m. But $f(b_m) > 0$ which is a contradiction.

Hence, $f(c) = 0$.

If $f(a) > 0$ and $f(b) < 0$, then $g = -f$ satisfies the hypotheses of Bolzano's theorem and therefore $g(c) = -f(c) = 0$ for some $c \in\,]a,b[$, and hence $f(c) = 0$.

Theorem 32 (Intermediate-value theorem).
If f is continuous on the interval $[a,b]$ and if s is a number between $f(a)$ and $f(b)$, then there exists a number $c \in\,]a,b[$ such that $f(c) = s$.

Proof. Let s be a number between $f(a)$ and $f(b)$.

The function $g(x) = f(x) - s$ is continuous and satifies the hypotheses of Bolzano's theorem, so there exists some $c \in\,]a,b[$ such that $g(c) = f(c) - s = 0$.

Hence, $f(c) = s$.

Example. Show that the equation $x^3 - x - 1 = 0$ has a solution in the interval [1,2].

The function $f(x) = x^3 - x - 1$ is a polynomial and is therefore continuous everywhere.

Now $f(1) = -1$ and $f(2) = 5$. Since 0 lies between -1 and 5, the Intermediate-value theorem assures us that there must be a number $c \in [1,2]$ such that $f(c) = 0$.

One method for finding a zero of a function that is continuous and changes sign on an interval involves bisecting the interval many times, each time determining which half of the previous interval must contain the root, because the function has opposite signs at the two ends of that half.

Example. Solve the equation $x^3 - x - 1 = 0$ correct to 3 decimal places by successive bisection.

i	a_i	b_i	$\dfrac{a_i + b_i}{2}$	$f\left(\dfrac{a_i + b_i}{2}\right)$
0	1	2	1.5	0.875
1	1	1.5	1.25	−0.2969
2	1.25	1.5	1.375	0.2246
3	1.25	1.375	1.3125	−0.0515
4	1.3125	1.375	1.34375	0.0826
5	1.3125	1.34375	1.328125	0.0146
6	1.3125	1.328125	1.3203125	−0.0187
7	1.3203125	1.328125	1.32421875	−0.0021
8	1.32421875	1.328125	1.326171875	0.0062
9	1.32421875	1.326171875	1.3251953125	0.002
10	1.32421875	1.3251953125	1.32470703125	−0.0
11	1.32470703125	1.3251953125	1.324951171875	0.001

The root is 1.325 rounded to 3 decimal places.

This method is slow. For example, if the original interval has length 1, it will take 11 bisections to cut down to an interval of length less than 0.0005 (because $2^{11} > 2000 = \frac{1}{0.0005}$, and thus to ensure that we have found the root correct to 3 decimal places.

In chapter 5, calculus will provide us with much faster methods of solving equations such as the one in the example above.

3–10 The Extreme-Value Theorem

The second of the properties states that a function f that is continuous on a closed, finite interval $[a, b]$ must have an absolute maximum value and an absolute minimum value. This means that the values of $f(x)$ at all points of the interval lie between the values of $f(x)$ at two particular points in the interval; the graph of f has a highest point and a lowest point.

To prove the extreme-value theorem, we will first show that a continuous function on a closed interval is bounded; that is, there exists a constant K such that $|f(x)| \leq K$ if $a \leq x \leq b$.

Theorem 33 (Boundedness theorem). *If f is continous on $[a, b]$, then f is bounded on $[a, b]$.*

Proof.
For each $x \in [a, b]$, since f is continuous at x, there exists an open interval $I_x =]x - \delta_x, x + \delta_x[$ such that for all $y \in I_x \cap [a, b]$ we have $|f(y) - f(x)| < 1$.

The set of all such open intervals I_x forms an open cover of $[a, b]$.

By the Heine-Borel theorem there exists a finite subcover such that $[a, b] \subset \bigcup_{i=0}^{n} I_{x_i}$.

For any point $y \in [a, b]$, y must be in at least one of the intervals $I_{x_i} = \left]x_i - \delta_{x_i}, x_i + \delta_{x_i}\right[$. Let's say $y \in \left]x_j - \delta_{x_j}, x_j + \delta_{x_j}\right[$.

By the property of the open interval $\left]x_j - \delta_{x_j}, x_j + \delta_{x_j}\right[$, we know that $\left|f(y) - f(x_j)\right| < 1$, i.e. $f(y) < f(x_j) + 1$ and $f(y) > f(x_j) - 1$.

Let $M = \max\{f(x_i) | 0 \leq i \leq n\} + 1$ and $m = \min\{f(x_i) | 0 \leq i \leq n\} - 1$.

We can conclude that for all $y \in [a, b]$ we have $m < f(y) < M$.

Therefore, f is bounded on $[a, b]$, with m as a lower bound and M as an upper bound.

Finally, we can prove the Extreme-value theorem.

Theorem 34 (Extreme-value theorem).
A continuous function on $[a, b]$ *attains both an absolute maximum and an absolute minimum on* $[a, b]$.

Proof (by contradiction). We prove f has a maximum on $[a, b]$.

Since f is continuous on $[a, b]$, by the Boundedness theorem f is bounded on $[a, b]$.

Since f is bounded, its image set is a nonempty subset of $\mathbb{R}$ which is bounded above, so by the Completeness axiom it has a least upper bound.

Let $M = \sup f([a, b])$. By definition of M, $f(x) \leq M$ for all $x \in [a, b]$.

Suppose that $f(x) < M$ for all $x \in [a, b]$.

Then $g(x) = \dfrac{1}{M - f(x)}$ is continuous on $[a, b]$ and hence bounded on $[a, b]$ by the Boundedness theorem.

So, there exists $K > 0$ such that $\dfrac{1}{M - f(x)} \leq K$ for all $x \in [a, b]$.

It follows that $f(x) \leq M - \dfrac{1}{K}$ for all $x \in [a, b]$, which says that $M - \dfrac{1}{K}$ is an upper bound for $f([a, b])$.

Since $K > 0$, $M - \dfrac{1}{K} < M$. This contradicts the fact that $M = \sup f([a, b])$.

Hence, there must exist $c \in [a, b]$ such that $f(c) = M$.

To see that f has a minimum on $[a, b]$, just note that $h = -f$ is continuous on $[a, b]$ and therefore has a maximum of $[a, b]$.

Exercise.
Complete the proof that a continuous function f on a closed interval $[a, b]$ attains its absolute minimum.

The conclusion of the Extreme-value theorem may fail if the function f is not continuous or if the interval is not closed.

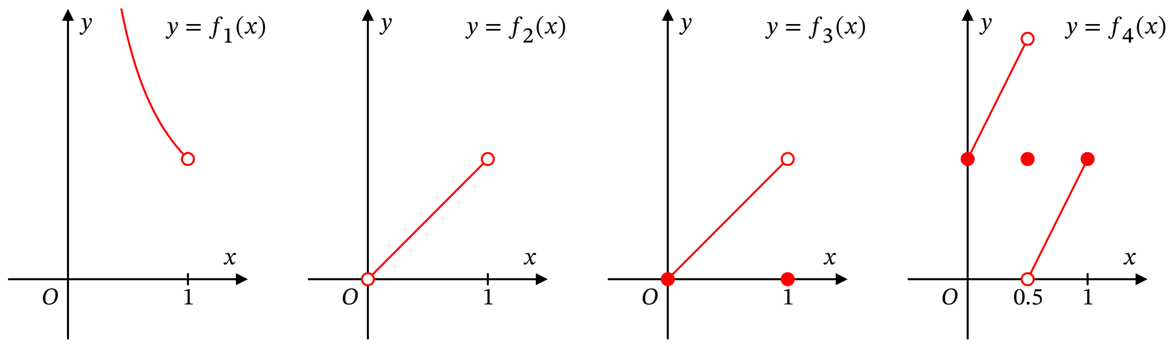

Figure 3–14:

- $f_1(x) = \dfrac{1}{x}$ *is continuous on the open interval* $]0,1[$. *It is not bounded and has neither a maximum nor a minimum value.*
- $f_2(x) = x$ *is continuous on the open interval* $]0,1[$. *It is bounded but neither has a maximum nor a minimum value.*
- f_3 *is defined on the closed interval* $[0,1]$ *but is discontinuous at the endpoint* $x = 1$. *It has a minimum value but no maximum value.*
- f_4 *is discontinuous at an interior point of its domain, the closed interval*$[0,1]$. *It is bounded but has neither maximum nor minimum values.*

The Intermediate-value theorem and the Extreme-value theorem are examples of what mathematicians call *existence theorems*. Such theorems assert that something exists without telling you how to find it.

By the Intermediate-value theorem and the Extreme-value theorem, a continuous function defined on a closed interval takes on all values between its minimum value m and its maximum value M, so its range is also a closed interval, $[m, M]$.

This is the reason why the graph of a function that is continuous on an interval cannot have any breaks. It must be *connected*, a single, unbroken curve with no jumps.

CHAPTER 4

Differentiation

The first of the fundamental problems that are considered in calculus, is the problem of slopes, that is finding the slope of (the tangent line to) a given curve at a given point on the curve. The solution of the problem of slopes is the subject of differential calculus. As we will see, it has many applications in mathematics and other disciplines.

4–1 Tangent Lines and their Slopes

This section deals with the problem of finding a straight line L that is tangent to a curve C at a point P. As is often the case in mathematics, the most important step in the solution of such a fundamental problem is making a suitable definition.

For simplicity, and to avoid certain problems best postponed until later, we will not deal with the most general kinds of curves now, but only with those that are the graphs of continuous functions. Let C be the graph of $y = f(x)$ and let P be the point (x_0, y_0) on C, so that $y_0 = f(x_0)$. We assume that P is not an endpoint of C. Therefore, C extends some distance on both sides of P.

What do we mean when we say that the line L is tangent to C at P?

A reasonable definition of tangency can be stated in terms of limits. If Q is a point on C different from P, then the line through P and Q is called a secant line to the curve. This line rotates around P as Q moves along the curve. If L is a line through P whose slope is the limit of the slopes of these secant lines PQ as Q approaches P along C, then we will say that L is tangent to C at P.

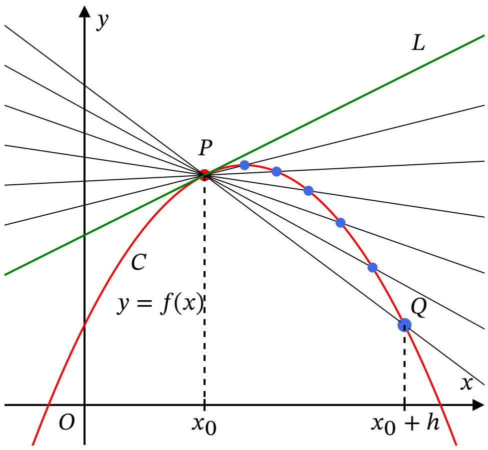

Figure 4–1: Secant lines PQ approach tangent line L as Q approaches P along the curve C.

Since C is the graph of the function $y = (x)$, then vertical lines can meet C only once. Since $P = \left(x_0, y_0\right)$, a different point Q on the graph must have a different x-coordinate, say $x_0 + h$, where $h \neq 0$. Thus $Q = (x_0 + h, f(x_0 + h))$, and the slope of the line PQ is

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

This expression is called the *Newton quotient* or *difference quotient* for f at x_0. Note that h can be positive or negative, depending on whether Q is to the right or left of P.

Definition 28 (Nonvertical tangent lines).
Suppose that the function f is continuous at $x = x_0$ and that

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

exists. Then the straight line having slope m and passing through the point $P = (x_0, f(x_0))$ is called the *tangent line* (or simply the *tangent*) to the graph of $y = f(x)$ at P. An equation of this tangent is

$$y = m(x - x_0) + y_0\,,$$

where $y_0 = f(x_0)$.

Example. Find an equation of the tangent line to the curve $y = x^2$ at the point (1,1).

Here $f(x) = x^2$, $x_0 = 1$, and $y_0 = f(1) = 1$. The slope of the required tangent is

$$\begin{aligned} m = \lim_{h \to 0} \frac{f(1 + h) - f(1)}{h} &= \lim_{h \to 0} \frac{(1 + h)^2 - 1}{h} \\ &= \lim_{h \to 0} \frac{1 + 2h + h^2 - 1}{h} \\ &= \lim_{h \to 0} \frac{2h + h^2}{h} = \lim_{h \to 0} (2 + h) = 2\,. \end{aligned}$$

Accordingly, the equation of the tangent line at (1,1) is $y = 2(x - 1) + 1$, or $y = 2x - 1$.

Vertical Tangent Lines

This definition deals only with tangents that have finite slopes and are, therefore, not vertical. It is also possible for the graph of a function to have a *vertical* tangent line.

Example. Consider the graph of the function $f(x) = \sqrt[3]{x} = x^{\frac{1}{3}}$, which is shown below.

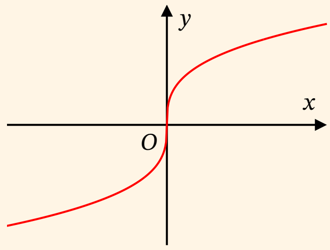

Figure 4–2: The y-axis is tangent to $y = x^{\frac{1}{3}}$ at the origin.

The graph is a smooth curve, and it seems evident that the y-axis is tangent to this curve at the origin. Let us try to calculate the limit of the Newton quotient for f at $x = 0$:

$$\lim_{h\to 0} \frac{f(0+h) - f(0)}{h} = \lim_{h\to 0} \frac{h^{\frac{1}{3}}}{h} = \lim_{h\to 0} \frac{1}{h^{\frac{2}{3}}} = \infty\,.$$

Although the limit does not exist, the slope of the secant line joining the origin to another point Q on the curve approaches infinity as Q approaches the origin from either side.

We extend the definition of tangent line to allow for vertical tangents as follows:

Definition 29 (Vertical tangents).
If f is continuous at $P = (x_0, f(x_0))$, where $y_0 = f(x_0)$, and if either

$$\lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h} = \infty \quad \text{or} \quad \lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h} = -\infty$$

then the vertical line $x = x_0$ is tangent to the graph $y = f(x)$ at P. If the limit of the Newton quotient fails to exist in any other way than by being ∞ or $-\infty$, the graph $y = f(x)$ has no tangent line at P.

Example. Does the graph of $y = |x|$ have a tangent line at $x = 0$?

The Newton quotient here is

$$\frac{|0+h| - |0|}{h} = \frac{|h|}{h} = \operatorname{sgn}(h) = \begin{cases} 1\,, & \text{if } h > 0 \\ -1\,, & \text{if } h < 0\,. \end{cases}$$

Since $\operatorname{sgn}(h)$ has different right and left limits at 0 (namely, 1 and -1), the Newton quotient has no limit as $h \to 0$, so $y = |x|$ has no tangent line at (0,0).

Curves have tangents only at points where they are smooth.

Definition 30.
The slope of a curve C at a point P is the slope of the tangent line to C at P if such a tangent line exists. In particular, the slope of the graph of $y = (x)$ at the point x_0 is

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

Example. Find the slope of the curve $y = \dfrac{x}{3x+2}$ at the point $x = -2$.

If $x = -2$, then $y = \frac{1}{2}$, so the required slope is

$$\begin{aligned} m &= \lim_{h\to 0} \frac{\frac{-2+h}{3(-2+h)+2} - \frac{1}{2}}{h} \\ &= \lim_{h\to 0} \frac{-4+2h-(-6+3h+2)}{2(-6+3h+2)h} \\ &= \lim_{h\to 0} \frac{-h}{2h(-4+3h)} = \lim_{h\to 0} \frac{-1}{2h(-4+3h)} = \frac{1}{8} \end{aligned}$$

Normals

If a curve C has a tangent line L at point P, then the straight line N through P perpendicular to L is called the *normal* to C at P. If L is horizontal, then N is vertical; if L is vertical, then N is horizontal. If L is neither horizontal nor vertical, then the slope of N is the negative reciprocal of the slope of L; that is,

$$\text{slope of the normal} = \frac{-1}{\text{slope of the tangent}}$$

Example.
Find equations of the straight lines that are tangent and normal to the curve $y = \sqrt{x}$ at the point (4,2).

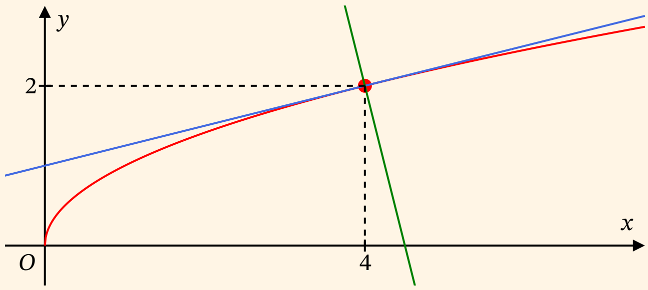

Figure 4–3: The tangent (blue) and normal (green) to $y = \sqrt{x}$ at (4,2).

The slope of the tangent at (4,2) is

$$\begin{aligned} m = \lim_{h\to 0} \frac{\sqrt{4+h}-2}{h} &= \lim_{h\to 0} \frac{\left(\sqrt{4+h}-2\right)\left(\sqrt{4+h}+2\right)}{h\left(\sqrt{4+h}+2\right)} \\ &= \lim_{h\to 0} \frac{4+h-4}{h\left(\sqrt{4+h}+2\right)} \\ &= \lim_{h\to 0} \frac{1}{\sqrt{4+h}+2} = \frac{1}{4} \end{aligned}$$

The tangent line has equations

$$y = \frac{1}{4}(x-4)+2 \quad \text{or} \quad x-4y+4=0$$

and the normal has slope -4 and, therefore, equation

$$y = -4(x-4)+2 \quad \text{or} \quad 4x+y-18=0$$

4–2 *The Derivative*

A straight line has the property that its slope is the same at all points. For any other graph, however, the slope may vary from point to point. Thus, the slope of the graph of $y = f(x)$ at the point x is itself a function of x. At any point x where the graph has a finite slope, we say that f is *differentiable*, and we call the slope the *derivative* of f. The derivative is therefore the limit of the Newton quotient.

Definition 31. The derivative of a function f is another function f' defined by

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

at all points x for which the limit exists (i.e., is a finite real number). If f' exists, we say that f is differentiable at x.

The domain of the derivative f' (read "f prime") is the set of numbers x in the domain of f where the graph of f has a nonvertical tangent line, and the value $f'(x_0)$ of f' at such a point x_0 is the slope of the tangent line to $y = f(x)$ there. Thus, the equation of the tangent line to $y = (x)$ at $f(x_0)$ is

$$y = f(x_0) + f'(x_0)(x - x_0)$$

The domain f' may be smaller than the domain of f because it contains only those points in the domain of f at which f is differentiable. Values of x in the domain of f where f is not differentiable and that are not endpoints of the domain of f are *singular points* of f.

The value of the derivative of f at a particular point x_0 can be expressed as a limit in either of two ways:

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

In the second limit $x_0 + h$ is replaced by x, so that $h = x - x_0$ and $h \to 0$ is equivalent to $x \to x_0$.

The process of calculating the derivative f' of a given function f is called *differentiation*.

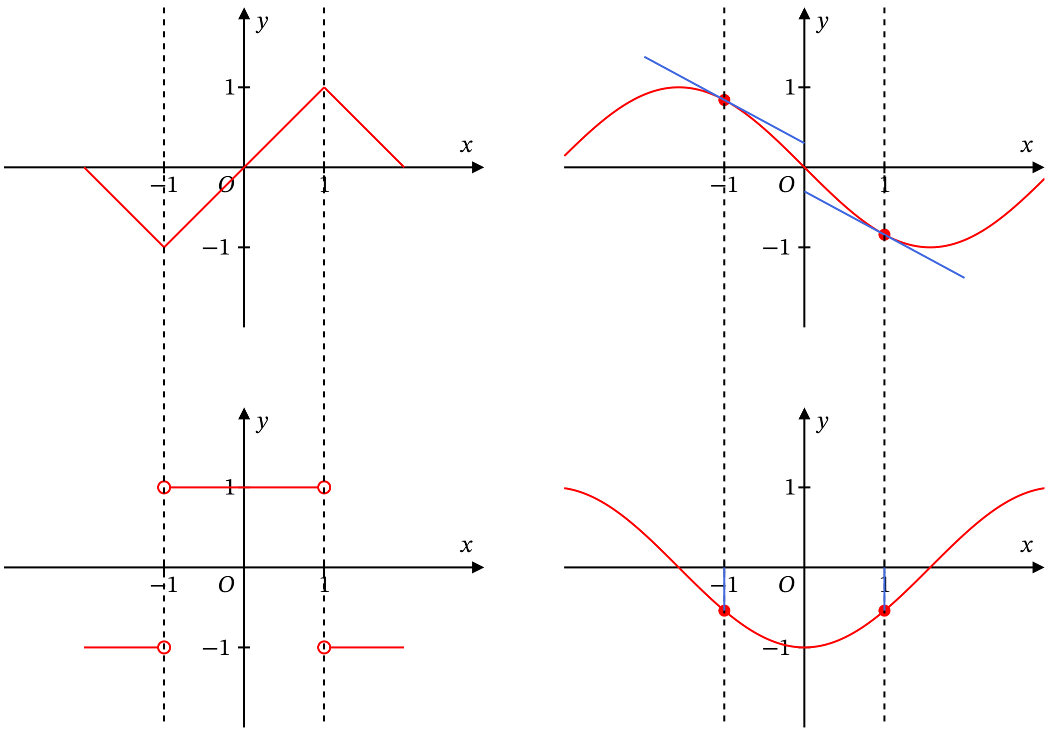

Figure 4–4: Graphical differentiation.

A function is differentiable on a set S if it is differentiable at every point x in S. Typically, the functions we encounter are defined on intervals or unions of intervals. If f is defined on a closed interval $[a,b]$, the definition does not allow for the existence of a derivative at the endpoints $x = a$ or $x = b$. (Why?) As we did for continuity, we extend the definition to allow for a right derivative at $x = a$ and a left derivative at $x = b$:

$$f'_+(a) = \lim_{h\to 0^+} \frac{f(a+h)-f(x)}{h}\ , \quad f'_-(b) = \lim_{h\to 0^-} \frac{f(b+h)-f(x)}{h}\ .$$

We now say that f is differentiable on $[a,b]$ if $f'(x)$ exists for all $x \in [a,b]$ and $f'_+(a)$ and $f'_-(b)$ both exist.

Example. Show that if $f(x) = ax + b$, then $f'(x) = a$.

The result is apparent from the graph of f, but we will do the calculation using the definition:

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \\ &= \lim_{h\to 0} \frac{a(x+h)+b-(ax+b)}{h} \\ &= \lim_{h\to 0} \frac{ah}{h} = a\,. \end{aligned}$$

An important special case of this example says that the derivative of a constant function is the zero function: If $g(x) = c$, a constant, then $g'(x) = 0$.

Example.

Use the definition of the derivative to calculate the derivatives of $f(x) = x^2$, $g(x) = \dfrac{1}{x}$ and $k(x) = \sqrt{x}$.

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \\ &= \lim_{h\to 0} \frac{(x+h)^2-x^2}{h} = \lim_{h\to 0} \frac{x^2+2xh+h^2-x^2}{h} \\ &= \lim_{h\to 0} \frac{2xh+h^2}{h} = \lim_{h\to 0}\ 2x+h = 2x\,. \end{aligned}$$

$$\begin{aligned} g'(x) &= \lim_{h\to 0} \frac{g(x+h)-g(x)}{h} \\ &= \lim_{h\to 0} \frac{\frac{1}{x+h}-\frac{1}{x}}{h} = \lim_{h\to 0} \frac{\frac{x-(x+h)}{(x+h)x}}{h} \\ &= \lim_{h\to 0} \frac{-h}{h(x+h)x} = \lim_{h\to 0} -\frac{1}{(x+h)x} = -\frac{1}{x^2}\,. \end{aligned}$$

$$\begin{aligned} k'(x) &= \lim_{h\to 0} \frac{k(x+h)-k(x)}{h} \\ &= \lim_{h\to 0} \frac{\sqrt{x+h}-\sqrt{x}}{h} = \lim_{h\to 0} \frac{\sqrt{x+h}-\sqrt{x}}{h}\,\frac{\sqrt{x+h}+\sqrt{x}}{\sqrt{x+h}+\sqrt{x}} \\ &= \lim_{h\to 0} \frac{x+h-x}{h\left(\sqrt{x+h}+\sqrt{x}\right)} = \lim_{h\to 0} \frac{1}{\sqrt{x+h}+\sqrt{x}} = \frac{1}{2\sqrt{x}}\,. \end{aligned}$$

The three derivative formulas calculated in this example are special cases of the following *General Power Rule*: If $f(x) = x^r$, then $f'(x) = rx^{r-1}$ for all values of r and x for which x^{r-1}makes sense as a real number. Eventually, we will prove all appropriate cases of the General Power Rule.

Example. Verify that: if $f(x) = |x|$, then $f'(x) = \frac{x}{|x|} = \operatorname{sgn} x$.

We have

$$f(x) = \begin{cases} x\,, & \text{if } x \geq 0\,, \\ -x\,, & \text{if } x < 0\,. \end{cases}$$

Thus, from the first example, $f'(x) = 1$ if $x > 0$ and $f'(x) = -1$ if $x < 0$. Also f is not differentiable at $x = 0$ (why?), which is a singular point of f. Therefore,

$$f'(x) = \begin{cases} 1\,, & \text{if } x > 0 \\ -1\,, & \text{if } x < 0 \end{cases} = \frac{x}{|x|} = \operatorname{sgn}\ x$$

Notation

Because functions can be written in different ways, it is useful to have more than one notation for derivatives. If $y = f(x)$, we can use the dependent variable y to represent the function, and we can denote the derivative of the function with respect to x in any of the following ways:

$$\mathsf{D}_x y = y' = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x} f(x) = f'(x) = \mathsf{D}_x f(x) = \mathsf{D} f(x)\,.$$

(In the forms using "D_x," we can omit the subscript x if the variable of differentiation is obvious.) Often the most convenient way of referring to the derivative of a function given explicitly as an expression in the variable x is to write $\frac{\mathrm{d}}{\mathrm{d}x}$ in front of that expression. The symbol $\frac{\mathrm{d}}{\mathrm{d}x}$ is a differential operator and should be read "the derivative with respect to x of …".

The value of the derivative of a function at a particular number x_0 in its domain can also be expressed in several ways:

$$\mathsf{D}_x y\big|_{x=x_0} = y'\big|_{x=x_0} = \frac{\mathrm{d}y}{\mathrm{d}x}\bigg|_{x=x_0} = \frac{\mathrm{d}}{\mathrm{d}x} f(x)\bigg|_{x=x_0} = f'(x_0) = \mathsf{D}_x f(x_0)\,.$$

The symbol $\Big|_{x=x_0}$ is called an *evaluation symbol*. It signifies that the expression preceding it should be evaluated at $x = x_0$. Do not confuse $\frac{\mathrm{d}}{\mathrm{d}x} f(x)$ and $\frac{\mathrm{d}}{\mathrm{d}x} f(x)\Big|_{x=x_0}$. The first expression represents a *function*, $f'(x)$. The second represents a *number* $f'(x_0)$.

The notation $\frac{\mathrm{d}y}{\mathrm{d}x}$ are called Leibniz notations for the derivative; Newton used notations similar to the prime y' notations we use here.

The Leibniz notation is suggested by the definition of derivative. The Newton quotient $\frac{f(x+h)-f(x)}{h}$, whose limit we take to find the derivative $\frac{\mathrm{d}y}{\mathrm{d}x}$, can be written in the form $\frac{\Delta y}{\Delta x}$, where $\Delta y = f(x+h) - f(x)$ is the increment in y, and $\Delta x = h$ is the corresponding increment in x as we pass from the point $(x, f(x))$ to the point $(x+h, f(x+h))$ on the graph of f. Δ is the uppercase Greek letter Delta. Using symbols:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}\,.$$

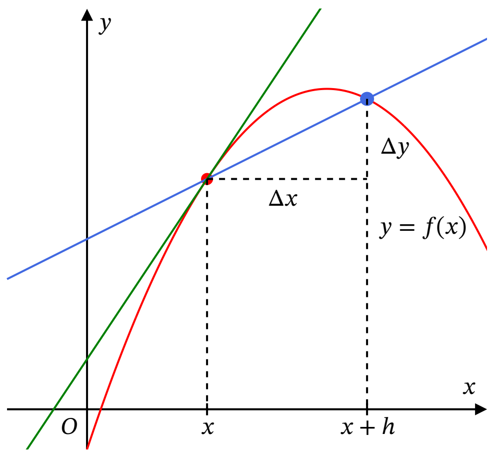

Figure 4–5: $\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$

The Newton quotient $\frac{\Delta y}{\Delta x}$ is actually the quotient of two quantities, Δy and Δx. It is not at all clear, however, that the derivative $\frac{dy}{dx}$, the limit of $\frac{\Delta y}{\Delta x}$ as Δx approaches zero, can be regarded as a quotient. If y is a continuous function of x, then Δy approaches zero when Δx approaches zero, so $\frac{dy}{dx}$ appears to be the meaningless quantity $\frac{0}{0}$. Nevertheless, it is sometimes useful to be able to refer to quantities dy and dx in such a way that their quotient is the derivative $\frac{dy}{dx}$. We can justify this by regarding dx as a new independent variable (called the *differential of* x) and defining a new dependent variable dy (the *differential of* y) as a function of x and dx by

$$dy = \frac{dy}{dx}\,dx = f'(x)\,dx$$

For example, if $y = x^2$, we can write $dy = 2x\,dx$ to mean the same thing as $\frac{dy}{dx} = 2x$. Similarly, if $f(x) = \frac{1}{x}$, we can write $dy = -\frac{1}{x^2}\,dx$ as the equivalent differential form of the assertion that $\frac{dy}{dx} = -\frac{1}{x^2}$. This *differential notation* is useful in applications, and especially for the interpretation and manipulation of integrals. Note that, defined as above, differentials are merely variables that may or may not be small in absolute value.

Approximation by Differentials

If one quantity, say y, is a function of another quantity x, that is,

$$y = f(x),$$

we sometimes want to know how a change in the value of x by an amount Δx will affect the value of y. The exact change Δy in y is given by

$$\Delta y = f\big(x + \Delta x\big) - f(x),$$

but if the change Δx is small, then we can get a good approximation to Δy by using the fact that $\frac{\Delta y}{\Delta x}$ is approximately the derivative $\frac{dy}{dx}$. Thus,

$$\Delta y = \frac{\Delta y}{\Delta x}\Delta x \approx \frac{dy}{dx}\Delta x = f'(x)\Delta x\,.$$

It is often convenient to represent this approximation in terms of differentials; if we denote the change in x by dx instead of Δx, then the change Δy in y is approximated by the differential dy, that is

$$\Delta y \approx dy = f'(x)\,dx\,.$$

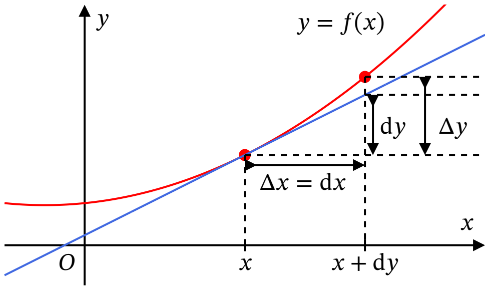

Figure 4–6: dy*, the change in height to the tangent line, approximates* Δy*, the change in height to the graph of* f

Some Questions

Is a function f defined on an interval I necessarily the derivative of some other function defined on I? The answer is no; some functions are derivatives and some are not. Although a derivative need not be a continuous function, it must, like a continuous function, have the intermediate-value property: on an interval $[a, b]$, a derivative $f'(x)$ takes on every value between $f'(a)$ and $f'(b)$. We will prove this in a next section.

If $g(x)$ is continuous on an interval I, then $g(x) = f'(x)$ for some function f that is differentiable on I . We will discuss this fact further in Chapter 6.

4–3 Differentiation Rules

If every derivative had to be calculated directly from the definition of derivative as in the examples of the previou section, calculus would indeed be a painful subject. Fortunately, there is an easier way. We will develop several general differentiation rules that enable us to calculate the derivatives of complicated combinations of functions easily if we already know the derivatives of the elementary functions from which they are constructed.

Before developing these differentiation rules we need to establish one obvious but very important theorem which states, roughly, that the graph of a function cannot possibly have a break at a point where it is smooth.

Theorem 35. *If f is differentiable at x, then f is continuous at x.*

Proof. Since f is differentiable at x, we know that

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

exists. Using the limit rules, we have

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

This is equivalent to $\lim_{h\to 0} f(x+h) = f(x)$, which says that f is continuous at x.

Sums and Constant Multiples

The derivative of a sum (or difference) of functions is the sum (or difference) of the derivatives of those functions. The derivative of a constant multiple of a function is the same constant multiple of the derivative of the function.

Theorem 36.

If functions f and g are differentiable at x, and if C is a constant, then the functions $f+g$, $f-g$, and Cf are all differentiable at x and

$$(f+g)'(x) = f'(x) + g'(x)$$
$$(f-g)'(x) = f'(x) - g'(x)$$
$$(Cf)'(x) = Cf'(x)\,.$$

Proof. The proofs of all three assertions are straightforward, using the corresponding limit rules.

For the sum, we have

$$\begin{aligned}(f+g)'(x) &= \lim_{h\to 0} \frac{(f+g)(x+h)-(f+g)(x)}{h} \\ &= \lim_{h\to 0} \frac{(f(x+h)+g(x+h))-(f(x)+g(x))}{h} \\ &= \lim_{h\to 0} \left(\frac{f(x+h)-f(x)}{h} + \frac{g(x+h)-g(x)}{h}\right) \\ &= f'(x)+g'(x)\,,\end{aligned}$$

because the limit of a sum is the sum of the limits. The proof for the difference $f-g$ is similar.

For the constant multiple, we have

$$\begin{aligned}(Cf)'(x) &= \lim_{h\to 0} \frac{Cf(x+h)-Cf(x)}{h} \\ &= C \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = Cf'(x)\,.\end{aligned}$$

The rule for differentiating sums extends to sums of any finite number of terms

$$\left(f_1+f_2+\cdots+f_n\right)' = f_1'+f_2'+\cdots+f_n'\,.$$

To see this we can use mathematical induction. theorem 2 shows that the case $n=2$ is true; this is STEP 1. For STEP 2, we must show that if the formula holds for some integer $n=h\geq 2$, then it must also hold for $n=k+1$. Therefore, *assume* that

$$\left(f_1+f_2+\cdots+f_k\right)' = f_1'+f_2'+\cdots+f_k'\,.$$

Then we have

$$\begin{aligned}\left(f_1+f_2+\cdots+f_k+f_{k+1}\right)' &= \Bigg(\underbrace{f_1+f_2+\cdots+f_k}_{\text{Let this function be } f}+f_{k+1}\Bigg)' \\ &= \left(f+f_{k+1}\right)' \\ &= f'+f_{k+1}' \\ &= f_1'+f_2'+\cdots+f_k'+f_{k+1}'\,.\end{aligned}$$

With both steps verified, we can claim that the formula holds for any $n>2$; by induction. In particular, therefore, the derivative of any polynomial is the sum of the derivatives of its terms.

Product Rule

The rule for differentiating a product of functions is a little more complicated than that for sums. It is **not** true that the derivative of a product is the product of the derivatives.

Theorem 37.
If functions f and g are differentiable at x, then their product fg is also differentiable at x, and

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x)$$

Proof.
We set up the Newton quotient for fg and then add 0 to the numerator in a way that enables us to involve the Newton quotients for f and g separately:

$$\begin{aligned}
(fg)'(x) &= \lim_{h\to 0} \frac{(fg)(x+h) - (fg)(x)}{h} \\
&= \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} \\
&= \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x+h) + f(x)g(x+h) - f(x)g(x)}{h} \\
&= \lim_{h\to 0} \left(\frac{f(x+h) - f(x)}{h} g(x+h) + f(x) \frac{g(x+h) - g(x)}{h} \right) \\
&= f'(x)g(x) + f(x)g'(x)\,,
\end{aligned}$$

To get the last line, we have used the fact that f and g are differentiable and the fact that g is therefore continuous, as well as limit rules.

Example. Use mathematical induction to verify the formula $\frac{\mathrm{d}}{\mathrm{d}x} x^n = nx^{n-1}$ for all positive integers n.

For $n = 1$ the formula says that $\frac{\mathrm{d}}{\mathrm{d}x} x^1 = 1x^0$, so the formula is true in this case. We must show that if the formula is true for $n = k \geq 1$, then it is also true for $n = k+1$. Therefore, assume that

$$\frac{\mathrm{d}}{\mathrm{d}x} x^k = kx^{k-1}$$

Using the Product Rule we calculate

$$\frac{\mathrm{d}}{\mathrm{d}x} x^{k+1} = \frac{\mathrm{d}}{\mathrm{d}x}\left(x^k x\right) = kx^{k-1}x + x^k 1 = (k+1)x^k = (k+1)x^{(k+1)-1}$$

Thus, the formula is true for $n = k+1$ also. The formula is true for all integers $n \geq 1$ by induction.

The Product Rule can be extended to products of any number of factors; for instance,

$$\begin{aligned}
(fgh)'(x) &= f'(x)(gh)(x) + f(x)(gh)'(x) \\
&= f'(x)g(x)h(x) + f'(x)g'(x)h(x) + f'(x)g(x)h'(x)\,.
\end{aligned}$$

In general, the derivative of a product of n functions will have n terms; each term will be the same product but with one of the factors replaced by its derivative:

$$\left(f_1 f_2 f_3 \cdots f_n\right)' = f_1' f_2 f_3 \cdots f_n + f_1 f_2' f_3 \cdots f_n + f_1 f_2 f_3' \cdots f_n + \cdots + f_1 f_2 f_3 \cdots f_n'\,.$$

Exercise. Prove this formula by mathematical induction.

Reciprocal Rule

Theorem 38. *If f is differentiable at x and $f(x) \neq 0$, then $\frac{1}{f}$ is differentiable at x, and*

$$\left(\frac{1}{f}\right)'(x) = \frac{-f'(x)}{(f(x))^2}\,.$$

Proof. Using the definition of the derivative, we calculate

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{f(x)} &= \lim_{h\to 0}\frac{\dfrac{1}{f(x+h)}-\dfrac{1}{f(x)}}{h} \\
&= \lim_{h\to 0}\frac{f(x)-f(x+h)}{hf(x+h)f(x)} \\
&= \lim_{h\to 0}\left(\frac{-1}{f(x+h)f(x)}\right)\left(\frac{f(x+h)-f(x)}{h}\right) \\
&= \frac{-1}{(f(x))^2}f'(x)
\end{aligned}$$

Again we have to use the continuity of f and the limit rules.

Example. Confirm the General Power Rule for negative integers:

$$\frac{\mathrm{d}}{\mathrm{d}x}x^{-n} = -nx^{-n-1}\,.$$

Since we have already proved the rule for positive integers, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}x^{-n} = \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x^n} = \frac{-nx^{n-1}}{(x^n)^2} = -nx^{-n-1}\,.$$

Quotient Rule

The Product Rule and the Reciprocal Rule can be combined to provide a rule for differentiating a quotient of two functions. Observe that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{f(x)}{g(x)}\right) = \frac{\mathrm{d}}{\mathrm{d}x}\left(f(x)\frac{1}{g(x)}\right) &= f'(x)\frac{1}{g(x)} + f(x)\left(-\frac{g'(x)}{(g(x))^2}\right) \\
&= \frac{g(x)f'(x)-f(x)g'(x)}{(g(x))^2}\,.
\end{aligned}$$

Thus, we have proved the following Quotient Rule.

Theorem 39.

If f and g are differentiable at x, and if $g(x) \neq 0$, then the quotient $\frac{f}{g}$ is differentiable at x and

$$\left(\frac{f}{g}\right)' = \frac{g(x)f'(x)-f(x)g'(x)}{(g(x))^2}\,.$$

Chain Rule

Although we can differentiate $\sqrt{x}$ and x^2+1, we cannot yet differentiate $\sqrt{x^2+1}$. To do this, we need a rule that tells us how to differentiate composites of functions whose derivatives we already know. This rule is known as the Chain Rule and is the most often used of all the differentiation rules.

Theorem 40.
If $f(u)$ is differentiable at $u = g(x)$, and $g(x)$ is differentiable at x, then the composite function $f \circ g(x) = f(g(x))$ is differentiable at x, and

$$(f \circ g)' = f'(g(x))g'(x)$$

In terms of Leibniz notation, if $y = f(u)$ where $u = g(x)$, then $y = f(g(x))$ and:

- at u, y is changing $\dfrac{\mathrm{d}y}{\mathrm{d}u}$ times as fast as u is changing;
- at x, u is changing $\dfrac{\mathrm{d}u}{\mathrm{d}x}$ times as fast as x is changing.

Therefore, at x, $y = f(u) = f(g(x))$ is changing $\dfrac{\mathrm{d}y}{\mathrm{d}u}\dfrac{\mathrm{d}u}{\mathrm{d}x}$ times as fast as x is changing. That is,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}u}\frac{\mathrm{d}u}{\mathrm{d}x}\,, \quad \text{where } \frac{\mathrm{d}y}{\mathrm{d}u} \text{ is evaluated at } u = g(x)\,.$$

It appears as though the symbol $\mathrm{d}u$ cancels from the numerator and denominator, but this is not meaningful because $\dfrac{\mathrm{d}y}{\mathrm{d}u}$ was not defined as the quotient of two quantities, but rather as a single quantity, the derivative of y with respect to u.

We would like to prove this theorem by writing

$$\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u}\frac{\Delta u}{\Delta x}$$

and taking the limit as $\Delta x \to 0$. Such a proof is valid for most composite functions but not all. The next demonstration is valid in all cases.

Proof. Suppose that f is differentiable at the point $u = g(x)$ and that g is differentiable at x.

Let the function $E(k)$ be defined by

$$E(k) = \begin{cases} 0 & \text{if } k = 0\,. \\ \frac{f(u+k)-f(u)}{k} - f'(u) & \text{if } k \neq 0\,. \end{cases}$$

By the definition of derivative

$$\lim_{k\to 0} E(k) = f'(u) - f'(u) = 0 = E(0)\,,$$

so $E(k)$ is continuous at $k = 0$. Also, whether $k = 0$ or not, we have

$$f(u+k) - f(u) = \big(f'(u) + E(k)\big)k\,.$$

Now put $u = g(x)$ and $k = g(x+h) - g(x)$, so that $u + k = g(x+h)$, and obtain

$$f(g(x+h)) - f(g(x)) = \big(f'(g(x)) + E(k)\big)(g(x+h) - g(x))\,.$$

Since g is differentiable at x,

$$\lim_{h\to 0} \frac{g(x+h) - g(x)}{h} = g'(x)\,.$$

Also, g is continuous at x, so

$$\lim_{h\to 0} k = \lim_{h\to 0} (g(x+h) - g(x)) = 0\,.$$

Since E is continuous at 0,

$$\lim_{h\to 0} E(k) = \lim_{k\to 0} E(k) = E(0) = 0\,.$$

Hence,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} f(g(x)) &= \lim_{h\to 0} \frac{f(g+h(x)) - f(g(x))}{h} \\ &= \lim_{h\to 0} \big(f'(g(x)) + E(k)\big)\frac{g(x+h) - g(x)}{h} \\ &= \big(f'(g(x)) + 0\big)g'(x) = f'(g(x))g'(x)\,, \end{aligned}$$

which was to be proved.

Example. Find the derivative of $y = \sqrt{x^2+1}$.

Here $y = f(g(x))$, where $f(u) = \sqrt{u}$ and $g(x) = x^2 + 1$. Since the derivatives of f and g are

$$f'(u) = \frac{1}{2\sqrt{u}} \quad \text{and} \quad g'(x) = 2x\,,$$

the Chain Rules gives

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}x} &= \frac{\mathrm{d}}{\mathrm{d}x} f(g(x)) = f'(g(x))g'(x) \\ &= \frac{1}{2\sqrt{g(x)}}\, g'(x) = \frac{1}{2\sqrt{x^2+1}}\, 2x = \frac{x}{\sqrt{x^2+1}}\,. \end{aligned}$$

4–4 *Derivatives of Trigonometric Functions*

In this section we will calculate the derivatives of the six trigonometric functions. We only have to work hard for one of them, sine; the others then follow from known identities and the differentiation rules

First, we have to establish some trigonometric limits that we will need to calculate the derivative of sine. It is assumed throughout that the arguments of the trigonometric functions are measured in radians.

Theorem 41.
The functions $\sin\,\theta$ *and* $\cos\,\theta$ *are continuous at every value of* θ*. In particular, at* $\theta = 0$ *we have:*

$$\lim_{\theta\to 0}\ \sin\,\theta = \sin\,0 = 0 \quad \text{and} \quad \lim_{\theta\to 0}\ \cos\,\theta = \cos\,0 = 1$$

This result seems obvious from the graphs of sine and cosine, but we will still prove it.

Proof. Let A, P and Q be the points as shown in the figure.

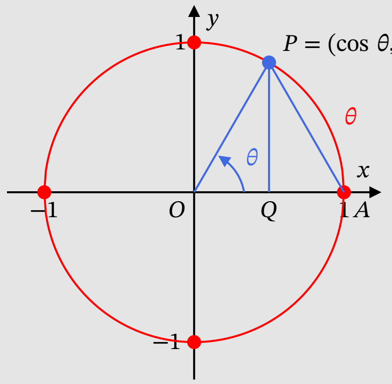

Figure 4–7: Continuity of cosine and sine.

Since the length of chord AP is less than the length of arc AP, we have

$$\sin^2\,\theta + (1-\cos\,\theta)^2 < \theta^2 .$$

Squaring yields positive values and the square root is an increasing function, so

$$0 \le |\sin\,\theta| < |\theta| \quad \text{and} \quad 0 \le |1-\cos\,\theta| < |\theta| .$$

Using the squeeze theorem, we find

$$0 \le \lim_{\theta\to 0} |\sin\,\theta| < \lim_{\theta\to 0} |\theta| = 0 \quad \text{and} \quad 0 \le \lim_{\theta\to 0} |1-\cos\,\theta| < \lim_{\theta\to 0} |\theta| = 0 .$$

Thus,

$$\lim_{\theta\to 0}\ \sin\,\theta = 0 = \sin\,0 \quad \text{and} \quad \lim_{\theta\to 0}\ \cos\,\theta = 1 = \cos\,0$$

and sine and cosine are continuous at $\theta = 0$.

To prove the continuity at another point, we use the addition formulas and the rules to evaluate limits:

$$\lim_{\theta\to\theta_0}\ \sin\,\theta = \lim_{\phi\to 0}\ \sin(\theta_0+\phi) = \sin\,\theta_0 \lim_{\phi\to 0}\ \cos\,\phi + \cos\,\theta_0 \lim_{\phi\to 0}\ \sin\,\phi = \sin\,\theta_0$$

and

$$\lim_{\theta\to\theta_0}\ \cos\,\theta = \lim_{\phi\to 0}\ \cos(\theta_0+\phi) = \cos\,\theta_0 \lim_{\phi\to 0}\ \cos\,\phi - \sin\,\theta_0 \lim_{\phi\to 0}\ \sin\,\phi = \cos\,\theta_0 .$$

The graph of the function $\dfrac{\sin\,\theta}{\theta}$ is shown in the next figure. Although it is not defined at $\theta = 0$, this function appears to have limit 1 as θ approaches 0.

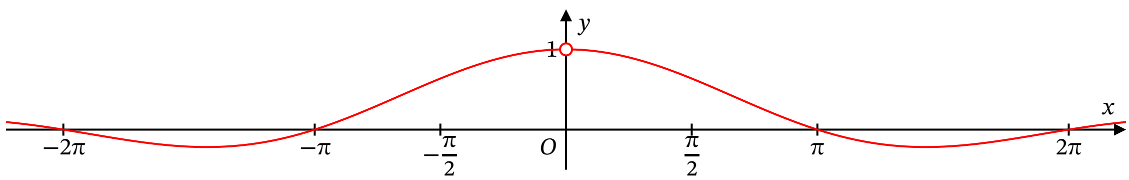

Figure 4–8: $\lim_{\theta\to 0} \dfrac{\sin\,\theta}{\theta} = 1$

Theorem 42. $\lim_{\theta \to 0} \dfrac{\sin \theta}{\theta} = 1$.

Proof. Let $0 < \theta < \frac{\pi}{2}$, and represent θ as shown in the figure.

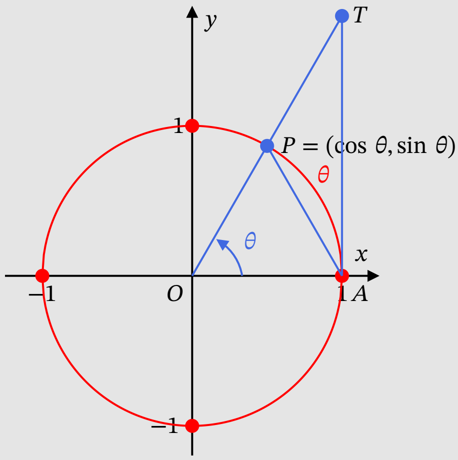

Figure 4–9: Area $\triangle OAP$ < Area sector OAP < Area $\triangle OAT$

Points $A(1,0)$ and $P(\cos \theta, \sin \theta)$ lie on the unit circle $x^2 + y^2 = 1$. The area of the circular sector OAP lies between the areas of triangles $\triangle OAP$ and $\triangle OAT$:

$$\text{Area } \triangle OAP < \text{Area sector } OAP < \text{Area } \triangle OAT\,.$$

As shown in Chapter 2, the area of a circular sector having central angle θ and radius 1 is $\frac{\theta}{2}$. The area of a triangle is $\frac{1}{2} \times \text{base} \times \text{height}$, so

$$\text{Area } \triangle OAP = \frac{1}{2}(1)\sin \theta = \frac{\sin \theta}{2} \quad \text{and} \quad \text{Area } \triangle OAT = \frac{1}{2}(1)\tan \theta = \frac{\tan \theta}{2}\,.$$

Thus

$$\frac{\sin \theta}{2} < \frac{\theta}{2} < \frac{\sin \theta}{2\cos \theta} \quad \text{or} \quad 1 < \frac{\theta}{\sin \theta} < \frac{1}{\cos \theta}\,.$$

Now take the reciprocals, thereby reversing the inequalities:

$$1 > \frac{\sin \theta}{\theta} > \cos \theta\,.$$

Since $\lim_{\theta \to 0^+} \cos \theta = 1$, the squeeze theorem gives $\lim_{\theta \to 0^+} \dfrac{\sin \theta}{\theta} = 1$.

Finally, note that $\sin \theta$ and θ are odd functions. Therefore, $f(\theta) = \dfrac{\sin \theta}{\theta}$ is an even function: $f(\theta) = f(-\theta)$. This symmetry implies that the left limit at 0 must have the same value as the right limit:

$$\lim_{\theta \to 0^-} \frac{\sin \theta}{\theta} = 1 = \lim_{\theta \to 0^+} \frac{\sin \theta}{\theta}\,,$$

so $\lim_{\theta \to 0} \dfrac{\sin \theta}{\theta} = 1$.

This theorem can be combined with limit rules and known trigonometric identities to yield other trigonometric limits.

Example. Show that $\lim_{h \to 0} \dfrac{\cos h - 1}{h} = 0\,.$

Using the half-angle formula $\cos h = 1 - 2\sin^2 \frac{h}{2}$, we calculate

$$\begin{aligned}\lim_{h\to 0} \frac{\cos h - 1}{h} &= \lim_{h \to 0} \frac{-2\sin^2 \frac{h}{2}}{h} \quad \text{let } \theta = \frac{h}{2}\,,\\ &= -\lim_{\theta \to 0} \frac{\sin \theta}{\theta} \sin \theta = -(1)(0) = 0\,.\end{aligned}$$

To calculate the derivative of $\sin x$, we need the addition formula for sine

$$\sin(x+h) = \sin x \cos h + \cos x \sin h$$

Theorem 43. $\dfrac{\mathrm{d}}{\mathrm{d}x} \sin x = \cos x\,.$

Proof.
We use the definition of derivative, the addition formula for sine, the rules for combining limits, and the previous theorem:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\sin x &= \lim_{h\to 0} \frac{\sin(x+h) - \sin x}{h}\\ &= \lim_{h\to 0} \frac{\sin x \cos h + \cos x \sin h - \sin x}{h}\\ &= \lim_{h\to 0} \frac{\sin x(\cos h - 1) + \cos x \sin h}{h}\\ &= \lim_{h\to 0} \sin x \lim_{h\to 0} \frac{\cos h - 1}{h} + \lim_{h\to 0} \cos x \lim_{h\to 0} \frac{\sin h}{h}\\ &= (\sin x)(0) + (\cos x)(1) = \cos x\,.\end{aligned}$$

Theorem 44. $\dfrac{\mathrm{d}}{\mathrm{d}x} \cos x = -\sin x\,.$

Proof. We make use of the complementary angle identities, and the Chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}x}\cos x = \frac{\mathrm{d}}{\mathrm{d}x}\sin\left(\frac{\pi}{2} - x\right) = (-1)\cos\left(\frac{\pi}{2} - x\right) = -\sin x\,.$$

Because $\sin x$ and $\cos x$ are differentiable everywhere, the functions

$$\tan x = \frac{\sin x}{\cos x}\,, \quad \cot x = \frac{\cos x}{\sin x}\,, \quad \sec x = \frac{1}{\cos x}\,, \quad \csc x = \frac{1}{\sin x}$$

are differentiable at every value of x at which they are defined (i.e., where their denominators are not zero). Their derivatives can be calculated by the Quotient and Reciprocal Rules and are as follows:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\tan(x) &= \sec^2 x\,, &\qquad \frac{\mathrm{d}}{\mathrm{d}x}\cot x &= -\csc^2 x\,,\\ \frac{\mathrm{d}}{\mathrm{d}x}\sec(x) &= \sec x \tan x\,, &\qquad \frac{\mathrm{d}}{\mathrm{d}x}\csc x &= -\csc x \cot x\,.\end{aligned}$$

Example. Verify the formula for $\tan x$.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\tan(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\sin x}{\cos x}\right) &= \frac{\cos x\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) - \sin x\frac{\mathrm{d}}{\mathrm{d}x}\cos(x)}{\cos^2 x} \\ &= \frac{\cos x\cos(x) - \sin x(-\sin x)}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} \\ &= \frac{1}{\cos^2 x} = \sec^2 x\,.\end{aligned}$$

Exercise. Verify the other formulas.

4–5 *Higher Order Derivatives*

If the derivative $y' = f'(x)$ of a function $y = f(x)$ is itself differentiable at x, we can calculate its derivative, which we call the second derivative of f and denote by $y'' = f''(x)$. As is the case for first derivatives, second derivatives can be denoted by various notations depending on the context. Some of the more common ones are

$$y'' = f''(x) = \frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = \frac{\mathrm{d}}{\mathrm{d}x}\frac{\mathrm{d}}{\mathrm{d}x}f(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}f(x) = \mathsf{D}_x^2 y = \mathsf{D}_x^2 f(x)\,.$$

Similarly, you can consider third-, fourth-, and in general nth-order derivatives. The prime notation is inconvenient for derivatives of high order, so we denote the order by a superscript in parentheses (to distinguish it from an exponent): the nth derivative of $y = f(x)$ is

$$y^{(n)} = f^{(n)}(x) = \frac{\mathrm{d}^n y}{\mathrm{d}x^n} = \frac{\mathrm{d}^n}{\mathrm{d}x^n}f(x) = \mathsf{D}_x^n y = \mathsf{D}_x^n f(x)\,,$$

and it is defined to be the derivative of the $(n-1)$st derivative. For $n = 1,2,3$ primes are still normally used: $f^{(3)} = f'''$. It is convenient to denote $f^{(0)} = f$, that is, to regard a function as its own zeroth-order derivative.

Example.
The *velocity* of a moving object is the (instantaneous) rate of change of the position of the object with respect to time; if the object moves along the x-axis and is at position $x = f(t)$ at time t, then its velocity at that time is

$$v = \frac{\mathrm{d}x}{\mathrm{d}t} = f'(t)\,.$$

Similarly, the *acceleration* of the object is the rate of change of the velocity. Thus, the acceleration is the second derivative of the position:

$$a = \frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = f''(t)\,.$$

Example. If $y = x^3$, then $y' = 3x^2$, $y'' = 6x$, $y''' = 6$, $y^{(4)} = 0$, and all higher derivatives are zero.

In general, if $f(x) = x^n$ (where n is a positive integer), then

$$\begin{aligned}f^{(k)} &= n(n-1)(n-2)\cdots(n-(k-1))x^{n-k} \\ &= \begin{cases}\frac{n!}{(n-k)!}x^{n-k} & \text{if } 0 \le k \le n \\ 0 & \text{if } k > n\,,\end{cases}\end{aligned}$$

where $n!$ is called the *factorial* and is recursively defined by:

$$n! = \begin{cases} 1 & \text{if } n = 0 \\ n(n-1)! & \text{if } n > 0\,. \end{cases}$$

It follows that if P is a polynomial of degree n,

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0\,,$$

where $a_n, a_{n-1}, \ldots, a_1, a_0$ are constants, then $P^{(k)}(x) = 0$ for $k > n$. For $k \leq n$, $P^{(k)}$ is a polynomial of degree $n - k$; in particular, $P^{(n)}(x) = n! a_n$, a constant function.

4–6 Implicit Differentiation

We know how to find the slope of a curve that is the graph of a function $y = f(x)$ by calculating the derivative of f. But not all curves are the graphs of such functions. To be the graph of a function $f(x)$, the curve must not intersect any vertical lines at more than one point.

Curves are generally the graphs of equations in two variables. Such equations can be written in the form

$$F(x, y) = 0\,,$$

where $F(x, y)$ denotes an expression involving the two variables x and y. For example, a circle with centre at the origin and radius 5 has equation

$$x^2 + y^2 = 25$$

so $F(x, y) = x^2 + y^2 - 25$ for that circle.

Sometimes we can solve an equation $F(x, y) = 0$ for y and so find explicit formulas for one or more functions $y = f(x)$ defined by the equation. Usually, however, we are not able to solve the equation. However, we can still regard it as defining yas one or more functions of x implicitly, even it we cannot solve for these functions explicitly. Moreover, we still find the derivative $\frac{\mathrm{d}y}{\mathrm{d}x}$ of these implicit solutions by a technique called *implicit differentiation*. The idea is to differentiate the given equation with respect to x, regarding y as a function of x having derivative $\frac{\mathrm{d}y}{\mathrm{d}x}$, or y'.

Example. Find $\frac{\mathrm{d}y}{\mathrm{d}x}$ if $y^2 = x$.

The equation $y^2 = x$ defines two differentiable functions of x; in this case we know them explicitly. They are $y_1 = \sqrt{x}$ and $y_2 = -\sqrt{x}$, having derivatives defined for $x > 0$ by

$$\frac{\mathrm{d}y_1}{\mathrm{d}x} = \frac{1}{2\sqrt{x}} \quad \text{and} \quad \frac{\mathrm{d}y_2}{\mathrm{d}x} = -\frac{1}{2\sqrt{x}} .$$

However, we can find the slope of the curve $y^2 = x$ at any point (x, y) satisfying that equation without first solving the equation for y. To find $\frac{\mathrm{d}y}{\mathrm{d}x}$, we simply differentiate both sides of the equation $y^2 = x$ with respect to x, treating y as a differentiable function of x and using the Chain Rule to differentiate y^2:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}(y^2) &= \frac{\mathrm{d}}{\mathrm{d}x}(x) \\ 2y\frac{\mathrm{d}y}{\mathrm{d}x} &= 1 \\ \frac{\mathrm{d}y}{\mathrm{d}x} &= \frac{1}{2y} . \end{aligned}$$

Observe that this agrees with the derivatives we calculated above for both of the explicit solutions $y_1 = \sqrt{x}$ and $y_2 = -\sqrt{x}$:

$$\frac{\mathrm{d}y_1}{\mathrm{d}x} = \frac{1}{2y_1} = \frac{1}{2\sqrt{x}} \quad \text{and} \quad \frac{\mathrm{d}y_2}{\mathrm{d}x} = \frac{1}{2y_2} = -\frac{1}{2\sqrt{x}} .$$

Example. Find y'' if $xy + y^2 = 2x$.

Twice differentiate both sides of the given equation with respect to x:

$$\begin{aligned} y + xy' + 2yy' &= 2 \\ y' + y' + xy'' + 2(y')^2 + 2yy'' &= 0 . \end{aligned}$$

Now solve these equations for y' and y''.

$$\begin{aligned} y' &= \frac{2-y}{x+2y} \\ y'' &= -\frac{2y' + 2(y')^2}{x+2y} = -\frac{8}{(x+2y)^3} . \end{aligned}$$

Until now, we have only proven the General Power Rule

$$\frac{\mathrm{d}}{\mathrm{d}x}x^r = rx^{r-1}$$

for integer exponents r and a few special rational exponents such as $r = \frac{1}{2}$. Using implicit differentiation, we can give the proof for any rational exponent $r = \frac{m}{n}$, where m and n are integers, and $n \neq 0$.

If $y = x^{\frac{m}{n}}$, then $y^n = x^m$. Differentiating implicitly with respect to x, we obtain

$$ny^{n-1}\frac{\mathrm{d}y}{\mathrm{d}x} = mx^{m-1},$$

so

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{m}{n}x^{m-1}y^{1-n} = \frac{m}{n}x^{m-1}\left(x^{\frac{m}{n}}\right)^{1-n} = \frac{m}{n}x^{m-1+\frac{m}{n}-m} = \frac{m}{n}x^{\frac{m}{n}-1} .$$

4–7 *Derivatives of Inverse Functions*

Suppose that the function f is differentiable on an interval $]a, b[$ and that either $f'(x) > 0$ for $a < x < b$, so that f is increasing on $]a, b[$, or $f'(x) < 0$ for $a < x < b$, so that f is decreasing on $]a, b[$. In either case f is bijective on $]a, b[$ and has an inverse, f^{-1} there. Differentiating the cancellation identity

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

with respect to x, using the chain rule, we obtain

$$f'\big(f^{-1}(x)\big)\frac{\mathrm{d}}{\mathrm{d}x}f^{-1}(x) = \frac{\mathrm{d}}{\mathrm{d}x}x = 1\,.$$

Thus,

$$\frac{\mathrm{d}}{\mathrm{d}x}f^{-1}(x) = \frac{1}{f'\big(f^{-1}(x)\big)}\,.$$

In Leibniz notation, if $y = f^{-1}(x)$, we have

$$\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_x = \frac{1}{\left.\frac{\mathrm{d}x}{\mathrm{d}y}\right|_{y=f^{-1}(x)}}\,.$$

The slope of the graph of f^{-1} at (x, y) is the reciprocal of the slope of the graph of f at (y, x).

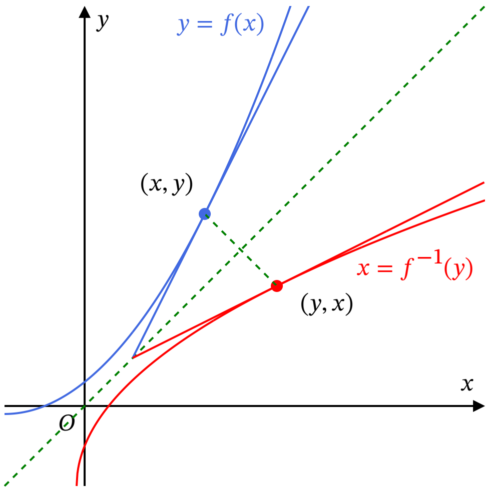

Figure 4–10: Tangents to the graphs of f and f^{-1}.

Example.

Show that $f(x) = x^3 + x$ is bijective on the whole real line, and, nothing that $f(2) = 10$, find $\big(f^{-1}\big)'(10)$.

Since $f'(x) = 3x^2 + 1 > 0$ for all real numbers x, f is increasing and therefore bijective and invertible. If $y = f^{-1}(x)$, then

$$\begin{aligned} x = f(y) = y^3 + y &\implies 1 = (3y^2 + 1)y' \\ &\implies y' = \frac{1}{3y^2 + 1}\,. \end{aligned}$$

Now $x = f(2) = 10$ implies $y = f^{-1}(10) = 2$. Thus,

$$\big(f^{-1}\big)'(10) = \left.\frac{1}{3y^2 + 1}\right|_{y=2} = \frac{1}{13}\,.$$

4–8 The Mean-Value Theorem

If you set out in a car at 1:00 p.m. and arrive in a town 150 km away from your starting point at 3:00 p.m., then you have travelled at an average speed of $\frac{150}{2} = 75$ km/h. Although you may not have travelled at constant speed, you must have been going 75 km/h at at least one instant during your journey, for if your speed was always less than 75 km/h you would have gone less than 150 km in 2 h, and if your speed was always more than 75 km/h, you would have gone more than 150 km in 2 h. In order to get from a value less than 75 km/h to a value greater than 75 km/h, your speed, which is a continuous function of time, must pass through the value 75 km/h at some intermediate time.

The conclusion that the average speed over a time interval must be equal to the instantaneous speed at some time in that interval is an instance of an important mathematical principle. In geometric terms it says that if A and B are two points on a smooth curve, then there is at least one point C on the curve between A and B where the tangent line is parallel to the chord line AB.

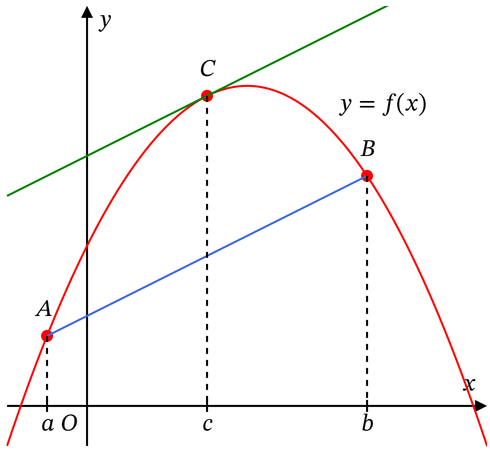

Figure 4–11: There is a point C on the curve where the tangent (green) is parallel to the chord AB (blue).

This principle is stated more precisely in the following theorem.

> ***Theorem 45 (The Mean-Value theorem).***
> *Suppose that the function f is continuous on the closed, finite interval $[a, b]$ and that it is differentiable on the open interval $]a, b[$. Then there exists a point c in the open interval $]a, b[$ such that*
> $$\frac{f(b) - f(a)}{b - a} = f'(c)\,.$$

This says that the slope of the chord line joining the points $(a, f(a))$ and $(b, f(b))$ is equal to the slope of the tangent line to the curve $f = f(x)$ at the point $(c, f(c))$, so the two lines are parallel.

We can make several observations:

1. The hypotheses of the Mean-Value theorem are all necessary for the conclusion; if f fails to be continuous at even one point of $[a, b]$ or fails to be differentiable at even one point of $]a, b[$, then there may be no point where the tangent line is parallel to the secant line AB.
2. The Mean-Value theorem gives no indication of how many points C there may be on the curve between A and B where the tangent is parallel to AB. If the curve is itself the straight line AB, then every point on the line between A and B has the required property. In general, there may be more than one point; the Mean-Value theorem asserts only that there must be at least one.
3. The Mean-Value theorem gives us no information on how to find the point c, which it says must exist. For some simple functions it is possible to calculate c (see the following example), but doing so is usually of no practical value. As we shall see, the importance of the Mean-Value theorem lies in its use as a theoretical tool. It belongs to a class of theorems called existence theorems

The Mean-Value theorem is one of those deep results that is based on the completeness of the real number system via the fact that a continuous function on a closed, finite interval takes on a maximum and minimum value (Extreme-Value theorem). Before giving the proof, we establish two preliminary results.

Theorem 46.
If f is defined on an open interval $]a, b[$ and achieves a maximum (or minimum) value at the point c in $]a, b[$, and if $f'(c)$ exists, then $f'(c) = 0$. (Values of x where $f'(x) = 0$ are called critical points of the function f.)

Proof. Suppose that f has a maximum value at c. Then $f(x) - f(c) \leq 0$ whenever x is in $]a, b[$.

If $c < x < b$, then

$$\frac{f(x) - f(c)}{x - c} \leq 0\,, \quad \text{so } f'(c) = \lim_{x \to c+} \frac{f(x) - f(c)}{x - c} \leq 0\,.$$

If $a < x < c$, then

$$\frac{f(x) - f(c)}{x - c} \geq 0\,, \quad \text{so } f'(c) = \lim_{x \to c-} \frac{f(x) - f(c)}{x - c} \geq 0\,.$$

Thus $f'(c)$. The proof for a minimum value at c is similar.

Theorem 47 (Rolle's theorem).
Suppose that the function g is continuous on the closed, finite interval $[a, b]$ and that it is differentiable on the open interval $]a, b[$. If $g(a) = g(b)$, then there exists a point c in the open interval $]a, b[$ such that $g'(c) = 0$.

Proof.
If $g(x) = g(a)$ for every x in $[a, b]$, then g is a constant function, so $g'(c) = 0$ for every c in $]a, b[$. Therefore, suppose there exists x in $]a, b[$ such that $g(x) \neq g(a)$.

Let us assume that $g(x) > g(a)$. (If $g(x) < g(a)$, the proof is similar.)

By the Extreme-Value theorem, being continuous on $[a, b]$, g must have a maximum value at some point c in $[a, b]$.

Since $g(c) \geq g(x) > g(a) = g(b)$, c cannot be either a or b. Therefore, c is in the open interval $]a, b[$, so g is differentiable at c.

By the previous theorem, c must be a critical point of g: $g'(c) = 0$.

Rolle's theorem is a special case of the Mean-Value theorem in which the chord line has slope 0, so the corresponding parallel tangent line must also have slope 0. We can deduce the Mean-Value theorem from this special case.

Proof (of the Mean-Value theorem). Suppose f satisfies the conditions of the Mean-Value theorem. Let

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

The function g is also continuous on $[a, b]$ and differentiable on $]a, b[$ because f has these properties. In addition, $g(a) = g(b) = 0$. By Rolle's theorem, there is some point c in $]a, b[$ such that $g'(c) = 0$. Since

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

it follows that

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

Many of the applications we will make of the Mean-Value theorem will actually use the following generalized version of it.

Theorem 48 (The Generalized Mean-Value theorem).
If functions f and g are both continuous on $[a, b]$ and differentiable on $]a, b[$, and if $g'(x) \neq 0$ for every $x \in\,]a, b[$, then there exists a number $c \in\,]a, b[$ such that

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

Proof.
Note that $g(b) \neq g(a)$ otherwise, there would be some number in $]a, b[$ where $g'(x) = 0$. Hence, neither denominator above can be zero. Apply the Mean-Value theorem to

$$h(x) = (f(b) - f(a))(g(x) - g(a)) - (g(b) - g(a))(f(x) - f(a)).$$

Since $h(a) = h(b) = 0$, there exists $c \in\,]a, b[$ such that $h'(c) = 0$. Thus,

$$(f(b) - f(a))g'(c) - (g(b) - g(a))f'(c) = 0,$$

and the result follows on division by the g factors.

4–9 Indeterminate Forms

We showed that

$$\lim_{x \to 0} \frac{\sin x}{x} = 1$$

We could not readily see this by substituting $x = 0$ into the function $\frac{\sin x}{x}$ because both $\sin x$ and x are zero at $x = 0$. We call $\frac{\sin x}{x}$ an indeterminate form of type $\left[\frac{0}{0}\right]$ at $x = 0$. The limit of such an indeterminate form can be any number. There are other types of indeterminate forms:

$$\left[\frac{\infty}{\infty}\right], \quad [0 \cdot \infty], \quad [\infty - \infty], \quad \left[0^0\right], \quad \left[\infty^0\right], \quad \left[1^\infty\right].$$

Indeterminate forms of type $\left[\frac{0}{0}\right]$ are the most common. You can evaluate many indeterminate forms of type $\left[\frac{0}{0}\right]$ with simple algebra, typically by cancelling common factors. We will now develop another method called l'Hôpital's Rules for evaluating limits of indeterminate forms of the types $\left[\frac{0}{0}\right]$ and $\left[\frac{\infty}{\infty}\right]$. The other types of indeterminate forms can usually be reduced to one of these two by algebraic manipulation and the taking of logarithms.

Theorem 49 (The first l'Hôpital Rule).
Suppose the functions f and g are differentiable on the interval $]a,b[$, and $g'(x) \neq 0$ there. Suppose also that

1. $\lim\limits_{x\to a^+} f(x) = lim_{x\to a^+} g(x) = 0$ *and*
2. $\lim\limits_{x\to a^+} \dfrac{f'(x)}{g'(x)} = L$ *where L is finite or ∞ or $-\infty$.*

Then

$$lim_{x\to a^+} \frac{f(x)}{g(x)} = L\,.$$

Similar results hold if every occurrence of $\lim_{x\to a^+}$ is replaced by $\lim_{x\to b^-}$ or even $\lim_{x\to c}$ where $a < c < b$. The cases $a = -\infty$ and $b = \infty$ are also allowed.

Proof. We prove the case involving $\lim_{x\to a^+}$ for finite a.

Define

$$F(x) = \begin{cases} f(x) & \text{if } a < x < b; \\ 0 & \text{if } x = a \end{cases} \quad \text{and} \quad G(x) = \begin{cases} g(x) & \text{if } a < x < b; \\ 0 & \text{if } x = a \end{cases}.$$

Then, F and G are continuous on the interval $[a,x]$ and differentiable on the interval $]a,x[$ for every $x \in\,]a,b[$.

By the Generalized Mean-Value theorem there exists a number $c \in\,]a,x[$ such that

$$\frac{f(x)}{g(x)} = \frac{F(x)}{G(x)} = \frac{F(x)-F(a)}{G(x)-G(a)} = \frac{F'(c)}{G'(c)} = \frac{f'(c)}{g'(c)}$$

Since $a < c < x$, if $x \to a^+$, then necessarily $c \to a^+$, so we have

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

The case involving $\lim_{x\to b^-}$ for finite b is proved similarly. The cases where $a = -\infty$ or $b = \infty$ follow from the cases already considered via the change of variable $x = \frac{1}{t}$.

Example. Evaluate $\lim\limits_{x\to \frac{\pi}{2}^-} \dfrac{2x-\pi}{\cos^2 x}$.

$$\lim_{x\to \frac{\pi}{2}^-} \frac{2x-\pi}{\cos^2 x} = \lim_{x\to \frac{\pi}{2}^-} \frac{2}{-2\sin x \cos x} = -\infty\,.$$

Evaluate $\lim\limits_{x\to 0^+} \left(\dfrac{1}{x} - \dfrac{1}{\sin x}\right)$.

$$\lim_{x\to 0^+} \left(\frac{1}{x} - \frac{1}{\sin x}\right) = \lim_{x\to 0^+} \frac{\sin x - x}{x\sin x} = \lim_{x\to 0^+} \frac{\cos x - 1}{\sin x + x\cos x} = \lim_{x\to 0^+} \frac{-\sin x}{2\cos x - x\sin x} = \frac{-0}{2} = 0\,.$$

Theorem 50 (The second l'Hôpital Rule).
Suppose the functions f and g are differentiable on the interval $]a, b[$, and $g'(x) \neq 0$ there. Suppose also that

1. $\lim_{x \to a+} g(x) = \pm\infty$ *and*
2. $\lim_{x \to a+} \dfrac{f'(x)}{g'(x)} = L$ *where L is finite or ∞ or $-\infty$.*

Then

$$\lim_{x \to a+} \frac{f(x)}{g(x)} = L\,.$$

Again, similar results hold if every occurrence of $\lim_{x\to a+}$ is replaced by $\lim_{x\to b-}$ or even $\lim_{x\to c}$ where $a < c < b$. The cases $a = -\infty$ and $b = \infty$ are also allowed.

Do not try to use l'Hôpital's Rules to evaluate limits that are not indeterminate of type $\left[\frac{0}{0}\right]$ or $\left[\frac{\infty}{\infty}\right]$; such attempts will almost always lead to false conclusions.

4–10 Increasing and Decreasing Functions

Intervals on which the graph of a function f has positive or negative slope provide useful information about the behaviour of f. The Mean-Value theorem enables us to determine such intervals by considering the sign of the derivative f'.

Definition 32.
Suppose that the function f is defined on an interval I and that x_1 and x_2 are two points of I.

- If $f(x_2) > f(x_1)$ whenever $x_2 > x_1$, we say f is *increasing* on I.
- If $f(x_2) < f(x_1)$ whenever $x_2 < x_1$, we say f is *decreasing* on I.
- If $f(x_2) \geq f(x_1)$ whenever $x_2 > x_1$, we say f is *nondecreasing* on I.
- If $f(x_2) \leq f(x_1)$ whenever $x_2 > x_1$, we say f is *nonincreasing* on I.

Note the distinction between increasing and nondecreasing. If a function is increasing (or decreasing) on an interval, it must take different values at different points. (Such a function is called *bijective*.) A nondecreasing function (or a nonincreasing function) may be constant on a subinterval of its domain, and may therefore not be bijective. An increasing function is nondecreasing, but a nondecreasing function is not necessarily increasing.

Theorem 51.
Let J be an open interval, and let I be an interval consisting of all the points in J and possibly one or both of the endpoints of J: Suppose that f is continuous on I and differentiable on J.

- *If $f'(x) > 0$ for all $x \in J$, then f is increasing on I.*
- *If $f'(x) < 0$ for all $x \in J$, then f is decreasing on I.*
- *If $f'(x) \geq 0$ for all $x \in J$, then f is nondecreasing on I.*
- *If $f'(x) \leq 0$ for all $x \in J$, then f is nonincreasing on I.*

Proof. Let x_1 and x_2 be points in I with $x_2 > x_1$.

By the Mean-Value theorem there exists a point $c \in]x_1, x_2[$ (and therefore in J) such that

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'(c);$$

hence, $f(x_2) - f(x_1) = (x_2 - x_1)f'(c)$. Since $x_2 - x_1 > 0$, the difference $f(x_2) - f(x_1)$ has the same sign as $f'(c)$ and may be zero if $f'(c)$ is zero. Thus, all four conclusions follow from the corresponding parts of the previous definition.

Example.
On what intervals is the function $f(x) = x^3 - 12x + 1$ increasing? On what intervals is it decreasing?

We have $f'(x) = 3x^2 - 12 = 3(x-2)(x+2)$. Observe that $f'(x) > 0$ if $x < -2$ or $x > 2$ and $f'(x) < 0$ if $-2 < x < 2$. Therefore, f is increasing on the intervals $]-\infty, -2[$ and $]2, \infty[$ and is decreasing on the interval $]-2, 2[$.

If a function is constant on an interval, then its derivative is zero on that interval. The Mean-Value theorem provides a converse of this fact.

Theorem 52.
If f is continuous on an interval I, and $f'(x) = 0$ every interior point of I (i.e., at interior point of I that is not an endpoint of I), then $f(x) = C$, a constant, on I.

Proof. Pick a point x_0 in I and let $C = f(x_0)$.

If x is any other point of I, then the Mean-Value theorem says that there exists a point c between x_0 and x such that

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

The point c must belong to I because an interval contains all points between any two of its points, and c cannot be an endpoint of I since $c \neq x_0$ and $c \neq x$.

Since $f'(c) = 0$ for all such points c, we have $f(x) - f(x_0) = 0$ for all $x \in I$, and $f(x) = f(x_0) = C$ as claimed.

4–11 Extreme Values

The first derivative of a function is a source of much useful information about the behaviour of the function.

Maximum and Minimum Values

Recall that a function has a maximum value at $x0$ if $f(x) \leq f(x_0)$ for all x in the domain of f. The maximum value is $f(x_0)$. To be more precise, we should call such a maximum value an *absolute* or *global* maximum because it is the largest value that f attains anywhere on its entire domain.

Definition 33.
Function f has an *absolute maximum value* $f(x_0)$ at the point x_0 in its domain if $f(x) \leq f(x_0)$ holds for every x in the domain of f.

Similarly, f has an *absolute minimum value* $f(x_1)$ at the point x_1 in its domain if $f(x) \geq f(x_1)$ holds for every x in the domain of f.

Maximum and minimum values of a function are collectively referred to as *extreme values*. The following theorem is a restatement (and slight generalization) of the Extreme-Value theorem. It will prove very useful in some circumstances when we want to find extreme values.

Theorem 53.
If the domain of the function f is a closed, finite interval or a union of finitely many such intervals, and if f is continuous on that domain, then f must have an absolute maximum value and an absolute minimum value.

In addition to these extreme values, f can have several other "local" maximum and minimum values corresponding to points on the graph that are higher or lower than neighbouring points. The absolute maximum is the highest of the local maxima; the absolute minimum is the lowest of the local minima.

Definition 34.
Function f has a *local maximum value* $f(x_0)$ at the point x_0 in its domain provided there exists a number $h > 0$ such that $f(x) \leq f(x_0)$ whenever x is in the domain of f and $|x - x_0| < h$.

Similarly, f has a *local minimum value* $f(x_1)$ at the point x_1 in its domain provided there exists a number $h > 0$ such that $f(x) \geq f(x_1)$ whenever x is in the domain of f and $|x - x_1| < h$.

Critical Points, Singular Points, and Endpoints

A function f can have local extreme values only at points x of three special types:

1. critical points of f (points x in the domain of f where $f'(x) = 0$),
2. singular points of f (points x in the domain of f where $f'(x)$ is not defined), and
3. endpoints of the domain of f (points x in the domain of f that do not belong to any open interval contained in the domain of f).

Theorem 54.
If the function f is defined on an interval I and has a local maximum (or local minimum) value at point $x = x_0$ in I, then x_0 must be either a critical point of f; a singular point of f; or an endpoint of I.

Proof.
Suppose that f has a local maximum value at x_0 and that x_0 is neither an endpoint of the domain of f nor a singular point of f.

Then for some $h > 0$, $f(x)$ is defined on the open interval $]x_0 - h, x_0 + h[$ and has an absolute maximum (for that interval) at x_0. Also, $f'(x_0)$ exists. By the first preliminary result of the Mean-value theorem, $f'(x_0) = 0$.

The proof for the case where f has a local minimum value at x_0 is similar.

Although a function cannot have extreme values anywhere other than at endpoints, critical points, and singular points, it need not have extreme values at such points.

Finding Absolute Extreme Values

If a function f is defined on a closed interval or a union of finitely many closed intervals, the Extreme-value theorem assures us that f must have an absolute maximum value and an absolute minimum value. The last theorem tells us how to find them. We need only check the values of f at any critical points, singular points, and endpoints.

Example.
Find the maximum and the minimum values of the function $g(x) = x^2 - 3x^2 - 9x + 2$ on the interval $[-2, 2]$.

Since g is a polynomial, it can have no singular points. For critical points, we calculate

$$g'(x) = 3x^2 - 6x - 9 = 3\left(x^2 - 2x - 3\right) = 3(x+1)(x-3) = 0$$

if $x = -1$ or $x = 3$.

However, $x = 3$ is not in the domain of g, so we can ignore it. We need to consider only the values of g at the critical point $x = -1$ and at the endpoints $x = -2$ and $x = 2$:

$$g(-2) = 0 \quad g(-1) = 7 \quad g(2) = -20\,.$$

The maximum value of g on $[-2, 2]$ is 7, at the critical point $x = -1$, and the minimum value is -20, at the endpoint $x = 2$.

Example.

Find the maximum and the minimum values of the function $h(x) = 2x^{\frac{2}{3}} - 2x$ on the interval $[-1, 1]$.

The derivative of h is

$$h'(x) = 3\left(\frac{2}{3}\right)x^{-\frac{1}{3}} - 2 = 2\left(x^{-\frac{1}{3}} - 1\right)$$

Note that $x^{-\frac{1}{3}}$ is not defined at the point $x = 0$ in the domain of h, so $x = 0$ is a singular point of h. Also, h has a critical point where $x^{-\frac{1}{3}} = 1$, that is, at $x = 1$, which also happens to be an endpoint of the domain of h. We must therefore examine the values of h at the points $x = 0$ and $x = 1$, as well as at the other endpoint $x = -1$. We have

$$h(-1) = 5\,, \quad h(0) = 0\,, \quad h(1) = 1\,.$$

The function h has maximum value 5 at the endpoint point $x = -1$ and minimum value 0 at the singular point $x = 0$.

The First Derivative Test

Most functions you will encounter in elementary calculus have nonzero derivatives everywhere on their domains except possibly at a finite number of critical points, singular points, and endpoints of their domains. On intervals between these points the derivative exists and is not zero, so the function is either increasing or decreasing there. If f is continuous and increases to the left of x_0 and decreases to the right, then it must have a local maximum value at x_0. The following theorem collects several results of this type together.

Theorem 55 (The First Derivative Test). *PART I. Testing interior critical points and singular points.*

Suppose that f is continuous at x_0, and x_0 is not an endpoint of the domain of f:

- *If there exists an open interval $]a,b[$ containing x_0 such that $f'(x_0) > 0$ on $]a,x_0[$ and $f'(x_0) < 0$ on $]x_0,b[$, then f has a local maximum value at x_0.*
- *If there exists an open interval $]a,b[$ containing x_0 such that $f'(x_0) < 0$ on $]a,x_0[$ and $f'(x_0) > 0$ on $]x_0,b[$, then f has a local minimum value at x_0.*

PART II. Testing endpoints of the domain.

Suppose a is a left endpoint of the domain of f and f is right continuous at a.

- *If $f'(x_0) > 0$ on some interval $]a,b[$, then f has a local minimum value at a.*
- *If $f'(x_0) <$ on some interval $]a,b[$, then f has a local maximum value at a.*

Suppose b is a right endpoint of the domain of f and f is left continuous at b.

- *If $f'(x_0) >$ on some interval $]a,b[$, then f has a local maximum value at b.*
- *If $f'(x_0) < 0$ on some interval $]a,b[$, then f has a local mainimum value at b.*

If f' is positive (or negative) on both sides of a critical or singular point, then f has neither a maximum nor a minimum value at that point.

Functions Not Defined on Closed, Finite Intervals

If the function f is not defined on a closed, finite interval, then the extended Extreme-value theorem cannot be used to guarantee the existence of maximum and minimum values for f. Of course, f may still have such extreme values. In many applied situations we will want to find extreme values of functions defined on infinite and/or open intervals.

Theorem 56. *If f is continuous on the open interval $]a,b[$, and if*

$$\lim_{x\to a^+} f(x) = L \quad \text{and} \quad \lim_{x\to b^-} f(x) = M\,,$$

then the following conclusions hold:

- *If $f(u) > L$ and $f(u) > M$ for some $u \in\ ]a,b[$, then f has an absolute maximum value on $]a,b[$.*
- *If $f(u) < L$ and $f(u) < M$ for some $u \in\ ]a,b[$, then f has an absolute minimum value on $]a,b[$.*

In this theorem a may be $-\infty$, in which case $\lim_{x\to a^+}$ should be replaced with $\lim_{x\to-\infty}$, and b may be ∞, in which case $\lim_{x\to b^-}$ should be replaced with $\lim_{x\to\infty}$.

Also, either or both of L and M may be either ∞ or $-\infty$.

4–12 Convexity/Concavity and Inflections

Like the first derivative, the second derivative of a function also provides useful information about the behaviour of the function and the shape of its graph: it determines whether the graph is *bending upward* (i.e., has increasing slope) or *bending downward* (i.e., has decreasing slope) as we move along the graph toward the right.

Definition 35.
We say that the function f is convex on an open interval I if it is differentiable there and the derivative f' is an increasing function on I. Similarly, f is concave I if f' exists and is decreasing on I.

The terms “convex” and “concave are used to describe the graph of the function as well as the function itself.

Note that convexity/concavity is defined only for differentiable functions, and even for those, only on intervals on which their derivatives are not constant. According to the above definition, a function is neither concave up nor concave down on an interval where its graph is a straight line segment. We say the function has no convexity/concavity on such an interval. We also say a function has opposite convexity/ concavity on two intervals if it is convex on one interval and concave the other.

Definition 36.
We say that the point $(x_0, f(x_0))$ is an *inflection point* of the curve $y = f(x)$ (or that the function f has an inflection point at x_0) if the following two conditions are satisfied:

1. the graph of $y = f(x)$ has a tangent line at $x = x_0$, and
2. the convexity/concavity of f is opposite on opposite sides of x_0.

Note that 1. implies that either f is differentiable at x_0 or its graph has a vertical tangent line there, and 2. implies that the graph crosses its tangent line at x_0. An inflection point of a function f is a point on the graph of a function, rather than a point in its domain like a critical point or a singular point. A function may or may not have an inflection point at a critical point or singular point. In general, a point P is an inflection point (or simply an inflection) of a curve C (which is not necessarily the graph of a function) if C has a tangent at P and arcs of C extending in opposite directions from P are on opposite sides of that tangent line.

If a function f has a second derivative f'', the sign of that second derivative tells us whether the first derivative f' is increasing or decreasing and hence determines the concavity of f.

Theorem 57.

1. *If $f''(x) > 0$ on interval I, then f is convex on I.*
2. *If $f''(x) < 0$ on interval I, then f is concave on I.*
3. *if f has an inflection point at x_0 and $f''(x_0)$ exists, then $f''(x_0) = 0$.*

Proof. Part 1. and 2. follow from applying theorem 52 to the derivative f'.

If f has an inflection point at x_0 and $f''(x_0)$ exists, then f must be differentiable in an open interval containing x_0. Since f' is increasing on one side of x_0 and decreasing on the other side, it must have a local maximum or minimum at x_0. By theorem 46, $f''(x_0) = 0$.

This theorem tells us that to find (the x-coordinates of) inflection points of a twice differentiable function f , we need only look at points where $f''(x) = 0$. However, not every such point has to be an inflection point. For example, $f(x) = x^4$, does not have an inflection point at $x = 0$.

A function f will have a local maximum (or minimum) value at a critical point if its graph is concave down (or up) in an interval containing that point. In fact, we can often use the value of the second derivative at the critical point to determine whether the function has a local maximum or a local minimum value there.

Theorem 58 (The Second Derivative Test).

1. *If $f'(x_0) = 0$ and $f''(x) < 0$, then f has a local maximum value at x_0.*
2. *If $f'(x_0) = 0$ and $f''(x) > 0$, then f has a local minimum value at x_0.*
3. *If $f'(x_0) = 0$ and $f''(x) = 0$, no conclusion can be drawn; f may have a local maximum at x_0 or a local minimum, or it may have an inflection point instead.*

Proof. Suppose that $f'(x_0) = 0$ and $f''(x) < 0$.

Since

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

it follows that $f'(x_0+h) < 0$ for all sufficiently small positive h, and $f'(x_0+h) > 0$ for all sufficiently small negative h. By the first derivative test, f must have a local maximum value at x_0. The proof of the local minimum case is similar.

4–13 Taylor Polynomials

Linear Approximations

Many problems in applied mathematics are too difficult to be solved exactly—that is why we resort to using computers, even though in many cases they may only give approximate answers. However, not all approximation is done with machines. Linear approximation can be a very effective way to estimate values or test the plausibility of numbers given by a computer.

The tangent to the graph $y = f(x)$ at $x = a$ describes the behaviour of that graph near the point $P = (a, f(a))$ better than any other straight line through P, because it goes through P in the same direction as the curve $y = f(x)$. We exploit this fact by using the height to the tangent line to calculate approximate values of $f(x)$ for values of x near a. The tangent line has equation $y = f(a) + f'(a)(x-a)$. We call the right side of this equation the *linearization* of f about a.

Definition 37. The linearization of the function f about a is the function L defined by

$$L(x) = f(a) + f'(a)(x-a)\,.$$

We say that $f(x) \approx L(x) = f(a) + f'(a)(x-a)$ provides *linear approximations* for values of f near a.

We have already made use of linearization in a previous section, where it was disguised as the formula

$$\Delta y = \frac{\mathrm{d}y}{\mathrm{d}x}\Delta y$$

and used to approximate a small change $\Delta y = f\big(a + \Delta x\big) - f(a)$ in the values of function f corresponding to the small change in the argument of the function from a to $a + \Delta x$. This is just the linear approximation

$$f\big(a + \Delta x\big) \approx L(x) = f(a) + f'(a)(x-a)\,.$$

In any approximation, the *error* is defined by

$$\text{error} \;=\; \text{true value} \;-\; \text{approximate value}\,.$$

If the linearization of f about a is used to approximate $f(x)$ near $x = a$, then the error $E(x)$ in this approximation is

$$E(x) = f(x) - L(x) = f(x) - f(a) - f'(a)(x-a)\,.$$

It is the vertical distance at x between the graph of f and the tangent line to that graph at $x = a$. Observe that if x is “near” a, then $E(x)$ is small compared to the horizontal distance between x and a.

The following theorem gives us a way to estimate this error if we know bounds for the second derivative of f:

Theorem 59.
If $f''(t)$ exists for all t in an interval containing a and x, then there exists some point s between a and x such that the error $E(x) = f(x) - L(x)$ in the linear approximation $f(x) \approx L(x) = f(a) + f'(a)(x-a)$ satisfies

$$E(x) = \frac{f''(s)}{2}(x-a)^2 .$$

Proof. Let us assume that $x > a$. (The proof for $x < a$ is similar.)

Since

$$E(t) = f(t) - f(a) - f'(a)(t-a) ,$$

we have $E'(t) = f'(t) - f'(a)$. We apply the Generalized Mean-Value theorem to the two functions $E(t)$ and $(t-a)^2$ on $[a,x]$. Noting that $E(a) = 0$, we obtain a number $u \in]a,x[$ such that

$$\frac{E(x)}{(x-a)^2} = \frac{E(x)-E(a)}{(x-a)^2-(a-a)^2} = \frac{E'(u)}{2(u-a)} = \frac{f'(u)-f'(a)}{2(u-a)} = \frac{1}{2}f''(s)$$

for some $s \in]a,u[$; the latter expression is a consequence of applying the Mean-Value theorem again, this time to f' on $[a,u]$.

Thus,

$$E(x) = \frac{f''(s)}{2}(x-a)^2 .$$

as claimed.

Example.
Use the linearization for $\sqrt{x}$ about $x = 25$ to find an approximate value for $\sqrt{26}$ and estimate the size of the error. Use these to give a small interval that you can be sure contains $\sqrt{26}$.

If $f(x) = \sqrt{x}$, then $f'(x) = \frac{1}{2\sqrt{x}}$. Since we now that $f(25) = 5$ and $f'(25) = \frac{1}{10}$, the linearization of $f(x)$ about $x = 25$ is

$$L(x) = 5 + \frac{1}{10}(x-25)$$

Putting $x = 26$, we get

$$\sqrt{26} = f(26) \approx L(26) = 5 + \frac{1}{10}(26-25) = 5.1 .$$

We have also $f''(x) = -\frac{1}{4}x^{-\frac{3}{2}}$. For $25 < x < 26$, $f''(x) < 0$, so $\sqrt{26} = f(26) < L(26) = 5.1$. Also, $|f''(x)| < \frac{1}{4}\frac{1}{125} = \frac{1}{500}$ and

$$|E(26)| < \frac{1}{2}\frac{1}{500}(26-25)^2 = \frac{1}{1000} = 0.001 .$$

Therefore, $f(26) > L(26) - 0.001 = 5.099$, and $\sqrt{26}$ is in the interval $]5.099, 5.1[$.

Higher Order Approximations

The linearization of a function $f(x)$ about $x = a$, namely, the linear function

$$P_1(x) = L(x) = f(a) + f'(a)(x-a) ,$$

describes the behaviour of f near a better than any other polynomial of degree 1 because both P_1 and f have the same value and the same derivative at a:

$$P_1(a) = f(a) \quad \text{and} \quad P_1'(a) = f'(a)\,.$$

We can obtain even better approximations to $f(x)$ by using quadratic or higherdegree polynomials and matching more derivatives at $x = a$. For example, if f is twice differentiable near a, then the polynomial

$$P_2(x) = L(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2\,,$$

satisfies $P_2(a) = f(a)$, $P_2'(a) = f'(a)$ and $P_2''(a) = f''(a)$ and describes the behaviour of f near a better than any other polynomial of degree at most 2.

In general, if $f^{(n)}(x)$ exists in an open interval containing $x = a$, then the polynomial

$$P_n(x) = L(x) = \sum_{i=0}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i\,,$$

matches f and its first n derivatives at $x = a$, and so describes $f(x)$ near $x = a$ better than any other polynomial of degree at most n. P_n is called the nth-order *Taylor polynomial* for f about a. Taylor polynomials about 0 are usually called *Maclaurin polynomials*.

Example. Find the Taylor polynomials $P_1(x)$, $P_2(x)$ and $P_3(x)$ for $f(x) = \sqrt{x}$ about $x = 25$.

We have

$$f'(x) = \frac{1}{2}x^{-\frac{1}{2}}\,, \quad f''(x) = -\frac{1}{4}x^{-\frac{3}{2}}\,, \quad f^{(3)}(x) = \frac{3}{8}x^{-\frac{5}{2}}\,.$$

Thus,

$$\begin{aligned}
P_1(x) &= f(25) + f'(25)(x-25)\\
&= 5 + \frac{1}{10}(x-25)\\
P_2(x) &= f(25) + f'(25)(x-25) + \frac{f''(25)}{2}(x-25)^2\\
&= 5 + \frac{1}{10}(x-25) - \frac{1}{1000}(x-25)^2\\
P_2(x) &= f(25) + f'(25)(x-25) + \frac{f''(25)}{2}(x-25)^2 + \frac{f^{(3)}(25)}{3!}(x-25)^3\\
&= 5 + \frac{1}{10}(x-25) - \frac{1}{1000}(x-25)^2 + \frac{1}{50000}(x-25)^3
\end{aligned}$$

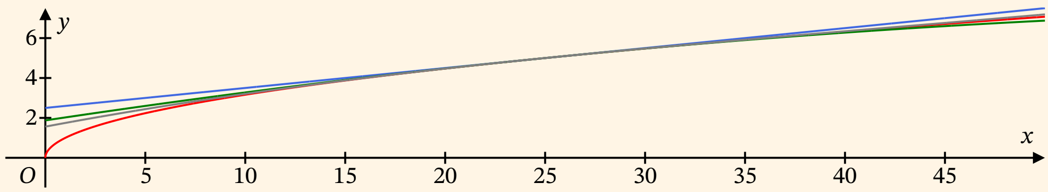

Figure 4–12: Taylor polynomials (blue, green, gray) for $f(x) = \sqrt{x}$ (red).

Theorem 60 (Taylor's theorem).
If the $(n+1)$st-order derivative, $f^{(n+1)}(t)$, exists for all t in an interval containing a and x, and if $P_n(x)$ is the nth order Taylor polynomial for f about a, then the error $E_n(x) = f(x) - p_n(x)$ is given by

$$E_n(x) = \frac{f^{(n+1)}(s)}{(n+1)!}(x-a)^{n+1},$$

where s is some number between a and x. The resulting formula

$$f(x) = \sum_{i=0}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i + \frac{f^{(n+1)}(s)}{(n+1)!}(x-a)^{n+1},$$

is called Taylor's formula with Lagrange remainder.

Proof (by induction). Observe that the case $n = 0$ of Taylor's formula, namely,

$$f(x) = P_0(x) + E_0(x) = f(a) + f'(x)(x-a)\,.$$

is just the Mean-Value theorem

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

for some s between a and x.

Suppose that we have proved the case $n = k-1$, where $k \geq 1$. Thus, we are assuming that if f is any function whose kth derivative exists on an interval containing a and x, then

$$E_{k-1}(x) = \frac{f^{(k)}(s)}{k!}(x-a)^k,$$

where s is some number between a and x.

Let us consider the next higher case: $n = k$. We assume $x > a$ (the case $x < a$ is similar) and apply the generalized Mean-Value theorem to the functions $E_k(t)$ and $(t-a)^{k+1}$ on $[a,x]$. Since $E_k(a) = 0$, we obtain a number $u \in\,]a,x[$ such that

$$\frac{E_k(x)}{(x-a)^{k+1}} = \frac{E_k(x) - E_k(a)}{(x-a)^{k+1} - (a-a)^{k+1}} = \frac{E_k'(u)}{(k+1)(u-a)^k}\,.$$

Now,

$$\begin{aligned}
E_k'(u) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(f(t) - f(a) - f'(a)(t-a) - \frac{f''(a)}{2}(t-a)^2 - \cdots - \frac{f^{(k)}(a)}{k!}(t-a)^k\right)\Bigg|_{t=u} \\
&= f'(t) - f'(a) - f''(a)(u-a) - \cdots - \frac{f^{(k)}(a)}{(k-1)!}(u-a)^{k-1}.
\end{aligned}$$

This last expression is just $E_{k-1}(u)$ for the function f' instead of f. By the induction assumption it is equal to

$$\frac{(f')^{(k)}(s)}{k!}(u-a)^k = \frac{f^{(k+1)}(s)}{k!}(u-a)^k$$

for some s between a and u. Therefore,

$$E_k(x) = \frac{f^{(k+1)}(s)}{(k+1)!}(x-a)^{k+1}.$$

We have shown that the case $n = k$ of Taylor's theorem is true if the case $n = k-1$ is true, and the inductive proof is complete.

4–14 *Antiderivatives*

Throughout this chapter we have been concerned with the problem of finding the derivative f' of a given function f. The reverse problem—given the derivative f', find f—is also interesting and important. It is the problem studied in integral calculus and is generally more difficult to solve than the problem of finding a derivative. We will take a preliminary look at this problem in this section and will return to it in more detail in Chapter 6.

We begin by defining an *antiderivative* of a function f to be a function F whose derivative is f: It is appropriate to require that $F'(x) = f(x)$ on an interval.

Definition 38. An antiderivative of a function f on an interval I is another function F satisfying

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

for $x \in I$.

Antiderivatives are not unique; since a constant has derivative zero, you can always add any constant to an antiderivative F of a function f on an interval and get another antiderivative of f on that interval. More importantly, all antiderivatives of f on an interval can be obtained by adding constants to any particular one. If F and G are both antiderivatives of f on an interval I, then

$$\frac{\mathrm{d}}{\mathrm{d}x}(G(x) - F(x)) = f(x) - f(x) = 0$$

on I, so $G(x) - F(x) = C$ (a constant) on I. Thus, $G(x) = F(x) + C$ on I.

Note that this conclusion is not valid over a set that is not an interval. For example, the derivative of

$$\operatorname{sgn}(x) = \begin{cases} 1\,, & \text{if } x > 0 \\ -1\,, & \text{if } x < 0\,. \end{cases}$$

is 0 for all $x \neq 0$, but $\operatorname{sgn}(x)$ is not constant for all $x \neq 0$.

The general antiderivative of a function $f(x)$ on an interval I is $F(x) + C$, where $F(x)$ is any particular antiderivative of $f(x)$ on I and C is a constant. This general antiderivative is called the *indefinite integral* of $f(x)$ on I and is denoted $\int f(x)\,\mathrm{d}x$.

Definition 39. The indefinite integral of $f(x)$ on interval I is

$$\int f(x)\,\mathrm{d}x = F(x) + C$$

on I, provided $F'(x) = f(x)$ for all $x \in I$.

The symbol $\int$ is called an *integral sign*. It is shaped like an elongated “S” for reasons that will only become apparent when we study the definite integral in Chapter 6. Just as you regard $\frac{\mathrm{d}y}{\mathrm{d}x}$ as a single symbol representing the derivative of y with respect to x, so you should regard $\int f(x)\,\mathrm{d}x$ as a single symbol representing the indefinite integral (general antiderivative) of f with respect to x. The constant C is called a *constant of integration*.

Finding antiderivatives is generally more difficult than finding derivatives; many functions do not have antiderivatives that can be expressed as combinations of finitely many elementary functions. However, every formula for a derivative can be rephrased as a formula for an antiderivative. For instance,

$$\frac{\mathrm{d}}{\mathrm{d}x}\sin x = \cos x\,; \quad \text{therefore,} \quad \int \cos x\,\mathrm{d}x = \sin x + C\,.$$

We will develop several techniques for finding antiderivatives in later chapters. Until then, we must content ourselves with being able to write a few simple antiderivatives based on the known derivatives of elementary functions:

$$
\begin{aligned}
&\int \mathrm{d}x = \int 1\,\mathrm{d}x = x + C && \int x\,\mathrm{d}x = \frac{x^2}{2} + C \\
&\int x^2\,\mathrm{d}x = \frac{x^3}{3} + C && \int \frac{1}{x^2}\,\mathrm{d}x = \frac{1}{x} + C \\
&\int \frac{1}{\sqrt{x}}\,\mathrm{d}x = 2\sqrt{x} + C && \int x^r\,\mathrm{d}x = \frac{x^{r+1}}{r+1} + C \quad (r \neq 1) \\
&\int \sin\, x\,\mathrm{d}x = -\cos\, x + C && \int \cos\, x\,\mathrm{d}x = \sin\, x + C \\
&\int \sec^2\, x\,\mathrm{d}x = \tan\, x + C && \int \csc^2\, x\,\mathrm{d}x = -\cot\, x + C
\end{aligned}
$$

For the moment, r must be rational, but this restriction will be removed in the next chapter.

The rule for differentiating sums and constant multiples of functions translates into a similar rule for antiderivatives

The graphs of the different antiderivatives of the same function on the same interval are vertically displaced versions of the same curve. In general, only one of these curves will pass through any given point, so we can obtain a unique antiderivative of a given function on an interval by requiring the antiderivative to take a prescribed value at a particular point x.

Example.
Find the function $f(x)$ whose derivative is $f'(x) = 6x^2 - 1$ for all real x and for which $f(2) = 10$.

Since $f'(x) = 6x^2 - 1$, we have

$$f(x) = \int \left(6x^2 - 1\right)\mathrm{d}x = 2x^3 - x + C$$

for some constant C. Since $f(2) = 10$, we have

$$10 = f(2) = 16 - 2 + C\,.$$

Thus, $C = -4$ and $f(x) = 2x^3 - x - 4$.

CHAPTER 5

TRANSCENDENTAL FUNCTIONS

With the exception of the trigonometric functions, all the functions we have encountered so far have been of three main types: polynomials, rational functions (quotients of polynomials), and algebraic functions (fractional powers of rational functions). On an interval in its domain, each of these functions can be constructed from real numbers and a single real variable x by using finitely many arithmetic operations (addition, subtraction, multiplication, and division) and by taking finitely many roots (fractional powers). Functions that cannot be so constructed are called *transcendental functions*. The only examples of these that we have seen so far are the trigonometric functions.

This chapter is devoted to developing other transcendental functions, including *exponential* and *logarithmic functions* and the *inverse trigonometric functions*.

5–1 Exponentials and Logarithms

Exponentials

An *exponential function* is a function of the form $f(x) = a^x$, where the *base* a is a positive constant and the *exponent* x is the variable. Do not confuse such functions with power functions such as $f(x) = x^a$, where the base is variable and the exponent is constant. The exponential function a^x can be defined for integer and rational exponents x as follows:

Definition 40. If $a > 0$, then

$$\begin{aligned} a^0 &= 1 \\ a^n &= \underbrace{a \cdot a \cdot a \cdots a}_{n \text{ factors}} \quad \text{if } n = 1,2,3,\ldots \\ a^{-n} &= \frac{1}{a^n} \quad \text{if } n = 1,2,3,\ldots \\ a^{\frac{m}{n}} &= \sqrt[n]{a^m} \quad \text{if } n = 1,2,3,\ldots \text{ and } m = \pm 1, \pm 2, \pm 3, \ldots \end{aligned}$$

In this definition, $\sqrt[n]{a}$ is the number $b > 0$ that satisfies $b^n = a$.

How should we define a^x if x is not rational? For example, what does 2^π mean? In order to calculate a derivative of a^x, we will want the function to be defined for all real numbers x, not just rational ones.

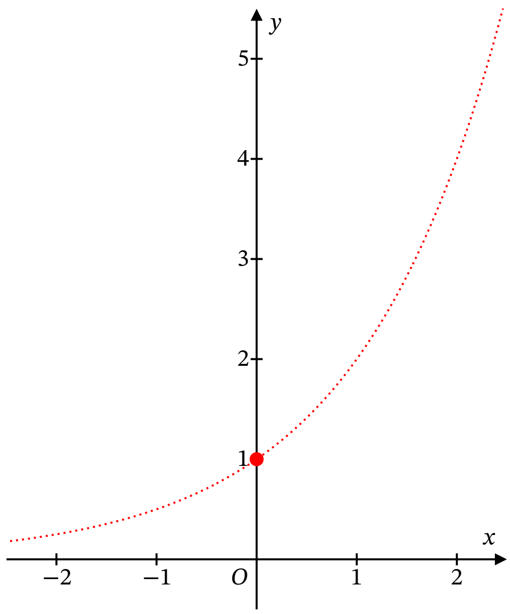

Figure 5–1: $y = 2^x$ for rational x.

In the previous figure we plot points with coordinates $(x, 2^x)$ for many closely spaced rational values of x. They appear to lie on a smooth curve. The definition of a^x can be extended to irrational x in such a way that a^x becomes a differentiable function of x on the whole real line. We will do so in the next section. For the moment, if x is irrational we can regard a^x as being the limit of values a^r for rational numbers rapproaching x.

Exponential functions satisfy several identities called *laws of exponents*:

Theorem 61. *If $a > 0$ and $b > 0$, and x and y are any real numbers, then*

$$a^0 = 1 \qquad a^{x+y} = a^x a^y$$

$$a^{-x} = \frac{1}{a^x} \qquad a^{x-y} = \frac{a^x}{a^y}$$

$$(a^x)^y = a^{xy} \qquad (ab)^x = a^x b^x$$

These identities can be proved for rational exponents using the definitions above. They remain true for irrational exponents, but we can't show that until the next section.

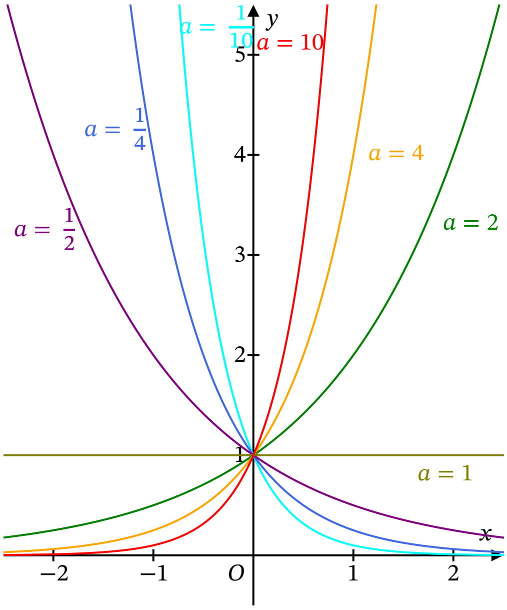

Figure 5–2: Graphs of some exponential functionsy $= a^x$.

The graphs of some typical exponential functions are shown in the previous figure. They all pass through the point (0,1) since $a^0 = 1$ for every $a > 0$. Observe that $a^x > 0$ for all $a > 0$ and all real x and that:

- If $a > 1$, then $\lim\limits_{x \to -\infty} a^x = 0$ and $\lim\limits_{x \to \infty} a^x = \infty$.
- If $0 < a < 1$, then $\lim\limits_{x \to -\infty} a^x = \infty$ and $\lim\limits_{x \to \infty} a^x = 0$.

Logarithms

The function $f(x) = a^x$ is a bijective function provided that $a > 0$ and $a \neq 1$. Therefore, f has an inverse which we call a *logarithmic function*.

Definition 41.
If $a > 0$ and $a \neq 1$, the function $\log_a x$, called the logarithm of x to the base a, is the inverse of the bijective function a^x:

$$\forall a > 0, a \neq 1 : x = a^y \implies y = \log_a x\,.$$

Since a^x has domain $]-\infty, \infty[$, $\log_a x$ has range $]-\infty, \infty[$. Since a^x has range $]0, \infty[$, $\log_a x$ has domain $]0, \infty[$. Since a^x and $\log_a x$ are inverse functions, the following cancellation identities hold:

$$\log_a(a^x) = x \text{ for all real } x \quad \text{and} \quad a^{\log_a x} = x \text{ for all } x > 0\,.$$

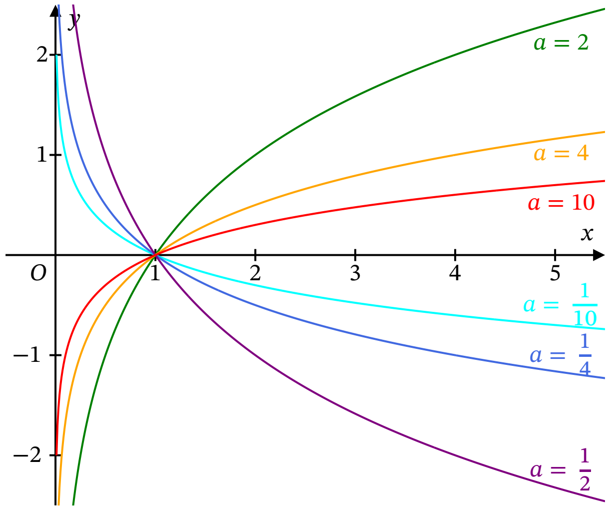

Figure 5–3: Graphs of some logarithmic functionsy $= \log_a x$.

The graphs of some typical logarithmic functions are shown in the previous figure. They all pass through the point $(1, 0)$. Each graph is the reflection in the line $y = x$ of the corresponding exponential graph.

From the laws of exponents we can derive the following laws of logarithms:

Theorem 62. *If* $x > 0$, $y > 0$, $a > 0$, $b > 0$, $a \neq 1$ *and* $b \neq 1$, *then*

$$\begin{array}{ll} \log_a 1 = 0 & \log_a (xy) = \log_a x + \log_a y \\ \log_a \left(\frac{1}{x}\right) = -\log_a x & \log_a \left(\frac{x}{y}\right) = \log_a x - \log_a y \\ \log_a (x^y) = y \log_a x & \log_a x = \dfrac{\log_b x}{\log_b a} \end{array}$$

Corresponding to the asymptotic behaviour of the exponential functions, the logarithmic functions also exhibit asymptotic behaviour. Their graphs are all asymptotic to the y-axis as $x \to 0$ from the right:

- If $a > 1$, then $\lim_{x \to 0^+} \log_a x = -\infty$ and $\lim_{x \to \infty} \log_a x = \infty$.
- If $0 < a < 1$, then $\lim_{x \to 0^+} \log_a x = \infty$ and $\lim_{x \to \infty} \log_a x = -\infty$.

5–2 *The Natural Logarithm and the Exponential Function*

In this section we are going to define a function $\ln x$, called the natural logarithm of x, in a way that does not at first seem to have anything to do with the logarithms considered in the previous section. We will show, however, that it has the same properties as those logarithms, and in the end we will see that $\ln x = \log_e$, the logarithm of x to a certain specific base e. We will show that $\ln x$ is a bijective function, defined for all positive real numbers. It must therefore have an inverse, e^x, that we will call the exponential function. Our final goal is to arrive at a definition of the exponential functions a^x (for any $a > 0$) that is valid for any real number x instead of just rational numbers, and that is known to be continuous and even differentiable without our having to assume those properties.

The Natural Logarithm

We do not yet know a function whose derivative is $x^{-1} = \frac{1}{x}$. We are going to remedy this situation by defining a function $\ln x$ in such a way that it will have derivative $\frac{1}{x}$.

To get a hint as to how this can be done, remember the relationship you have seen in high school between definite integral (area) and derivative. The definite integral between 2 points is the signed area between the 2 points bounded by the graph of the function and the x-axis. We will call this the *Fundamental Theorem of Calculus*, which we will explore in the next chapter.

Definition 42.
For $x > 0$, let A_x be the area of the plane region bounded by the curve $y = \frac{1}{t}$, the t-axis, and the vertical lines $t = 1$ and $t = x$. The function $\ln\ x$ is defined by

$$\ln\ x = \begin{cases} A_x & \text{if } x \geq 1\,, \\ -A_x & \text{if } 0 < x < 1\,, \end{cases}$$

as shown in the next figure.

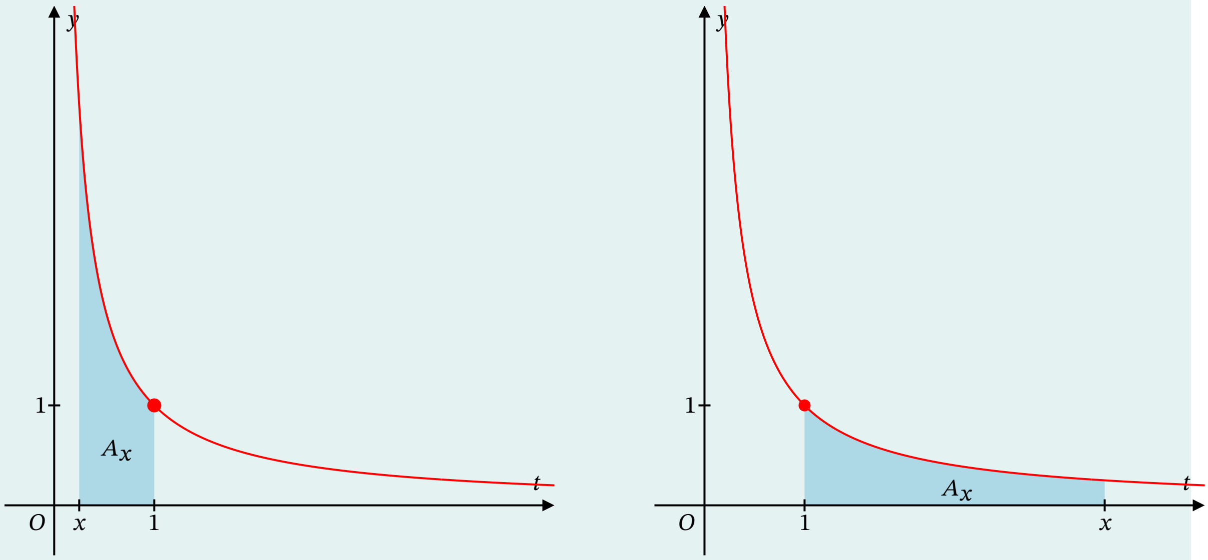

Figure 5–4: Definition of $\ln\ x$*.*

The definition implies that $\ln\ 1 = 0$, that $\ln\ x > 0$ if $x > 1$, that $\ln\ x < 0$ if $0 < x < 1$, and that ln is a bijective function. We now show that if $y = \ln\ x$, then $y' = \frac{1}{x}$.

Theorem 63. *If* $x > 0$*, then*

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln\ x = \frac{1}{x}\,.$$

Proof.
For $x > 0$ and $h > 0$, $\ln(x+h) - \ln(x)$ is the area of the plane region bounded by $y = \frac{1}{t}$, $y = 0$, and the vertical lines $t = x$ and $t = x + h$; it is the shaded area in the next figure.

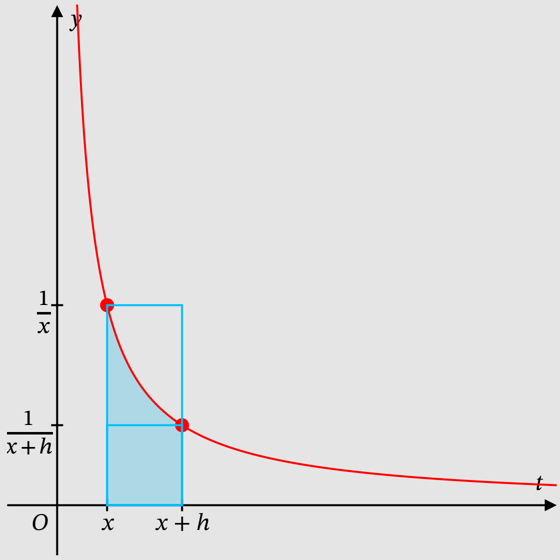

Figure 5–5: Derivative of $\ln x$.

Comparing this area with that of two rectangles, we see that

$$\frac{h}{x+h} < \text{shaded area} = \ln(x+h) - \ln x < \frac{h}{x}\,.$$

Hence, the Newton quotient for $\ln x$ satisfies

$$\frac{1}{x+h} < \frac{\ln(x+h) - \ln x}{h} < \frac{1}{x}\,.$$

Letting h approach 0 from the right, we obtain (by the Squeeze Theorem applied to one-sided limits)

$$\lim_{h \to 0^+} \frac{\ln(x+h) - \ln x}{h} = \frac{1}{x}\,.$$

A similar argument shows that if $0 < x + h < x$, then

$$\frac{1}{x} < \frac{\ln(x+h) - \ln x}{h} < \frac{1}{x+h}\,,$$

so that

$$\lim_{h \to 0^-} \frac{\ln(x+h) - \ln x}{h} = \frac{1}{x}\,.$$

Combining these two one-sided limits we get the desired result:

$$\lim_{h \to 0} \frac{\ln(x+h) - \ln x}{h} = \frac{1}{x}\,.$$

The two properties $\frac{\mathrm{d}}{\mathrm{d}x} \ln x = \frac{1}{x}$ and $\ln 1 = 0$ are sufficient to determine the function $\ln x$ completely. We can deduce from these two properties that $\ln x$ satisfies the appropriate laws of logarithms:

Theorem 64.

$$\ln(xy) = \ln x + \ln y \qquad \ln\left(\frac{1}{x}\right) = -\ln x$$

$$\ln\left(\frac{x}{y}\right) = \ln x - \ln y \qquad \ln(x^r) = r \ln x$$

Because we do not want to assume that exponentials are continuous, we should regard the last property for the moment as only valid for exponents r that are rational numbers.

Proof.
We will only prove the first property because the other parts are proved by the same method. If $y > 0$ is a constant, then by the Chain Rule,

$$\frac{\mathrm{d}}{\mathrm{d}x}(\ln(xy) - \ln x) = \frac{y}{xy} - \frac{1}{x} = 0$$

for all $x > 0$. So, $\ln(xy) - \ln x = C$ (a constant). Putting $x = 1$, we get $C = \ln y$ and the first property follows.

The last property shows that $\ln(2^n) = n \ln 2 \to \infty$ as $n \to \infty$. Therefore, we also have $\ln\left(\frac{1}{2}\right)^n = -n \ln 2 \to -\infty$ as $n \to \infty$. Since $\frac{\mathrm{d}}{\mathrm{d}x} \ln x = \frac{1}{x} > 0$ for $x > 0$, it follows that $\ln x$ is increasing, so we must have

$$\lim_{x\to\infty} \ln x = \infty \quad \text{and} \quad \lim_{x\to-\infty} \ln x = -\infty\,.$$

Example. Show that $\frac{\mathrm{d}}{\mathrm{d}x} \ln|x| = \frac{1}{x}$ for any $x \neq 0$. Hence find $\int \frac{1}{x}\,\mathrm{d}x$.

If $x > 0$, then

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln|x| = \frac{\mathrm{d}}{\mathrm{d}x} \ln x = \frac{1}{x}\,.$$

If $x < 0$, then

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln|x| = \frac{\mathrm{d}}{\mathrm{d}x} \ln(-x) = \frac{1}{-x}(-1) = \frac{1}{x}\,.$$

Therefore, $\frac{\mathrm{d}}{\mathrm{d}x} \ln|x| = \frac{1}{x}$, and on any interval not containing $x = 0$,

$$\int \frac{1}{x}\,\mathrm{d}x = \ln|x| + C\,.$$

The Exponential Function

The function $\ln x$ is bijective on its domain, the interval $]0, \infty[$, so it has an inverse there. For the moment, let us call this inverse $\exp x$. Thus,

$$\forall y > 0 : x = \ln y \implies y = \exp x\,.$$

Since $\ln 1 = 0$, we have $\exp 0 = 1$. The domain of $\exp$ is $]-\infty, \infty[$, the range of $\ln$. The range of $\exp$ is $]0, \infty[$, the domain of $\ln$. We have cancellation identities

$$\forall x \in \mathbb{R} : \ln(\exp x) = x \quad \text{and} \quad \forall x > 0 : \exp(\ln x) = x\,.$$

We can deduce various properties of $\exp$ from corresponding properties of $\ln$. Not surprisingly, they are properties we would expect an exponential function to have.

Theorem 65.

$$(\exp\ x)^r = \exp(rx) \qquad \exp(x+y) = \exp\ x \exp\ y$$
$$\exp(-x) = \frac{1}{\exp\ x} \qquad \exp(x-y) = \frac{\exp\ x}{\exp\ y}$$

For the moment, the first property is asserted only for rational numbers r.

Proof. We prove only the first property, the rest are done similarly.

If $u = (\exp\ x)^r$, then, $\ln u = r \ln(\exp\ x) = rx$. Therefore, $u = \exp(rx)$.

Now we make an important definition!

Definition 43. Let $e = \exp(1)$.

The number e satisfies $\ln e = 1$, so the area bounded by the curve $y = \frac{1}{t}$, the t-axis, and the vertical lines $t = 1$ and $t = e$ must be equal to 1 square unit.

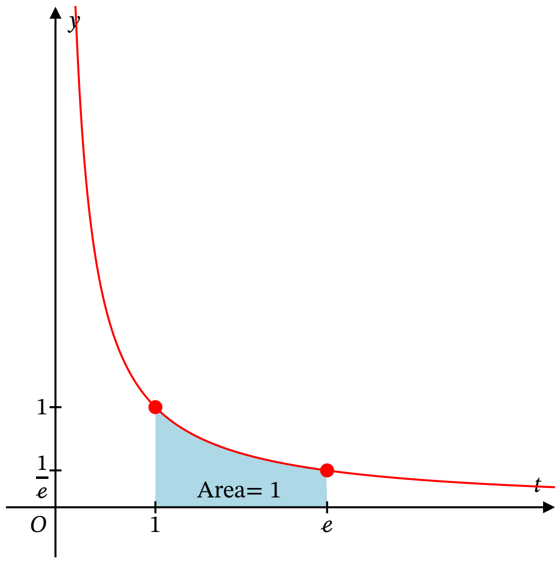

Figure 5–6: The definition of e.

The number e is one of the most important numbers in mathematics. Like π, it is irrational and not a zero of any polynomial with rational coefficients. Its value is between 2 and 3 and begins

$$e = 2.718281828459045\ldots.$$

Later on we will learn that

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots,$$

a formula from which the value of e can be calculated to any desired precision.

The first property of the previous theorem shows that $\exp r = \exp(1r) = (\exp\ 1)^r = e^r$ holds for any rational number r. Now here is a crucial observation. We only know what e^r means if r is a rational number (if $r = \frac{m}{n}$, then $e^r = \sqrt[n]{e^m}$). But $\exp x$ is defined for all real x, rational or not. Since $e^r = \exp r$ when r is rational, we can use $\exp\ x$ as a definition of what e^x means for any real number x, and there will be no contradiction if x happens to be rational.

Definition 44. Let

$$e^{x} = \exp x$$

for all real x.

The last theorem can now be restated in terms of e^x:

$$(e^x)^y = e^{xy} \qquad e^{x+y} = e^x e^y$$
$$e^{-x} = \frac{1}{e^x} \qquad e^{x-y} = \frac{e^x}{e^y}$$

The graph of e^x is the reflection of the graph of its inverse, $\ln x$, in the line $y = x$.

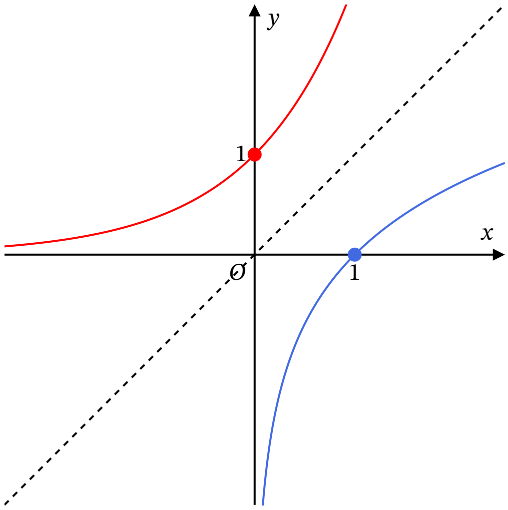

Figure 5–7: The graph of e^x and $\ln x$.

Observe that the x-axis is a horizontal asymptote of the graph of $y = e^x$ as $x \to -\infty$. We have

$$\lim_{x\to-\infty} e^x = 0 \quad \text{and} \quad \lim_{x\to\infty} e^x = \infty\,.$$

Since $\exp x = e^x$ actually is an exponential function, its inverse must actually be a logarithm:

$$\ln x = \log_e x\,.$$

The derivative of $y = e^x$ is calculated by implicit differentiation:

$$\begin{aligned} y = e^x &\implies x = \ln y \\ &\implies 1 = \frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x} \\ &\implies \frac{\mathrm{d}y}{\mathrm{d}x} = y = e^x\,. \end{aligned}$$

Thus, the exponential function has the remarkable property that it is its own derivative and, therefore, also its own antiderivative:

$$\frac{\mathrm{d}}{\mathrm{d}x} e^x = e^x \quad \text{and} \quad \int e^x \,\mathrm{d}x = e^x + C\,.$$

General Exponentials and Logarithms

We can use the fact that e^x is now defined for all real x to define the arbitrary exponential a^x (where $a > 0$) for all real x. If r is rational, then $\ln(a^r) = r \ln a$; therefore, $a^r = e^{r \ln a}$. However, $e^{x \ln a}$ is defined for all real x, so we can use it as a definition of a^x with no possibility of contradiction arising if x is rational.

Definition 45. Let

$$\forall x \in \mathbb{R}, \forall a > 0 : a^x = e^{x \ln a}.$$

The laws of exponents for a^x can now be obtained from those for e^x, as can the derivative:

$$\frac{\mathrm{d}}{\mathrm{d}x} a^x = \frac{\mathrm{d}}{\mathrm{d}x} e^{x \ln a} = e^{x \ln a} \ln a = a^x \ln a.$$

We can also verify the General Power Rule for x^a, where a is any real number, provided $x > 0$:

$$\frac{\mathrm{d}}{\mathrm{d}x} x^a = \frac{\mathrm{d}}{\mathrm{d}x} e^{a \ln x} = e^{a \ln x} \frac{a}{x} = \frac{a x^a}{x} = a x^{a-1}.$$

Example. Find the critical point of $y = x^x$.

We can't differentiate x^x by treating it as a power (like x^a) because the exponent varies. We can't treat it as an exponential (like a^x) because the base varies. We can differentiate it if we first write it in terms of the exponential function, $x^x = e^{x \ln x}$, and then use the Chain Rule and the Product Rule:

$$\frac{\mathrm{d}}{\mathrm{d}x} x^x = \frac{\mathrm{d}}{\mathrm{d}x} e^{x \ln x} = e^{x \ln x}\left(\ln x + x \frac{1}{x}\right) = x^x(1 + \ln x).$$

Now x^x is defined only for $x > 0$ and is itself never 0. (Why?) Therefore, the critical point occurs where $1 + \ln x = 0$; that is, $\ln x = -1$, or $x = \frac{1}{e}$.

Finally, observe that $\frac{\mathrm{d}}{\mathrm{d}x} a^x = a^x \ln a$ is negative for all x if $0 < a < 1$ and is positive for all x if $a > 1$. Thus, a^x is bijective and has an inverse function, $\log_a x$, provided $a > 0$ and $a \neq 1$. If $y = \log_a x$, then $x = a^y$ and, differentiating implicitly with respect to x, we get

$$1 = a^y \ln a \frac{\mathrm{d}y}{\mathrm{d}x} = x \ln a \frac{\mathrm{d}y}{\mathrm{d}x}.$$

Thus, the derivative of $\log_a x$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}x} \log_a x = \frac{1}{x \ln a}.$$

Since $\log_a x$ can be expressed in terms of logarithms to any other base, say e,

$$\log_a x = \frac{\ln x}{\ln a},$$

we normally use only natural logarithms. Exceptions are found in chemistry, acoustics, and other sciences where "logarithmic scales" are used to measure quantities for which a one-unit increase in the measure corresponds to a tenfold increase in the quantity. Logarithms to base 10 are used in defining such scales. In computer science, where powers of 2 play a central role, logarithms to base 2 are often encountered.

Logarithmic Differentiation

Suppose we want to differentiate a function of the form

$$y = (f(x))^{g(x)} \quad \text{for } f(x) > 0.$$

Since the variable appears in both the base and the exponent, neither the general power rule, $\frac{\mathrm{d}}{\mathrm{d}x} x^a = a x^{a-1}$, nor the exponential rule, $\frac{\mathrm{d}}{\mathrm{d}x} a^x = a^x \ln a$ can be directly applied. One method for finding the derivative of such a function is to express it in the form

$$y = e^{g(x) \ln f(x)}$$

and then differentiate, using the Product Rule to handle the exponent. This is the method used in previous example.

The derivative in the example can also be obtained by taking natural logarithms of both sides of the equation $y = x^x$ and differentiating implicitly:

$$\begin{aligned}\ln\, y &= x \ln\, x\\ \frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x} &= \ln\, x + \frac{x}{x} = 1 + \ln\, x\\ \frac{\mathrm{d}y}{\mathrm{d}x} &= y(1 + \ln\, x) = x^x(1 + \ln\, x)\end{aligned}$$

This latter technique is called *logarithmic differentiation*.

Logarithmic differentiation is also useful for finding the derivatives of functions expressed as products and quotients of many factors. Taking logarithms reduces these products and quotients to sums and differences. This usually makes the calculation easier than it would be using the Product and Quotient Rules, especially if the derivative is to be evaluated at a specific point.

Example. Find $\left.\frac{\mathrm{d}u}{\mathrm{d}x}\right|_{x=1}$ if $u = \sqrt{(x+1)(x^2+1)(x^3+1)}$.

$$\begin{aligned}\ln\, u &= \frac{1}{2}\left(\ln(x+1) + \ln\left(x^2+1\right) + \ln\left(x^3+1\right)\right)\\ \frac{1}{u}\frac{\mathrm{d}u}{\mathrm{d}x} &= \frac{1}{2}\left(\frac{1}{x+1} + \frac{2x}{x^2+1} + \frac{3x^2}{x^3+1}\right).\end{aligned}$$

At $x = 1$ we have $u = \sqrt{8} = 2\sqrt{2}$. Hence

$$\left.\frac{\mathrm{d}u}{\mathrm{d}x}\right|_{x=1} = \sqrt{2}\left(\frac{1}{2} + 1 + \frac{3}{2}\right) = 3\sqrt{2}\,.$$

5–3 The Inverse Trigonometric Functions

The six trigonometric functions are periodic and, hence, not one-to-one. However, as we did with the function x^2, we can restrict their domains in such a way that the restricted functions are one-to-one and invertible.

The Inverse Sine (or Arcsine) Function

Let us define a function Sin x (note the capital letter) to be sin x, restricted so that its domain is the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$:

Definition 46.

$$\operatorname{Sin}\, x = \sin\, x \quad \text{if } -\frac{\pi}{2} \leq x \leq \frac{\pi}{2}\,.$$

Since its derivative cos x is positive on the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, the function Sin x is increasing on its domain, so it is a bijective function. It has domain $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and range $[-1,1]$.

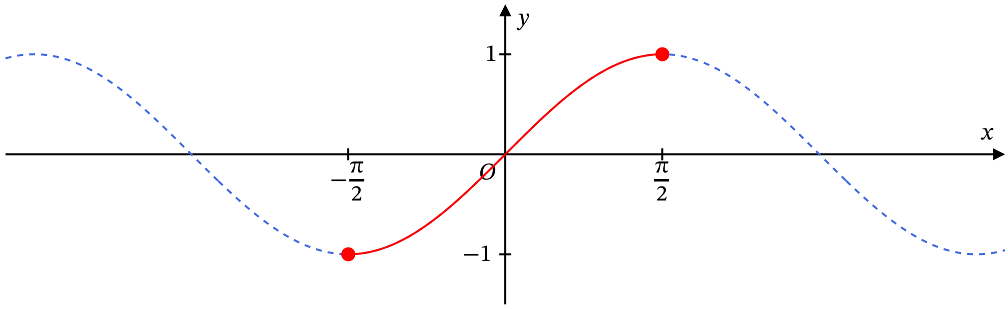

Figure 5–8: The graph of Sin x.

Being bijective, Sin has an inverse function which is denoted Arcsin (or, in some books and computer programs, by arcsin , asin , or Sin^{-1}) and which is called the *inverse sine* or *arcsine function*.

Definition 47.

$$\forall y \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] : x = \sin\, y = \mathrm{Sin}\; y \implies y = \mathrm{Arcsin}\; x\,.$$

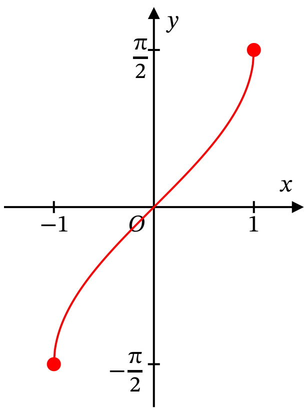

Figure 5–9: The graph of Arcsin x.

The graph of Arcsin is the reflection of the graph of Sin in the line $y = x$. The domain of Arcsin is $[-1,1]$(the range of Sin), and the range of Arcsin is $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ (the domain of Sin).

Example. Simplify the expression tan(Arcsin x).

We want the tangent of an angle whose sine is x. Suppose first that $0 \le x < 1$. We draw a right triangle with one angle θ, and label the sides so that $\theta = \mathrm{Arcsin}\; x$. The side opposite θ is x, and the hypotenuse is 1. The remaining side is $\sqrt{1-x^2}$, and we have

$$\tan(\mathrm{Arcsin}\; x) = \tan\,\theta = \frac{x}{\sqrt{1-x^2}}\,.$$

Because both sides of the above equation are odd functions of x, the same result holds for $0 > x > -1$.

Now let us use implicit differentiation to find the derivative of the inverse sine function. If$y = \mathrm{Arcsin}\; x$, then $x = \sin\, y$ and $-\frac{\pi}{2} \le y \le \frac{\pi}{2}$. Differentiating with respect to x, we obtain

$$1 = \cos\, y \frac{\mathrm{d}y}{\mathrm{d}x}\,.$$

Since $-\frac{\pi}{2} \le y \le \frac{\pi}{2}$, we know that $\cos\, y \ge 0$. Therefore,

$$\cos\, y = \sqrt{1-\sin^2\, y} = 1 - x^2 \quad \text{and} \quad \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\cos\, y} = \frac{1}{\sqrt{1-x^2}}\;;$$

Note that the inverse sine function is differentiable only on the open interval $]-1{,}1[$; the slope of its graph approaches infinity as $x \to -1^+$ or as $x \to 1^-$.

Example. Let $f(x) = \text{Arcsin}(\sin\ x)$ for all real numbers x.

1. Calculate and simplify $f'(x)$.

   Using the Chain Rule and the Pythagorean identity we calculate

$$f'(x) = \frac{1}{\sqrt{1-\sin^2 x}} \cos\ x$$

$$= \frac{\cos\ x}{\sqrt{\cos^2 x}} = \frac{\cos\ x}{|\cos\ x|} = \begin{cases} 1 & \text{if } \cos\ x > 0 \\ -1 & \text{if } \cos\ x < 0\,. \end{cases}$$

2. Where is f differentiable? Where is f continuous?

   f is differentiable at all points where $\cos\ x \neq 0$, that is, everywhere except at odd multiples of $\frac{\pi}{2}$. Since sin is continuous everywhere and has values in $[-1{,}1]$, and since Arcsin is continuous on$[-1{,}1]$, we have that f is continuous on the whole real line.

3. Use the previous results to sketch the graph of f.

   Since f is continuous, its graph has no breaks. The graph consists of straight line segments of slopes alternating between 1 and -1 on intervals between consecutive odd multiples of $\frac{\pi}{2}$. Since $f'(x) = 1$ on the interval $]-1{,}1[$, the graph must be as shown below

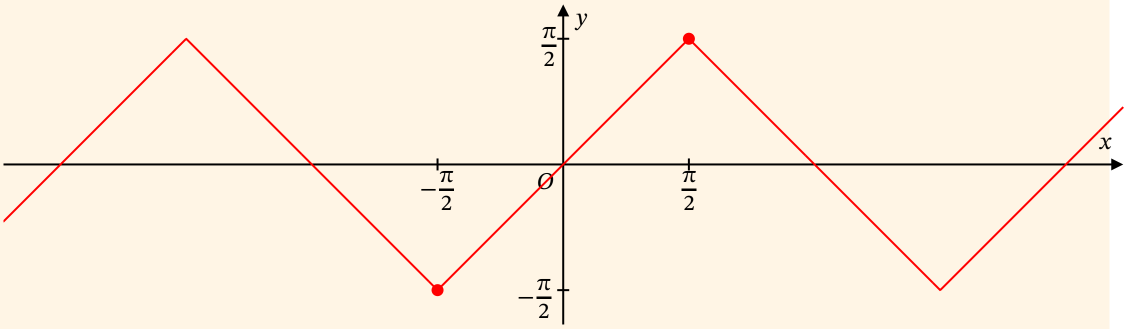

Figure 5–10: The graph of Arcsin(sin x).

Other Inverse Trigonometric Functions

The inverse tangent function is defined in a manner similar to the inverse sine. We begin by restricting the tangent function to an interval where it is bijective.

Definition 48.

$$\text{Tan}\ x = \tan\ x \quad \text{if } -\frac{\pi}{2} \leq x \leq \frac{\pi}{2}\,.$$

The inverse of the function Tan is called the inverse tangent function and is denoted Arctan (or arctan , atan , or Tan^{-1}). The domain of Arctan is the whole real line (the range of Tan). Its range is the open interval $\left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$.

Definition 49.

$$\forall y \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[: x = \tan\ y = \text{Tan}\ y \implies y = \text{Arctan}\ x\,.$$

The derivative of the inverse tangent function is also found by implicit differentiation: if $y = \operatorname{Arctan} x$, then $x = \operatorname{Tan} y$ and

$$1 = \sec^2 y \frac{dy}{dx} = \left(1 + \tan^2 y\right) \frac{dy}{dx} = \left(1 + x^2\right) \frac{dy}{dx} .$$

The function $\cos x$ is bijective on the interval $[0, \pi]$, so we could define the inverse cosine function, Arccos (or arccos , acos , or Cos^{-1}), so that

$$\forall y \in [0, \pi] : x = \cos y \implies y = \operatorname{Arccos} x .$$

However, $\cos y = \sin\left(\frac{\pi}{2} - y\right)$, and $\frac{\pi}{2} - y$ is in the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ when $y \in [0, \pi]$.

Definition 50.

$$\operatorname{Arccos} x = \frac{\pi}{2} - \operatorname{Arcsin} x \quad \text{for } -1 \leq x \leq 1$$

The derivative of $\operatorname{Arccos} x$ is the negative of that of $\operatorname{Arcsin} x$.

5–4 Hyperbolic Functions

Definition 51.
For any real x the hyperbolic cosine, $\cosh x$, and the hyperbolic sine, $\sinh x$, are defined by

$$\cosh x = \frac{e^x + e^{-x}}{2} \quad \text{and} \quad \sinh x = \frac{e^x - e^{-x}}{2} .$$

Recall that cosine and sine are called circular functions because, for any t, the point $(\cos t, \sin t)$ lies on the circle with equation $x^2 + y^2 = 1$. Similarly, cosh and sinh are called hyperbolic functions because the point $(\cosh t, \sinh t)$ lies on the rectangular hyperbola with equations $x^2 - y^2 = 1$,

$$\cosh^2 t - \sinh^2 t = 1 \quad \text{for any real } t .$$

There is no interpretation of t as an arc length or angle as there was in the circular case; however, the area of the hyperbolic sector bounded by $y = 0$, the hyperbola $x^2 - y^2 = 1$, and the ray from the origin to $(\cosh t, \sinh t)$ is $\frac{t}{2}$ square units, just as is the area of the circular sector bounded by $y = 0$, the circle $x^2 + y^2 = 1$, and the ray from the origin to $(\cos t, \sin t)$.

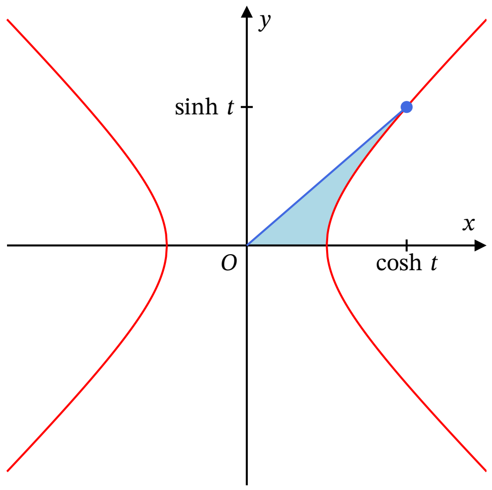

Figure 5–11: The shaded area is $\frac{t}{2}$ square units.

Observe that, similar to the corresponding values of $\cos x$ and $\sin x$, we have

$$\cosh 0 = 1 \quad \text{and} \quad \sinh 0 = 0 ,$$

and $\cosh\ x$, like $\cos\ x$, is an even function, and $\sinh\ x$, like $\sin\ x$, is an odd function:

$$\cosh(-x) = \cosh\ x \quad \text{and} \quad \sinh(-x) = -\sinh\ x\,.$$

The graph $y = \cosh\ x$ is called a *catenary*. A chain hanging by its ends will assume the shape of a catenary.

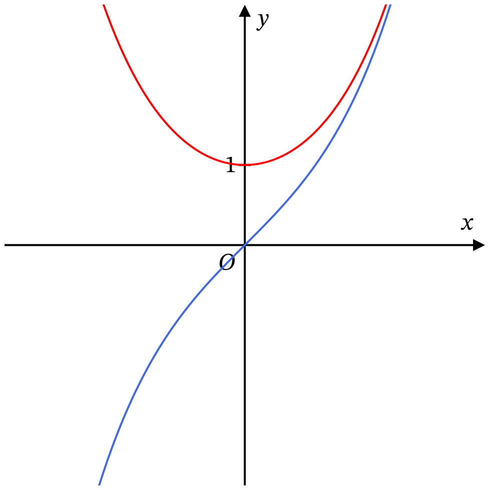

Figure 5–12: The graphs of $\cosh\ x$ *(red) and* $\sinh\ x$ *(blue).*

Many other properties of the hyperbolic functions resemble those of the corresponding circular functions, sometimes with signs changed.

Example. Show that

$$\frac{\mathrm{d}}{\mathrm{d}x}\cosh\ x = \sinh\ x \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}x}\sinh\ x = \cosh\ x$$

We have

$$\frac{\mathrm{d}}{\mathrm{d}x}\cosh\ x = \frac{\mathrm{d}}{\mathrm{d}x}\frac{e^x + e^{-x}}{2} = \frac{e^x + e^{-x}(-1)}{2} = \sinh\ x$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\sin\ x = \frac{\mathrm{d}}{\mathrm{d}x}\frac{e^x - e^{-x}}{2} = \frac{e^x - e^{-x}(-1)}{2} = \cosh\ x\,.$$

The following addition formulas and double-angle formulas can be checked algebraically by using the definition of cosh and sinh and the laws of exponents:

$$\begin{aligned}
\cosh(x+y) &= \cosh\ x \cosh\ y + \sinh\ x \sinh\ y\,,\\
\sinh(x+y) &= \sinh\ x \cosh\ y + \cosh\ x \sinh\ y\,,\\
\cosh(2x) &= \cosh^2\ x + \sinh^2\ x = 1 + 2\sinh^2\ x = 2\cosh^2\ x - 1\,,\\
\cosh(2x) &= 2\sinh\ x \cosh\ x\,.
\end{aligned}$$

By analogy with the trigonometric functions, other hyperbolic functions can be defined in terms of cosh and sinh .

Definition 52.

$$\tanh\ x = \frac{\sinh\ x}{\cosh\ x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}} = \frac{e^{2x} - 1}{e^{2x} + 1}\,.$$

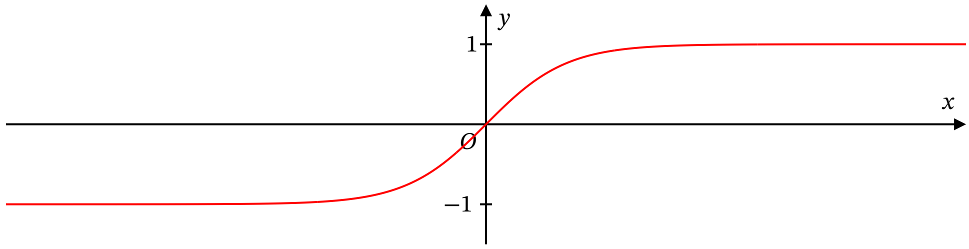

Figure 5–13: The graphs of tanh x.

The functions *sinh* and *tanh* are increasing and therefore bijective and invertible on the whole real line. Their inverses are denoted arcsinh and arctanh, respectively.

Since the hyperbolic functions are defined in terms of exponentials, it is not surprising that their inverses can be expressed in terms of logarithms.

Example. Express the functions arcsinh and arctanh in terms of natural logarithms.

Let $y = \operatorname{arcsinh} x$. Then

$$x = \sinh y = \frac{e^y - e^{-y}}{2} = \frac{e^{2y} - 1}{2e^{2y}}.$$

Therefore,

$$(e^y)^2 - 2xe^y - 1 = 0.$$

This is a quadratic equation in e^y, and it can be solved by the quadratic formula:

$$e^y = \frac{2x \pm \sqrt{4x^2 + 4}}{2} = x \pm \sqrt{x^2 + 1}.$$

Note that $\sqrt{x^2 + 1} > x$. Since e^y cannot be negative, we need to use the positive sign:

$$e^y = x + \sqrt{x^2 + 1}.$$

Hence, $y = \ln\left(x + \sqrt{x^2 + 1}\right)$, and we have

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

Now let $y = \operatorname{arctanh} x$. Then

$$x = \tanh y = \frac{e^y - e^{-y}}{e^y + e^{-y}} = \frac{e^{2y} - 1}{e^{2y} + 1} \quad \text{for } -1 < x < 1$$

or

$$e^{2y} = \frac{1 + x}{1 - x}.$$

Thus,

$$\operatorname{arctanh} x = \frac{1}{2} \ln\left(\frac{1 + x}{1 - x}\right) \quad \text{for } -1 < x < 1.$$

Since cosh is not bijective, its domain must be restricted before an inverse can be defined. Let us define the principal value of cosh to be

$$\forall x \geq 0 : \operatorname{Cosh} x = \cosh x.$$

The inverse, Arccosh x, is then defined by

$$\forall y \geq 0 : x = \cosh y = \operatorname{Cosh} y \implies y = \operatorname{Arccosh} x.$$

As we did for arcsinh, we can obtain the formula

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

Exercise. Find the derivatives of the inverse hyperbolic functions.

CHAPTER 6

INTEGRATION

The second fundamental problem addressed by calculus is the problem of areas, that is, the problem of determining the area of a region of the plane bounded by various curves. Like the problem of tangents considered in Chapter 2, many practical problems in various disciplines require the evaluation of areas for their solution, and the solution of the problem of areas necessarily involves the notion of limits. On the surface the problem of areas appears unrelated to the problem of tangents. However, we will see that the two problems are very closely related; one is the inverse of the other. Finding an area is equivalent to finding an antiderivative or, as we prefer to say, finding an integral. The relationship between areas and antiderivatives is called the Fundamental theorem of Calculus.

6–1 Areas as Limits of Sums

We began the study of derivatives in Chapter 4 by defining what is meant by a tangent line to a curve at a particular point. We would like to begin the study of integrals by defining what is meant by the area of a plane region, but a definition of area is much more difficult to give than a definition of tangency. Let us assume that we know intuitively what area means and list some of its properties.

1. The area of a plane region is a nonnegative real number of square units.
2. The area of a rectangle with width w and height h is $A = wh$.
3. The areas of congruent plane regions are equal.
4. If region S is contained in region R, then the area of S is less than or equal to that of R.
5. If region R is a union of (finitely many) nonoverlapping regions, then the area of R is the sum of the areas of those regions.

Using these five properties we can calculate the area of any *polygon* (a region bounded by straight line segments). First, we note that properties (2) and (3) show that the area of a parallelogram is the same as that of a rectangle having the same base width and height. Any triangle can be butted against a congruent copy of itself to form a parallelogram, so a triangle has area half the base width times the height. Finally, any polygon can be subdivided into finitely many nonoverlapping triangles so its area is the sum of the areas of those triangles.

We can't go beyond polygons without taking limits. If a region has a curved boundary, its area can only be approximated by using rectangles or triangles; calculating the exact area requires the evaluation of a limit. We showed how this could be done for a circle in Section 3.1.

We are going to consider how to find the area of a region R lying under the graph $y = f(x)$ of a nonnegative-valued, continuous function f, above the x-axis and between the vertical lines $x = a$ and $x = b$, where $a < b$. To accomplish this, we proceed as follows. Divide the interval $[a, b]$ into n subintervals by using division points:

$$a = x_0 < x_1 < x_2 < x_3 < \cdots < x_{n-1} < x_n = b$$

Denote by Δx_i the length of the ith subinterval $[x_{i-1}, x_i]$:

$$\Delta x_i = x_i - x_{i-1}, \quad \forall i \in \{1,2,3, \ldots, n\}.$$

Vertically above each subinterval $[x_{i-1}, x_i]$ build a rectangle whose base has length Δx_i and whose height is $f(x_i)$. The area of this rectangle is $f(x_i)\Delta x_i$. Form the sum of these areas:

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

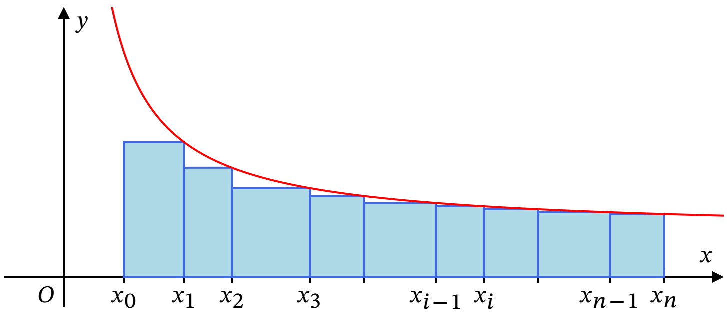

Figure 6–1: Approximating the area under the graph of a decreasing function using rectangles.

The rectangles are shown shaded in the figure for a decreasing function f. For an increasing function, the tops of the rectangles would lie above the graph of f rather than below it. Evidently, S_n is an approximation to the area of the region R, and the approximation gets better as n increases, provided we choose the points $a = x_0 < x_1 < x_2 < x_3 < \cdots < x_{n-1} < x_n = b$ in such a way that the width Δx_i of the widest rectangle approaches zero.

Subdividing a subinterval into two smaller subintervals reduces the error in the approximation by reducing that part of the area under the curve that is not contained in the rectangles. It is reasonable, therefore, to calculate the area of R by finding the limit of S_n as n with the restriction that the largest of the subinterval widths Δx_i must approach zero:

$$\text{Area of } R = \lim_{\substack{n \to \infty \\ \max \Delta x_i \to 0}} S_n .$$

6–2 *The Definite Integral*

We generalize and make more precise the procedure used for finding areas developed in the previous section, and we use it to define the *definite integral* of a function f on an interval I:

Let a partition P of $[a, b]$ be a finite, ordered set of points $P = \{x_0, x_1, x_2, \ldots, x_n\}$, where $a = x_0 < x_1 < x_2 < x_3 < \cdots < x_{n-1} < x_n = b$. Such a partition subdivides $[a, b]$ into n subintervals $[x_{i-1}, x_i]$, where $n = n(P)$ depends on the partition. The length of the ith subinterval $[x_{i-1}, x_i]$ is $\Delta x_i = x_i - x_{i-1}$.

Suppose that the function f is bounded on $[a, b]$. Given any partition P, the n sets $S_i = \{f(x) \mid x_{i-1} \le x \le x_i\}$ have least upper bounds M_i and greatest lower bounds m_i, so that

$$m_j \le f(x) \le M_i \quad \forall x \in [x_{i-1}, x_i].$$

We define *upper* and *lower Riemann sums* for f corresponding to the partition P to be

$$U(f,P)=\sum_{i=1}^{n(P)} M_i\Delta x_i \quad \text{and} \quad L(f,P)=\sum_{i=1}^{n(P)} m_i\Delta x_i\,.$$

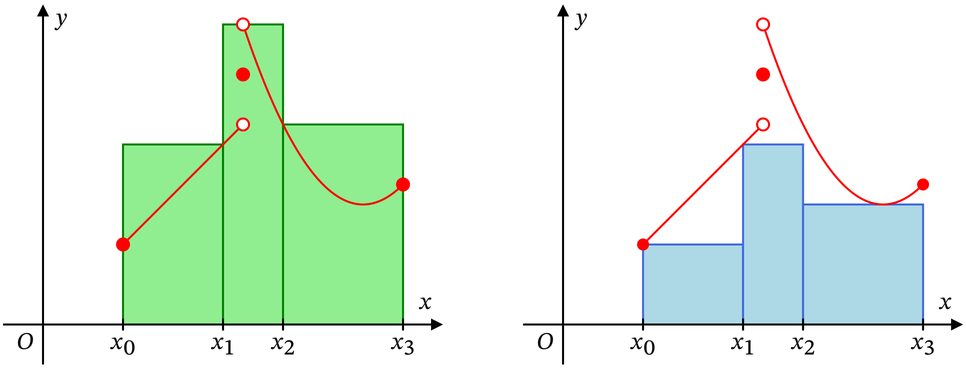

Figure 6–2: Upper and lower sums corresponding to the partition $P=\{x_0,x_1,x_2,x_3\}$.

Note that if f is continuous on $[a,b]$, then m_i and M_i are, in fact, the minimum and maximum values of f over $[x_{i-1},x_i]$ by the Extreme-Value theorem, that is, $m_i=f(l_i)$ and $M_i=(u_i)$, where $f(l_i)\le f(x)\le f(u_i)$ for $x\in[x_{i-1},x_i]$.

Example.
Calculate the lower and upper Riemann sums for the function $f(x)=x^2$ on the interval $[0,a]$ (where $a>0$), corresponding to the partition P_n of $[0,a]$ into n subintervals of equal length.

Each subinterval of P_n has length $\Delta x=\frac{a}{n}$, and the division points are given by $x_i=\frac{ia}{n}$ for $i=0,1,2,\ldots,$ n. Since x^2 is increasing on $[0,a]$, its minimum and maximum values over the ith subinterval $[x_{i-1},x_i]$ occur at $l_i=x_{i-1}$ and $u_i=x_i$, respectively. Thus, the lower Riemann sum of f for P_n is

$$\begin{aligned}L(f,P_n)&=\sum_{i=1}^{n} x_{i-1}^2\Delta x=\frac{a^3}{n^3}\sum_{i=1}^{n}(i-1)^2\\&=\frac{a^3}{n^3}\sum_{j=0}^{n-1} j^2=\frac{a^3}{n^3}\frac{(n-1)n(2(n-1)+1)}{6}=\frac{(n-1)(2n-1)a^3}{6n^2}\end{aligned}$$

where we have used

$$\sum_{i=1}^{n} i^2=\frac{n(n+1)(2n+1)}{6}$$

to evaluate the sum of squares. Similarly, the upper Riemann sum is

$$\begin{aligned}U(f,P_n)&=\sum_{i=1}^{n} x_i^2\Delta x\\&=\frac{a^3}{n^3}\sum_{i=1}^{n} i^2=\frac{a^3}{n^3}\frac{n(n+1)(2n+1)}{6}=\frac{(n+1)(2n+1)a^3}{6n^2}\,.\end{aligned}$$

If P is any partition of $[a,b]$ and we create a new partition $P^\star$ by adding new subdivision points to those of P, thus subdividing the subintervals of P into smaller ones, then we call $P^\star$ a *refinement* of P.

Theorem 66. *If $P^\star$ is a refinement of P, then $L\big(f,P^\star\big)\ge L(f,P)$ and $U\big(f,P^\star\big)\le U(f,P)$.*

Proof.
If S and T are sets of real numbers, and $S \subset T$, then any lower bound (or upper bound) of T is also a lower bound (or upper bound) of S. Hence, the greatest lower bound of S is at least as large as that of T; and the least upper bound of S is no greater than that of T.

Let P be a given partition of $[a,b]$ and form a new partition P' by adding one subdivision point to those of P, say, the point k dividing the ith subinterval $[x_{i-1}, x_i]$ of P into two subintervals $[x_{i-1}, k]$ and $[x_{i-1}, k]$.

Let m_i, m_i' and m_i'' be the greatest lower bounds of the sets of values of $f(x)$ on the intervals $[x_{i-1}, x_i]$, $[x_{i-1}, k]$ and $[k, x_i]$, respectively. Then $m_i \le m_i'$ and $m_i \le m_i''$.

Thus,

$$m_i(x_i - x_{i-1}) \le m_i'(k - x_{i-1}) + m_i''(x_i - k),$$

so $L(f,P) \le L\big(f,P'\big)$.

If $P^\star$ is a refinement of P, it can be obtained by adding one point at a time to those of P and thus $L(f,P) \le L\big(f,P^\star\big)$.

We can prove that $U\big(f,P^\star\big) \le U(f,P)$ in a similar manner.

Theorem 67. *If P and P' are two partitions of $[a,b]$, then $L(f,P) \le U\big(f,P^\star\big)$.*

Proof.
Combine the subdivision points of P and P' to form a new partition $P^\star$, which is a refinement of both P and P'. Then by the previous theorem,

$$L(f,P) \le L\big(f,P^\star\big) \le U\big(f,P^\star\big) \le U(f,P').$$

No lower sum can exceed any upper sum.

The last theorem shows that the set of values of $L(f,P)$ for fixed f and various partitions P of $[a,b]$ is a bounded set; any upper sum is an upper bound for this set. By completeness, the set has a least upper bound, which we shall denote $I_\star$. Thus, $L(f,P) \le I_\star$ for any partition P. Similarly, there exists a greatest lower bound $I^\star$ for the set of values of $U(f,P)$ corresponding to different partitions P.

Theorem 68. $I_\star \le I^\star$.

Proof. Let S and T be sets of real numbers, such that for any $x \in S$ and any $y \in T$ we have $x \le y$.

Because every x is a lower bound of T, we have $x \le \inf\ T$.

Because $\inf\ T$ is an upper bound of S, we have $\sup\ S \le \inf\ T$.

Therefore,

$$x \le \sup\ S \le \inf\ T \le y.$$

Applying this result to the set of values of $L(f,P)$ and $U(f,P)$ for fixed f and various partitions P, we find using the previous theorem

$$\sup\ L(f,P) = I_\star \le I^\star = \inf\ U(f,P).$$

Definition 53.
If f is bounded on $[a, b]$ and $I_\star = I^\star$, then we say that f is *Darboux integrable*, or simply integrable on $[a, b]$, and denote by

$$\int_a^b f(x)\,\mathrm{d}x = I_\star = I^\star$$

the (*Darboux*) *integral* of f on $[a, b]$.

The definite integral of $f(x)$ over $[a, b]$ is a number; it is not a function of x. It depends on the numbers a and b and on the particular function f, but not on the variable x (which is a *dummy variable* like the variable i in the sum $\sum_{i=1}^n f(i)$). Replacing x with another variable does not change the value of the integral:

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b f(t)\,\mathrm{d}t\,.$$

The various parts of the symbol $\int_a^b f(x)\,\mathrm{d}x$ have their own names:

- $\int$ is called the *integral sign*; it resembles the letter S since it represents the limit of a sum.
- a and b are called the *limits of integration*; a is the *lower limit*, b is the *upper limit*.
- The function f is the *integrand*; x is the *variable of integration*.
- $\mathrm{d}x$ is the *differential of* x. It replaces x in the Riemann sums. If an integrand depends on more than one variable, the differential tells you which one is the variable of integration.

Example. Show that $f(x) = x^2$ is integrable over the interval $[0, a]$ (where $a > 0$), and evaluate

$$I = \int_0^a f(x)\,\mathrm{d}x$$

We evaluate the limits as $n \to \infty$ of the lower and upper sums of f over $[0, a]$ obtained in the previous example.

$$\begin{aligned} \lim_{n\to\infty} L(f, P_n) &= \lim_{n\to\infty} \frac{(n-1)(2n-1)a^3}{6n^2} = \frac{a^3}{3}\,, \\ \lim_{n\to\infty} U(f, P_n) &= \lim_{n\to\infty} \frac{(n+1)(2n+1)a^3}{6n^2} = \frac{a^3}{3}\,. \end{aligned}$$

Since $L(f, P_n) \le I \le U(f, P_n)$, we must have $I = \frac{a^3}{3}$. Thus, $f(x) = x^2$ is integrable over $[0, a]$, and

$$\int_0^a f(x)\,\mathrm{d}x = \int_0^a x^2\,\mathrm{d}x = \frac{a^3}{3}\,.$$

Let $P = \{x_0, x_1, x_2, \ldots, x_n\}$, where $a = x_0 < x_1 < x_2 < \cdots < x_n = b$, be a partition of $[a, b]$. In each subinterval $[x_{i-1}, x_i]$ of P pick a point c_i (called a *tag*). Let $c = (c_1, c_2, \ldots, c_n)$ denote the list of these tags. The sum

$$\begin{aligned} R(f, P, c) &= \sum_{i=1}^n f(c_i)\Delta x \\ &= f(c_1)\Delta x_1 + f(c_2)\Delta x_2 + \cdots + f(c_n)\Delta x_n \end{aligned}$$

is called the *Riemann sum* of f on $[a, b]$ corresponding to partition P and tags c.

For any choice of the tags c, the Riemann sum $F(f, P, c)$ satisfies

$$L(f, P) \le R(f, P, c) \le U(f, P)$$